INFERENCE ON THE DIET COMPOSITION OF PREDATORS USING FATTY ACID SIGNATURES: AN APPLICATION OF BAYESIAN INFERENCE ON LINEAR MIXING MODELS

by

J. Wade Blanchard

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
April 2011

DALHOUSIE UNIVERSITY

DEPARTMENT OF MATHEMATICS AND STATISTICS

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled “INFERENCE ON THE DIET COMPOSITION OF PREDATORS USING FATTY ACID SIGNATURES: AN APPLICATION OF BAYESIAN INFERENCE ON LINEAR MIXING MODELS” by J. Wade Blanchard in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Dated: April 4, 2011

External Examiner: ____________________________
Peter Guttorp

Research Supervisors: ____________________________
Bruce Smith

____
Don Bowen

Examining Committee: ____________________________
Sara Iverson

____
David Hamilton

____
Michael Dowd

DALHOUSIE UNIVERSITY

DATE: April 4, 2011

AUTHOR: J. Wade Blanchard

TITLE: INFERENCE ON THE DIET COMPOSITION OF PREDATORS USING FATTY ACID SIGNATURES: AN APPLICATION OF BAYESIAN INFERENCE ON LINEAR MIXING MODELS

DEPARTMENT OR SCHOOL: Department of Mathematics and Statistics

DEGREE: PhD CONVOCATION: May YEAR: 2011

Signature of Author

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Determining the diet composition of predators is an important ingredient in many areas of ecology: understanding predator prey relationships, foraging behaviour of predators and consumption models to name a few. Iverson et al. (2004) developed a method based on the fatty acid signatures known as quantitative fatty acid signature analysis (QFASA). Fatty acids are the basic building blocks of most lipids and are indicative of diet, in the sense that higher level predators have limited ability to modify the fatty acids they ingest.

Billheimer (2001) introduced a Bayesian compositional receptor model, where he apportioned the air pollution recorded a receptor site in Juneau Alaska into two components, woodstove smoke and automobile emissions. Building on this model we add components to allow for predator biosynthesis and differential fat content and also introduce a model which allows for design effects.

Additionally we give some interesting results on the multi–modality of the logistic normal distribution. We also generalize the test of stationarity proposed by Priestley and Subba Rao (1969), based on evolutionary spectral ideas, as an alternative way of assessing when a MCMC sampler has reached its stationary distribution.

LIST OF ABBREVIATIONS AND SYMBOLS USED

Symbol	Description
a	the number of fatty acids
$\mathcal{C}$	compositional closure operator a mapping $\mathcal{R}_+^D \rightarrow \mathcal{S}^d$
D	dimension of a generic compositional vector $D = d + 1$
$\mathcal{D}^d(\boldsymbol{\delta})$	Dirichlet distribution
$\underset{(a\times nr)}{\mathbf{E}}$	an $a\times nr$ matrix of the compositional errors for the predators, note the columns consist of $\boldsymbol{\epsilon}_{ir}$.
$\underset{(a\times n_j)}{\mathbf{E}_{\mathbf{X}_j}}$	an $a \times n_j$ matrix of the compositional errors for the jth prey type's fatty acid profile.
$\underset{(2\times n_j)}{\mathbf{E}_{\mathbf{Z}_j}}$	an $2 \times n_j$ matrix of the compositional errors for the jth prey type's fat composition profile.
$\underset{(a\times L)}{\mathbf{E}_{\mathbf{U}}}$	an $a\times L$ matrix of the compositional errors for the calibration prey fatty acid profile.
$\underset{(a\times L)}{\mathbf{E}_{\mathbf{V}}}$	an $a \times M$ matrix of the compositional errors for the calibration prey fatty acid profile.
$\mathbf{F}$	a $d \times D$ matrix $\mathbf{F} = [I_d : -\mathbf{j}_d]$
$\mathbf{I}_d$	a $d \times d$ dimensional identity matrix
$\mathbf{j}_D$	D–dimensional vector of 1's.
$\mathcal{J}_D$	the zero element of the simplex, $\mathcal{J}_D = \mathbf{j}_D/D$
$\mathbf{J}_d$	a $d \times d$ dimensional matrix of ones $J_d = \mathbf{j}\mathbf{j}'$
$\mathcal{L}^d(.\|\boldsymbol{\mu}, \Sigma)$	the d-dimensional additive logistic normal distribution
L	the number of prey in the calibration experiment
M	the number of predators in the calibration experiment
$\mathcal{N}^d(.\|\boldsymbol{\mu}, \Sigma)$	the d–dimensional multivariate normal distribution
$\mathbf{N}_d$	$\mathbf{N}_d = I_d + J_d$
$\mathbf{N}_d^{-1}$	$\mathbf{N}_d^{-1} = I_d - \frac{1}{D}\mathbf{J}_d$
n	the number of predators

Symbol	Description
n_j	the number of samples from the jth prey type
p	the number of prey types
$\mathcal{R}^D$	D–dimensional real space
$\mathcal{R}_+^D$	D–dimensional positive real quadrant, that is, $u_i > 0; \forall i$
$\mathcal{S}^d$	d–dimensional simplex
QFASA	Quantitative Fatty Acid Signature Analysis
$\mathbf{u}_l$	the fatty acid profile of the prey in the calibration experiment (a vector of length a)
$\underset{(p\times w}{\mathbf{T}_j}$	an $p \times w$ matrix of population diet compositions
$\underset{(a\times L}{\mathbf{U}_j}$	an $a \times L$ matrix with the individual samples, $\mathbf{u}_j$, forming the columns
$\mathbf{v}_l$	the fatty acid profile of the predator in the calibration experiment (a vector of length a)
$\underset{(a\times L}{\mathbf{V}_j}$	an $a \times L$ matrix with the individual samples, $\mathbf{v}_j$, forming the columns
$\underset{(w\times n)}{\mathbf{W}}$	the design matrix for the population level effects
$\underset{(1\times n)}{\mathbf{W}_{\mathbf{X}_j}}$	a $1 \times n_j$ matrix of ones or a row vector.
$\underset{(1\times n)}{\mathbf{W}_{\mathbf{Z}_j}}$	a $1 \times n_j$ matrix of ones or a row vector.
$\underset{(1\times n)}{\mathbf{W}_{\mathbf{U}_j}}$	a $1 \times L$ matrix of ones or a row vector.
$\underset{(1\times n)}{\mathbf{W}_{\mathbf{V}_j}}$	a $1 \times M$ matrix of ones or a row vector.
$\mathbf{x}_{jk}$	the fatty acid profile of the kth prey of type j (a vector of length a)($j = 1, \ldots, p$ and $k = 1, \ldots, n_j$)
$\underset{(a\times nr)}{\mathbf{X}_j}$	an $a \times n_j$ matrix with the individual samples, $\mathbf{x}_{jk}$, of the j prey type forming the columns
$\mathbf{y}_i$	the fatty acid profile of the ith predator (a vector of length a) $(i = 1, \ldots, n)$
$\underset{(a\times nr)}{\mathbf{Y}}$	the collection of all fatty acids concatenated column–wise.
$\mathbf{z}$	D–dimensional compositional vector

Symbol	Description
$\mathbf{z}_{jk}$	the proportion of fat and non–fat of the kth prey of type j (a vector of length 2)($j = 1, \ldots, p$ and $k = 1, \ldots, n_j$)
$\underset{(2\times n_j)}{\mathbf{Z}_j}$	a $2 \times n_j$ matrix with the individual samples, $\mathbf{z}_{jk}$, of the $_j$ predator forming the columns
$\boldsymbol{\alpha}_i$	the p–vector of convex mixing coefficients for the ith predator
$\delta_{\boldsymbol{\tau}}$	degrees of freedom for the inverse Wishart prior on the mixing covariance matrix $\Sigma_{\boldsymbol{\tau}}$
$\delta_{\mathbf{x}_j}$	degrees of freedom for the inverse Wishart prior on the covariance matrix $\Sigma_{\mathbf{x}_j}$ for the jth prey type's fatty acid profiles
$\delta_{\mathbf{z}_j}$	degrees of freedom for the inverse Wishart prior on the covariance matrix $\Sigma_{\mathbf{x}_j}$ for the jth prey type's fat composition
$\delta_{\boldsymbol{\epsilon}}$	degrees of freedom for the inverse Wishart prior on $\Sigma_{\boldsymbol{\epsilon}}$
$\delta_{\mathbf{u}}$	degrees of freedom for the inverse Wishart prior on the covariance matrix $\Sigma_{\mathbf{u}}$ for the fatty acid profiles of the prey from the calibration experiment
$\delta_{\mathbf{v}}$	degrees of freedom for the inverse Wishart prior on the covariance matrix $\Sigma_{\mathbf{x}}$ for the fatty acid profiles of the predator from the calibration experiment
$\boldsymbol{\epsilon}_i$	the compositional error for the ith predator
$\underset{(p\times n)}{\Gamma}$	a $p \times n$ matrix consisting of the mixing distribution and the population level effects and adjusted for fat content.
$\underset{(p\times n)}{\Gamma^m}$	a $p \times n$ matrix of samples from the multilevel distribution with zero mean.
$\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)'$	the p–dimensional vector of fat contents for each prey type
$\boldsymbol{\lambda}_j^v = (\lambda_j, 1 - \lambda_j)'$	the fat content profile consisting of the proportion of fat for the jth prey type and the proportion of non–fat
$\boldsymbol{\eta}_r$	prior location parameter for $\boldsymbol{\mu}_{\boldsymbol{\tau}_r}$
$\boldsymbol{\mu}_{\boldsymbol{\theta}_j}$	prior location parameter for $\boldsymbol{\theta}_j$
$\boldsymbol{\mu}_{\lambda_j}$	prior location parameter for λ_j
$\boldsymbol{\mu}_{\boldsymbol{\theta}_{\mathbf{u}}}$	prior location parameter for $\boldsymbol{\theta}_{\mathbf{u}}$

Symbol	Description
$\boldsymbol{\mu}_{\boldsymbol{\theta}_{\mathbf{v}}}$	prior location parameter for $\boldsymbol{\theta}_{\mathbf{v}}$
ϕ	the additive logratio transformation
ϕ_c	the column–wise additive logratio transformation
$\Psi_{\boldsymbol{\tau}}$	scale matrix for the inverse Wishart prior on the mixing covariance matrix $\Sigma_{\boldsymbol{\tau}}$
$\Psi_{\mathbf{x}_j}$	scale matrix for the inverse Wishart prior on the covariance matrix $\Sigma_{\mathbf{x}_j}$ for the jth prey type's fatty acid profiles
$\Psi_{\mathbf{z}_j}$	scale matrix for freedom for the inverse Wishart prior on the covariance matrix $\Sigma_{\mathbf{x}_j}$ for the jth prey type's fat composition
$\Psi_{\boldsymbol{\epsilon}}$	scale matrix for the inverse Wishart prior on $\Sigma_{\boldsymbol{\epsilon}}$
$\Psi_{\mathbf{u}}$	scale matrix for the inverse Wishart prior on the covariance matrix $\Sigma_{\mathbf{u}}$ for the fatty acid profiles of the prey from the calibration experiment
$\Psi_{\mathbf{v}}$	scale matrix for the inverse Wishart prior on the covariance matrix $\Sigma_{\mathbf{x}}$ for the fatty acid profiles of the predator from the calibration experiment
$\Sigma_{\boldsymbol{\mu}_{\boldsymbol{\tau}_r}}$	prior covariance matrix for $\boldsymbol{\mu}_{\boldsymbol{\tau}_r}$
$\Sigma_{\boldsymbol{\theta}_j}$	prior covariance matrix for $\boldsymbol{\theta}_j$
Σ_{λ_j}	prior covariance matrix for λ_j^v
$\Sigma_{\boldsymbol{\theta}_{\mathbf{u}}}$	prior covariance matrix for $\boldsymbol{\theta}_{\mathbf{u}}$
$\Sigma_{\boldsymbol{\theta}_{\mathbf{v}}}$	prior covariance matrix for $\theta_{\mathbf{v}}$
$\Sigma_{\boldsymbol{\tau}}$	the covariance matrix of the mixing distribution
$\Sigma_{\mathbf{x}_j}$	the covariance matrix of the fatty profiles for the samples of the jth prey type
$\Sigma_{\mathbf{z}_j}$	the covariance matrix of the fat compositions for the samples of the jth prey type
$\Sigma_{\boldsymbol{\epsilon}}$	the error covariance matrix
$\Sigma_{\mathbf{u}}$	the covariance matrix of the fatty profiles for the samples of the prey in the calibration experiment
$\Sigma_{\mathbf{v}}$	the covariance matrix of the fatty profiles for the samples of the predator in the calibration experiment

Symbol	Description
$\boldsymbol{\tau}$	the actual diet consumed by the predator
$\boldsymbol{\theta}_j$	the a–dimensional location vector for the j prey type
$\boldsymbol{\theta}_{\mathbf{u}}$	the a–dimensional location vector for the calibration experiment prey
$\boldsymbol{\theta}_{\mathbf{v}}$	the a–dimensional location vector for the calibration experiment predator
$\underset{(a\times p)}{\boldsymbol{\Theta}} = (\boldsymbol{\theta}_1\vert\ldots\vert\boldsymbol{\theta}_p)$	the $a \times p$ matrix of location vectors
$\oplus$	perturbation operator, the compositional equivalent to addition on Euclidean spaces
$\oplus_c$	the column–wise perturbation operator
$\ominus$	inverse perturbation operator, the compositional equivalent to subtraction on Euclidean spaces

ACKNOWLEDGEMENTS

Writing a dissertation is a long and difficult process and is not a task that one accomplishes without tremendous support from a number of individuals. I am certainly no exception in this regard and feel compelled to thank numerous people for their support and encouragement.

Firstly, I would like to thank my supervisor Bruce Smith. Without his patience and encouragement the road would have been much more difficult. We had numerous discussions about a number of issues that were not directly relevant to my dissertation but were extremely important to my profession development. My co–supervisor Don Bowen gave great insights in the many biological issues that were pertinent to estimating diets of predators. He also invited me along for three unforgettable field seasons on Sable Island, which gave me valuable insights into the issues involved in collection biological data. Not many statisticians have the opportunity to get their hands dirty with the collection of real biological data. I owe a huge debt of gratitude to Sara Iverson, as it was her original idea that the diet of a predator can be inferred from the fatty acid profiles. The numerous discussions we have had over the past 10 years on these issues were very valuable in the formulation of many of the ideas of the thesis. I would also like to thank David Hamilton and Michael Dowd, for their valuable input as committee members and readers of my thesis. I would also like to thank my external Peter Guttorp for taking the time out of his busy schedule to read my thesis and provide valuable feedback.

I would also like to thank George Gabor for being so enthusiastic about the Bayesian approach to statistical inference and introducing me to the objective Bayesian approach advocated by E.T. Jaynes and Richard Cox. I thank Chris Field, my long time supervisor and collaborator on numerous projects, for hiring me those many years ago as a statistical consultant and providing many stimulating conversations and discussions over the years. Our interactions formed the basis of expertise as an applied statistician.

I am privileged to have a large number of very supportive friends. For fear of leaving out anyone inadvertently I will not attempt to list them, you know who you are and I thank you for your support and friendship.

I would also like to thank my family for their support during my slightly premature

mid–life crises.

the Iverson lab and Myers/Worm/Lotze lab. And also my

CHAPTER 1

INTRODUCTION

The diet composition of predators is crucial in understanding many complex ecological systems, however, for most predators direct observation of feeding is difficult, if not impossible, particularly for marine predators. Iverson et al. (2004) gives details of a relatively new method, quantitative fatty acid signature analysis (QFASA) , of estimating the diet composition based on the fatty acid profiles of the predator and their potential prey. The method is based on a variant of mass balance model, that is, the fatty acid profile of the prey is essentially laid down in the fat stores (adipose tissue) of the predator in a very predictable fashion. In a nutshell, you are what you eat.

Fatty acids are the basic building blocks of most lipids and typically are not degraded during digestion and those not used for energy are deposited in adipose tissue. We refer to fatty acids by the standard nomenclature of the carbon chain length:number of double bonds, and the location (n-x) of the double bond nearest the terminal methyl group (see Iverson et al., 2004; Budge et al., 2006, and the references therein for more details). For example, 22:1n-11 would be a fatty acid with 22 carbons and one double bond located 11 carbon atoms from the methyl group. There are well over 70 distinct fatty acids that can be identified depending on the analytical methods used and the gas chromatograph (GC) column used, however, the details of which are beyond the scope of the present work. The interested reader is referred to Iverson et al. (2004); Budge et al. (2006).

Cook (1991) states that a relatively limited number of fatty acids can be biosynthesized by animals, this allows one to separate fatty acids into dietary and non–dietary components. Thus, we only consider those fatty acids that are indicative of diet, however, we also consider the possibility that some components maybe be biosynthesized in small amounts or may exhibit some type of modified deposition . With this in mind Iverson et al. (2004) developed the concept of calibration coefficients to account for potential predator biosynthesis. This was done through several controlled feeding studies of captive animals that were fed the same diet for extended periods of time. The calibration coefficient is then defined as the ratio of the fatty acid signature of the predator, at the end of feeding, to the fatty acid signature of the diet. Thus, calibration coefficients differing from one would indicate predator biosynthesis and less than one would indicate reduced deposition.

There are two other indirect methods of diet reconstruction used in ecology: digestive resistant hard part analysis and stable isotope analysis. We discuss each briefly.

Digestive hard part methods (Gaston and Noble, 1985; Peirce and Boyle, 1991) have several potential biases: soft bodied prey are difficult to identify, the so called diagnostic hard parts of prey may not be eaten by the predator, the hard parts maybe eroded during digestion and finally these methods only give a snapshot of the most recent kill (see Iverson et al., 2004). A particular difficulty with stomach hard parts is that the animal has to be sacrificed making longitudinal studies of diet impossible. A practical difficulty with fecal analysis is that they can only be collected on land. Therefore, to be useful for marine predators, the species has to spend some duration of its life cycle in a terrestrial setting. Additionally, the hard parts may be further damaged by erosion process compared to those obtained from stomach samples as the hard parts have to travel the whole length of the digestive track.

Stable isotope methods typically use carbon and nitrogen isotopes and are also based on a chemical mass balance model. Phillips (2001); Phillips and Gregg (2001); Ben-David and Schell (2001); Phillips and Koch (2001); Phillips and Gregg (2003); Lubetkin and Simenstad (2004); Phillips et al. (2005) plus numerous other authors have developed statistical approaches that are very similar in flavour to Iverson et al. (2004), which is not surprising due to their similar underpinnings. Moore and Semmens (2008); Jackson et al. (2009); Semmens et al. (2009) develop a Bayesian linear mixing model for stable isotopes which has some similarities to the methods proposed here. However, stable isotopes usually only give the trophic level of the predator and not the species composition of the diet (see Hobson, 1993; Gilmore et al., 1995; Koch et al., 1995; Iverson et al., 2004).

Figure 1.1 gives mean fatty acid profiles for the 28 species thought to be potential prey items of Harbour seals on Sable Island for a subset of 37 fatty acids of the approximately 70 that can be identified by gas chromatography. The fatty acid profiles of the species are somewhat similar, that is, species tend to have large concentrations of similar fatty acids. This will make the problem of apportioning the diet of the predator to the individual species difficult. Figure 1.2 gives the fatty acid profiles of 23 adult male Harbour seals. The Harbour seals are remarkably similar in their fatty acid profiles, which could be indicative of similar foraging strategies among individual males. The Harbour seal data will be used as an illustrative example (see chapter 6).

The goal of the thesis is to develop a Bayesian approach to the QFASA method developed by Iverson et al. (2004) as an alternative method of inference to the bootstrapping approach developed there and further refined by the PhD work of Connie Stewart (see Stewart, 2005).

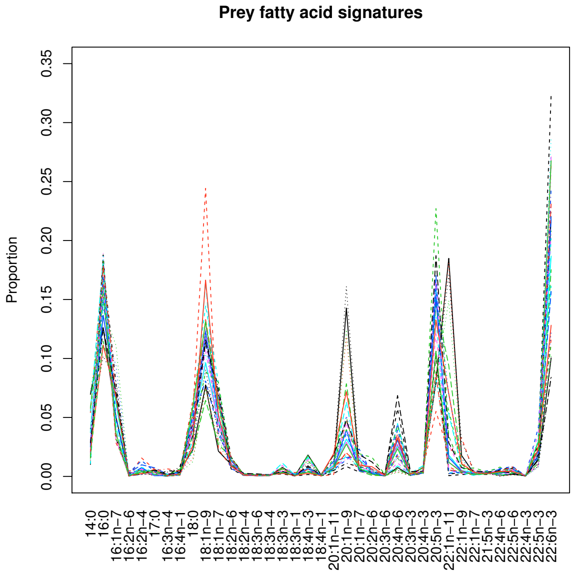

Figure 1.1: The mean fatty acid profiles for the 28 species of potential prey of the Harbour seals. See chapter 6 for details on the species names.

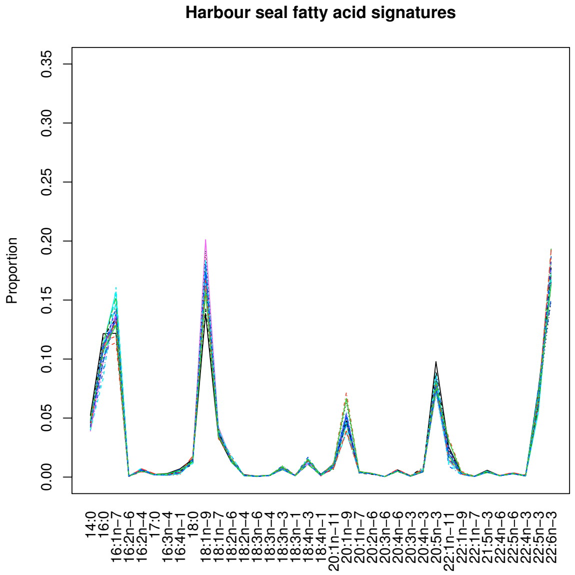

Figure 1.2: The fatty acid profiles of the 23 adult male Harbour seals that were equipped with a National geographic critter cams (see Bowen et al., 2002).

We begin with a brief description of the previous work on QFASA given in Iverson et al. (2004); Stewart (2005).

1.1 QFASA Methods

In this section we briefly present the work of Iverson et al. (2004) which outlines the basic QFASA model and the PhD work of Stewart (2005) which considers traditional statistical inference for composition, specifically confidence intervals and hypothesis tests.

1.1.1 Diet Point Estimates: Review of Iverson et al. (2004)

The Iverson et al. (2004) model is described below:

Let $\mathbf{y}_i$ denote the a–vector of fatty acids, or the fatty acid profile of the ith predator, let $\mathbf{x}_{jk}$ denote the fatty acid profile from the kth prey of the jth prey type, let p represent the number of prey types and let n_j be the number of samples from the j type. The fitted value for the diet of the ith predator is given by

$$\hat{\mathbf{y}}_i = \sum_{j=1}^{p} \hat{\alpha}_{ij} \overline{\mathbf{x}}_j$$

where $\boldsymbol{\alpha}_i = (\alpha_{i1}, \ldots, \alpha_{ip})$ is the diet composition for the ith individual, $\hat{\boldsymbol{\alpha}}$ is the estimated diet composition and $\overline{\mathbf{x}}_j$ is the mean vector for the jth prey type given by

$$\overline{\mathbf{x}}_j = \frac{1}{n_j} \sum_{k=1}^{n_j} \mathbf{x}_{jk}$$

where the summation is done over the fatty acid profiles (vectors). The primary goal is the estimation of $\boldsymbol{\alpha}$ which is carried out conditional on the observed prey fatty acid profiles. This formulation can be seen as a variant of chemical mass balance models (see Henry et al., 1984). Iverson et al. (2004) considered distance based methods, that is, they chose the composition $\hat{\boldsymbol{\alpha}}$ which minimized the distance between the actual predator and the predicted predator subject to the constraints that components must be between zero and one and also sum to one. They considered several distance measures but eventually settled on the following

$$KL(\mathbf{x}, \mathbf{y}) = \sum_j (y_j - x_j) \log(y_j / x_j)$$

which is a variant of the Kulback-Liebler (KL) distance (actually a combination of the forward and backward Kulback–Liebler) Encyclopedia of Statistics (1983). They also considered the squared distance, the squared relative error and the squared error distance of the logs. Thus for each distance, they find the estimates of $\boldsymbol{\alpha}_i$ that minimize the distance, denoted by $\hat{\boldsymbol{\alpha}}_i$ using a constrained nonlinear optimizer, since each of the components of $\hat{\boldsymbol{\alpha}}_i$ must lie between zero and one and also sum to one.

To capture the large amount of variability in the prey profiles, a bootstrapping procedure was implemented to compute standard errors for their point estimates of the diet composition vectors $\hat{\boldsymbol{\alpha}}_i$.

As previously mentioned the fatty acid composition of the diet isn't deposited directly in the fat stores of the predator, that is, biosynthesis and deposition effects happen on some fatty acids no matter their trophic level. To assess the effect of predator metabolism Iverson et al. (2004) carried out a long term captive feeding experiment on juvenile grey(n=8) and harp seals(n=5) fed a stable diet of herring for at least 5 months. At the end of the sampling period a blubber biopsy was taken from each of the predators and 30 herring were saved for subsequent analysis. The samples were analyzed according to the methods described in Iverson et al. (2004). The effect of predator biosynthesis was computed by taking the 10% trimmed mean of the ratio of each predator to each prey, giving a calibration coefficient vector $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_a)$ as shown in the following equation

$$\kappa_k = \text{trimmed mean}_k \mathbf{y}^c_{ik}/\mathbf{x}^c_{jk}$$

where the ratio is taken over all possible (i, j) pairs.

The fat content of the potential prey items was also taken into account, but we delay discussion of this till the next section.

In order to investigate the performance of their estimation procedure a large simulation study was carried out to determine the effect of the following factors: fatty acid set (dietary vs extended), choice of distance(KL, squared distance, etc), amount of noise, diet composition and sample size.

They found that the method performed quite well in the simulation studies, in that, it was able to reconstruct the true signature diet quite well and also decided that the Kulback–Liebler distance performed the best under most conditions. They also verified their method with two captive experiments and 23 free ranging harbour seals equipped with critter cams.

1.1.2 Interval Estimates of Diet: Stewart (2005)

The major thrust of Stewart (2005) was to provide interval estimates for the QFASA method proposed by Iverson et al. (2004) and to investigate alternative estimation strategies. This section will outline this work.

Iverson et al. (2004) used the mean of each prey type to represent the prey fatty acid profile. Stewart (2005) explored alternative ways of summarizing the prey, including the following:

- The random sampling method replaces the mean for the jth prey type with a randomly selected sample of that type. Rather than fit all possible random selections, a smaller number of possible random configurations are run and the estimates are recorded for each.

- Considering each prey type as a random sample from a multivariate distribution, Stewart (2005) defines multivariate quantiles for each prey type based on the work of Chakraborty (2001). The distance minimization is then performed for a random selection of the quantiles. The following quantiles were used for each prey type 0.25, 0.5, 0.75.

- The KL quantile method, consists of computing the Kulback-Liebler distance from each predator to each individual prey of a given type, then using a univariate quantile of this distribution and then choosing appropriate quantiles as above. The estimation is carried out on a random sample of the selected quantiles. Note that unlike the other methods, the actual predator is used in determining the chosen quantile.

Billheimer et al. (2001) and Stewart (2005), among others, discuss some difficulties in interpreting log–ratios on the transformation scale (see section on compositions). Stewart (2005) goes to great lengths to justify the choice of an appropriate measure of location, however, Billheimer et al. (2001) gives a cleaner interpretation provided work is performed on the simplex (see section on compositions).

Stewart (2005) makes the following statement which gives their rationale for confidence interval construction:

> Developing confidence interval (CI) methods for the true diet of a predator or common diet of a group of predators based on QFASA essentially required

an examination into ways of parametric modeling the QFASA diet estimates, “parametrizing” the true diet, and and estimating the standard error of the diet estimates.

With this in mind, one of the challenges faced by Stewart (2005) was dealing with the essential zeros (see Martin-Fernández et al., 2003) encountered in the diet estimates. She based her confidence intervals on marginal distributions of each diet component rather than confidence regions since the number of components can be rather large. She used mixture distributions, as suggested by Martin-Fernández et al. (2003), to sub–divide the diet estimates into populations and then modeled the populations. This was then followed by modeling of the non-zero components as advocated by Aitchison (2003).

Stewart (2005) considers four broad classes of confidence intervals: large sample intervals, which were based on asymptotic normality of the diet estimates; parametric and semi–parametric intervals which were based on the properties of the additive logistic and the multiplicative logistic distributions; finally, non-parametric intervals, based on the bootstrap, one based on the percentile method and one based on inverting a test statistic. The predators were generated under the null hypothesis and bootstrap p–values were then computed.

Stewart (2005) carried out a simulation study with two known diets, and two distances, Kulback-Liebler and the Aitchison distance (see section 2.1). On the basis of these simulations the basic percentile confidence intervals perform best.

Stewart (2005) also examined a goodness of fit statistic, that was loosely based on the usual regression R^2 statistic, called PVE (proportion of variability explained). The sum of squares of error is the final distance (AIT or KL), while the total sum of squares, was based on randomly assigning prey fatty acid signatures a species label. Simulation studies verified that the PVE had desirable properties. She also studied a “backward elimination” procedure based on the bootstrap.

Stewart (2005) considered permutation tests for changes in diet or changes in fatty acid signature of groups of independent samples and also for changes in fatty acid signature over time. She also considered the power of such tests via simulation.

1.2 Thesis Outline

Compositions are central to the development of the thesis. We say a a–dimensional vector $\mathbf{x}$ is a composition if $\mathbf{x}_i > 0$ and $\sum_{i=1}^{a} x_i = 1.0$, in other words, it lies in the a–dimensional simplex. John Aitchison spent much of his long career warning of the follies that can arise if one ignores the unit sum constraint inherent in data of this type. He stresses that at best compositions give you relative amounts and not absolute amounts, motivating the consideration of ratios and their logarithms. Chapter 2 gives a brief summary of the vast literature of compositional data taken mostly from the work of Aitchison (2003) and Billheimer (2001); Billheimer et al. (2001).

One difficulty associated with the Bayesian approach is that the computations are often more difficult than their classical counterpart, when they both exist. This difficulty has plagued the application of Bayesian approaches for many decades, however, during the 1990's the Bayesian approach gained much ground with the advent of Markov Chain Monte Carlo (MCMC) methods. Chapter 3 gives a short review of the basic MCMC algorithms and gives some references to more theoretical results.

Chapter 4 gives the details on a test for stationarity of a time series that we present as an approach to determine if the MCMC outputs have, in fact, reached their stationary distributions which is crucial for valid inference to be drawn.

Chapter 5 gives a Bayesian approach to the linear mixing models that are used as the building blocks of a slightly more complicated model which we apply to the diet composition problem. Of note is the effect of collinearity among the sources in the mixing models and how the Bayesian approach deals with this problem.

Chapter 6 gives the Bayesian approach to the diet reconstruction problem. We adapt the linear mixing model developed in Chapter 5 to account for multiple populations, predator biosynthesis, and prey fat content. We illustrate the methods with several synthetic data sets and one captive study on birds and 23 free ranging Harbour seals that were equipped with National Geographic critter cams.

Chapter 7 presents some concluding remarks and some directions for further work.

A brief introduction to the Bayesian paradigm and Directed Acyclic Graphs (DAGs) are given in Appendix A . The MCMC algorithms used for the constant and individual diet models are given in Appendix B. Appendix C gives a list of the statistical distributions used in the thesis.

CHAPTER 2

INTRODUCTION TO THE THEORY OF COMPOSITIONAL DATA

2.1 Introduction

In many fields of scientific study, the data collected are known to satisfy certain constraints. For example, the mass of an object must be positive or the temperature of an object must be greater than -273° Celsius. Compositional data, by their nature being parts of some whole are subject to a positivity constraint and a constant sum constraint. More formally, let $\mathbf{z}$ be a D dimensional vector subject to $z_i > 0$ and $\sum_{i=1}^{D} z_i = c$ (where c is some arbitrary constant), then we say that $\mathbf{z}$ is a composition. Without loss of generality we assume that $c = 1$.

Data of this sort, arise quite naturally in a number of applied settings, for example, in geochemistry $\mathbf{z}$ might represent the chemical composition of rocks; in economics, $\mathbf{z}$ might represent the proportional household expenditure on a number of different commodity types. Fatty acid profiles are compositional in nature as each profile represents the proportional contribution of each fatty acid. Much of the early work on compositional data analysis was set out by John Aitchison (Aitchison, 1982, 2003).

The sample space for compositional vectors is the d–dimensional unit simplex given by

$$\mathcal{S}^d = \left\{ (z_1, \ldots, z_D) : z_i > 0 \; (i = 1, \ldots, D), \sum_{i=1}^{D} z_i = 1 \right\},$$

where $d = D - 1$. Thus, the unit sum constraint forces the compositional vector to lie in smaller dimensional surface. Aitchison (2001) notes that many analyzes are still carried out ignoring the distinction between $\mathcal{R}^d$ and $\mathcal{S}^d$, which leads to many spurious results, particularly correlational results. Aitchison makes it clear in much of his work that compositional data should be interpreted with caution and stresses that compositions give only relative magnitudes of the components.

2.1.1 Ternary Diagrams

Consider a compositional vector, $\mathbf{z} = (z_1, z_2, z_3)$, how would one depict this? Ternary diagrams (Aitchison, 2003) allow one to graphically display 3–part compositions as a equilateral triangle, by virtue of the unit sum constraint. Consider the following three points: $z_1 = (1/3, 1/3, 1/3)$, $z_2 = (0.1, 0.1, 0.8)$ and $z_3 = (0.01, 0.01, 0.99)$, which are depicted in figure 2.1. The labeling of the ternary diagram is a bit non–traditional, typically the vertices of the equilateral triangle are label, however, we have labeled the edges. The

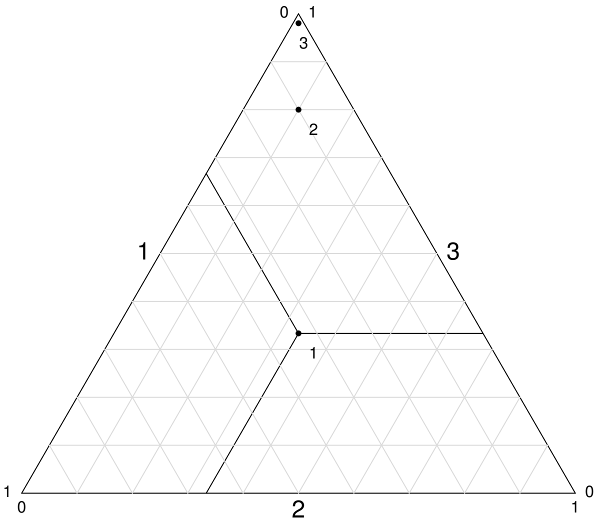

Figure 2.1: Ternary diagram with 3 points with increasing values of the third component, specifically the points are $z_1 = (1/3, 1/3, 1/3)$, $z_2 = (0.1, 0.1, 0.8)$, $z_3 = (0.01, 0.01, 0.99)$. The point $z_1 = (1/3, 1/3, 1/3)$ is the "center" of the simplex also known as the compositional zero.

traditional labeling scheme would correspond to shifting the labels clockwise to the nearest vertex. Note that the zero for the first coordinate corresponds to the edge labeled 3, and the one is the vertex given by edges 1 and 2.

With the help of the grid, which divides each axes in to 10 equally spaced divisions, we can read the coordinates of any point in the ternary diagram, for example the point labeled 2, as follows: locate the fraction of the distance from the base of the opposite the appropriate vertex and locate the fraction of the distance between the base and the vertex. This operation, only needs to be done for two of the three dimensions as the third is determined by the unit sum constraint.

Ternary diagrams very quickly lose their ability to help us depict compositions of higher dimensions as is the case in more general Euclidean spaces. However, they still have a role to play when depicting sub–compositions.

2.2 Compositional Algebra

This section gives some of the basic compositional algebra results, for more details see Aitchison (2003); Billheimer et al. (2001). The closure operator, for a positive vector $\mathbf{u} = (u_1, \ldots, u_D)$, such that $u_j > 0$ for all j is defined as follows:

$$\mathbf{z} = \mathcal{C}(\mathbf{u}) = \left(\frac{u_1}{\sum_{j=1}^{D} u_j}, \ldots, \frac{u_D}{\sum_{j=1}^{D} u_j} \right).$$

This transforms a vector $\mathbf{u}$ defined on $\mathcal{R}_+^D$ to a vector $\mathbf{z}$ defined on the simplex $\mathcal{S}^d$. For example, assume that u_j represents the amount of prey of type j in kilograms that a predator consumed in a given year, then $\mathbf{z} = \mathcal{C}(\mathbf{u})$ would then represent the diet composition of said predator for that year.

The perturbation operator, denoted by $\oplus$, for composition $\mathbf{u}$ perturbed by a composition $\mathbf{v}$ is defined as:

$$\mathbf{z} = \mathbf{u} \oplus \mathbf{v} = \mathcal{C}(\mathbf{u} \cdot \mathbf{v})$$

where $\mathbf{z}$, $\mathbf{u}$ and $\mathbf{z} \in \mathcal{S}^d$ and $\mathbf{u} \cdot \mathbf{v} \equiv (u_1 v_1, \ldots, u_D v_D)$, is the usual elementwise multiplication of two positive vectors. Note that the perturbation vector $\mathbf{v}$ is not strictly required to lie in the simplex, however, we require $v_i > 0$ for all $i = 1, \ldots, D$. Perturbation on the simplex is analogous to translation (addition/subtraction) on Euclidean spaces. It is quite easy to show that the perturbation operator obeys the usual laws of addition: commutative ($\mathbf{u} \oplus \mathbf{v} = \mathbf{v} \oplus \mathbf{u}$) and associative ($(\mathbf{u} \oplus \mathbf{v}) \oplus \mathbf{w} = \mathbf{u} \oplus (\mathbf{v} \oplus \mathbf{w})$). The inverse perturbation operator is defined as follows:

$$\mathbf{u} \ominus \mathbf{v} \equiv \mathbf{u} \oplus \mathbf{v}^{-1},$$

where $\mathbf{v}^{-1} = (1/v_1, \ldots, 1/v_D)$ and the zero perturbation element

$$\mathcal{J}_D = (\mathbf{j}_D)/D$$

where $\mathbf{j}_D$ is a D–dimensional column vector of ones.

Billheimer et al. (2001) also defines the scalar multiplication of a composition $\mathbf{u}$ by a scalar b, through the power transformation, given by

$$\mathbf{z} = \mathcal{C}(\mathbf{u}^b) = \mathcal{C}(u_1^b, \ldots, u_D^b)$$

which plays the role of scalar multiplication in the simplex. The power transformation is useful in defining linear models on the simplex introduced later. Also, we can show that the power transformation and the perturbation operator form a complete inner product space.

Consider two compositional vectors $\mathbf{u}, \mathbf{v}$ residing in $\mathcal{S}^d$ and define the following matrix

$$\mathbf{N}_d = I_d + J_d$$

where I_d is a $d \times d$ dimensional identity matrix and $J_d = \mathbf{j}_d\mathbf{j}_d^{'}$ and $d = D - 1$. The inverse is given by

$$\mathbf{N}_d^{-1} = I_d - \frac{1}{D}\mathbf{J}_d.$$

Billheimer et al. (2001) define the inner product as follows:

$$\langle \mathbf{u}, \mathbf{v} \rangle = \phi(\mathbf{u})^{'}\mathbf{N}^{-1}\phi(\mathbf{v})$$

where ϕ is the additive logratio transformation defined in the next section . The norm is given by

$$\|\mathbf{u}\| = \langle \mathbf{u}, \mathbf{u} \rangle^{1/2}.$$

Billheimer et al. (2001) notes that the inclusion of $\mathbf{N}^{-1}$ ensures that the inner product and norm are invariant to permutations of the components of $\mathbf{u}$ and $\mathbf{v}$. It is also noted that this function satisfies the requirements of compositional metric given by Aitchison (1992). The distance, ∇, between two compositional vectors $\mathbf{u}$ and $\mathbf{v}$ is given by

$$\nabla(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} \oplus \mathbf{v}\|.$$

For illustration consider the three points plotted in the ternary diagram of figure 2.1. The norms for each of the compositions $\mathbf{z}_1$, $\mathbf{z}_2$, $\mathbf{z}_3$ are 0, 1.698 and 3.752 respectively. Since $\mathbf{z}_1$ is the center or zero of the simplex the distance between it and the other points is identical to their norms and the distance between points 2 and 3 is 3.752. Thus, we can see that even though the euclidean distance between points $\mathbf{z}_1$ and $\mathbf{z}_2$ is greater than euclidean distance between $\mathbf{z}_2$ and $\mathbf{z}_3$, in the simplex, the opposite is true.

2.2.1 Transformations

Aitchison (2003) defines several transformations from $\mathcal{R}^d$ to $\mathcal{S}^d$ which we discuss in this section. The central idea behind these transformations is to provide a one–to–one and onto (bijective) mapping between the simplex $\mathcal{S}^d$ and the real space $\mathcal{R}^d$. Consider the additive logistic transform denoted by, $\phi^{-1}(.)$

$$\mathbf{z}_D = \phi^{-1}(\mathbf{y}) = \begin{cases} \frac{\exp y_i}{1+\sum_{j=1}^{d} \exp y_i} & i = 1, \dots, d \\ \frac{1}{1+\sum_{j=1}^{d} \exp y_i} & i = D \end{cases}$$

The inverse of this transformation is the additive logratio (alr) transformation, which we denote by $\phi(.)$

$$\mathbf{y} = \phi(\mathbf{z}) = \log \frac{\mathbf{z}_{-D}}{z_D}$$

where $\mathbf{y} \in \mathcal{R}^d$ and $\mathbf{z} \in \mathcal{S}^d$ and $\mathbf{z}_{-i}$ is vector $\mathbf{z}$ with the ith component deleted. We can also write this transformation in matrix notation as follows

$$\mathbf{z} = F \log \mathbf{z}$$

where F is a $d \times D$ dimensional matrix given by the following

$$F = [I_d : -\mathbf{j}_d].$$

The multiplicative logistic transformation, $\mathbf{z} = \phi_M^{-1}(\mathbf{y})$ with the inverse transform $\mathbf{y} = \phi_M(z)$ with elements given by

$$z_i = \begin{cases} \exp y_i / \left(\prod_{j=1}^{i} (1 + \exp y_i) \right) & (i = 1, \dots, d) \\ 1 / \left(\prod_{j=1}^{i} (1 + \exp y_i) \right) & (i = D). \end{cases}$$

and

$$y_i = \log \left(\frac{z_i}{1 - \sum_{j=1}^{i} z_j} \right).$$

In matrix notation we have

$$\mathbf{x} = G \log \mathbf{z}$$

where

$$G = I_D - D^{-1} J_D.$$

Aitchison (2003) also introduces the centered logratio given by

$$\phi_C(\mathbf{z}) = \log\left(\frac{\mathbf{z}}{g(\mathbf{z})}\right)$$

where $g(\mathbf{z}) = (z_1 z_2 \dots z_D)^{1/D}$, the usual geometric mean. This mapping is from simplex $\mathcal{S}^d$ to $\mathcal{R}^D$ and isn't bijective as the other transformations are. The reason for introducing this transformation is the apparent lack of symmetry in the additive and multiplication logratio transformations. Egozcue et al. (2003) introduced an isometric logratio transformation to deal with the non-symmetry issue. This transformation was not pursued further in the sequel.

The following properties can be established for the additive logratio transformation, where $\mathbf{u}, \mathbf{v} \in \mathcal{S}^d$, $a \in \mathcal{R}$

$$\begin{aligned} \phi(\mathbf{u} \oplus \mathbf{v}) &= \phi(\mathbf{u}) + \phi(\mathbf{v}) \\ \phi(\mathbf{u}^a) &= a\phi(\mathbf{u}) \\ \phi(\mathbf{j}_d/d) &= \mathbf{0}_d \end{aligned}$$

and $\mathbf{0}_d$ is a d vector of zeros.

2.2.2 Vector and Hilbert Spaces

Billheimer et al. (2001) prove the following results which demonstrate that the operations on the simplex defined in the previous section form a Hilbert space.

Theorem 2.1. *$\mathcal{S}^d$ is a vector space with addition defined by the perturbation operator and scalar multiplication by the scalar a.*

Theorem 2.2. *Let $\mathbf{u}$ and $\mathbf{z}$ be elements of the d dimensional simplex $\mathcal{S}^d$. Then, $\langle \mathbf{u}, \mathbf{v} \rangle = \phi(\mathbf{u})' \mathcal{N}^{-1} \phi(\mathbf{v})$ is an inner product.*

Theorem 2.3. *$\mathcal{S}^d$ is a Hilbert space (a complete, inner product space).*

The proofs are given in the appendix of Billheimer et al. (2001). Also note that Pawlowsky-Glahn and Egozcue (2002) showed similar results.

2.2.3 Parametrizations of the Covariance Matrix

Aitchison (2003) gives three equivalent covariance relationships for the simplex, relying on logratios of components($\log{(z_i/z_j)}$).

Definition 2.1. *The covariance structure of a D–part composition* $\mathbf{z}$ *is the set of all*

$$\sigma_{ij.kl} = \mathit{cov}\{\log(z_i/z_k), \log(z_j/z_l)\}$$

as i,j,k,l run through the values $1, \ldots, D$*, generating* D^4 *covariances.*

Clearly, these values are dependent on each other and Aitchison (2003) gives several dependencies among the covariances, such as

$$\sigma_{ij.il} = \sigma_{ij.kj} = \sigma_{ij.ij} = 0.$$

Definition 2.2. *For a D–part composition* $\mathbf{z}$ *the* $D \times D$ *variation matrix*

$$T = [\tau_{ij}] = [\mathit{var}\{\log(z_i/z_j)\} : i, j = 1, \ldots, D] .$$

The elements of the covariance structure $\sigma_{ij.kl}$ *can be expressed in terms of the* τ_{ij} *as follows:*

$$\sigma_{ij.kl} = \frac{1}{2}(\tau_{il} + \tau_{jk} - \tau_{ij} - \tau_{kl})$$

The logratio variances are trivially zero on the diagonal and also have a symmetry property ($\tau_{ij} = \tau_{ji}$), which leads to a reduction in the effective number of parameters required from D^2 to $\frac{1}{2}dD$.

Definition 2.3. *For a D–part composition* $\mathbf{z}$ *the* $d \times d$ *logratio covariance matrix*

$$\Sigma = [\sigma_{ij}] = [\mathit{cov}\{\log(z_i/z_D), \log(z_j/z_D)\}; i, j = 1, \ldots, d].$$

The elements of the covariance structure $\sigma_{ij.kl}$ *can be expressed in terms of the* σ_{ij} *as follows:*

$$\sigma_{ij.kl} = (\sigma_{ij} + \sigma_{kl} - \sigma_{il} - \sigma_{jk})$$

This parametrization requires $\frac{1}{2}dD$ parameters rather than d^2 due to the symmetry considerations ($\sigma_{ij} = \sigma_{ji}$).

Definition 2.4. *For a D–part composition* $\mathbf{z}$ *the* $D \times D$ *centered logratio covariance matrix*

$$\Gamma = [\gamma_{ij}] = [cov\{\log(z_i/g(\mathbf{z})), \log(z_j/g(\mathbf{z}))\}; i, j = 1, \ldots, D]$$

where $g(\mathbf{z}) = (z_1 z_2 \ldots z_D)^{1/D}$*. The elements of the covariance structure* $\sigma_{ij.kl}$ *can be expressed in terms of the* γ_{ij} *as follows:*

$$\sigma_{ij.kl} = (\gamma_{ij} + \gamma_{kl} - \gamma_{il} - \gamma_{jk})$$

This parametrization requires $\frac{1}{2}D(D+1)$ parameters rather than D^2 due to the symmetry considerations ($\gamma_{ij} = \gamma_{ji}$). However, this matrix is singular, and requires D more parameters than the logratio covariance matrix.

Aitchison (2003), gives expressions for relating the various matrix formulations, for example

$$T \to \Sigma : \Sigma = -\frac{1}{2}\mathbf{F}T\mathbf{F}'.$$

and

$$\Gamma \to \Sigma : \Sigma = \mathbf{F}\Gamma\mathbf{F}'.$$

The information contained in the different covariance matrices are equivalent, however, the logratio covariance matrix has the properties that are most similar to the usual covariance matrices. Additionally the logratio covariance matrix will be used in the parametrization of the logistic normal distribution.

2.3 Permutational Invariance

As we noticed in the previous section, the logratio variance specifies a component (ie the last) and uses it as the divisor in all subsequent definitions of the variance. In this section, we determine if the determinant of the variance covariance matrix and the quadratic form (to be used in the definition of the Logistic Normal Distribution) are invariant to the order of the components of the composition. Consider a permutation matrix P , a square matrix that has exactly one 1 in each row and column and 0's elsewhere. For example, consider a vector $\mathbf{x} = (1, 2, 3)$ and the following permutation matrix

$$P = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

thus $\mathbf{x}_p = P\mathbf{x} = (3, 1, 2)$ Permutation matrices have some very simple properties

- permutation matrices are orthogonal $PP^{'} = I = P^{'}P$
- the inverse $P^{-1} = P^{'}$
- $P = P_1 \ldots P_m$ where each of theP_i consist of elementary permutations, that is, only two off-diagonal elements are 1's.
- $\det P = (-1)^m$ where m is the number of elementary permutations.

Recall the additive logratio transformation for a composition $\mathbf{z}$

$$\mathbf{y} = \log(\mathbf{z}_{-D}/z_D) = F \log \mathbf{z},$$

and the centered logratio transformation

$$\mathbf{x} = \log(\mathbf{z}/g(\mathbf{z})) = G \log \mathbf{z},$$

where the matrices F and G are defined in section (2.2.1).

Aitchison (2003) shows the following relationship between the additive and multiplicative logratio transformations:

$$\begin{aligned} \mathbf{y} &= F\mathbf{x} \\ \mathbf{x} &= F^{'}(FF^{'})^{-1}y = F^{'}\mathcal{P}\mathbf{y} \end{aligned}$$

where $N = FF' = I_d + \mathbf{j}_d\mathbf{j}_d^{'} = I_d + J_d$.

First consider the centered logratio transformation $\mathbf{x} = G\mathbf{z}$ and consider a general permutation, P, applied to this vector.

$$\begin{aligned} \mathbf{z}_P &= P\mathbf{z} \\ \mathbf{x}_P &= P\mathbf{x} \end{aligned}$$

Now

$$\begin{aligned} \mathbf{x}_P &= P\mathbf{x} \\ &= PG\log\mathbf{z} \\ &= P[I_D - J_D/D]\log\mathbf{z} \\ &= [P - J_D/D]\log\mathbf{z} \\ &= [I_D - J_D/D]P\log\mathbf{z} \\ &= [I_D - J_D/D]\log P\mathbf{z} \\ &= G\log\mathbf{z}_P \end{aligned}$$

Now consider the additive logratio transformation. As before we let $\mathbf{y}$ be the non–permuted vector and $\mathbf{y}_P$ represent the permuted vector.Then

$$\begin{aligned} \mathbf{y} &= F\mathbf{x}_P \\ &= FP\mathbf{x} \\ &= FPF^{'}\mathcal{P}\mathbf{y} \end{aligned}$$

Aitchison (2003) proves the following proposition

Proposition 2.1. *Define $Q_P = FPF^{'}\mathcal{P}$. If P is any $D \times D$ permutation matrix and a $\mathbf{z}_P = P\mathbf{z}$ then $\mathbf{y}_P$ and Σ_P are given by*

$$\begin{aligned} \mathbf{y} &= Q_P\mathbf{y} \\ \Sigma_P &= Q_P\Sigma Q_P^{'} \end{aligned}$$

The following proposition due to Aitchison (2003) is key to our results in section (2.5) for the logistic normal distribution.

Proposition 2.2. *For any permutation P of the parts of a D–part composition $\mathbf{z}$ the following relationships hold*

1. $|\Sigma_P| = |\Sigma|$
2. $\mathbf{y}'_P \Sigma_P^{-1} \mathbf{y}_P = \mathbf{y}' \Sigma^{-1} \mathbf{y}$

That is, what we choose as a divisor does not affect the quadratic form of the logistic normal distribution that will be introduced in a subsequent section.

2.4 Sub–compositional Invariance

Given a D–dimensional compositional vector, $\mathbf{z}$, it is often of interest to examine a subset of the components. Following Aitchison (2003) we define the following.

Definition 2.5. *If S is any subset of the parts,* $1, \ldots, D$ *of a D–part composition* $\mathbf{z}$*, and* $\mathbf{z}_S$ *is the sub–vector formed from the corresponding components of* $\mathbf{z}$*, then* $\mathcal{C}(\mathbf{z}_S)$ *is termed the sub–composition of the parts S.*

The following definition gives a matrix definition of a sub–composition.

Definition 2.6. *A selection matrix* $\mathbf{S}$ *is any matrix or order* $C \times D$*, with* C *elements equal to 1, one each row and at most 1 in each column, with the remaining elements 0.*

Therefore, the selection matrix S gives rise to a new composition as follows

$$\mathbf{z}_S = \mathbf{S}\mathbf{z}$$

Application of the these two definitions leads to a mapping from the d-dimensional simplex to the $c = C - 1$ dimensional simplex.

$$\mathcal{C}\mathbf{S} : \mathcal{S}^d \to \mathcal{S}^c.$$

The ratio of any two components of a sub–composition is the same as the ratio of the corresponding two components in the full composition. If $\mathbf{s} = \mathcal{C}(\mathbf{z}_S)$ then

$$s_i/s_j = z_i/z_j \ \ (i, j \in S).$$

The above result is known as sub–compositional invariance.

2.5 Compositional Distributions

2.5.1 Dirichlet

Given the nature of the simplex, the most obvious choice of a distribution would be the Dirichlet, $\mathcal{D}^d(\boldsymbol{\delta})$, which has density given by

$$\frac{\Gamma(\delta_1 + \ldots + \delta_D)}{\Gamma(\delta_1)\ldots\Gamma(\delta_D)} \prod_{i=1}^{D} z_i^{\delta_i - 1} \ (\mathbf{z} \in \mathcal{S}^d) \tag{2.1}$$

where $\boldsymbol{\delta} \in \mathcal{R}_+^D$.

Aitchison (2003) states the following properties of the Dirichlet distribution.

Proposition 2.3 (Aitchison)**.** *If* $\mathbf{z}$ *has a* $\mathcal{D}^d(\boldsymbol{\delta})$ *distribution and* $\delta_+ = \sum_{i=1}^{D} \delta_i$ *then*

(a) $E(z_i) = \delta_i/\delta_+$,

(b) $var(z_i) = \delta_i(\delta_+ - \delta_i)/\left\{\delta_+^2(\delta_+ + 1)\right\}$,

(c) $cov(z_i, z_j) = -\delta_i\delta_j/\left\{\delta_+^2(\delta_+ + 1)\right\} \ (i \neq j)$,

(d) $corr(z_i, z_j) = -\sqrt{\delta_i\delta_j}/\sqrt{(\delta_+ - \delta_i)(\delta_+ - \delta_j)} \ (i \neq j)$,

Thus, all covariances between components in the Dirichlet class are negative and hence are not appropriate for compositional data where there are positive associations.

Definition 2.7. *A variable* w *is said to have a Gamma*(δ, β) *distribution if its density function is given by*

$$\frac{\beta^\delta w^{\delta-1} \exp -\beta w}{\Gamma(\delta)} \ \ w > 0$$

where $\delta > 0$ *and* $\beta > 0$

Proposition 2.4 (Aitchison)**.** *Every Dirichlet composition may be visualized as the composition of independent, equally scaled gamma-distributed components. That is, if* $w_i (i = 1, \ldots, D)$ *are independently distributed as Gamma*(δ_i, β) *then* $\mathbf{z} = \mathcal{C}(\mathbf{w})$ *has a* $\mathcal{D}^d(\boldsymbol{\delta})$ *distribution.*

Definition 2.8. *Let*

$$0 = a_0 < a_1 < \ldots < a_c < a_C = D.$$

The partition based on the separation of the D*–part composition* $\mathbf{z}$ *into* C *subsets by the divisions*

$$(z_{a_0+1}, \ldots, z_{a_1} | z_{a_1+1}, \ldots, z_{a_2} | \ldots | z_{a_c+1}, \ldots, z_{a_C})$$

is termed a general partition of $\mathbf{z}$*. If*

$$t_i = z_{a_{i-1}+1} + \ldots + z_{a_i} \; (i = 1, \ldots, C)$$
$$s_i = \mathcal{C}(z_{a_i+1}, \ldots, z_{a_i}) \; (i = 1, \ldots, C)$$

then we denote the general partition as

$$\mathcal{P}(\mathbf{z}) = (\mathbf{t}; \mathbf{s}_1, \ldots, \mathbf{s}_C)$$

Proposition 2.5. *Let* $\mathcal{P}(\mathbf{z}) = (\mathbf{t}; \mathbf{s}_1, \ldots, \mathbf{s}_C)$ *be a general partition, as specified in the previous definition, of order* c *of a* $\mathcal{D}^d(\boldsymbol{\delta})$ *composition. Then* $\mathbf{t}$*,* $\mathbf{s}_1, \ldots, \mathbf{s}_C$ *are independent and distributed as* $\mathcal{D}^c(\boldsymbol{\gamma})$*,* $\mathcal{D}^{d_1}(\boldsymbol{\beta}_1), \ldots, \mathcal{D}^{d_C}(\boldsymbol{\beta}_C)$*, respectively, where* $\mathcal{P}(\boldsymbol{\delta}) = (\boldsymbol{\delta}; \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_C)$

These propositions imply very strong independence for the Dirichlet class and hence it is not very practical for modeling data with more structure, as is apparent in actual compositional data (Aitchison, 1982, 2003)

2.5.2 Logistic Normal

The multivariate normal distribution has a very rich history and is the most commonly used multivariate distribution (Anderson, 1984; Morrison, 1976; Srivastava and Khatri, 1979). The multivariate normal density is given by

$$p(\mathbf{y}|\boldsymbol{\mu}, \Sigma) = \mathcal{N}(\boldsymbol{\mu}, \Sigma) = \left(\frac{1}{2\pi}\right)^{p/2} |\Sigma|^{-1/2} exp\left\{-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{y} - \boldsymbol{\mu})\right\}$$

where $\mathbf{y}$ is a d–dimensional vector, $E(\mathbf{y}|\boldsymbol{\mu},\Sigma)=\boldsymbol{\mu}$ and var$(\mathbf{y}|\boldsymbol{\mu},\Sigma)=\Sigma$.

Consider the the additive logistic transformation ϕ^{-1} which maps a point in $\mathbf{y}\in\mathcal{R}^d$ into a point $\mathbf{z}\in\mathcal{S}^d$. The Jacobian of this transformation is given by $1/\prod_{i=1}^{D} z_i$. Thus if we assume that $\mathbf{y}\sim\mathcal{N}(\boldsymbol{\mu},\Sigma)$ then the transformation $\phi^{-1}(\mathbf{y})$ induces the following density on the vector $\mathbf{z}\in\mathcal{S}^d$

$$p(\mathbf{z}|\boldsymbol{\mu},\Sigma)=\left(\frac{1}{2\pi}\right)^{d/2}|\Sigma|^{-1/2}\left(\frac{1}{\prod_{i=1}^{D} z_i}\right)exp\left\{-\frac{1}{2}(\phi(\mathbf{z})-\boldsymbol{\mu})'\Sigma^{-1}(\phi(\mathbf{z})-\boldsymbol{\mu})\right\}$$

which we call the additive logistic normal, denoted by $\mathcal{L}^d(\boldsymbol{\mu},\Sigma)$. Proposition (2.2) demonstrates that the additive logistic normal is permutation invariant, that is, the density function is invariant to reorderings of the elements of the composition. Billheimer et al. (2001) gives the following interpretation of the location parameter $\boldsymbol{\mu}$. Consider applying the additive logistic transformation $\phi^{-1}(.)$, specifically

$$\boldsymbol{\xi}=\phi^{-1}(\boldsymbol{\mu}),\quad \text{where } \boldsymbol{\xi}\in\mathcal{S}^d$$

Billheimer et al. (2001) states that the interpretation of $\boldsymbol{\xi}$ is more direct on the simplex than for $\boldsymbol{\mu}$ on the multivariate logit scale. However, some of the statistical properties are lost. They state that $\boldsymbol{\mu}$ is the mean and mode of the multivariate normal logit, however, the $\phi^{-1}(.)$ does not preserve these properties. Since $\phi^{-1}(.)$ is monotone in each of the d components of $\boldsymbol{\mu}$, the transformation is order preserving. Thus, they interpret $\boldsymbol{\xi}=\phi^{-1}(\boldsymbol{\mu})$ as a component–wise multivariate median for the Logistic Normal distribution in $\mathcal{S}^d$.

There is a close relationship between the logistic normal class and the additive logistic normal class. We have the following definition of a multivariate lognormal distribution

Definition 2.9. *A D dimensional vector $\mathbf{w}\in\mathcal{R}_+^D$ has a multivariate lognormal distribution with mean $\boldsymbol{\xi}$ and covariance matrix Ω denoted by $\boldsymbol{\Lambda}^D(\boldsymbol{\xi},\Omega)$ if $\log\mathbf{w}$ has a $\mathcal{MN}^D(\boldsymbol{\xi},\Omega)$.*

The connection between the additive logistic normal distribution and multivariate lognormal is established by the following proposition proved in Aitchison (2003).

Proposition 2.6. *If a vector $\mathbf{w}$ has a $\Lambda^D(\boldsymbol{\xi},\Omega)$ distribution then its composition $\mathbf{x}=\mathcal{C}(\mathbf{w})$ has a $\mathcal{L}^d(\boldsymbol{\mu},\Sigma)$ distribution, where*

$$\boldsymbol{\mu}=F\boldsymbol{\xi},\quad \Sigma=F\Omega F^{'}$$

We will have occasion to use the following special case of this general result. Take $\boldsymbol{\xi} = c$ and $\Omega = \sigma^2 I$ which corresponds to uncorrelated lognormal distributions with the same mean, c, and the same variance, σ^2, this gives a logistic normal distribution with mean $\boldsymbol{\mu} = \mathbf{0}$ and covariance matrix $\Sigma = \sigma^2 FF' = \sigma^2\mathbf{N}$. We use this result in the Metropolis–Hastings algorithms developed in subsequent chapters.

2.5.3 Multi–Modality

There is an interesting relationship between the covariance matrix of the logistic normal distribution and the number of modes of the distribution. This has implications for assigning prior distributions as we shall discuss later. The logistic normal density is given by

$$\mathcal{L}^d(\mathbf{z}|\boldsymbol{\mu}, \Sigma) = (2\pi)^{-d/2} |\Sigma|^{-1/2} \left(\frac{1}{\prod_{i=1}^{D} z_i}\right) \exp\left\{-\frac{1}{2}(\phi(\mathbf{z}) - \boldsymbol{\mu})'\Sigma^{-1}(\phi(\mathbf{z}) - \boldsymbol{\mu})\right\}.$$

We consider the special case where the logistic normal distribution arises from the closure of independent log–normal distributions with zero means and constant variances σ^2 (see proposition 2.6). This implies the mean is $\boldsymbol{\mu} = \mathbf{0}_d$ and the covariance matrix is

$$\Sigma = \sigma^2 \left(I_d + \mathbf{J}_d\right)$$

where I_d is the $d \times d$ dimensional identity matrix and $\mathbf{J}_d$ is a matrix of ones. The determinant of this matrix is $\sigma^{2d}D$. The precision matrix or Σ^{-1} also has a simple form given by

$$\Sigma^{-1} = \frac{1}{\sigma^2}\left(I_d - \frac{1}{D}\mathbf{J}_d\right).$$

The exponent in the numerator is given by

$$\begin{aligned}
-\frac{1}{2}\phi(\mathbf{z})'\Sigma\phi(\mathbf{z}) &= -\frac{1}{2\sigma^2}\phi(\mathbf{z})'\left(I_d - \frac{1}{D}\mathbf{j}\mathbf{j}'\right)\phi(\mathbf{z}) \\
&= -\frac{1}{2\sigma^2}\left[\sum_{i=1}^{d}\phi(\mathbf{z})_i^2 - \frac{1}{D}\sum_{i=1}^{d}\phi(\mathbf{z})_i^2 - \frac{2}{D}\sum_{\substack{1\le i,j\le d,\\ i<j}}\phi(\mathbf{z})_i\phi(\mathbf{z})_j\right] \\
&= -\frac{1}{2D\sigma^2}\left[d\sum_{i=1}^{d}\phi(\mathbf{z})_i^2 - 2\sum_{\substack{1\le i,j\le d,\\ i<j}}\phi(\mathbf{z})_i\phi(\mathbf{z})_j\right].
\end{aligned}$$

The logistic normal under consideration can be written as follows

$$f(\mathbf{z}) = (2\pi)^{-d/2}(\sigma^{2d}D)^{-1/2}$$
$$\exp\left\{-\frac{1}{2D\sigma^2}\left[d\sum_{i=1}^{d}\phi(\mathbf{z})_i^2 - 2\sum_{\substack{1\le i,j\le d,\\ i<j}}\phi(\mathbf{z})_i\phi(\mathbf{z})_j\right] - \sum_{i=1}^{D}\log z_i\right\}$$

The modes of the distribution are located at the critical points of the function, $f(\mathbf{z})$. Note the following partial derivatives

$$\frac{\partial \log \phi(\mathbf{z})_i}{\partial z_i} = \frac{z_D + z_i}{z_1 z_D},$$

$$\frac{\partial \log \phi(\mathbf{z})_i}{\partial z_j} = \frac{1}{z_D},$$

and

$$\frac{\partial - \sum_{i=1}^{D}\log(z_i)}{\partial z_i} = \frac{z_i - 1 + \sum_{i=1}^{d} z_i}{z_i z_D}$$

Now taking partial derivatives of $\log f(\mathbf{z})$ with respect to each of the $z_i, i = 1, \ldots, d$ and substituting $z_D = 1 - \sum_{j=1}^{d} z_i$ gives:

$$\frac{\partial \log f(\mathbf{z})}{\partial z_i} = \frac{d\phi(\mathbf{z})_i - d\sum_{j\neq i}(z_i\phi(\mathbf{z})_j - z_j\phi(\mathbf{z})_i) + (d-1)z_i\phi(\mathbf{z})_i}{D\sigma^2 z_i z_D}$$
$$+ \frac{\sum_{j\neq i}\sum_{k\neq i} z_j\phi(\mathbf{z})_k - \sum_{j\neq i}(\phi(\mathbf{z})_j + z_i\phi(\mathbf{z})_j) - z_D + z_i}{D\sigma^2 z_i z_D}$$

To find the critical points we solve the system of equations $\frac{\partial \log f(\mathbf{z})}{\partial z_i} = 0$ for z_i. This system has no closed form solutions with the exception of $\mathbf{z} = \mathbf{j_D}/D$. However, numerical methods can be employed to find approximate solutions of any other modes to arbitrary accuracy. This is very similar to Aitchison (2003)'s statement that closed form solutions for the moments of the logistic normal distribution do not exist.

Figures 2.2 and 2.3 give graphical representations of the logistic normal distribution for $D = 2$ and $D = 3$ respectively for $\sigma^2 = (0.5, 1.0, 1.5, 2.0)$. For $\sigma^2 = 0.5$ both densities are uni–modal with a peak at the compositional zero. When $\sigma^2 = 1.0$ the two-dimensional logistic normal has a less well defined peak and the three dimensional logistic normal has 3 modes located away from the compositional zero. For the larger variances both distributions

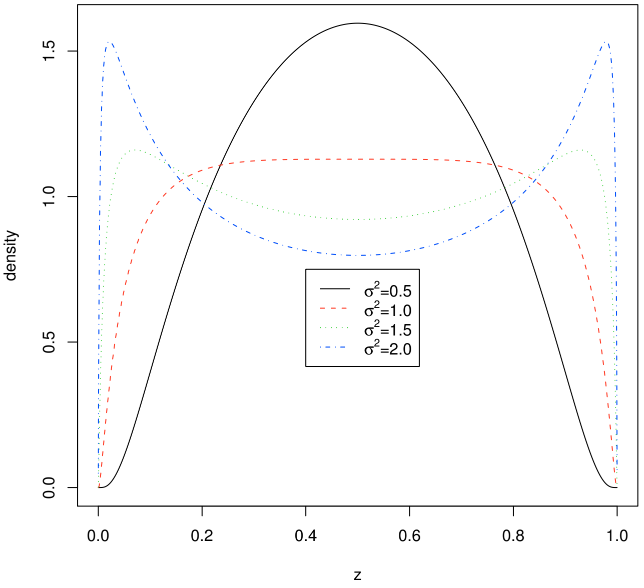

Figure 2.2: Probability density plots of the 2 dimensional logistic normal distribution for $\sigma^2 = 0.5, 1.0, 1.5, 2.0$ and $\mu = 0$.

are clearly multi–modal with the modes moving further away from the compositional zero. This indicates that in order for the variance to increase beyond some threshold the logistic normal distribution has to move its mass away from zero and this results in multi–modality.

The form of the density suggests this, in that, for larger variances σ^2 the density is dominated by term $\prod_{i=1}^{D} z_i$ which appears in the denominator compared to the term

$$\exp\left\{-\frac{1}{2\sigma^2}\phi(\mathbf{z})' \left(I_d - \frac{1}{D}\mathbf{J}_d\right) \phi(\mathbf{z})\right\}.$$

Thus while the term in the denominator dominates, the density behaves like $(\prod_{i=1}^{D} z_i)^{-1}$ which increases as one of the z_i goes to one and the remainder go to zero. Or more formally, we conjecture that d terms go to ϵ and the remaining term goes to $1 - d\epsilon$ and the density behaves like $\epsilon^d(1-d\epsilon)$ for large σ^2 . The logistic normal distribution is a proper distribution, this implies that eventually the exponential term dominates and the tails of the distribution decay to zero. Unfortunately, the distribution is sufficiently complex that analytically solutions are not possible. So we resort to numerical solutions or approximations. Figures 2.2 and 2.3 and the functional form of the logistic normal distribution indicate that there is

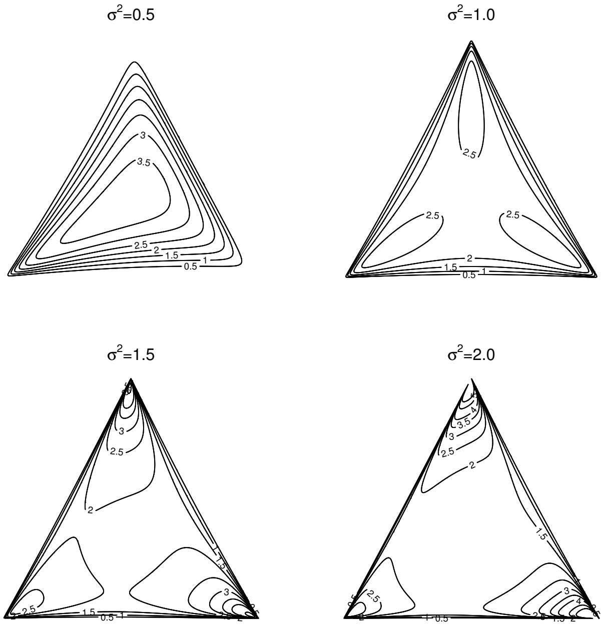

Figure 2.3: Probability density plots of the 3 dimensional logistic normal distribution for $\sigma^2 = 0.5, 1.0, 1.5, 2.0$ and $\mu = 0$. The distortion in the lower left corner of the plots is due the method of plotting.

a critical σ^2 that controls the uni–modality of the distribution. Logistic normal distributions, of the restricted class under consideration, with σ^2 smaller than this critical value are uni–modal while distributions with σ^2's larger than the critical value would be multi–modal. The goal is to show this fact numerically, to this end we solve the first order conditions $\frac{\partial \log f(\mathbf{z})}{\partial z_i} = 0$, for z_i.

Are there values of σ^2 such that the logistic normal distribution is uni–modal a critical value of σ^2 to indicate uni-modality, in fact, there is no such break point due to the continuity of the logistic normal distribution with respect to σ^2. We are using the term break–point in a heuristic sense not in a formal mathematical one.

Definition 2.10. *The critical σ^2, denoted by σ^2_c, is the largest σ^2 such that the critical point, denoted by $z^c_{\sigma^2}$, is within δ of 0, for some δ . More formally,*

$$\sigma^2_c = \arg\max_{\sigma^2} \left(\arg\max_i |\mathbf{z}^c_{\sigma^2}| < \delta \right)$$

where $\arg\max_i |\mathbf{x}|$ is the largest element of the vector $\mathbf{x}$ in absolute value.

We solve for the critical points for dimensions ranging from 2 to 32 (32 is the largest dimension of interest in the application), σ^2 varies from 0.001 to 1 by 0.001 and from 1.01 to 5 by .01. Starting values of 0.0, -0.5, -1.0, -2.0, -4.0 and -8.0 were used for of the d components. The results of applying definition 2.10 with $\delta = 0.00001$ are given in table 2.1. By our heuristic definition its apparent that as the dimension increases, uni–modality requires a much smaller σ^2 of the generating log–normal distribution for the induced logistic normal distribution to be uni–modal. In fact, it appears to decrease at an exponential rate.

For logistic normal distributions with $\sigma^2 > \sigma^2_c$, it is of interest to locate the additional modes. Figure 2.4 plots the modes, on the additive logistic scale, by dimension for 5 selected values of σ^2 (0.5,1.0,1.5,2.0,2.5). The second panel of the plot shows that the compositional center is dwarfed by the modes as they approach the boundaries of the simplex as measured by log density units. In fact, the compositional center actually becomes a local minimum of the function.

The logistic normal distribution when generated as the closure of independent log–normal distributions with common σ^2 displays some very interesting multi-modal behaviour. This is an artifact of how a distribution with limited support, in this case confined to the simplex, must change its shape to accommodate larger variation. The modes move further from the

dimension	σ^2_{max}	dimension	σ^2_{max}	dimension	σ^2_{max}
2	0.998	13	0.378	24	0.241
3	0.915	14	0.358	25	0.234
4	0.802	15	0.340	26	0.228
5	0.712	16	0.324	27	0.222
6	0.639	17	0.309	28	0.217
7	0.580	18	0.296	29	0.212
8	0.531	19	0.285	30	0.207
9	0.490	20	0.274	31	0.203
10	0.455	21	0.265	32	0.199
11	0.426	22	0.256		
12	0.400	23	0.248		

Table 2.1: Values of σ_c^2 by dimension for the logistic normal distribution.

compositional center proportional to both the variance and the dimension as evidenced by our numerical results.

2.5.4 Generalizations

Azzalini and DallaValle (1996) introduced a generalization of the multivariate normal distribution known as the multivariate skew–normal distribution. We say that the d dimensional vector $\mathbf{y}$ follows a multivariate skew–normal distribution, denoted by $\mathcal{SN}^d(\boldsymbol{\mu}, \Sigma, \boldsymbol{\beta})$, if its density function is given by

$$2(2\pi)^{-d/2}|\Sigma|^{-1/2} \exp\left[-\frac{1}{2}\left(\mathbf{y}-\boldsymbol{\mu}\right)' \Sigma^{-1}\left(\mathbf{y}-\boldsymbol{\mu}\right)\right] \Phi(\boldsymbol{\beta}'\Omega^{-1}(\mathbf{y}-\boldsymbol{\mu})),$$

where $\Phi(.)$ is the standard normal distribution function and Ω is the square root of diag(Σ) and diag(Σ) is the diagonal matrix of Σ. The vector $\boldsymbol{\beta}$ controls the shape of the distribution and determines the direction of maximum skewness. If $\boldsymbol{\beta} = \mathbf{0}$ the skew–normal reduces to the Multivariate normal distribution.

For illustration, consider the univariate skew–normal given by

$$2(2\pi)^{-1/2}\sigma^{-1/2} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y}-\mu)^2\right] \Phi(\beta\sigma^{-1}(\mathbf{y}-\mu)),$$

Figure 2.5 shows the univariate skew-normal density for selected values of β.

Azzalini and Capitanio (1999) gives a number of properties of the multivariate skew–normal distribution. For example, it is closed under linear transformation and quadratic

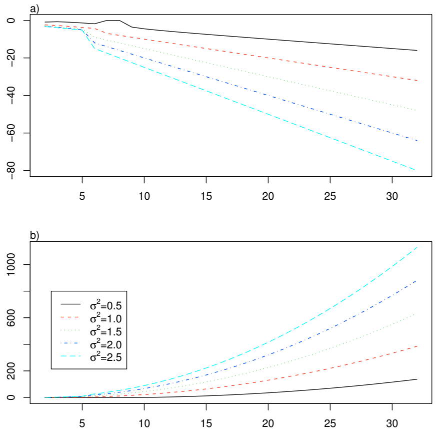

Figure 2.4: Logistic normal modes a) Location of the modes on the additive logistic scale as a function of dimension and variance. b) The difference in log density units at the mode in graph a) and the compositional center (local mode).

forms of skew–normal vectors follow χ^2 distributions.

Subsequently (Mateu-Figueras et al., 2005; Mateu-Figueras and Pawlowsky-Glahn, 2007) introduced the additive logistic skew–normal distribution(alsn), denoted by $\mathcal{SL}$. A D–part composition $\mathbf{z}$ is said to have alsn distribution when the vector $\mathbf{y} = \phi(\mathbf{z})$ has an $\mathcal{SN}^d(\boldsymbol{\mu}, \Sigma, \boldsymbol{\beta})$ distribution. We denote this distribution by $\mathcal{SL}^D(\boldsymbol{\mu}, \Sigma, \boldsymbol{\beta})$ and its density function is given by

$$2(2\pi)^{-d/2}|\Sigma|^{-1/2}\left(\prod_{i=1}^{D} z_i\right)^{-1} \exp\left[-\frac{1}{2}\left(\phi(\mathbf{z})-\boldsymbol{\mu}\right)'\Sigma^{-1}\left(\phi(\mathbf{z})-\boldsymbol{\mu}\right)\right]\Phi(\boldsymbol{\beta}'\Omega^{-1}(\phi(\mathbf{z})-\boldsymbol{\mu}))$$

Mateu-Figueras et al. (2005) gives a number of properties of this distribution. Specifically the class of alsn distributions is closed under perturbation and scalar multiplication, sub–compositions remain in the alsn class, it is permutation invariant, plus numerous others.

2.5.5 *Zero components*

Strictly speaking a compositional vector $\mathbf{z}$ must have all non-zero components, ie $z_j > 0, \forall j$. Martin-Fernández et al. (2003) discuss this problem at length, and describe two types of zeros, random zeros and essential zeros. Random zeros are ones that might be below detection limits while essential zeros would represent the complete absence of a component.

In order to overcome the random zero problem, Martin-Fernández et al. (2003), discuss three approaches: The first method described, first described by Aitchison (2003), the

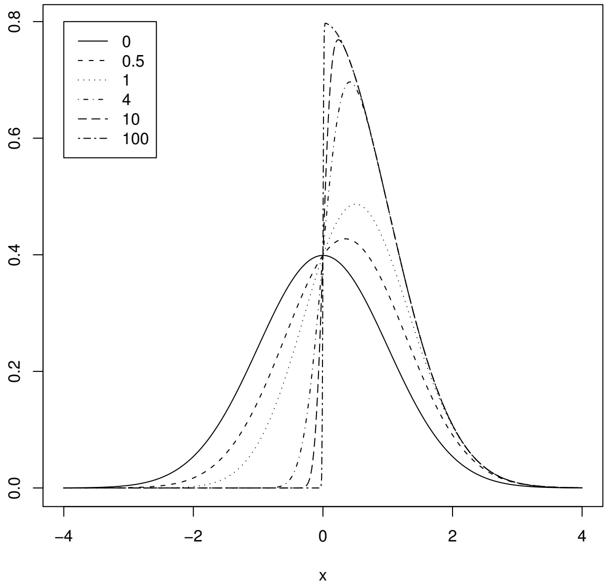

Figure 2.5: Density functions for 6 different skew normal distributions ranging from $\beta = 0$ the standard normal distribution to $\beta = 100$, essentially the half normal distribution.

so called additive zero-replacement strategy, consists of the following. Let $\mathbf{z}$ be the p dimensional composition with zeros and the new zero corrected composition be, $\mathbf{r}$, defined as follows:

$$r_j = \begin{cases} \frac{\delta(q+1)(p-q)}{p^2} & : \quad x_j = 0 \\ x_j - \frac{\delta(q+1)q}{p^2} & : \quad x_j > 0 \end{cases}$$

where, q is the number of zeros, and δ is a small value, less than a given threshold.

The second method discussed is the simple replacement strategy. In this strategy the zeros are replaced by some small non-zero constant and then the whole vector is subjected to the closure operator $\mathcal{C}$. The strategy is described by:

$$r_j = \begin{cases} \frac{1}{1+\sum_{k|x_k=0}\hat{\delta}_k}\hat{\delta}_j & : \quad x_j = 0 \\ \frac{1}{1+\sum_{k|x_k=0}\hat{\delta}_k}x_j & : \quad x_j > 0 \end{cases}$$

where, $\hat{\delta}_j$ is the imputed value on the part x_j. This method was employed by Iverson et al. (2004).

Finally, the method that Martin-Fernández et al. (2003) recommend is the multiplicative replacement strategy, which can be described as follows:

$$r_j = \begin{cases} \delta_j & : \quad x_j = 0 \\ \left(1 - \sum_{k|x_k=0} \delta_k\right) x_j & : \quad x_j > 0 \end{cases}$$

There doesn't appear to be a generally accepted strategy for dealing with essential zeros but see proposed methods.

2.5.6 Other Compositional Results

Pawlowsky-Glahn and Egozcue (2001, 2002) work along similar lines to that of Billheimer et al. (2001) and define a vector space on the simplex. They use this to define best linear unbiased estimates on the simplex and also the metric variance and metric center. In fact, they show that the metric center is the Closure of the geometric mean on each component of the composition and the metric variance as defined previously.

Egozcue et al. (2003) also consider the simplex space as a metric space and define the isometric transformation to be the one which preserves all metric properties. They also show its relationship to the additive log–ratio (alr) and centered log–ratio transformations.

2.6 Compositional Models

2.6.1 Billheimer

Billheimer et al. (2001) considers modeling the effect of co–variates, motivated by the work given in Aitchison (2003). To fix ideas, let $\boldsymbol{\mu}$, be the location parameter and consider a scalar covariate x_j, where j indexes the observations. Consider the following linear model for $\boldsymbol{\mu}_j$

$$\boldsymbol{\mu}_j = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1(x_j - \bar{x})$$

where $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ are vectors in $\mathcal{R}^d$. Thus, we can interpret $\boldsymbol{\beta}_0$ as the location when $x_j = \bar{x}$ and $\boldsymbol{\beta}_1$ as the change in location for a unit change in x. Apply the additive logistic transformation, $\phi^{-1}(.)$, to both sides of the above equation to give

$$\phi^{-1}(\boldsymbol{\mu}_j) = \phi^{-1}(\boldsymbol{\beta}_0) \oplus \phi^{-1}(\boldsymbol{\beta}_1)^{(x_j - \bar{x})}.$$

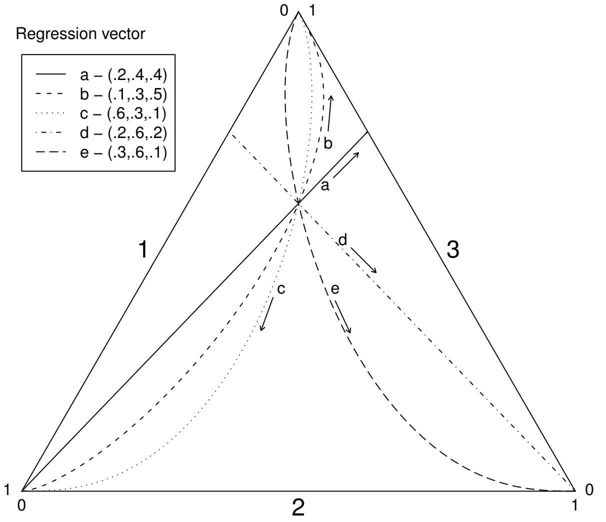

Figure 2.6: Regression curves for various regression parameter compositions. The curves are of the form $\boldsymbol{\xi}_j = \boldsymbol{\xi} \oplus \boldsymbol{\gamma}^{u_j}$, for five different values of $\boldsymbol{\gamma}$ and $-\infty < u_j < \infty$. The arrows indicate the direction of the perturbation. The initial location, *i.e. when* $u_j = 0$, $\boldsymbol{\xi} = (.2, .2, .6)$

Which we can write as follows

$$\boldsymbol{\xi}_j = \boldsymbol{\xi} \oplus \boldsymbol{\gamma}^{u_j}.$$

where $\boldsymbol{\xi}_j = \phi^{-1}(\boldsymbol{\mu}_j)$, $\boldsymbol{\xi} = \phi^{-1}(\boldsymbol{\beta}_0)$, $\boldsymbol{\gamma} = \phi^{-1}(\boldsymbol{\beta}_1)$ and u_j is the centered covariate. $\boldsymbol{\xi}$ is the location in the simplex when the covariate is at its mean value ($x_j = \bar{x}$). The regression parameter, $\boldsymbol{\gamma}$, indicates the amount the location, $\boldsymbol{\xi}$, is perturbed for a unit change in the covariate $u_j = 1$. Also, deviations in $\boldsymbol{\gamma}$ from the identity composition, indicate direction and the magnitude of the change of the perturbation. Figure 2.6,gives a graphical depiction of the curves $\boldsymbol{\xi}_j = \boldsymbol{\xi} \oplus \boldsymbol{\gamma}^{u_j}$ for five different values of the regression composition $\boldsymbol{\gamma}$.

Note that the full regression model would be written as follows:

$$\begin{aligned} \phi(\mathbf{y}_j) &= \boldsymbol{\mu}_j + \epsilon_j \\ &= \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1(x_j - \bar{x}) + \mathbf{e}_j \end{aligned}$$

and now applying the additive logistic transformation to both sides yields

$$\begin{aligned}\mathbf{y}_j &= \phi^{-1}\boldsymbol{\beta}_0 \oplus \phi^{-1}(\boldsymbol{\beta}_1)^{(x_j-\bar{x})} \oplus \mathbf{e}_j \\ &= \boldsymbol{\xi} \oplus \boldsymbol{\gamma}^{u_j} \oplus \boldsymbol{\epsilon}_j.\end{aligned}$$

The interpretation of the parameters would be the same as above, however, we now see that the actual observation is modeled as a perturbed version of $\boldsymbol{\xi}_j = \boldsymbol{\xi} \oplus \boldsymbol{\gamma}^{u_j}$.

2.6.2 Bandeen–Roche

Aldershof and Ruppert (1987) describes an EPA study on the atmospheric effects of wood stove smoke and vehicle emissions. Specifically, they collected 50 ambient air samples near Juneau, Alaska. Each observation is time averaged daily air composition vector of five chemicals (fluoranthene, benzoanthracene, chrysene, benzofluoranthene and pryene). The goal of the study was to estimate the fraction of wood stove smoke in each of the air samples.

Bandeen-Roche (1994) uses this problem as a motivation for a more general model which is described below. Of a known number m of source components of dimension p, r are predetermined and the remaining $m-r$ are of unknown composition. The predetermined compositions are denoted by $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r$ and the unknown compositions $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots \boldsymbol{\theta}_{m-r}$. Bandeen-Roche (1994) defines the following model:

$$E[\mathbf{Y}|\mathbf{y}] = \mathbf{y} \tag{2.2}$$
$$\mathbf{Y} = \sum_{k=1}^{r} \alpha_k \mathbf{x}_k + \sum_{k=r+1}^{m} \alpha_k \boldsymbol{\theta}_{k-r}, \quad \boldsymbol{\alpha} \sim G$$

with, $x_{kj} \geq 0, \sum_{j=1}^{p} x_{kj} = 1$ for all j, k and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_m) \sim G$ for some distribution G defined on the simplex. Also,$\mathbf{Y}$ represents the unobserved compositions, while $\mathbf{y}$ represents what they actually observed, ie the 50 samples actually collected. More specifically, $G(a) = P(\alpha_1 \leq a_1, \ldots, \alpha_{m-1} \leq a_{m-1})$. For a given sample they assume that $\{\boldsymbol{\alpha}\}$ are independent and identically distributed (iid), it then follows that the collection $\{\mathbf{y}\}$ are iid random vectors with a distribution denoted by H. The parameter space associated with H is the product space $\boldsymbol{\theta} \times \boldsymbol{\Gamma}$ where $\boldsymbol{\theta} = \{(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{m-r}\} : \theta_{lj} \geq 0, \sum_{j=1}^{p} \theta_{lj} = 1, \text{for each } l, j\}$ and $\boldsymbol{\Gamma} = \{G : \text{G is a distribution with support constrained on the m simplex}\}$

Note that the number of parameters increases with the sample size n, which was first discussed by Neymann and Scott (1948) and further by Kiefer and Wolfowitz (1956). However, in the present context, it is of equal importance to estimate the relative amounts of each source component, the $\boldsymbol{\alpha}$'s, as the unknown source components $\boldsymbol{\theta}$. That is, the incidental parameters as described by Neymann and Scott (1948) and Kiefer and Wolfowitz (1956) are on equal footing to the structural parameters.

The above assumptions give the following distribution for the observations:

$$F_Y(\mathbf{z}) = \int F_0(\mathbf{z}|\mathbf{y})dH(\mathbf{y}) \tag{2.3}$$

where, F_0 denotes the conditional distribution of $\mathbf{Y}$ given $\mathbf{y}$ and satisfies $\int \mathbf{z}dF_0(\mathbf{z}|\mathbf{y}) = \mathbf{y}$. Hence it captures the measurement error that leads us to observe $\mathbf{Y}$ instead of $\mathbf{y}$ (Bandeen-Roche, 1994).

Bandeen-Roche (1994) give some formal conditions for the identifiability issues addressed by Neymann and Scott (1948) and Kiefer and Wolfowitz (1956), which they show hold for their particular example.

Bandeen-Roche (1994) chose the Dirichlet class of distributions to model compositions, despite the problems associated with this distribution as noted by Aitchison (2003). Bandeen-Roche (1994) favours the Dirichlet class for the chief reason that the parameters of interest are modeled directly, not as transformations. Vectors formed by permuting the data, restandardizing a subset of components, or aggregation all have Dirichlet distributions.

Combining 2.2 with the Dirichlet distribution given in equation 2.1 with some additional assumptions gives

$$f_Y(\mathbf{z}) = \int f_0(\mathbf{z}|\boldsymbol{\delta}(\boldsymbol{\alpha}))\frac{\Gamma(\Lambda)}{\prod_{k=1}^{m}\Gamma(\lambda_k)}\prod_{k=1}^{m}\alpha_k^{\lambda_k-1}d\boldsymbol{\alpha} \tag{2.4}$$

where

$$\boldsymbol{\delta}(\boldsymbol{\alpha}) = \left[\sum_{k=1}^{r}\alpha_k\mathbf{x}_k + \sum_{k=r+1}^{m}\alpha_k\boldsymbol{\theta}_{k-r}\right]$$

and $f_0(.|.)$ is the density defined by equation 2.3 assuming it exists.

2.6.3 Aitchison and Bacon–Shone

Aitchison and Bacon-Shone (1999) consider the following problem: to determine the distribution of the D–part composition $\mathbf{y}$, formed as a convex linear combination

$$\mathbf{y} = \mathrm{cvx}(\mathbf{x}, \boldsymbol{\pi}) = \pi_1\mathbf{x}_1 + \ldots + \pi_C\mathbf{x}_C$$

where $\mathbf{x} = [\mathbf{x}_1, \ldots, \mathbf{x}_C]$ is the $C \times D$ matrix of independent D–part compositions with known distributions and $\boldsymbol{\pi}$ is a vector of non–negative mixing proportions. This model is very similar to the one proposed by Bandeen-Roche (1994), however, in this model no source components are known exactly. The issue of identifiability isn't addressed.

Aitchison and Bacon-Shone (1999) consider the following example: A Scottish loch is supplied by 3 rivers. At the mouth of each river, 10 water samples have been taken at random times and analyzed into 4–part compositions of pollutants. Also taken are 20 samples at each of three fishing locations. The aim of the study is to determine if the fishing locations can be modeled as convex mixtures of compositions of the 3 sources. The independence of the sources is assumed since the sampling was done randomly and separately in each of the 3 rivers.

Aitchison and Bacon-Shone (1999) are interested in finding the distribution of the D–part composition $\mathbf{y}$, where the D–part compositions $\mathbf{x}_1, \ldots, \mathbf{x}_C$ are independently distributed as $\mathcal{L}^D(\boldsymbol{\xi}_1, \mathbf{T}_1), \ldots, \mathcal{L}^D(\boldsymbol{\xi}_C, \mathbf{T}_C)$, where $\mathbf{T}_i = [\tau_{jk}]$. However, they state that there is no simple closed form solution to this problem and then proceed to consider 3 approximations which are given below.

Approximation 1. The distribution $\mathbf{y} = \pi_1\mathbf{x}_1 + \ldots + \pi_C\mathbf{x}_C$ where $\mathbf{x}_1, \ldots, \mathbf{x}_C$ are independently distributed as $\mathcal{L}^D(\boldsymbol{\xi}_1, \mathbf{T}_1), \ldots, \mathcal{L}^D(\boldsymbol{\xi}_C, \mathbf{T}_C)$, is approximately $\mathcal{L}^D(\boldsymbol{\eta}, \boldsymbol{\Theta})$, $\boldsymbol{\Theta} = [\theta_{ij}]$ and

$$\boldsymbol{\eta} = \sum_{b=1}^{C} \pi_b \boldsymbol{\xi}_b, \quad \theta_{ij} = -\frac{1}{2} \sum_{b=1}^{C} \sum_{k=1}^{D} \sum_{l=1}^{D} G_{bijk} G_{bijl} \tau_{bkl}$$

where,

$$G_{bijk} = \rho_{bi}(\delta_{ik} - \xi_{bk}) - \rho_{bj}(\delta_{jk} - \xi_{bk}), \quad \rho_{bi} = \pi_b \xi_{bi} / \eta_i$$

and δ_{ik} is the Kronecker delta, equal to 1 when $k = i$ and 0 otherwise.

Approximation 2: The distribution of $\mathrm{cvx}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\Omega})$, that is of $\mathbf{y} = \pi_1\mathbf{x}_1 + \ldots + \pi_C\mathbf{x}_C$, where $\mathbf{x}_1, \ldots, \mathbf{x}_C$ are independently distributed as $\mathcal{L}^D(\boldsymbol{\xi}_1, \mathbf{T}_1), \ldots, \mathcal{L}^D(\boldsymbol{\xi}_C, \mathbf{T}_C)$, and $\boldsymbol{\pi}$ is distributed as $\mathcal{L}^C(\boldsymbol{\alpha}, \boldsymbol{\Omega})$, is approximately $\mathcal{L}^D(\boldsymbol{\kappa}, \boldsymbol{\Lambda})$ where

$$\boldsymbol{\kappa} = \sum_{b=1}^{C} \alpha_b \boldsymbol{\xi}_b, \quad \lambda_{ij} = -\frac{1}{2}\sum_{b=1}^{C}\sum_{k=1}^{D}\sum_{l=1}^{D} H_{bijk}H_{bijl}\tau_{bkl} - \frac{1}{2}\sum_{a=1}^{C}\sum_{b=1}^{C} B_{aij}B_{bij}\omega_{ab},$$

where

$$H_{bijk} = \chi_{bi}(\delta_{ik} - \xi_{bk}) - \chi_{bj}(\delta_{jk} - \xi_{bk}), \quad \chi_{bi} = \alpha_b \xi_{bi}/\kappa_i, \quad B_{bij} = \chi_{bi} - \chi_{bj}.$$

Consider the following alternative model

$$\mathbf{y} = \mathrm{cvx}(\mathbf{x}, \boldsymbol{\pi}) \oplus \mathbf{u}$$

where, $\mathbf{u}$ is a compositional perturbation distributed as $\mathcal{L}^D(\mathbf{e}, \boldsymbol{\Psi})$, where $\mathbf{e}$ is the identity perturbation. This leads to the third approximation:

Approximation 3. The distribution of $\mathrm{cvx}(\mathbf{x}, \boldsymbol{\pi}) \oplus \mathbf{u}$ is approximately $\mathcal{L}^D(\boldsymbol{\eta}, \boldsymbol{\Theta} + \boldsymbol{\Psi})$, where $\boldsymbol{\Theta}$ and $\boldsymbol{\Psi}$ are as given in Approximation 1.

2.6.4 Billheimer – Mixing

Billheimer (2001) considers a Bayesian approach to the problem considered by Bandeen-Roche (1994), which allows the inclusion of prior knowledge of the pollution sources.

The model he proposes is the following (assuming a fixed number of sources, p):

$$E[\mathbf{Y}_i] = \sum_{j=1}^{p} \alpha_{ji}\boldsymbol{\theta}_j = [\boldsymbol{\theta}_1|\boldsymbol{\theta}_2|\ldots|\boldsymbol{\theta}_p]\begin{bmatrix}\alpha_{1i}\\ \alpha_{2i}\\ \vdots\\ \alpha_{pi}\end{bmatrix} = \Theta\boldsymbol{\alpha}_i$$

where, $\mathbf{Y}_j$ is a vector of concentrations of k chemical species, $\boldsymbol{\alpha}_i$ is a p–vector of mixing coefficients, and $\boldsymbol{\theta}_j$ ($j = 1, 2, \ldots, p$) is a k–vector describing the chemical profile for source j. Note that $\mathbf{Y}_j$ and $\boldsymbol{\theta}_j$ are compositional vectors as defined previously.

The complete model specification is given by:

$$\mathbf{Y}_i = \boldsymbol{\Theta}\boldsymbol{\alpha}_i \oplus \boldsymbol{\epsilon}_i$$

Billheimer (2001) casts the specification of $\boldsymbol{\theta}_j$ and $\boldsymbol{\alpha}_i$ in a Bayesian hierarchical framework.

He assumes that each source profile $\boldsymbol{\theta}_j$ is described by an informative prior distribution describing the partial knowledge of the relative likelihoods of its chemical components. He assigns a diffuse prior to the mixing proportions, $\boldsymbol{\alpha}_i$.

$$\pi(\boldsymbol{\Theta}, \boldsymbol{\alpha}_i, \boldsymbol{\epsilon}_i, \boldsymbol{\mu}_\alpha, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}_\epsilon) = \pi(\boldsymbol{\alpha}_i|\boldsymbol{\mu}_\alpha, \boldsymbol{\Gamma})\pi(\boldsymbol{\epsilon}_i|\boldsymbol{\Sigma}_\epsilon)\pi(\boldsymbol{\mu}_\alpha)\pi(\boldsymbol{\Gamma})\pi(\boldsymbol{\Sigma}_\epsilon)\pi(\boldsymbol{\Theta})$$

where

$$\begin{aligned}
\boldsymbol{\epsilon}_i|\boldsymbol{\Sigma}_\epsilon &\sim \mathcal{L}^{k-1}(\mathbf{0}_{k-1}, \boldsymbol{\Sigma}_\epsilon);\ \boldsymbol{\Sigma}_\epsilon^{-1} \sim \mathcal{IW}(a\mathcal{N}, \rho) \\
\boldsymbol{\theta}_j &\sim \mathcal{L}^{k-1}(\boldsymbol{\mu}_{\theta_j}, \boldsymbol{\Sigma}_{\theta_j}) \\
\boldsymbol{\alpha}_i &\sim \mathcal{L}^{p-1}(\boldsymbol{\mu}_\alpha, \boldsymbol{\Gamma});\ \boldsymbol{\mu}_\alpha \sim \mathcal{N}^{p-1}(\boldsymbol{\eta}, \Psi);\ \boldsymbol{\Gamma}^{-1} \sim \mathcal{IW}(b\mathcal{N}, \delta)
\end{aligned}$$

The joint posterior distribution is given by the following:

$$\begin{aligned}
&\pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_p, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_n, \boldsymbol{\mu}_\alpha, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}_\epsilon|\mathbf{y}) \propto \\
&\prod_{i=1}^{n}\left\{|\boldsymbol{\Sigma}_\epsilon|^{-1/2}\prod_{t=1}^{k}[\mathbf{y}_i]_t^{-1}\exp\left\{-\frac{1}{2}[\phi(\mathbf{y}_i)-\phi(\boldsymbol{\Theta}\boldsymbol{\alpha}_i)]'\boldsymbol{\Sigma}_\epsilon^{-1}[\phi(\mathbf{y}_i)-\phi(\boldsymbol{\Theta}\boldsymbol{\alpha}_i)]\right\}\right. \\
&\left.\times\prod_{j=1}^{p}|\boldsymbol{\Gamma}|^{-1/2}\alpha_{ij}^{-1}\exp\left\{-\frac{1}{2}[\phi(\boldsymbol{\alpha}_i)-\boldsymbol{\mu}_\alpha]'\boldsymbol{\Gamma}^{-1}[\phi(\boldsymbol{\alpha}_i)-\boldsymbol{\mu}_\alpha]\right\}\right\} \\
&\times\exp\left\{-\frac{1}{2}(\boldsymbol{\mu}_\alpha-\boldsymbol{\eta})'\Psi^{-1}(\boldsymbol{\mu}_\alpha-\boldsymbol{\eta})\right\}\prod_{j=1}^{p}\exp\left\{-\frac{1}{2}[\phi(\boldsymbol{\theta}_j)-\boldsymbol{\mu}_{\theta_j}]'\boldsymbol{\Sigma}_{\theta_j}^{-1}[\phi(\boldsymbol{\theta}_j)-\boldsymbol{\mu}_{\theta_j}]\right\} \\
&\times\pi(\boldsymbol{\Gamma})\times\pi(\boldsymbol{\Sigma}_\epsilon)
\end{aligned}$$

Variants of this model form the basis of the predator diet model.

Chapter 3

Markov Chain Monte Carlo: Basic Algorithms and Adaption

Bayesian inference has long struggled to reach the mainstream of statistical practice despite its inherent theoretical advantages. The major stumbling block has largely been computational. Specifically how to evaluate high dimensional non-analytically tractable integrals that constitute posterior distributions resulting from most statistical applications. For many decades the Bayesian analyst had to rely on restricting attention to conjugate families of prior and likelihoods or using numerical approximations that broke down in high dimensional problems.

We focus our attention on Monte Carlo methods, however, several alternative methods have been suggested. Numerical integration techniques have been successfully applied to low dimensional posterior distributions, however they suffer from the so called curse of dimensionality. That is, the number of function evaluations increases exponentially with the dimension of the problem. Evans and Swartz (2000) give an excellent description of these techniques. Laplace approximations (Robert and Casella, 2004) are essentially second order Taylor series approximations to the posterior distribution. They tend to work well in moderate dimensional problems and when the posterior is log–concave but also don't perform well in higher dimensional situations.

We begin with a short review of several Monte Carlo methods before turning to Markov Chain Monte Carlo methods. The main distinction between these two simulation based approaches is that Monte Carlo methods typically give uncorrelated samples, while Markov chain methods give correlated samples. Therefore, more care must be taken to process the resulting MCMC samples.

3.1 Monte Carlo Methods

3.1.1 Monte Carlo Integration

Consider the following integral

$$E_\pi[f(x)] = \int f(x)\pi(x)dx, \tag{3.1}$$

where $f(x)$ is a deterministic function, and $\pi(x)$ is a continuous probability density function. The integral represents the expected value of the function $f(x)$. Assume that the distribution of $f(x)$ is such that the integral is not available in closed form, however, we can sample directly from $\pi(x)$. The method of Monte Carlo integration approximates $E_\pi[f(x)]$ by the following

$$\bar{f}_n = \sum_{i=1}^{n} f(x_i)$$

where$(x_1, \ldots, x_n)$ is a sample from $\pi(x)$. The strong law of large numbers gives

$$\lim_{n\to\infty} \bar{f}_n \overset{a.s.}{\to} E_\pi[f(x)]$$

One can also establish a central limit theorem, when $E_\pi[f(x)^2] < \infty$, since we can calculate the variance of $\bar{f}_n$

$$\text{var}(\bar{f}_n) = \frac{1}{n}\int (f(x) - E_\pi[f(x)])^2 \pi(x)dx$$

and therefore,

$$\frac{(\bar{f}_n - E_\pi[f(x)])}{\sqrt{\text{var}(\bar{f}_n)}} \overset{\mathcal{L}}{\to} N(0,1)$$

and we can estimate var$(\bar{f}_n)$ by

$$v_n = \frac{1}{n^2}\sum_{i=1}^{n}(f(x_i) - \bar{f}_n)^2$$

We can extend this method to vector valued random variables $\mathbf{x}$ without difficulty provided we can generate samples from $\pi(\mathbf{x})$. Robert and Casella (2004) give several generic methods for generating samples from an arbitrary density function $\pi(x)$, such as the inverse probability transform, accept–reject methods.

3.1.2 Importance Sampling

Consider the original integral given by equation (3.1) and consider the following modification

$$E_\pi[f(x)] = \int f(x)\frac{\pi(x)}{h(x)}h(x)dx.$$

provided the support of $\pi(x)$ is subset of the support of $h(x)$. This is a necessary condition otherwise we introduce singularities into the integration process. We then consider the following approximation:

$$\bar{f}^*_n = \frac{1}{n}\sum_{i=1}^{n}\frac{\pi(x_i)}{h(x_i)}f(x_i)$$

We can show that $\bar{f}_n^*$ converges almost surely to $E_\pi[f(x)]$ by the Law of Large Numbers. There is great flexibility in the choice of the importance function $h(x)$, however, can we say that some choices are better than others? Robert and Casella (2004) give some results about the choice of importance function which minimize the variance of $\bar{f}_n^*$.

The crux of this method, is that the density $h(x)$ is easier to generate random samples from and also the support of $h(x)$ is greater than $\pi(x)$. This can be difficult to find in practice, particularly in higher dimensional problems.

3.1.3 Random Sample Generation

We have not mentioned how one would generate samples from the distribution of interest, $\pi(x)$. There are numerous methods that are employed, for example, the inverse cdf method, transformation methods, ratio of uniforms and accept–reject methods to name but a few of the many methods. Robert and Casella (2004) give an introduction to several of these methods. In addition there are several classic texts on the subject (Hammersley and Handscomb, 1964; Knuth, 1964; Ripley, 1987, and the references therein).

We briefly mention a few of the methods as they have connections to the MCMC methods to be introduced in the next section. Accept–reject methods can be seen as a special case of the so called fundamental theorem of simulation, which states

Theorem 3.1. *(Fundamental Theorem of Simulation) Simulating*

$$X \sim f(x)$$

is equivalent to simulating

$$(X, U) \sim \mathcal{U}\{(x, u) : 0 < u < f(x)\},$$

where U is a uniform over the $[0, 1]$ interval. To see this consider the following

$$f(x) = \int_0^{f(x)} du.$$

We have seemingly made the simulation problem more difficult by introducing an auxiliary variable U, however, it turns out the simulation is more tractable for a certain class of problems. This fundamental theorem is the basis of the slice sampler, which has some connections to the Gibbs Sampler as well.

The basic accept–reject algorithm is based on the following corollary of the fundamental theorem

Corollary 3.1. *Let $X \sim f(x)$ and let $g(x)$ be a density function that satisfies $f(x) \leq Mg(x)$ for some constant $M \geq 1$. Then, to simulate $X \sim f$, it is sufficient to generate*

$$Y \sim g \ \textit{ and } U|Y = y \sim \mathcal{U}(0, Mg(y)),$$

until $0 < u < f(y)$.

And the accept–reject algorithm is given by

1. Generate $X \sim g$ and $U \sim \mathcal{U}_{[0,1]}$
2. Accept $Y = X$ if $U \leq f(X)/Mg(X)$
3. Otherwise repeat step 1.

It is always possible to find a dominating distribution $Mg(x)$ for all x, however, if we have to choose M excessively large then a lot of steps of the algorithm must be performed to generate a single sample.

3.2 Markov Chain Monte Carlo (MCMC)

We now turn our attention to Markov Chain Monte Carlo methods, which have an interesting history which we briefly describe. Metropolis et al. (1953)'s paper on the statistical mechanics of particles introduced the method of Markov Chain Monte Carlo (MCMC) to the world of physics. Hammersley and Handscomb (1964) describes the method in a more rigorous statistical framework in terms of Markov chains. Hastings (1970) provided a generalization of the original Metropolis algorithm to allow for non-symmetric proposal distributions. While Geman and Geman (1984) used the Gibbs sampler on the Bayesian image restoration problem, the work of Gelfand and Smith (1990) illustrated how the Gibbs sampler could be used to help Bayesians solve a much wider class of problems. Till this point, most Bayesian approaches used modal approximations, numerical methods which tended to break down in higher dimensional problems, or restricted themselves to classes of problems with conjugate priors.

The key difference between importance sampling or the usual Markov simulation strategies and MCMC methods is that MCMC methods produce Markov chains and hence there is a form of dependence among the samples generated during the sampling. This is both a blessing and a curse as we shall see.

The following definition taken from Robert and Casella (2004) gives a general description of Markov Chain Monte Carlo algorithms

Definition 3.1. *A Markov Chain Monte Carlo (MCMC) method for the simulation of a distribution f is any method producing an ergodic Markov chain (X_t) whose stationary distribution is f.*

There are two main building blocks of most MCMC algorithms: the Metropolis–Hastings algorithm and the Gibbs sampler. They are often combined to form Metropolis–within-Gibbs. We describe these basic building blocks, discussing their basic structure and some of their convergence properties. We also give details on the Metropolis–within–Gibbs algorithm. Finally, we discuss the relatively new area of adaptive algorithms which have taken off in the past decade. The adaptive algorithms are typically applied in Metropolis–Hastings algorithms, but have also been applied in various other settings.

3.3 Metropolis–Hastings

The original Metropolis algorithm was applied in statistical mechanics to simulate the distribution of particles (Metropolis et al., 1953). We are concerned with its applications to statistical problems, more specifically the evaluation of analytically intractable posterior distributions. To make the development slightly more general, we assume a general continuous probability density function denoted by $f(x)$ from which we wish to generate samples, known as the target distribution. To fix ideas think of the target distribution as the posterior distribution of interest.

The Metropolis–Hastings algorithm requires the objective or target distribution $f(x)$ be known up to a constant of proportionality, that is, the functional form of the posterior must be known. The key ingredient, is the proposal distribution $q(.|x)$, a conditional density which is easy to simulate from and is also known up to a constant of proportionality that does not depend on x. If the proposal distribution is symmetric, that is $q(y|x) = q(x|y)$, then the algorithm simplifies to the Metropolis algorithm (Metropolis et al., 1953). The acceptance probability of a new point y given that the chain is currently in state x is given by

$$\rho(x,y) = \begin{cases} \min\left\{\frac{f(y)q(x|y)}{f(x)q(y|x)}, 1\right\}, & \text{if} f(x)q(y|x) > 0 \\ 1, & \text{otherwise} \end{cases}$$

The algorithm proceeds in the following manner.

0. Given a starting value x_t.

1. Generate $y_t \sim q(y|x_t)$.

2. $$x_{t+1} = \begin{cases} y_t & \text{with probability } \ \rho(x_t, y_t) \\ x_t & \text{with probability } \ 1 - \rho(x_t, y_t) \end{cases}$$

Rewriting the acceptance probability $\rho(x_t, y_t)$ as follows

$$\frac{f(y_t)/q(y_t|x_t)}{f(x_t)/q(x_t|y_t)}$$

reveals that y_t's that increase the numerator relative to the denominator will always be accepted. In addition, there is a probability that it will accept y_t's that lower this ratio and hence move to regions of lower support of the distribution $f(x)$. This allows the algorithm to potentially transverse lower probability regions. In contrast with the accept–reject methods, the Metropolis–Hastings algorithm stays in the current state if the proposed state is not accepted.

The algorithm in its current form is too general in that it applies for all f's and q's. Robert and Casella (2004) impose the following regularity conditions on f and q, to ensure that the limiting distribution of the constructed Markov chain is in fact f. We restrict our attention to distributions f with connected supports, denoted by $\mathcal{E}$ as unconnected supports can invalidate the Metropolis–Hastings algorithm.

In addition we need the following definition,

Definition 3.2. *We say the support of f is truncated by q, if there exits $A \subset \mathcal{E}$ such that*

$$\int_A f(x)dx > 0 \ \text{ and } \ \int_A q(y|x)dy = 0, \ \ \forall x \in \mathcal{E}.$$

If q truncates f, then the Markov chain induced by the MH algorithm cannot have f as its limiting distribution since the chain never visits the set A unless it happens to be initialized in A. We need the following minimally necessary condition

$$\bigcup_{x \in \text{ supp } f} \text{supp } q(.|x) \subset \text{ supp } f$$

to ensure that f is the limiting distribution for induced Markov chain.

The transition kernel of a Markov chain is given by the following definition

Definition 3.3. *If $K = \{K(x, A), x \in \mathcal{X}, A \in \mathcal{B}(\mathcal{X})\}$ is such that*

(i) for each $A \in \mathcal{B}(\mathcal{X})$, $K(., A)$ is a non–negative measurable function on $\mathcal{X}$ and

(ii) for each $x \in \mathcal{X}$, $K(x, .)$ is a probability measure on $\mathcal{B}(\mathcal{X})$, then we call P a transition probability kernel or Markov transition function.

Essentially, the transition kernel for a Markov chain gives the probability of being in a set A at the next time step given that the chain is at state x at the current state.

The transition kernel for the Metropolis–Hastings algorithm is given by

$$K(x, y) = \rho(x, y)q(y|x) + r(x)\delta_x(y)$$

where

$$\delta_x(y) = \begin{cases} 1 & \text{if } y = x \\ 0 & \text{otherwise} \end{cases}$$

is the usual Kronecker delta function and $r(x) = 1 - \int \rho(x, y)q(y|x)dy$. The first term $\rho(x, y)q(y|x)$ is the probability of transition out of the current state x and $r(x)$ is the probability that the chain remains in the current state x.

The detailed balance or reversibility condition is given by

$$f(x)\rho(x, y)q(y|x) = \rho(y, x)q(x|y)f(y)$$

or

$$K(y, x)f(y) = K(x, y)f(x).$$

We have the following theorem from Robert and Casella (2004)

Theorem 3.2. *Let (X_t) be the chain produced by the above algorithm. For every conditional distribution $q(y|x)$ whose support includes $\mathcal{E}$,*

1. *the kernel of the chain satisfies the detailed balance condition with f ;*
2. *f is a stationary distribution of the chain.*

For almost any proposal distribution q the resulting Markov chain will have f as its limiting distribution.

Chib and Greenberg (1995) give an excellent description of the Metropolis–Hastings algorithm. There are numerous other descriptions of the Metropolis–Hastings algorithm and the reader is referred to Robert and Casella (2004) and the references therein.

3.3.1 Variants of the Metropolis–Hastings Algorithm

Chib and Greenberg (1995) give the following five variants of the Metropolis–Hastings algorithm:

1. random walk chain originally proposed by Metropolis et al. (1953),
2. independence chain (Tierney, 1994),
3. Tierney (1994) suggests generating proposals from a vector auto–regressive order one process,
4. candidate–generating density by exploiting the form of $f(x)$ (Chib and Greenberg, 1994),
5. use an accept–reject method with a pseudo–dominating density (Tierney, 1994).

Each of these variants have had, to varying degrees, their theoretical properties studied. The reader is referred to Robert and Casella (2004) and the references therein for details of these algorithms. However, it is important to point out that we can establish that the independence chain is uniformly ergodic provided the proposal distribution matches the target distribution closely enough. Unfortunately we cannot establish the uniform ergodicity of the random walk chain, however, Mengersen and Tweedie (1996) have shown that for log–concave target distributions meeting certain conditions, geometric ergodicity can be established.

We implement the general Metropolis–Hastings algorithm as our proposal distribution do not match any of the above variants. There do not appear to be any results establishing the ergodicity of the general MH algorithm.

Hastings (1970) give the following example of a univariate MH algorithm for generation of univariate normal distribution.

Example 3.1. *Consider the generation of samples from a standard normal distribution using a random walk MH algorithm with q uniform on $[-\delta, \delta]$. The acceptance probability is given by*

$$\rho(x, y) = \min\left\{\exp\{(x^2 - y^2)/2\}, 1\right\}$$

Figure 3.1 gives trace plots, auto–correlation plots and histograms of the MH algorithm for three of the δ values used. Table 3.1 gives the estimated means and variances, also indicates the number of rejections and the integrated auto–correlation times (see equation 3.2) for the 20,000 replications of the MH algorithm. The table and figure clearly display the so called "Goldilocks principle", a term coined by Jeffrey Rosenthal. δ's that are too low lead to highly correlated chains by making very small proposals which are almost always accepted. While δ's that are too large lead to highly correlated chains by proposing larger moves that are not accepted very often and hence the chain stays in one state for longer periods of time.

δ	mean	variance	Number of acceptances	Integrated auto–correlation time
0.1	0.1996	0.736	19635	1113.1
0.5	0.0342	1.031	17972	58.70
1.0	-0.0223	1.053	16081	16.97
2.0	0.0266	0.990	12658	6.88
5.0	-0.0037	1.006	6317	4.65
10.0	0.0126	1.012	3123	10.07
20.0	-0.0104	0.9963	1524	19.10

Table 3.1: Estimates of the mean and variance of the 20,000 samples from the $\mathcal{N}(0, 1)$ generated using a random walk on $[-\delta, \delta]$. The number of acceptances and the integrated auto–correlation time are also given for each of the proposed samplers.

Chib and Greenberg (1995) consider the following bivariate normal example.

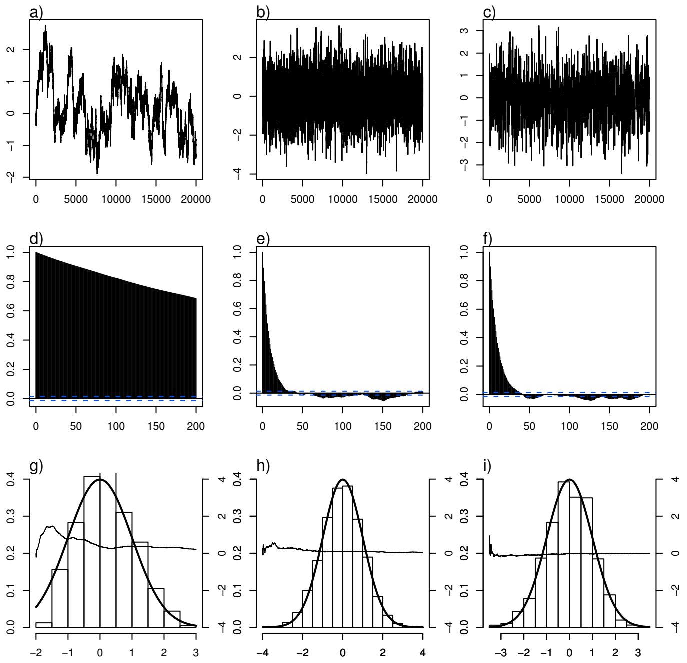

Figure 3.1: Trace plots (a-c), auto–correlation functions(d-f) and histograms(g-i) of the 20,000 samples produced by the MH algorithm using a random walk on $[-\delta, \delta]$ for three selected δ's. The first column corresponds to $\delta = 0.1$, the second column $\delta = 1.0$ and the last column $\delta = 20.0$. Overlayed on the histogram is the density of the standard normal distribution and also displayed are the cumulative mean estimates

Example 3.2. *Let $f(\mathbf{x})$ be a bivariate normal density with mean vector $\boldsymbol{\mu} = (1, 2)'$ and covariance matrix*

$$\Sigma = \begin{pmatrix} 1 & .9 \\ .9 & 1 \end{pmatrix}$$

Chib and Greenberg (1995) consider four proposal densities, we consider just their first. They use a random walk proposal distribution, $\mathbf{y} = \mathbf{x} + \mathbf{z}$, where $\mathbf{z}$ is a bivariate uniform distribution over $[-\delta, \delta] \times [-\delta, \delta]$. The acceptance probability can be written as follows

$$\rho(\mathbf{x}, \mathbf{y}) = \min\left\{ \frac{exp[-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{y} - \boldsymbol{\mu})]}{exp[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})]}, 1 \right\}, \ \mathbf{x}, \mathbf{y} \in \mathcal{R}^2$$

Table 3.2 gives the estimated means, variances, covariance, number of acceptances and the integrated auto–correlation times for the 20,000 samples generated using varying values of δ. It is clear that there is an optimal δ in the vicinity of $\delta = 2.0$. Figure 3.1 shows trace plots, auto–correlation functions and a bivariate scatter plot.

δ	mean1	mean2	var1	cov	var2	No. Accept	IACT1	IACT2
0.1	0.925	1.899	0.799	0.670	0.751	18640	1514.168	1344.215
0.5	1.086	2.088	0.986	0.902	1.015	13919	91.129	90.546
1.0	1.055	2.054	1.058	0.966	1.068	9429	41.378	41.372
2.0	0.987	1.987	1.063	0.958	1.040	4624	22.111	22.409
5.0	1.061	2.057	1.021	0.882	0.958	1139	34.818	41.052
10.0	1.140	2.095	0.768	0.695	0.831	260	117.987	128.837
20.0	0.903	1.986	1.476	1.327	1.472	84	525.504	512.207

Table 3.2: Estimates of the means, variances and covariance of the 20,000 samples from the bivariate normal distribution described in example 3.2 generated using a bivariate random walk on $[-\delta, \delta] \times [-\delta, \delta]$. The number of acceptances and the integrated auto–correlation times for each dimension are also given for each of the proposed samplers.

3.4 Acceptance Rates

The Metropolis–Hastings algorithms just considered all have great flexibility in the form of the proposal distribution and also the parameters of the chosen proposal distribution. However, this leads to the question of what form of the proposal to choose and also once the functional form of the proposal distribution is chosen how should we choose the parameters of that distribution. In practice, the proposal distribution is usually limited to distributions that are easy to simulate from, for example, the multivariate normal or multivariate t-distributions are common choices. However, there is still the issue of what covariance one should choose. Unfortunately, this is still a relatively unanswered question.

Roberts et al. (1997) showed that for a particular class of target distributions the “optimal” acceptance rate, as the dimension goes to infinity, is 0.234 for a random walk Metropolis–Hastings algorithm. Roberts and Rosenthal (2001) also consider the question of optimal scaling, however, they approach it from a efficiency perspective. They first define the integrated auto–correlation time, for an arbitrary square–integrable function g, by

$$\tau_g = 1 + 2\sum_{i=1}^{\infty} Corr(g(X_0), g(X_i)) \tag{3.2}$$

where X_0 is assumed to distributed according to the stationary distribution f. If generated samples were uncorrelated, then $\tau_g = 1$, thus τ_g is a measure of the dependence in the

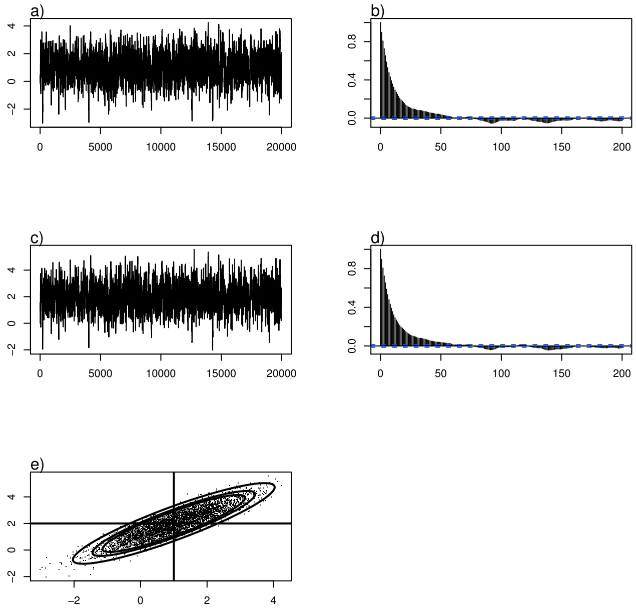

Figure 3.2: Trace plots of components 1 and 2 are shown in panels a and c respectively and auto–correlation functions in panels (b,d) for the bivariate normal described in example 3.2 for the 20,000 samples produced by the MH algorithm using a random walk on $[-2.0, 2.0] \times [-2.0, 2.0]$. Panel (e) shows a bivariate plot of the samples, along with 90%, 95% and 99% probability contours.

Markov chain. Roberts and Rosenthal (2001) state that if a central limit theorem for X and g exits then the variance of the estimator is

$$\sum_{i=1}^{n} g(X_i)/n.$$

They also define the acceptance rate and propose to estimate it by

$$\begin{aligned} \rho &= \int \rho(\mathbf{x}, \mathbf{y}) f(\mathbf{x}) q(\mathbf{x}|\mathbf{y}) d\mathbf{x} d\mathbf{y} \\ &= \lim_{n \to \infty} n^{-1} \#\{\text{accepted moves}\} \end{aligned}$$

They showed that for a random walk Metropolis–Hastings algorithm acceptance rates in the range of 0.1 to 0.4 are optimal for a smooth target distribution. In addition, they list several references that detail work of a similar nature.

Both Roberts et al. (1997) and Roberts and Rosenthal (2001) show that the optimal proposal variance is related to the Fisher's expected information.

3.5 Gibbs

The Gibbs sampler first appeared in the statistics literature with the work of Geman and Geman (1984) on image reconstruction. However, it wasn't till the seminal work of Gelfand and Smith (1990), that the ideas took hold and revolutionized Bayesian computational methods. Till then, the problems that could be solved were relatively low dimensional in nature, mainly due to the coarse approximation methods in use. Numerical integration methods were hampered by the curse of dimensionality, and it was difficult to assess the accuracy of the modal approximations (i.e. Laplacian approximation methods).

Robert and Casella (2004) develop the Gibbs sampler by the following progression: the slice sampler and the fundamental theorem of simulation, the two stage Gibbs Sampler and finally the general Gibbs sampler. Due to space constraints, we just present the general case and refer the reader to Robert and Casella (2004) for more details, however, we will indicate some of the connections and convergence results as required.

Suppose that for some $p > 1$ that the joint distribution $\mathbf{X} \in \mathcal{X}$ can be decomposed into p components as follows:

$$\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p),$$

where the length of the vector, $\mathbf{X}_i$, is $p_i \geq 1$, *i.e.*, the components can be potentially be multivariate. We assume the full conditional distributions $f_1, \ldots, f_p$ can be simulated from as follows

$$\mathbf{X}_i | \mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_p \sim f_i(\mathbf{x}_i | \mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_p),$$

for $i = 1, \ldots, p$. The Gibbs sampler consists of the following algorithm.

0. Initialize the vector $\mathbf{x}_0 = (\mathbf{x}_{1,0}, \ldots, \mathbf{x}_{p,0})$

1. Simulate $\mathbf{x}_{1,t+1} \sim f_1(\mathbf{x}_1 | \mathbf{x}_{2,t}, \ldots, \mathbf{x}_{p,t})$,

2. Simulate $\mathbf{x}_{2,t+1} \sim f_2(\mathbf{x}_2 | \mathbf{x}_{1,t+1} \mathbf{x}_{3,t}, \ldots, \mathbf{x}_{p,t})$,

p. Simulate $\mathbf{x}_{p,t+1} \sim f_p(\mathbf{x}_p | \mathbf{x}_{1,t+1}, \ldots, \mathbf{x}_{p-1,t+1})$.

To illustrate ideas we consider the following examples.

Example 3.3. *The is a continuation of Example 3.2, a bivariate normal with mean* $\boldsymbol{\mu} = (1, 2)'$ *and covariance matrix*

$$\Sigma = \begin{pmatrix} 1 & .9 \\ .9 & 1 \end{pmatrix}$$

The full conditionals are univariate normal and are given by

$$\begin{aligned} x_{1,t+1}|x_{2,t} &\sim \mathcal{N}(1 + 0.9(x_{2,t} - 2), 1 - 0.9^2) \\ x_{2,t+1}|x_{1,t+1} &\sim \mathcal{N}(2 + 0.9(x_{1,t+1} - 1), 1 - 0.9^2) \end{aligned}$$

The Gibbs sampler was run for 20,000 iterations and does a very good job of reproducing the means of the distribution, 1.003 and 2.001 respectively. It also does a reasonable job of recovering the variances and covariances, 0.965, 0.967 and 0.956 respectively, however, it under estimates the variances and inflates the covariance. The integrated auto–correlation times for the two dimensions are 182.8 and 183.4 respectively, which are substantially higher than the appropriately scaled two–dimensional MH algorithm. Figure 3.3 shows trace plots, auto–correlation functions and a scatter plot. The scatter plot reveals that the Gibbs sampler produces samples that are much more correlated than required ($\rho = 0.989$).

The example shows quite clearly the dependence issues that can occur when the Gibbs sampler is employed on problems with high dependence between the components. Alternative strategies such as blocking are typically used to avoid this problem (see Robert and Casella, 2004, and the references therein).

Example 3.4. *Normal distribution with unknown mean* μ *and unknown precision* ω *with samples given by* $\mathbf{y} = (y_1, \ldots, y_n)$ *and joint prior distribution* $\pi(\mu, \omega) = \omega$*, (the so called Jeffrey's prior). We have the following posterior,*

$$\begin{aligned} \pi(\mu, \omega|\mathbf{y}) &\propto \omega^{n/2+1} \exp\{-\frac{\omega}{2}\sum_{i=1}^{n}(y_i - \mu)^2\} \\ &\propto \omega^{(n+1)/2-1} \exp\{-\frac{\omega}{2}\sum_{i=1}^{n}(y_i - \mu)^2\} \end{aligned}$$

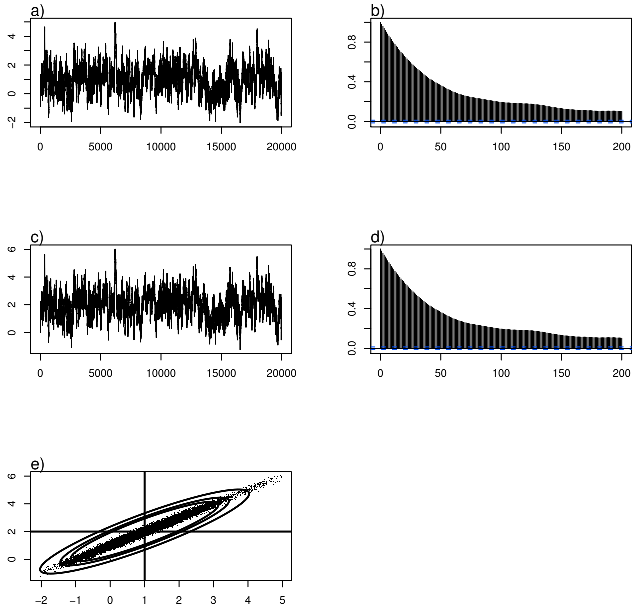

Figure 3.3: Trace plots of components 1 and 2 are shown in panels a and c respectively and auto–correlation functions in panels (b,d) for the bivariate normal described in example 3.3 for the 20,000 samples produced by the Gibbs sampler. Panel (e) shows a bivariate plot of the samples, along with 90%, 95% and 99% probability contours.

The conditional distributions are

$$\begin{aligned} p(\mu|\omega,\mathbf{y}) &\sim \mathcal{N}(\bar{y}, n\omega), \\ p(\omega|\mu,\mathbf{y}) &\sim \mathcal{G}\left((n+1)/2, \frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^2\right). \end{aligned}$$

The Gibbs sampling proceeds as follows:

1. *select starting values μ^0 and ω^0*

2. *generate $\mu_t \sim N(\bar{y}, n\omega_{t-1})$*

3. *generate $\omega_t \sim G(n/2, \frac{1}{2}\sum_{i=1}^{n}(y_i-\mu_t)^2)$*

Definition 3.4. *Given a probability density f, a density g that satisfies*

$$\int_Z g(x,z)dz = f(x)$$

is called a completion of f.

The idea of completion can, in certain situations, make the implementation of the Gibbs sampler more straightforward. This idea is related to the data augmentation ideas of Tanner and Wong (1987). For example, completion is used in Bayesian approaches to mixing, where with the addition of the group membership the Gibbs implementation becomes straightforward (see Robert and Casella, 2004; Diebolt and Robert, 1994; Celeux et al., 2000).

The Markov chain induced by the systematic scan (sweep) algorithm is not reversible, that is, it does not satisfy the detailed balance(reversibility) condition that the Metropolis–Hastings algorithm enjoys by construction as showed in Chib and Greenberg (1995). This lack of reversibility leads to some technical difficulties, most notably the lack of a central limit theorem. Fortunately, there is a relatively straightforward fix to the lack of reversibility. The following Reversible Scan Gibbs sampler is reversible

0. Initialize the vector $\mathbf{x}_0 = (\mathbf{x}_{1,0}, \dots, \mathbf{x}_{p,0})$

1. Simulate $\mathbf{x}_1^* \sim f_1(\mathbf{x}_1|\mathbf{x}_{2,t}, \dots, \mathbf{x}_{p,t})$,

2. Simulate $\mathbf{x}_2^* \sim f_2(\mathbf{x}_2|\mathbf{x}_1^*, \mathbf{x}_{3,t}, \dots, \mathbf{x}_{p,t})$,

p-1. Simulate $\mathbf{x}^*_{p-1} \sim f_{p-1}(\mathbf{x}_p|\mathbf{x}^*_1, \ldots, \mathbf{x}^*_{p-2}, \mathbf{x}_{p,t})$.

p. Simulate $\mathbf{x}_{p,t+1} \sim f_{p-1}(\mathbf{x}_p|\mathbf{x}^*_1, \ldots, \mathbf{x}^*_{p-1})$.

p+1. Simulate $\mathbf{x}_{p-1,t+1} \sim f_{p-1}(\mathbf{x}_p|\mathbf{x}^*_1, \ldots, \mathbf{x}^*_{p-2}, \mathbf{x}_{p,t+1})$.

2p-2. Simulate $\mathbf{x}_{2,t+1} \sim f_2(\mathbf{x}_2|\mathbf{x}^*_1, \mathbf{x}_{3,t+1}, \ldots, \mathbf{x}_{p,t+1})$,

2p-1. Simulate $\mathbf{x}_{1,t+1} \sim f_1(\mathbf{x}_1|\mathbf{x}_{2,t+1}, \ldots, \mathbf{x}_{p,t+1})$,

The major disadvantage of the reversible scan Gibbs algorithm is that the first $p-1$ steps of the algorithm don't get used, however, they do guarantee the reversibility of the resulting chain.

An alternative way to guarantee reversibility of the induced Markov chain is to use a random scan Gibbs proposed by Liu et al. (1995), however, it was also suggested by Metropolis et al. (1953). The most straightforward variant is given by the following algorithm where $\mathcal{G}_p$ is a distribution producing permutations of the integers $1 \ldots, p$.

1. Generate a permutation $\sigma \in \mathcal{G}_p$

2. Simulate $\mathbf{x}_{\sigma_1,t+1} \sim f_{\sigma_1}(\mathbf{x}_{\sigma_1}|\mathbf{x}_{j,t}, j \neq \sigma_1)$;

p+1. Simulate $\mathbf{x}_{\sigma_p,t+1} \sim f_{\sigma_p}(\mathbf{x}_{\sigma_p}|\mathbf{x}_{j,t}, j \neq \sigma_p)$.

Alternative algorithms have been suggested in the literature, where at each step only one of the conditional distributions is updated and also have a potentially non–uniform distribution $\mathcal{G}$. This modification allows more frequent updates for conditional distributions that are more difficult to sample from.

Geman and Geman (1984) introduced the Gibbs sampler to the image reconstruction literature, though its roots can be traced to the Hammersley-Clifford theorem, which states that the joint distribution is completely determined by its full conditional distributions. The theorem can be stated as follows

Theorem 3.3 (Hammersley—Clifford). *Under the positivity condition, the joint distribution g satisfies*

$$g(y_1, \ldots, y_p) \propto \prod_{j=1}^{p} \frac{g_{l_j}(y_{l_j}|y_{l_1}, \ldots, y_{l_j-1}, y'_{l_{j+1}}, \ldots, y'_{l_p})}{g'_{l_j}(y'_{l_j}|y_{l_1}, \ldots, y_{l_j-1}, y'_{l_{j+1}}, \ldots, y'_{l_p})}$$

for every permutation l on $\{1, 2, \ldots, p\}$ and every $y' \in \mathcal{Y}$.

Here positivity is defined as follows

Definition 3.5. *Let $(Y_1, \ldots, Y_p) \sim g(y_1, \ldots, y_p)$ and let $g^{(i)}$ denotes the marginal distribution of Y_i. If $g^{(i)}(y_i) > 0$ for every $i = 1, \ldots, p$, implies that $g(y_1, \ldots, y_p) > 0$, then g satisfies the positivity condition.*

Robert and Casella (2004) give a proof when the positivity condition holds. Besag (1994) and subsequently Hobert et al. (1997) give other versions of the theorem under more general conditions.

3.6 Metropolis–Hastings–Within–Gibbs

Chib and Greenberg (1995) draw attention to the fact that Hastings (1970) introduces the so called “block–at–a-time” or “variable–at–a–time” variants of his extension to the Metropolis algorithm. In essence, the original proposal consisted of updating the target conditionally rather than all the variables at once. The Gibbs sampler is a variant of this algorithm, where the full conditionals are actually available to sample directly. The so called Metropolis–Hastings–within–Gibbs (MHWG) algorithm is also a special case of this algorithm where some of the full conditionals are available to sample directly and others are only available to be updated by the Metropolis–Hastings algorithm. Chib and Greenberg (1995) argue that we should revert to the original terminology in Hastings (1970), however, we shall continue to used the the flawed terminology as it seems to have taken hold in the literature.

Müller (1991, 1993) proposed doing a Metropolis–Hastings step for each of the difficult conditional distributions. The MHWG algorithm is virtually identical to the Gibbs sampler already described. However, when the full conditional distribution isn’t available to be sampled a Metropolis–Hastings step is used. This isn’t the only approach, Gilks and Wild (1992) introduce adaptive rejection sampling for difficult to sample conditional distributions. This approach and its improvements are implemented in WinBUGS (Lunn et al., 2000).

The MH–within–Gibbs algorithm is given by the following

0. For $i = 1, \ldots, p$ given $(\mathbf{x}_{1,t+1}, \ldots, \mathbf{x}_{i-1,t+1}, \mathbf{x}_{i+1,t}, \ldots, \mathbf{x}_{p,t})$:

1. Generate $\mathbf{y}_i \sim q_i(\mathbf{y}_i | \mathbf{x}_{1,t+1}, \ldots, \mathbf{x}_{i-1,t+1}, \mathbf{x}_{i+1,t}, \ldots, \mathbf{x}_{p,t})$

2.

$$\mathbf{x}_{i,t+1} = \begin{cases} \mathbf{y}_i & \text{with probability } \rho_i \\ \mathbf{x}_{i,t} & \text{with probability } 1-\rho_i \end{cases}$$

where

$$\rho_i = 1 \wedge \left\{ \frac{\frac{f_i(\mathbf{y}_i|\mathbf{x}_{1,t+1},\ldots,\ldots,\mathbf{x}_{i-1,t+1},\mathbf{x}_{i+1,t},\ldots,\mathbf{x}_{p,t})}{q_i(\mathbf{y}_i|\mathbf{x}_{1,t+1},\ldots,\mathbf{x}_{i-1,t+1},\mathbf{x}_{i+1,t},\ldots,\mathbf{x}_{p,t})}}{\frac{f_i(\mathbf{x}_{i,t}|\mathbf{x}_{1,t+1},\ldots,\ldots,\mathbf{x}_{i-1,t+1},\mathbf{x}_{i+1,t},\ldots,\mathbf{x}_{p,t})}{q_i(\mathbf{x}_{i,t}|\mathbf{x}_{1,t+1},\ldots,\mathbf{x}_{i-1,t+1},\mathbf{x}_{i+1,t},\ldots,\mathbf{x}_{p,t})}} \right\}$$

The algorithm proceeds doing a single step of the Metropolis–Hastings algorithm for each full conditional that is difficult to sample from directly. The Metropolis–Hastings step isn't approximating the full conditional distribution $f_i(\mathbf{x}_i|\mathbf{x}_{-i})$, in fact, only a single step is used. Further details are given in Robert and Casella (2004) and Chen and Schmeiser (1998).

Müller (1993) suggests a further modification of the MHWG algorithm by running a single acceptance step after each of the p conditionals are simulated. This can be more time consuming as the full update may be rejected, however, it can be written as a simple Metropolis–Hastings algorithm. It also has the added advantage of producing a global approximation to the distribution of interest rather than local approximations of the conditional distributions.

3.6.1 Theoretical Details

The theoretical details of MCMC algorithms have received considerable attention over the past 20 years and it is beyond the scope of the current work to summarize these results. Rather we give some brief results and direct the reader to several classic references.

For many Markov chains much effort is spent deriving the stationary distribution of the Markov chain induced by a given transition kernel. In MCMC applications, by contrast, the stationary distribution is known and we construct transition kernels, that have desirable properties. The transition kernels are typically chosen from some small subset of algorithms that we have discussed. The challenge in MCMC applications is showing that the Markov chain induced by a particular kernel has certain convergence properties. The most important being, an ergodic theorem and a central limit theorem. The ergodic theorem is a generalization of the law of large numbers to dependence situations, and guarantees the convergence of sample averages to the required expectations.

Meyn and Tweedie (2009) and Nummelin (1984, 2002) give many of the required technical conditions and Robert and Casella (2004) summarize the area effectively and gives numerous references to other technical results. Numerous other authors give results for particular cases, most notably for our purposes, Roberts and Rosenthal (2006) show that Metropolis–within–Gibbs algorithms are Harris recurrent, a key ingredient to establish the various forms of ergodicity.

3.7 Adaptive MCMC

Examples 3.1 and 3.2 showed quite clearly that the properties of Markov chain resulting from a given Metropolis–Hastings algorithm depend greatly on having chosen the correct scale for the proposal distribution. The choice of the correct scale can become quite daunting in higher dimensional problems as there are a number of scale parameters to choose. Roberts and Rosenthal (2001) showed that the optimal scale factors depend on the shape of the target distribution. Effectively, a posterior modal approximation is carried out and the resulting Hessian is used to approximate the covariance matrix. However, modal approximations can be difficult, if not impossible, to carry out in high–dimensional problems.

Adaptive algorithms, in other words algorithms that learn from the past states they have visited, have a relatively long history with the first ones suggested by Gelfand and Sahu (1994); Gilks et al. (1994). Gilks et al. (1998) considered updating the Metropolis–Hastings at the regeneration times of the chain to avoid the potential problem of updating too frequently as Gelfand and Sahu (1994) caution against. Regeneration times, for discrete state spaces, occur when the chain returns to a state it has already visited. At the regeneration times the Markov chain can be thought of as starting over probabilistically. Nummelin (1984) and numerous others have generalized the concept of regeneration times to general state spaces, however, we did not pursue this method as they become difficult to find in high dimensional problems. Gilks et al. (1998) give several examples illustrating the theoretical basis of their method.

The two articles by Haario et al. (1999, 2001) sparked considerable interest in the area of adaptive algorithms. They considered random walk Metropolis–Hastings algorithms with multivariate normal proposal distributions, thus, the proposal distribution is completely determined by the covariance matrix. They suggested that the proposal covariance could be updated using the past values of the states. On the surface, this violates the Markovian nature of the chain, as the proposal variance depends on the complete past and not just the most immediate past.

Andrieu and Thoms (2008) give an excellent review of the theory and algorithms for adaptive MCMC and Roberts and Rosenthal (2009) give several examples of adaptive MCMC. To guarantee the asymptotically validity of adaptive algorithms a number of technical conditions must be met. The most important is the so called vanishing or diminishing adaption condition. Specifically, as stated in Roberts and Rosenthal (2009), the diminishing adaption condition is

$$\lim_{n\to\infty} \sup_{x\in\mathcal{X}} ||K_{\Gamma_{n+1}}(x,.) - K_{\Gamma_n}|| = 0 \text{ in probability}$$

where $||.||$ represents the total variance norm, $\mathcal{X}$ is the state space and K_{Γ_n} is the transition kernel from X_n to X_{n+1}, with the proviso that each of the transition kernels has the stationary distribution of interest. Note that K_{Γ_n} depends on the past states of the Markov chain. The diminishing adaption condition states that eventually the algorithm does not depend on the immediate past. Essentially, the adaptive algorithm behaves as though it was generated from a non-adaptive algorithm in the limit, roughly speaking.

Andrieu and Thoms (2008) give several adaptive algorithms which are variants of the Haario et al. (2001) adaptive metropolis algorithm. To fix ideas consider the following algorithm

Algorithm 1 Base Adaptive Algorithm

```
Input: Initialize $X_0$ $\mu_0$ and $\Sigma_0$
Output: An adaptive Markov chain $X_i, i = 1, \ldots, n$
At iteration $i+1$, given $X_i$, $\mu_i$ and $\Sigma_i$
Sample $X_{i+1} \sim q^{SRWM}_{\mu_i,\Sigma_i}(X_i, .)$
Update
    $\mu_{i+1} = \mu_i + \gamma_{i+1}(X_i - \mu_i)$
    $\Sigma_{i+1} = \Sigma_i + \gamma_{i+1}\Big((X_{i+1} - \mu_i)(X_{i+1} - \mu_i)' - \Sigma_i\Big)$
```

Andrieu and Thoms (2008) give guidelines for how one should chose, γ_i, the amount of adaption in the algorithm. We shall only discuss the deterministic choice of step sizes, specifically they recommend sequences γ_i that meet the following two conditions

$$\sum_{i=1}^{\infty} \gamma_i = \infty$$

and

$$\sum_{i=1}^{\infty} \gamma_i^{1+\lambda} < \infty$$

for some $\lambda > 0$. That is, the sequence should go to zero, but not too fast. They state that the first condition ensures that any point of Θ can be reached while the second ensures that the noise is contained and does not prevent convergence of the algorithm. They suggest using $\gamma_i = C/i^{\alpha}$ for $\alpha \in [(1+\lambda)^{-1}, 1]$. We follow their advice as well as advice given in Andrieu and Robert (2001) and chose $\alpha = 0.7$, however, very little guidance is given on the choice of the constant C, and take it as 1.

Roberts and Rosenthal (2009) suggest using the following adaptive algorithm, which is more in line with the original proposal of Haario et al. (2001). Use this proposal when $n \leq 2d$

$$Q_n(x, .)N(x, (0.1)^2 I_d/d)$$

and the following when $m > 2d$

$$Q_n(x, .)(1-\beta)N(x, 2.38^2 \Sigma_n/d) + \beta N(x, (0.1)^2 I_d/d).$$

Σ_n is then updated using recursions very similar to algorithm 1 with $\gamma_{i+1} = \frac{1}{n+1}$.

We give details of the adaptive algorithm used in the application chapter 6.

3.8 Convergence Diagnostics

A key aspect of any Markov Chain Monte Carlo run is determining if the chain has reached its stationary distribution and whether therefore one can use the samples of the chain as samples from the desired posterior distribution. As discussed in the previous chapter, there are several theoretical results that show that the MCMC's are ergodic and have central limit theorems. However, these methods are typically very conservative, and the conditions for the theorems are difficult to establish with reasonably complex models.

There are numerous methods that authors have proposed to assess the empirical convergence of MCMC output. However, they all leave us in an unsatisfactory position, in that, they can only tell us if the chain has not converged but not whether it actually has converged.

There are well over 30 diagnostic tests that have been suggested since the early 1980's. Note some methods were originally employed for other purposes, which were discovered before the explosion of MCMC methods, then found new uses in the MCMC literature. Additionally, there have been at least three review articles: Cowles and Carlin (1996), followed by Brooks and Roberts (1998) and most recently Mengersen et al. (1999). Robert and Casella (2004) include a comprehensive review chapter in their book on Monte Carlo methods. We do not attempt to conduct another review, however, we do propose a new method of assessing convergence in the next chapter based on evolutionary spectra theory and therefore present some of the basic information on empirical convergence diagnostics.

Robert and Casella (2004) categorize the convergence diagnostics into two broad categories: monitoring the convergence to the stationary distribution and monitoring the convergence of averages. Essentially the first class of methods considers when the chain has forgotten its initial conditions and has reached the stationary distribution. The second aspects considers how long we should run the chain once we have reached the stationary distribution to get reasonably good estimates of posterior moments, for example. Of course, this broad generalization is a gross simplification of the individual methods.

The test we propose, in the next chapter, would naturally fit in the first category, that is, determining if there is any evidence that the chain has not reached its stationary distribution as yet. In general these methods use time series based methods to determine if there is any evidence that the chain has not reached its stationary distribution. However, other methods are based on renewal theory and small sets which can be difficult to implement in large dimensional problems, as we consider here. Other methods are based on distance calculations but these are also difficult to implement in practice. One method that stands out is the so called missing mass method (Robert and Casella, 2004), which in theory, is ideally what we should strive for, however, it is also difficult to implement with large dimensional problems.

The second class of estimates are also useful and we employed several of the standard methods available in the R package, CODA (Plummer et al., 2006) to assess the convergence of averages.

CHAPTER 4

A TEST FOR STATIONARITY: ASSESSING MCMC CONVERGENCE

4.1 Univariate Test of Stationarity

Priestley and Subba Rao (1969) proposed a test for stationarity of a time series employing the so called evolutionary spectrum. Subba Rao and Tong (1972) consider a variation of this test for time–dependence of the transfer function in open loop systems. We propose using this test for monitoring the convergence of MCMC output to determine if the stationarity distribution of the chain has been reached. We also propose a multivariate generalization of the test for monitoring several parameters at once.

We begin with a short review of stationarity and spectral representations, taken from Priestley (1988), for more details the reader is referred to classic texts of Priestley (1981) or Brillinger (1981).

Let $\{X(t)\}$ be a real valued process (time series) measured at discrete time intervals ($t = 0, \pm 1, \pm 2, \ldots$). For example, $\{X(t)\}$ could represent the hourly temperature measured at a recording station. Let $\mu(t)$ represent the mean of the process at time t, that is,

$$\mu(t) = E[X(t)]$$

where the expectation is across all potential realizations of the process at time t. Similarly, we can define the variance of the process

$$\sigma^2(t) = E[(X(t) - \mu(t))^2]$$

and the auto–covariance function

$$c_{XX}(t + u, t) = E[(X(t + u) - \mu(t + u))(X(t) - \mu(t))].$$

With a single realization of the process, as is typical in most applications, we don't have enough information to estimate the time varying properties of the process. Thus, we make some further restrictions on the process, specifically consider the following definition:

Definition 4.1 (Strict Stationarity)**.** *The process $\{X(t)\}$ is to be strictly(completely) stationary if, for any admissible $t_1, t_2, \ldots, t_n$ and any k, the joint distribution of $\{X(t_1), X(t_2), \ldots, X(t_n)\}$ is the same as the joint probability distribution of $\{X(t_1 + k), X(t_2 + k), \ldots, X(t_n + k)\}$.*

A weaker form of stationary, which doesn't require equivalence in distribution is given in the following definition:

Definition 4.2 (Second Order Stationarity)**.** *The process $\{X(t)\}$ is said to be second order stationary (Weakly Stationary) if, for any admissible $t_1, t_2, \ldots, t_n$ and any k, all the joint second order moments of $\{X(t_1), X(t_2), \ldots, X(t_n)\}$ exist and equal the corresponding moments of $\{X(t_1+k), X(t_2+k), \ldots, X(t_n+k)\}$.*

Or equivalently:

- $\mu(t) = \mu \ \forall t$,
- $\sigma^2(t) = \sigma^2 \ \forall t$,
- $c_{XX}(t+u, t) = c_{XX}(u) \ \forall t$.

In other words,none of the first or second order moments of the distribution depend on the time origin.

Note that weakly stationary processes are not necessarily strictly stationary, however, the reverse is true.

We can establish the following properties of the auto–covariance function, $c_{XX}(u)$, of a second order stationary process

- $c_{XX}(0) = \sigma^2$.
- $|c_{XX}(u)| \leq c_{XX}(0), \forall u$.
- $c_{XX}(u) = c_{XX}(-u)$, provided $\{X(t)\}$ is real valued.
- $c_{XX}(u)$ is positive semi–definite meaning that for any set of points $t_1, t_2, \ldots, t_n$ and all real $z_1, z_2, \ldots, z_n$

$$\sum_{r=1}^{n}\sum_{s=1}^{n} c_{XX}(t_r - t_s) z_r z_s \geq 0.$$

4.2 Spectral Analysis of Univariate Process

Brillinger (1981) defines the power spectral density of a weakly stationary process $\{X(t)\}$ with auto–covariance function $c_{XX}(u)$ as follows:

Definition 4.3. *The power spectral density function, denoted by $h(\omega)$ is defined as the discrete Fourier transform of the auto-covariance function $c_{XX}(u)$ as follows*

$$h_{XX}(\omega) = \frac{1}{2\pi} \sum_{u=-\infty}^{\infty} e^{-i\omega u} c_{XX}(u), \quad -\pi \leq \omega \leq \pi,$$

provided that $\sum_{u=-\infty}^{\infty} |c_{XX}(u)| < \infty$.

Priestley (1981) defines the power spectral density to be

$$h(\omega) = \lim_{T\to\infty} \left[E \left\{ \frac{|G_T(\omega)|^2}{2T} \right\} \right]$$

where,

$$G_T(\omega) = \frac{1}{\sqrt{2\pi}} \sum_{t=-T}^{T} X(t) e^{-i\omega t}$$

$G_T(\omega)$ is the discrete Fourier transform of the process $X(t)$. Priestley (1981) gives an excellent interpretation of the spectral density $h(\omega)d\omega$ as the average (over all realizations) of the contribution of the total power from components in $X(t)$ with frequencies between ω and $\omega + d\omega$.

Also, consider the inverse Fourier Transform of $h_{XX}(\omega)$, that is, we can write

$$c_{XX}(u) = \int_{\pi}^{\pi} e^{i\omega u} h_{XX}(\omega) d\omega \quad u = 0, \pm 1, \pm 2, \ldots$$

Thus, the auto–covariance function of a stationary process can be written as the Fourier transform of the power spectral density function $h(\omega)$.

The Wiener–Khintchine theorem (Priestley, 1981) allows one to write $c_{XX}(u)$ as a generalized Fourier transform, even when $c_{XX}(u)$ does not decay fast enough for $h(\omega)$ to exist, as follows:

$$c_{XX}(u) = \int_{-\pi}^{\pi} e^{i\omega u} dH(\omega), \tag{4.1}$$

where $H(\omega)$ is a non–decreasing function with $H(-\pi) = 0$, $H(\pi) = \sigma^2$ and is called the integrated spectrum. If we invert the above equation (Priestley, 1988), we have:

$$H(\omega) = \sigma^2 \left(\frac{\omega + \pi}{2\pi} \right) + \frac{1}{2\pi} \left[\sum_{u=-\infty}^{-1} + \sum_{u=1}^{\infty} \right] \frac{e^{-iu\omega}}{-iu} c_{XX}(u).$$

When $h(\omega)$ exists we have $dH(\omega) = h(\omega)d\omega$ and

$$H(\omega) = \int_{-\pi}^{\omega} h(\theta)d\theta$$

There is a spectral representation for $\{X(t)\}$, as for $c_{XX}(u)$ given in (4.1), when $E[X(t)] = 0$, given by

$$X(t) = \int_{-\pi}^{\pi} e^{i\omega t} dZ(\omega), \quad t = 0, \pm 1, \pm 2, \ldots, \tag{4.2}$$

where $Z(\omega)$ is a (complex–valued) stochastic process with orthogonal increments, *i.e.*

$$E[dZ(\omega)dZ^*(\omega^{'})] = 0, \omega \neq \omega^{'}, \tag{4.3}$$

where (*) denotes complex conjugate.

It is instructive to consider the following. Define the auto–covariance function for complex–valued series by

$$c_{XX}(u) = E[X(t)^* X(t+u)].$$

and substitute the spectral representation for $X(t)$ (equation 4.2) into the above definition,

$$\begin{aligned} c_{XX}(u) &= E\left[\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} e^{-i\omega t} e^{i\omega(t+u)} dZ^*(\omega)dZ(\omega^{'})\right] \\ &= \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} e^{-i\omega t} e^{i\omega^{'}(t+u)} E\left[dZ^*(\omega)dZ(\omega^{'})\right]. \end{aligned}$$

Since, $\{X(t)\}$ is a stationary process, $c_{XX}(u)$ is a function of u only, thus the right hand side must be a function of u only. This is guaranteed by the condition given in equation 4.3. Thus, we now have

$$c_{XX}(u) = \int_{-\pi}^{\pi} e^{i\omega u} E\left[dZ^*(\omega)dZ(\omega)\right],$$

which gives the following relationship between $dZ(\omega)$ and $H(\omega)$ by equating the above equation with equation (4.1) as given below

$$E[|dZ(\omega)|^2] = dH(\omega) \tag{4.4}$$

Equation (4.2) is known as the spectral representation of $\{X(t)\}$ (see Brillinger, 1981;

Priestley, 1981, for more details). The spectral representation is a key result in that any stationary discrete parameter process $\{X(t)\}$ can be represented as a "sum" of sines and cosines over a continuous range of frequencies $(-\pi, \pi)$, with random amplitudes, $|dZ(\omega)|$, and phases, arg$\{dZ(\omega)\}$. Equation (4.4) provides us with a rough physical interpretation of $dH(\omega)$ as follows: $dH(\omega)$ is the mean–squared amplitude of the component of $\{X(t)\}$ with frequency ω. If we think of $\{X(t)\}$ as representing a stationary physical process, $H(\omega)$ is a measure of the average total power, where the averaging is done across realizations of the process. Additionally, $dH(\omega)(= h(\omega)d\omega)$ represents the average contribution to the total power from the components of $\{X(t)\}$ with frequencies between $\omega, \omega + d\omega$. Thus, $h(\omega)$, the power spectral density, represents the distribution of power over frequency.

4.2.1 Estimation of Univariate Spectra

Given a sample record of our discrete parameter process $\{X(t)\}$, measured at time points $t = 0, \ldots, T-1$, how does one estimate the power spectral density? Consider the finite Fourier transform, $d_X^{(T)}(\omega)$, given by

$$d_X^{(T)}(\omega) = \sum_{t=0}^{T-1} e^{-i\omega t} X(t).$$

Theorem 4.4.2 of Brillinger (1981) gives the following asymptotic sampling properties of $d_X^{(T)}(\omega)$

$$\begin{array}{rl} N_1^C(0, 2\pi T h_{XX}(\omega)) & \text{if } \omega \neq 0 \pmod{\pi} \\ N_1(T\mu, 2\pi T h_{XX}(\omega)) & \text{if } \omega = 0, \pm 2\pi, \ldots \\ N_1(0, 2\pi T h_{XX}(\omega)) & \text{if } \omega = \pm\pi, \pm, 3\pi, \ldots \end{array}$$

where μ is the mean of the series, N_1^C denotes a complex valued normal distribution and N_1 indicates a one dimensional normal distribution. This leads us to consider the following estimate of the spectrum $h_{XX}(\omega)$, known as the periodogram

$$\begin{aligned} I_{XX}^{(T)}(\omega) &= \frac{1}{2\pi T} |d_X^{(T)}(\omega)|^2 && (4.5) \\ &= \frac{1}{2\pi T} | \sum_{t=0}^{T-1} e^{-i\omega t} X(t)|^2 && (4.6) \end{aligned}$$

Brillinger (1981) shows that

$$E[I_{XX}^{(T)}(\omega)] = h_{xx}(\omega) + \frac{1}{2\pi T}\left[\frac{sinT\omega/2}{sin\omega/2}\right]\mu_x^2 + O(T^{-1})$$

where the $O(T^{-1})$ term is uniform in ω. Thus as $T \to \infty$ the second and third terms vanish, leading to the result that $I_{XX}^{(T)}(\omega)$ is asymptotically unbiased, provided $\omega \neq 0 \pmod{2\pi}$

Consider the following theorem given in Brillinger (1981)

Theorem 4.1. *Let $X(t), t = 0, \pm 1, \pm 2, \ldots,$ be a real valued series and let $I_{XX}^T(\omega)$ be the periodogram defined by equation (4.5). Let r, s be integers with $r, s, r \pm s \neq 0 \pmod{T}$.Let $\omega = 2\pi r/T$, $\lambda = 2\pi s/T$. Then*

$$\begin{aligned} var[I_{XX}^T(\omega)] &= h_{XX}(\omega)^2 + O(T^{-1}) \\ cov[I_{XX}^T(\omega), I_{XX}^T(\lambda)] &= O(T^{-1}) \end{aligned}$$

where the $O(T^{-1})$ terms are uniform in ω, λ of the indicated form.

Although, the periodogram is asymptotically unbiased, it is not consistent as the variance of the estimator does not vanish as $T \to \infty$. Thus, we cannot improve our estimates by taking a larger sample of the time series.

Also, of importance is the following theorem from Brillinger (1981)

Theorem 4.2. *Let $X(t), t = 0, \pm 1, \pm 2, \ldots,$ be a real valued series. Let $s_j(T)$ be an integer with $\omega_j(T) = 2\pi s_j(T)/T$ tending to ω_j as $T \to \infty$ for $j = 1, \ldots, J$. Suppose $2\omega_j(T), \omega_j(T) \pm \omega_k(T) \neq 0 \pmod{2\pi}$ for $1 \leq j < k \leq J$ and $T = 1, 2, \ldots$. Let $I_{XX}^T(\omega)$ be the periodogram defined by equation (4.5). Then $I_{XX}^T(\omega_j(T)), j = 1, \ldots, J$ are asymptotically independent $h_{XX}(\omega_j)\chi_2^2/2$ variates. Also, if $\omega = \pm\pi, \pm 3\pi, \ldots$, $I_{XX}^T(\omega)$ is asymptotically $h_{XX}(\omega)\chi_1^2$ independently of the previous variates.*

Following Brillinger (1981), let $s(T)$ be an integer with $2\pi s(T)/T$ near the frequency of interest ω. We have by the previous theorem, that the $(2m+1)$ adjacent periodogram ordinates $I_{XX}^T(2\pi(s(T)+j)/T)$, $j = 0, \pm 1, \ldots, \pm m$ are approximately independent $h_{XX}(\omega)\chi_2^2/2$ variates, if $2[s(T)+j] \neq 0 \pmod{T}, j = 1, \pm 1, \ldots, \pm m$. Thus, these $(2m+1)$ values all provide estimates of $h_{XX}(\omega)$, giving rise to an estimator of the following form

$$h_{XX}^{(T)}(\omega) = (2m+1)^{-1} \sum_{j=-m}^{m} I_{XX}^{T} \left(\frac{2\pi[s(T)+j]}{T} \right) \text{ if } \omega \neq 0 \pmod{\pi}.$$

We also need to consider the following special cases. Firstly, consider $\omega = 0, \pm 2\pi, \pm 4\pi, \ldots$ or if $\omega = 0, \pm\pi, \pm 3\pi, \ldots$ and T is even

$$h_{XX}^{(T)}(\omega) = m^{-1} \sum_{j=1}^{m} I_{XX}^{T} \left(\omega + \frac{2\pi + j}{T} \right),$$

and finally if $\omega = 0, \pm\pi, \pm 3\pi, \ldots$ and T is odd,

$$h_{XX}^{(T)}(\omega) = m^{-1} \sum_{j=1}^{m} I_{XX}^{T} \left(\omega - \frac{\pi}{T} + \frac{2\pi + j}{T} \right).$$

Brillinger (1981) shows that this estimate is also asymptotically unbiased and more importantly, the asymptotic variance of our estimator is given by

$$\begin{aligned} \text{var}\left(h_{XX}^{(T)}(\omega) \right) &= \frac{h_{XX}(\omega)^2}{2m+1} + O(T^{-1}) \text{ if } \omega \neq 0 \pmod{\pi} \\ &= \frac{h_{XX}(\omega)^2}{m} + O(T^{-1}) \text{ if } \omega \equiv 0 \pmod{\pi} \end{aligned}$$

and the covariance between spectral ordinates is asymptotically zero. Also, note that if we look at the variance of the $\log$ spectral estimate Brillinger (1981) shows via the delta method that

$$\text{var}\left(\log h_{XX}^{(T)}(\omega) \right) = \frac{1}{2m+1} \text{ if } \omega \neq 0 \pmod{\pi}$$

Brillinger (1981) considers the general class of consistent estimators

$$h_{XX}^{(T)}(\omega) = \frac{2\pi}{T} \sum_{s=1}^{T-1} W^{(T)} \left(\omega - \frac{2\pi s}{T} \right) I_{XX}^{(T)} \left(\frac{2\pi s}{T} \right)$$

where $W^{(T)}(\alpha), -\infty < \alpha < \infty, T = 0, 1, 2, \ldots$ is a family of weight functions that weight $2m_T + 1$ periodogram ordinates in the vicinity of ω. In order for variance of this estimator to diminish as $T \to \infty$, we also require $m_t \to \infty$. That is, we require an increasing number of periodogram ordinates to be averaged as $T \to \infty$. However, to remain unbiased, we also require $m_T/T \to 0$ as $T \to \infty$.

Brillinger (1981) considers a sequence of scale parameters $B_T, T = 1, 2, \ldots$ with the properties $B_T > 0$, $B_T \to 0$, $B_T T \to \infty$ as $T \to \infty$ and sets

$$W^{(T)}(\alpha) = \sum_{j=-\infty}^{\infty} B_T^{-1} W(B_T^{-1}[\alpha + 2\pi j]) \quad -\infty < \alpha < \infty,$$

where $W(\beta), -\infty < \beta < \infty$ is a fixed function satisfying

$$\int_{-\infty}^{\infty} W(\beta) d\beta = 1$$

and

$$\int_{-\infty}^{\infty} |W(\beta)| d\beta < \infty$$

If $W(\beta)$ is zero for $|\beta| > 2\pi$, then we can see that the estimate is a weighted average of $2B_T T + 1$ periodogram ordinates in the range $(\omega - 2\pi B_T, \omega + 2\pi B_T)$ and we make the connection $m_T = B_T T$.

Various choices for window, $W^{(T)}(\alpha)$, are available such as the Bartlett, Parzen, Tukey, Tukey–Hamming to name a few. However, for our purposes we consider the Daniel window, which weights each of the $2m + 1$ periodogram ordinates equally.

Tapering and pre–whitening are methods that are routinely used to improve spectral estimates, the reader is referred to Brillinger (1981) and Bloomfield (2000); Priestley (1981) for excellent descriptions. Pre-whitening consists, typically, of finding best fitting auto-regressive process (the order chosen by AIC) and the spectrum of the residuals from the best fitting auto–regressive is estimated using the methods already described. The spectral density of an auto–regressive process is completely determined by its coefficients. And using the properties of the spectral representation of stationary time series the "recolored" spectrum can then be reconstructed.

Tapering is a method to reduce leakage of the power from one frequency band to another. This happens when the spectra has several peaks and through the averaging process power from one of the peaks tends to leak into neighboring frequency bands. The split cosine bell window is the most commonly used taper and consists of the following

$$w_p(x) = \begin{cases} \frac{1}{2}(1 - \cos 2\pi x/p), & 0 \leq x < p/2, \\ 1, & p/2 \leq x < 1 - p/2, \\ \frac{1}{2}(1 - \cos 2\pi x/p), & 1 - p/2 \leq x \leq 1 \end{cases}$$

where $x = t/T$ and p is the proportion of data that is tapered, we use $p = 10\%$. Brillinger (1981); Bloomfield (2000) give excellent descriptions of tapering and its effects on the asymptotic variance of the resulting estimator. Tapers are also known as data windows which we employ in the estimation of the evolutionary spectra given below.

4.3 Evolutionary Spectra

Spectral theory decomposes stationary processes into combinations of sines and cosines. These functions also give us the physical concepts of frequency and amplitude or energy. However, the complex exponentials are themselves stationary and it is not surprising that these functions do not allow one to represent a non–stationary process. Thus the question becomes, is there a similar concept to the spectral representation theorem for non–stationary processes? Priestley (1965) introduces the concept of evolutionary spectra, which extends the spectral ideas to non–stationary processes, by using different non–stationary basis functions. Priestley (1996) discusses the relevance of an alternative set of basis functions to those considered here, namely wavelets.

Following Priestley (1988), let $\{X(t),\ t = 0, \pm 1, \pm 2, \ldots\}$ be a discrete parameter stochastic process, with

$$\begin{aligned} E[X(t)] &= 0 \\ E[|X(t)|^2] &< \infty, \forall t \\ c_{XX}(s,t) &= E[X^*(s)X(t)] \end{aligned}$$

If $c_{XX}(s,t)$ is function of $u = |s - t|$ only, then we have representations as before given by equations (4.1) and (4.2). If this is not the case, then these representations do not apply, however, as given in Priestley (1981), we can represent $c_{XX}(s,t)$ by a class of expansions called "general orthogonal expansions", provided the usual complex exponentials are replaced by a more general family of functions $\{\phi_t(\omega)\}$.

Let us restrict our attention to non–stationary processes that have covariance functions $c_{XX}(s,t)$ that can be represented in the following way,

$$c_{XX}(s,t) = \int_{\pi}^{\pi} \phi_s^*(\omega)\phi_t(\omega)d\mu(\omega) \tag{4.7}$$

where the measure $d\mu(\omega)$ is defined on the real line.

For the variance of $\{X(t)\}$ to be finite, $\phi_t(\omega)$ must be mean squared integrable with respect to μ for each t. Priestley (1981) shows that by the general orthogonal expansion theory, when the representation $c_{XX}(s,t)$ given by equation (4.7), the underlying process $\{X(t)\}$ has a representation given by

$$X(t) = \int_{-\pi}^{\pi} \phi_t(\omega)dZ(\omega)$$

where $Z(\omega)$ is an orthogonal increment process with $E[|dZ(\omega)|^2] = d\mu(\omega)$. Thus the measure $\mu(\omega)$ plays an analogous role as the integrated spectrum for stationary processes.

We further restrict our attention to families $\phi_t(\omega)$ that have the following form

$$\phi_t(\omega) = A_t(\omega)e^{i\theta(\omega)t}$$

where we assume that for a given ω, $\phi_t(\omega)$ has a generalized Fourier transform whose modulus has an absolute maximum at frequency $\theta(\omega)$ and thus $\phi_t(\omega)$ behaves like an amplitude modulated sine wave. Also, the modulating function $A_t(\omega)$ is assumed to have an absolute maximum at zero frequency, that is, it is slowly varying. Priestley (1981) calls $\phi_t(\omega)$ an oscillatory function if it can be written as above and if $A_t(\omega)$ is given by

$$A_t(\omega) = \int_{-\pi}^{\pi} e^{it\theta}dK_\omega(\theta)$$

with $|dK_\omega(\theta)|$ having an absolute maximum at $\theta = 0$. Provided we satisfy some further technical conditions (see Priestley (1988)), we can now write,

$$c_{XX}(s,t) = \int_{-\pi}^{\pi} A_s^*(\omega)A_t(\omega)e^{i\omega(t-s)}d\mu(\omega)$$

and a spectral representation is given by

$$X(t) = \int_{-\pi}^{\pi} A_t(\omega) e^{i\omega t} dZ(\omega), \quad t = 0, \pm 1, \pm 2, \ldots, \tag{4.8}$$

where

$$E\left[|dZ(\omega)|^2\right] = d\mu(\omega)$$

Any process which admits such a representation will be call an oscillatory process.

If $\{X(t)\}$ admits a representation given in equation (4.8) for the family of functions given by $\mathcal{F} = \{\phi_t(\omega)\} = \{A_t(\omega)e^{i\omega t}\}$, the evolutionary power spectrum at time t with respect to the family $\mathcal{F}$ is

$$dH_t(\omega) = |A_t(\omega)|^2 d\mu(\omega), \quad -\pi \leq \omega \leq \pi.$$

If the measure $d\mu(\omega)$ is absolutely continuous with respect to Lebesgue measure, as we are assuming throughout, we have

$$dH_t(\omega) = h_t(\omega) d\omega$$

where $h_t(\omega)$, exists for all $\omega \in (-\pi, \pi)$, and is called the evolutionary spectral density function.

Priestley (1988) goes into detail about the properties of the families, which are not directly applicable to the current discourse.

4.4 Evolutionary Power Spectra Estimation

For estimation of evolutionary power spectra, Priestley (1965) suggests using a so called "double–window" technique, which can be described as follows. Let $\{X(t)\}, t = 0, \ldots, T-1$ represent a sample from a discrete parameter process. Choose a filter (or window) $\{g(u)\}$ which is square summable and normalized such that

$$2\pi \sum_{u=-\infty}^{\infty} |g(u)|^2 = \int_{-\pi}^{\pi} |\Gamma(\omega)|^2 d\omega = 1,$$

where,

$$\Gamma(\omega) = \sum_{u=-\infty}^{\infty} g(u)e^{-iu\omega}$$

is the transfer function of the filter $\{g(u)\}$. Now, for any given frequency, ω, we define

$$U(t, \omega) = \sum_{u=-\infty}^{\infty} g(u)X(t-u)e^{-i\omega(t-u)}.$$

this is the complex demodulate of the series $\{X(t)\}$, see Bloomfield (2000)

We then choose a second filter (window), $w_{T'}(t)$, that satisfies the following conditions

1. $w_{T'}(t) \geq 0 \; \forall t, T^{'}$,

2. $w_{T'}(t)$ decays to zero as $|t| \to \infty, \forall T^{'}$,

3. $\sum_{t=-\infty}^{\infty} w_{T'}(t) = 1, \forall T^{'}$,

4. $\sum_{t=-\infty}^{\infty} w_{T'},(t)^2 < \infty, \forall T^{'}$,

5. there exists a constant C such that

$$\lim_{T^{'} \to \infty} \left\{ T^{'} \int_{-\pi}^{\pi} |W_{T'}(\lambda)|^2 d\lambda \right\} = C,$$

where

$$W_{T'}(\theta) = \sum_{u=-\infty}^{\infty} e^{-i\theta t} w_{T'}(u).$$

Then we may estimate $h_t(\omega)$ by

$$\hat{h}_t(\omega) = \sum_{v=-\infty}^{\infty} w_{T'}(v)|U(t-v, \omega)|^2.$$

Priestley (1965) derives the approximate mean of this estimator

$$E[\hat{h}_t(\omega)] \approx \int_{-\pi}^{\pi} \bar{h}_t(\omega + \theta)|\Gamma(\theta)|^2 d\theta$$

where

$$\bar{h}_t(\omega + \theta) = \sum_{u=-\infty}^{\infty} W_{T'}(u)h_{t-u}(\omega + \theta).$$

Priestley (1965) also derives the sampling variance of $\hat{h}_t(\omega)$

$$\text{var}(\hat{h}_t(\omega)) \approx \tilde{h}_t^2(\omega)\left\{\int_{-\pi}^{\pi}|W_{T'}(\theta)|^2 d\theta\right\}\left\{\int_{-\pi}^{\pi}|\Gamma(\theta)|^4 d\omega\right\}(1+\delta_{0,\pm\pi}(\omega))$$

where

$$\tilde{h}_t^2(\omega) = \frac{\sum_{u=-\infty}^{\infty} h_{t-u}^2(\omega) W_{T'}(u)^2}{\sum_{u=-\infty}^{\infty} W_{T'}(u)^2},$$

and

$$\delta_{0,\pm\pi}(\omega) = \begin{cases} 1 & \text{if } \omega = 0 \text{ or } \omega = \pm\pi \\ 0 & \text{otherwise} \end{cases}$$

We can further simplify the expression for the variance by making use of the limit given in assumption 5 above to give.

$$\text{var}(\hat{h}_t(\omega)) \approx (C/T^{'})\tilde{h}_t^2(\omega)\left\{\int_{-\pi}^{\pi}|\Gamma(\theta)|^4 d\omega\right\} \quad \omega \neq 0$$

Priestley and Subba Rao (1969) make one further approximation, by assuming that (i) if the "bandwidth" of $|\Gamma(\theta)|^2$ is small compared with the "frequency domain bandwidth" of $h_t(\omega)$ and (ii) if the "bandwidth" of $W_{T'}(u)$ is small compared with the "time–domain bandwidth" of $h_t(\omega)$, then it can be shown that

$$E\{\hat{h}_t(\omega)\} \approx h_t(\omega)$$

and

$$\text{var}(\hat{h}_t(\omega)) \approx (C/T^{'})h_t^2(\omega)\left\{\int_{-\pi}^{\pi}|\Gamma(\theta)|^4 d\omega\right\} \quad \omega \neq 0 \tag{4.9}$$

Priestley and Subba Rao (1969) give the following example of windows $\{g(u)\}$ and $W_T^{'}(t)$ in the continuous parameter process case,

$$g(u) = \begin{cases} \frac{1}{2\sqrt{h\pi}} & |u| \leq h \\ 0 & |u| > h \end{cases},$$

and

$$W_{T'}(t) = \begin{cases} \frac{1}{T^{'}}, & -\frac{1}{2}T^{'} \leq t \leq \frac{1}{2}T^{'}, \\ 0, & \text{otherwise} \end{cases}$$

which are both Daniel windows in the time domain.

Then

$$\begin{aligned}
\Gamma(\omega) &= \int_{-h}^{h} \frac{1}{2\sqrt{h\pi}} e^{-i\omega u} du \\
&= \frac{1}{2\sqrt{h\pi}} \int_{-h}^{h} e^{-i\omega u} du \\
&= \frac{1}{2\sqrt{h\pi}} \int_{-h}^{h} [\cos \omega u - i \sin \omega u] du \\
&= \frac{\sin h\omega}{\omega\sqrt{h\pi}}.
\end{aligned}$$

Therefore,

$$|\Gamma(\omega)|^2 = \frac{\sin^2 h\omega}{\pi h \omega^2}$$

which corresponds to the Bartlett frequency domain window. Also,

$$\begin{aligned}
\int_{-\infty}^{\infty} |\Gamma(\theta)|^4 d\theta &= \int_{-\infty}^{\infty} \frac{\sin^4 h\theta}{\pi^2 h^2 \theta^4} d\theta \\
&= \frac{1}{\pi^2 h^2} \int_{-\infty}^{\infty} \frac{\sin^4 h\theta}{\theta^4} d\theta \\
&= \frac{2h}{3\pi}
\end{aligned}$$

where the third line follows from the fact that $\int_{-\infty}^{\infty} \frac{sin^4 x}{x^4} dx = 2\pi/3$.

Then

$$\begin{aligned}
W_{T'}(\lambda) &= \int_{-\infty}^{\infty} e^{-i\lambda t} w_{T'}(t) dt \\
&= \frac{1}{T'} \int_{-T'}^{T'} e^{-i\lambda t} dt \\
&= \frac{1}{T'} \int_{-T'}^{T'} [\cos \lambda t - i \sin \lambda t] dt \\
&= \frac{2 \sin \lambda T'}{\lambda T'},
\end{aligned}$$

and

$$\begin{aligned}
\lim_{T'\to\infty}\left\{T^{'}\int_{-\infty}^{\infty}W_{T'}(\lambda)d\lambda\right\} &= \lim_{T'\to\infty}\left\{T^{'}\int_{-\infty}^{\infty}\frac{2sin^2\lambda T^{'}}{\lambda^2 T^{'2}}d\lambda\right\} \\
&= \lim_{T'\to\infty}\left\{\frac{1}{T^{'}}\int_{-\infty}^{\infty}\frac{2\sin^2\lambda T^{'}}{\lambda^2}d\lambda\right\} \\
&= \lim_{T'\to\infty}\left\{\frac{1}{T^{'2}}\int_{-\infty}^{\infty}\frac{2\sin^2 u}{(u/T^{'})^2}du\right\} \\
&= \lim_{T'\to\infty}\left\{\int_{-\infty}^{\infty}\frac{2\sin^2 u}{u^2}du\right\} \\
&= \int_{-\infty}^{\infty}\frac{2\sin^2 u}{u^2}du \\
&= 2\pi.
\end{aligned}$$

Plugging these two results into equation (4.9), gives

$$\text{var}\hat{h}_t(\omega) \approx \frac{4h}{3T^{'}}h_t^2(\omega)$$

Finally, Priestley (1966) gives expressions for the covariance between $\hat{h}_{t_1}(\omega_1)$ and $\hat{h}_{t_2}(\omega_2)$. However, for our purposes, we just need the following, that states when the covariance will be approximately 0 when .

- $|\omega_1 \pm \omega_2| \gg$ bandwidth of $|\Gamma(\omega)|^2$ or
- $|t_1 - t_2| \gg$ "width" of the function $W_{T'}(u)$

Priestley (1981), chapter 11, states that if $W_{T'}(u)$ has a rectangular form, then the double window technique is essentially equivalent to the "averaging of time blocks". For computational ease, we implemented the evolutionary spectral estimation in this manner. Specifically, we take non–overlapping blocks in the time domain and apply the spectral estimation methods mentioned in the subsection 4.2.1.

4.4.1 Form of the Test for Stationarity

Priestley and Subba Rao (1969) propose a test of stationarity based on the evolutionary spectral ideas presented in the previous section. The details of the test for a single time series are developed in the sequel and we will also generalize the test to multiple time series in the following section.

Specifically, let $X(t)$ denote the potentially non–stationary series of interest. We assume it comes from an oscillatory process as developed in section 4.3. Let $h_t(\omega)$ denote the evolutionary spectral density of the given time series and denote the estimate of this spectral density by $\hat{h}_t(\omega)$. As noted in Jenkins and Priestley (1957), the logarithmic transformation tends to stabilize the variance of the spectral estimates for a stationary process. We use this fact here and apply it to the locally stationary process.

We now let

$$Y_t(\omega) = \log_e \hat{h}_t(\omega).$$

Applying the delta method, it follows that

$$E[Y_t(\omega)] \approx \log_e h_t(\omega)$$

and

$$\text{var}Y_t(\omega) = \sigma^2(\omega \neq 0, \pi)$$

where

$$\sigma^2 = (C/T^{'}) \left\{ \int_{\infty}^{\infty} |\Gamma(\theta)|^4 d\theta \right\}$$

which does not depend on ω or t.

Consider writing the above in a slightly different form as follows,

$$Y_t(\omega) = \log_e h_t(\omega) + \epsilon_t(\omega)$$

where

$$\begin{aligned} E[\epsilon_t(\omega)] &\sim 0 \;\; \forall t, \omega \\ \text{var}[\epsilon_t(\omega)] &\sim \sigma^2 \;\; \forall t \text{ and } \omega \neq 0, \pi \end{aligned}$$

Now, assume that we have estimated the evolutionary spectrum $h_t(\omega)$, by computing $\hat{h}_t(\omega)$ from a sample from $t = 0, \ldots, T-1$. Now choose a set of times $t_1, t_2, \ldots, t_I$ and frequencies $\omega_1, \omega_2, \ldots, \omega_J$ which satisfy the conditions for zero covariance mentioned in the previous section. Now, if we let

$$\begin{aligned} Y_{ij} &= Y_{t_i}(\omega_j) \\ f_{ij} &= h_{t_i}(\omega_j) \\ \epsilon_{ij} &= \epsilon_{t_i}(\omega_j) \;\; i = 1, \ldots, I, \; j = 1, \ldots, J, \end{aligned}$$

we have a more traditional two–way analysis of variance model,

$$Y_{ij} = f_{ij} + \epsilon_{ij}, \;\; i = 1, \ldots, I, \; j = 1, \ldots, J.$$

Priestley and Subba Rao (1969) note that only the first two moments of the process have been used, however, they argue heuristically that the logarithmic transformation also makes the distribution approximately normal for stationary series and argue that a similar result should hold for the estimator $\hat{h}_t(\omega)$, at least approximately. They use this argument to justify that ϵ_{ij} has an approximately normal distribution. With this assumption the above two–way analysis of variance model is now complete with the the error distribution assumed to be approximately normal. Note that since the sampling variances $\hat{h}_t(0)$ and $\hat{h}_t(\pi)$ are equal to $2\sigma^2$, they should be omitted from the choice of frequencies, or an adjustment must be made to the usual analysis of variance.

Finally, the two–way analysis of variance is written more traditionally as

$$M : Y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ij} \;\; i = 1, \ldots, I, \; j = 1, \ldots, J.$$

Now, consider a stationary process $\{X(t)\}$. We know that the spectral density function does not depend on time t, therefore

$$E[\log_e \hat{h}_t(\omega)] \sim \log_e h(\omega)$$

Thus, the model for this process would be

$$M1 : Y_{ij} = \mu + \beta_j + \epsilon_{ij} \;\; i = 1, \ldots, I, \; j = 1, \ldots, J.$$

A comparison of models M and M_1 gives a test for stationarity of a process $\{X(t)\}$.

As a further test, consider the following model

$$M2 : Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \;\; i = 1, \ldots, I, \; j = 1, \ldots, J.$$

If we compare this to model M then we have a test for complete randomness, that is, a test that determines whether or not the spectra differs across frequency.

It is instructive to consider the interpretation of the interaction term γ_{ij}. Consider the following model

$$M3 : Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \;\; i = 1, \ldots, I, \; j = 1, \ldots, J,$$

where $\{\alpha_i\}$ and $\{\beta_j\}$ are the main effects of frequency and time respectively. This is a model which on the log scale is additive in terms of both frequency and time. Thus, we have

$$\begin{aligned} E[Y_{ij}] &= \mu + \alpha_i + \beta_j \\ \log_e h_t(\omega) &= \mu + \alpha_i + \beta_j \\ h_t(\omega) &= \exp\left(\mu + \alpha_i + \beta_j\right) \\ h_t(\omega) &= c_t^2 h(\omega) \end{aligned}$$

for some functions $c(t)$ and $h(\omega)$. Priestley and Subba Rao (1969) state that if $h_t(\omega)$ is of the form given above, then $\{X(t)\}$ must be of the form

$$X(t) = c(t) X_o(t)$$

where $X_o(t)$ is a stationary process with spectral density function $h(\omega)$. Such processes are known as uniformly modulated processes (Priestley, 1965). Therefore, the test for interaction corresponds to a test for whether or not the process under consideration is a uniformly modulated process.

To perform the tests mentioned above we construct the analysis of variance given in table 4.1 where as usual, the means are defined by

Item	Degrees of Freedom	Sum of Squares
Times	$I-1$	$SS_T = J\sum_{i=1}^{I}(\bar{Y}_{i.}-\bar{Y}_{..})^2$
Freqs	$J-1$	$SS_F = I\sum_{j=1}^{J}(\bar{Y}_{.j}-\bar{Y}_{..})^2$
Int+Res	$(I-1)(J-1)$	$SS_{I+R} = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-\bar{Y}_{i.}-\bar{Y}_{.j}+\bar{Y}_{..})^2$
Total	$IJ-1$	$SS_0 = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-\bar{Y}_{..})^2$

Table 4.1: Frequency by time analysis of variance

$$\begin{aligned}
\bar{Y}_{..} &= \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}Y_{ij}, \\
\bar{Y}_{i.} &= \frac{1}{J}\sum_{j=1}^{J}Y_{ij}, \\
\bar{Y}_{.j} &= \frac{1}{I}\sum_{i=1}^{J}Y_{ij}.
\end{aligned}$$

Priestley and Subba Rao (1969) then suggest the following testing order, which is analogous to the usual testing that is done in analysis of variance, with the caveat that σ^2 is known, and hence significance results are compared to χ^2 rather than F distributions.

1. First determine if the process is uniformly modulated, that is, does $\gamma_{ij} = 0 \forall i, j$. This is done using the following test statistic $SS_{I+R}/\sigma^2 \sim \chi_{(I-1)(J-1)^2}$. If this result is significant, we conclude the process is non–stationary and non–uniformly modulated and would then stop at this point.

2. If the result of step 1, is insignificant, we can conclude the process is uniformly modulated and then test for stationarity by testing $\beta_j = 0,\ \forall j$, using the test statistic $SS_T/\sigma^2 \sim \chi^2_{(I-1)}$.

The comparison of models M and M_1 is accomplished by the following test statistic.

$$(SS_T + SS_{I+R})/\sigma^2 \sim \chi^2_{(I-1)J}.$$

4.5 Simulation Study

In this section we describe the performance of the univariate test of stationarity proposed by Priestley and Subba Rao (1969) and described in the previous sections. Specifically, we examine the Neyman–Pearson properties of the test, that is, the type I error rate for known stationary processes and the power of the test under models that are non–stationary. Ten thousand realizations of the following models were generated for $t = 0, \ldots, 99, 999$.

- Gaussian White Noise

  $$X(t) = \epsilon(t), \;\; \epsilon_t \sim N(0, \sigma_W^2),$$

  where $\sigma_W^2 = 1$. The spectral density given by

  $$h_{XX}(\omega) = \frac{\sigma_W^2}{2\pi}$$

- Four auto–regressive order one models

  $$X(t) - \phi X(t-1) = \epsilon(t), \;\; \epsilon_t \sim N(0, \sigma_W^2)$$

  with $\phi = 0.3, 0.5, 0.7, 0.9$. The spectra of each of these series is given by

  $$h_{XX}(\omega) = \frac{\sigma_W^2}{2\pi(1 - \phi\cos\omega + \phi^2)}$$

- An auto–regressive order two model (discussed by Priestley and Subba Rao (1969))

  $$X(t) - 0.8X(t) + 0.4X(t-2) = \epsilon(t), \;\; \epsilon_t \sim N(0, \sigma_W^2).$$

  The spectra of this series is given by

  $$h_{XX}(\omega) = \frac{\sigma_X^2}{2\pi}\left\{\frac{0.792}{1.4 - 3.136\cos\omega + 2.24\cos^2\omega}\right\}$$

  where, $\sigma_X^2 = \frac{(1+\phi_2)\sigma_W^2}{(1-\phi_2)(1-\phi_1+\phi_2)(1+\phi_1+\phi_2)}$ and $\phi_1 = -0.8$ and $\phi_2 = 0.4$. For this particular case, $\sigma_X^2 = 1.33\sigma_W^2$.

The following non–stationary processes were used:

- Random walk process,

$$X(t) - X(t-1) = \epsilon(t), \ \ \epsilon_t \sim N(0, \sigma_W^2).$$

  which corresponds to an auto–regressive model with $\phi = 1$. However, we cannot simply apply the formula for the spectral density as in this case the auto–covariance function $c_{XX}(s,t) = |t-s|\sigma_W^2$ does not decay as $|t-s|$ goes to infinity, which is required.

- uniformly modulated version of the auto–regressive order two series discussed above.

$$Z(t) = \exp\left\{\left(\frac{t-50000}{20000\sqrt{2}}\right)^2\right\} X(t)$$

  The evolutionary spectral density of this series is given by

$$h_{ZZ}(t,\omega) = \exp\left\{\left(\frac{t-50000}{20000\sqrt{2}}\right)^2\right\} h_{ZZ}(\omega)$$

Figure 4.1 gives example time series plots of each of the eight time series. Visually, the first 6 processes look stationary, while the last 2 are obviously non–stationary, though of different character.

We employ the time blocking technique to estimate the evolutionary spectra for each of the eight series considered. The block length was chosen to be 1000 and the length of the spectral Daniel smoothing window was 101. We chose 5 Fourier frequencies that were essentially independent according to the previous sections. We also chose to investigate the impact of pre–whitening, which was done in each block of 1000. That is, the best fitting auto–regressive model was chosen via AIC and the resulting residuals were subjected to the usual spectral estimation procedure. The effect of tapering was also investigated by use of a 10% split cosine bell taper applied at the block level. Finally, the effect of detrending and demeaning within the block was also considered.

Table 4.2 gives the number of rejections in 10,000 trials of each model. As the effect of detrending and demeaning has very little effect on the level of the test, the following discussion applies equally to whether or not they are employed or not. Pre–whitening has some interesting effects on the level of the test, the effect of time is too liberal while the

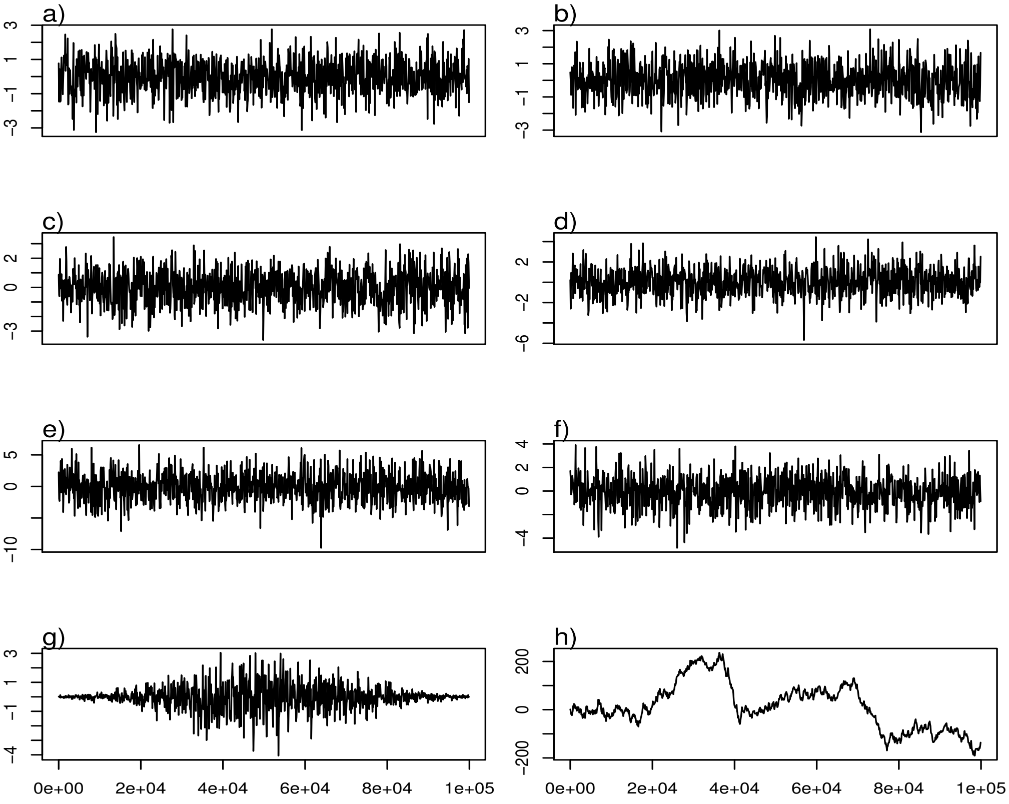

Figure 4.1: Time Series plots for each of the eight simulated series. a) White noise, b) AR1 ($\phi = 0.3$), c) AR1 ($\phi = 0.5$), d) AR1 ($\phi = 0.7$), e) AR1 ($\phi = 0.9$), f) AR2 ($\phi_1 = -0.8, \phi_2 = 0.4$) g) AR2 ($\phi_1 = -0.8, \phi_2 = 0.4$) uniformly modulated, h) Random Walk. The scaling on the y–axis is different for each plot.

	Pre		Demean=F & Detrend=F				Demean=T & Detrend=T			
Series	Whiten	Taper	Freq	Time	Int	T+I	Freq	Time	Int	T+I
WN	N	0	502	693	616	698	606	707	629	716
		10%	497	731	579	667	605	732	598	685
	Y	0	209	1507	0	0	170	1517	0	0
		10%	233	1171	0	0	196	1173	0	0
AR(1)	N	0	10000	782	728	875	10000	805	747	895
$\phi = 0.3$		10%	10000	774	667	811	10000	777	671	824
	Y	0	10000	1778	0	0	10000	1779	0	0
		10%	10000	1343	0	0	10000	1361	0	0
AR(1)	N	0	10000	965	1091	1308	10000	986	1113	1317
$\phi = 0.5$		10%	10000	960	999	1198	10000	979	1007	1213
	Y	0	10000	1733	0	0	10000	1741	0	0
		10%	10000	1380	0	0	10000	1365	0	0
AR(1)	N	0	10000	1604	2495	3101	10000	1629	2531	3172
$\phi = 0.7$		10%	10000	1484	2378	2918	10000	1500	2396	2938
	Y	0	10000	1744	0	0	10000	1751	0	0
		10%	10000	1359	0	0	10000	1367	0	0
AR(1)	N	0	10000	7709	9818	9954	10000	7783	9836	9953
$\phi = 0.9$		10%	10000	6565	9817	9930	10000	6528	9809	9933
	Y	0	10000	1788	0	0	10000	1794	0	0
		10%	10000	1403	0	0	10000	1406	0	0
AR(2)	N	0	10000	1598	2253	2828	10000	1630	2314	2890
		10%	10000	1467	2056	2574	10000	1471	2111	2624
	Y	0	10000	2028	0	0	10000	2066	0	0
		10%	10000	1508	0	0	10000	1525	0	0
AR(2)	N	0	10000	10000	2343	10000	10000	10000	2382	10000
UM		10%	10000	10000	2173	10000	10000	10000	2198	10000
	Y	0	10000	10000	0	10000	10000	10000	0	10000
		10%	10000	10000	0	10000	10000	10000	0	10000
RW	N	0	10000	10000	10000	10000	10000	10000	10000	10000
		10%	10000	10000	10000	10000	10000	10000	10000	10000
	Y	0	10000	970	5	21	10000	901	4	20
		10%	10000	868	7	15	10000	828	8	23

Table 4.2: The number of rejections in 10,000 trials of the univariate test procedure for each of the 8 processes described in the text. The factors considered are: pre–whitening, tapering and detrending/demeaning for each of the 4 factors of the basic model: frequency, time and their interaction, plus the combined effect of time and interaction.

test for interaction is too conservative. One possible explanation for this curious effect is that the order of the auto-regressive model chosen via AIC was very variable. For example, the following table gives the percentage of time each order was chosen for the white noise process

order	0	1	2	3	4+
	71.33%	11.23%	5.74%	3.43%	8.28%

The maximum order chosen was 30. The performance of the estimation of the order of the auto–regressive model could be improved if the block length was increased. However, we must assume that the process is stationary over a longer duration. Priestley (1965, 1966) make similar observations with regards to the estimation of the evolutionary process by the double window technique.

There are several other issues associated with pre–whitening, the uncertainty in the auto–regressive order and the estimation error in the model parameters is not reflected in the final estimate of the spectrum. Pre–whitening will not be considered further.

In the absence of pre–whitening, the level of the tests for the white noise series is approximately correct, however, it tends to be slightly liberal. This potentially due to the delta method to find the variance of the log periodogram, which is only asymptotically correct. The level of the tests get progressively more liberal as $\phi \to 1$, this is not surprising as the auto-regressive process does tend to exhibit more non–stationary like behaviour as $\phi \to 1$. However, for moderate ϕ's the processes are still stationary and should not result in increased levels.

One particular difficulty with testing for stationarity with auto–regressive processes, is that the ϕ parameter does not behave like a a usual effect size. In the sense that, strictly speaking, all AR(1) processes with $\phi < 1$ are stationary and should not be rejected. However, once $\phi = 1$ they are non-stationary, thus the power function should in theory be a step function.

For the two non–stationary processes, the test for Time behaves exactly as it should, however, for the uniformly modulated process there should not be any evidence of an interaction, however, the observed level is approximately 22%. Thus, we would mis–characterize the nature of the non–stationarity.

With the above problems in mind, we propose a slight variation on the previous tests. The modification is only possible in very specific situations, and isn't at all practical when

Item	DF	SS	MS	EMS
Times	$I-1$	SS_T	$SS_T/(I-1)$	$\sigma^2 + \frac{JK}{I-1}\sum_{i=1}^{I}\alpha_i^2$
Freqs	$J-1$	SS_F	$SS_F/(J-1)$	$\sigma^2 + \frac{IK}{J-1}\sum_{j=1}^{J}\beta_j^2$
Int	$(I-1)(J-1)$	SS_I	$SS_I/(I-1)(J-1)$	$\sigma^2 + \frac{K}{(I-1)(J-1)}\sum_{i=1}^{I}\sum_{j=1}^{J}\gamma_{ij}^2$
Res	IJ(K-1)	SS_R	$SS_R/(IJ(K-1))$	σ^2
Total	$IJK-1$	SS_0		

Table 4.3: Frequency by time analysis of variance with K replicate realizations.

one only has a single realization available. Specifically the modification is aimed at MCMC users where there are multiple chains available. The variant is considered in the next section.

4.6 Variant: Multiple Realizations

We consider the following variant of the test considered in the previous sections. Specifically, assume we have multiple realizations of the process under study available. Using the same notation as before, we have

$$M : Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk} \;\; i = 1, \ldots, I, \; j = 1, \ldots, J, k = 1, \ldots, K.$$

where K is the number of realizations(replicates) available. We generalize the other models under consideration in the obvious way. We are now led to the following ANOVA table.

where

$$\begin{aligned}
SS_T &= JK\sum_{i=1}^{I}(\bar{Y}_{i..} - \bar{Y}_{...})^2, \\
SS_F &= IK\sum_{j=1}^{J}(\bar{Y}_{.j.} - \bar{Y}_{...})^2, \\
SS_I &= K\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{Y}_{ij} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...})^2, \\
SS_R &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{j=1}^{K}(Y_{ijk} - \bar{Y}_{ij.})^2, \\
SS_0 &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{j=1}^{K}(Y_{ijk} - \bar{Y}_{...})^2
\end{aligned}$$

The asymptotic variances derived are, of course, large sample results and it is not entirely clear how large T needs to be for the results to apply. This is compounded in the non–stationary case as the block size is the determining factor for the sampling results. There is a trade off between choosing block lengths that are long enough for the large sample results to apply but on the other hand short enough for the assumption of stationarity to be reasonable. In the current case, where multiple realizations are available, we can now separate the interaction from the replication error and hence get an estimate of the variance which is appropriate for the particular sample of interest. Thus, we now replace the χ^2–tests with the more traditional F–tests used in ANOVA models.

4.6.1 Simulation Study

A very similar simulation study was performed as in the previous section. However, we only consider the effect of tapering, without pre–whitening and we consider the situation where we have two realization of each of the processes.

The results of the simulation study are given in Table 4.4. The results clearly show that having the second realization improves the level of the test, with the exception of the random walk process. However, if we use tapering the test has much better properties, rejecting the random walk as being stationary almost 45% of the time. Tapering has the added benefit of improving the level slightly for all other processes. Also, the power

Series	Taper	Freq	Time	Int	T+I
WN	0	501	627	471	513
	10%	499	633	481	506
AR(1)	0	10000	618	456	518
$\phi = 0.3$	10%	10000	579	443	481
AR(1)	0	10000	664	480	525
$\phi = 0.5$	10%	10000	640	474	511
AR(1)	0	10000	645	441	488
$\phi = 0.7$	10%	10000	617	469	459
AR(1)	0	10000	943	507	630
$\phi = 0.9$	10%	10000	608	568	617
AR(2)	0	10000	663	446	477
	10%	10000	651	450	483
AR(2)	0	10000	10000	428	10000
UM	10%	10000	10000	458	10000
RW	0	10000	10000	0	1931
	10%	10000	4052	4447	4492

Table 4.4: The number of rejections in 10,000 trials of the univariate test procedure with 2 replicates for each of the 8 processes described in the text. The effect of tapering is considered for each of the 4 factors of the basic model: frequency, time and their interaction, plus the combined effect of time and interaction.

function of the test resembles the idealized step function much more closely than with a single realization.

In conclusion, the numerical results seem to indicate that having the second realization and using tapering gives a test for stationarity with approximately the right level.

4.7 Multivariate Test of Stationarity

This section considers a multivariate extension of the test of stationarity developed by Priestley and Subba Rao (1969). The outline of this section is very similar to the previous section, where we will lay out some basic spectral theory for stationary multivariate processes and then turn our attention to evolutionary spectral methods. We discuss the form of the multivariate test and present a small simulation study.

Definition 4.4. *Suppose we have r discrete parameter processes$\{X_1(t)\}$, $\{X_2(t)\}$, $\ldots$, $\{X_r(t)\}$, $t = 0, \pm 1, \pm 2, \ldots$,, which we write in vector notation as*

$$\{\mathbf{X}(t)\} = (X_1(t), X_2(t), \ldots, X_r(t)).$$

We say that $\mathbf{X}(t)$ is jointly second order stationary if

- *$\{X_i(t)\}$ for $i = 1, \ldots r$ are each univariate stationary processes as defined by definition (4.2).*
- *cov$(X_i(t), X_j(s))$ is a function of $|t - s|$*

Definition 4.5. *The auto–covariance $c_{X_iX_i}(u)$ and cross–covariance functions $c_{X_iX_j}(u)$ for a vector–valued zero–mean stationary process, $\{\mathbf{X}(t)\}$, are given by*

$$c_{X_iX_i}(u) = \textit{cov}(X_i(t), X_i(t{+}u)) = E\left[X_i(t)X_i(t+u)\right], \quad i = 1, \ldots, R, \quad u = 0, \pm 1, \pm 2, \ldots$$

and

$$c_{X_iX_j}(u) = \textit{cov}(X_i(t), X_j(t{+}u)) = E\left[X_j(t)X_i(t+u)\right], \quad i, j = 1, \ldots, R, \quad i \neq j \quad u = 0, \pm 1, \pm 2, \ldots$$

respectively. Note that, $c_{X_iX_j}(u)$ means that that j is leading i.

For each, u we can represent the auto–covariance and cross–covariance functions in matrix form $\mathbf{c_{XX}}(u)$ (the covariance matrix)

$$\mathbf{c_{XX}}(u) = [c_{X_iX_j}(u)], \quad i = 1, \ldots, r, j = 1, \ldots, r$$

that is, the matrix has $c_{X_iX_j}(u)$ in the ith row and jth column.

Assume that for each i, $\sum_{u=-\infty}^{\infty} |c_{X_iX_i}(u)| < \infty$ and for all i and j $\sum_{u=-\infty}^{\infty} |c_{X_iX_j}(u)| < \infty$ we can introduce the univariate spectral density and cross–spectral density functions respectively as follows:

$$\begin{aligned} h_{X_iX_i}(\omega) &= \frac{1}{2\pi}\sum_{u=-\infty}^{\infty} e^{-i\omega u} c_{X_iX_i}(u), \quad i = 1, \ldots, r \\ h_{X_iX_j}(\omega) &= \frac{1}{2\pi}\sum_{u=-\infty}^{\infty} e^{-i\omega u} c_{X_iX_j}(u), \quad i, j = 1, \ldots, r \end{aligned}$$

The spectral matrix at frequency ω can be defined analogously as $\mathbf{h}_{\mathbf{XX}}(\omega) = [h_{ij}(\omega)], i, j = 1, \ldots, r$.

In order to simplify the above notation it is more convenient to express the series as a column vector for each time t as, $\mathbf{X}(t) = \{X_{1,t}, \ldots, X_{p,t}\}^{'}$, and we can then write $\mathbf{c}_{\mathbf{XX}}(u)$ as

$$\mathbf{c}_{\mathbf{XX}}(u) = E[\mathbf{X}(t)^{*}\mathbf{X}(t+u)] \tag{4.10}$$

where $E[\mathbf{X}(t)] = 0$ and * denotes both complex conjugate and transposition. Thus we can now write $\mathbf{h}(\omega)$ more compactly as follows:

$$\mathbf{h}_{\mathbf{XX}}(\omega) = \frac{1}{2\pi}\sum_{u=-\infty}^{\infty} e^{-i\omega u}\mathbf{c}_{\mathbf{XX}}(u)$$

and upon inversion we have

$$\mathbf{c}_{\mathbf{XX}}(u) = \int_{-\pi}^{\pi} e^{i\omega u}\mathbf{h}_{\mathbf{XX}}(\omega)d\omega. \tag{4.11}$$

Priestley (1988) states the following results:

- $\mathbf{c}^{*}_{\mathbf{XX}}(u) = \mathbf{c}_{\mathbf{XX}}(-u)$ for each u
- $\mathbf{h}^{*}_{\mathbf{XX}}(\omega) = \mathbf{h}_{\mathbf{XX}}(\omega)$ for each ω, that is, $\mathbf{h}(\omega)$ is a Hermitian matrix (the generalization of a symmetric matrix to complex values)

Now, since each series is stationary with $E[\mathbf{X}] = 0$ we know each exhibits a spectral representation given by

$$X_{i,t} = \int_{-\pi}^{\pi} e^{i\omega t} dZ_i(\omega) \tag{4.12}$$

or in vector form

$$\mathbf{X}(t) = \int_{-\pi}^{\pi} e^{i\omega t} d\mathbf{Z}(\omega) \tag{4.13}$$

where $d\mathbf{Z} = \{dZ_1(\omega), dZ_2(\omega), \ldots, dZ_p(\omega)\}^{'}$.

Now we can substitute equation (4.13) into equation (4.10) to give

$$\begin{aligned} \mathbf{c}_{\mathbf{XX}}(u) &= E\left[\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} e^{-i\omega t} e^{i\omega^{'}(t+u)} d\mathbf{Z}^{*}(\omega) d\mathbf{Z}(\omega^{'})\right] \\ &= \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} e^{-i\omega t} e^{i\omega^{'}(t+u)} E\left[d\mathbf{Z}^{*}(\omega) d\mathbf{Z}(\omega^{'})\right] \end{aligned}$$

As with the previous argument for univariate series, the left hand side is a function of u only, therefore, the right hand must also be a function of u only, and the only that can happen is if

$$E\left[d\mathbf{Z}^{*}(\omega) d\mathbf{Z}(\omega^{'})\right] = 0 \text{ when } \omega \neq \omega^{'}$$

thus, the increment process $\{d\mathbf{Z}(\omega)\}$ is orthogonal and cross–orthogonal. Now, we have the following expression

$$\mathbf{c}_{\mathbf{XX}}(u) = \int_{-\pi}^{\pi} e^{i\omega u} E\left[d\mathbf{Z}^{*}(\omega) d\mathbf{Z}(\omega)\right]$$

Comparing this expression with equation (4.11) we have the following

$$\mathbf{h}_{\mathbf{XX}}(\omega) d\omega = E\left[d\mathbf{Z}^{*}(\omega) d\mathbf{Z}(\omega)\right].$$

Thus, we may interpret $\mathbf{h}_{\mathbf{XX}}(\omega)d\omega$ as the variance–covariance matrix of the vector $d\mathbf{Z}(\omega)$.

Note that by writing the inversion formula

$$\mathbf{c}_{\mathbf{XX}}(u) = \int_{-\pi}^{\pi} e^{iu\omega} \mathbf{h}_{\mathbf{XX}}(\omega) d\omega$$

and let $u = 0$, the covariance matrix can be expressed as

$$\mathbf{c}_{\mathbf{XX}}(0) = \int_{-\pi}^{\pi} \mathbf{h}_{\mathbf{XX}}(\omega) d\omega.$$

The covariance matrix can be expressed as an integral over all frequencies of the power spectral density matrix. In other words, the power spectral density matrix can be seen as a decomposition of the covariance matrix into its constituent parts, a similar interpretation as in the univariate case.

4.7.1 Multivariate Evolutionary–Spectra

We next turn our discussion to evolutionary cross–spectra as laid out in Priestley and Tong (1973); Priestley (1988). We present results from Priestley (1988), based on original work in Priestley and Tong (1973). We first consider only two non–stationary oscillatory processes, denoted by $\{X(t), Y(t)\}$ that admit representations as follows

$$\begin{aligned} X(t) &= \int_{-\pi}^{\pi} A_X(t,\omega)e^{i\omega t}dZ_X(\omega) \\ Y(t) &= \int_{-\pi}^{\pi} A_Y(t,\omega)e^{i\omega t}dZ_Y(\omega) \end{aligned}$$

with

$$\begin{aligned} E[dZ_X^*(\omega)dZ_X(\omega^{'})] &= E[dZ_Y^*(\omega)dZ_Y(\omega^{'})] = E[dZ_X^*(\omega)dZ_Y(\omega^{'})] = 0 \;\; \omega \neq \omega^{'} \\ E[|dZ_X(\omega)|^2] &= d\mu_{XX}(\omega) \\ E[|dZ_Y(\omega)|^2] &= d\mu_{YY}(\omega) \\ E[dZ_X^* dZ_Y(\omega)] &= d\mu_{XY}(\omega) \end{aligned}$$

We have two different oscillatory families, represented by $\mathcal{F}_X = \{\phi_X(t,\omega) = A_X(t,\omega)e^{i\omega t}\}$ and $\mathcal{F}_Y = \{\phi_Y(t,\omega) = A_Y(t,\omega)e^{i\omega t}\}$. We next define the evolutionary power spectral densities as follows with respect to $\mathcal{F}_X$ and $\mathcal{F}_Y$ respectively as

$$\begin{aligned} dH_{XX}(t,\omega) &= |A_{XX}(t,\omega)|^2 d\mu_{XX}(\omega), \;\; -\pi \leq \omega \leq \pi. \\ dH_{YY}(t,\omega) &= |A_{YY}(t,\omega)|^2 d\mu_{YY}(\omega), \;\; -\pi \leq \omega \leq \pi. \end{aligned}$$

We assume that the measures $\mu_{XX}(\omega)$ and $\mu_{XX}(\omega)$ are absolutely continuous with respect to Lebesgue measure and hence we have evolutionary spectral densities given by

$$\begin{aligned} dH_{XX}(t,\omega) &= h_{XX}(t,\omega)d\omega \\ dH_{YY}(t,\omega) &= h_{YY}(t,\omega)d\omega \end{aligned}$$

We can similarly define the evolutionary power cross–spectrum at time t with respect to families$\mathcal{F}_X$ and $\mathcal{F}_Y$ by

$$dH_{t,XY}(\omega) = A_X(t,\omega)^* A_Y(t,\omega) d\mu_{XY}(\omega).$$

We again assume that the measure $\mu_{XY}(\omega)$ is absolutely continuous with respect to Lebesgue measure, and hence we assume that we have evolutionary cross–spectral densities, $h_{XY}(t,\omega)$ given by

$$dH_{XY}(t,\omega) = h_{XY}(t,\omega)d\omega$$

Note that we could also assume that the families $\mathcal{F}_X$ and $\mathcal{F}_Y$ are equivalent which would lead to simplifications, but we will not pursue this further.

We can extend these ideas in a straightforward manner to multiple series, and represent the evolutionary power spectral densities and cross–spectral densities in a matrix $h^t_{\mathbf{XX}}(\omega)$.

4.7.2 Estimation of the Spectral Density Matrix

Let $\mathbf{X}(t)$ be a stationary vector valued time series with r components, which we have sampled for $t = 0, \ldots, T-1$, with mean $\boldsymbol{\mu}_{\mathbf{X}}$, which we assume without loss of generality is zero and spectral density matrix $\mathbf{h}_{\mathbf{XX}}(\omega), -\pi \leq \omega \leq \pi$. Following, the development for the univariate case, consider the vector valued finite Fourier transform given by

$$\begin{aligned} \mathbf{d}^{(T)}_{\mathbf{X}}(\omega) &= [d^{(T)}_a(\omega)] \\ &= \left[\sum_t X_{at} e^{-i\omega t}\right] \quad -\pi \leq \omega \leq \pi, \end{aligned}$$

leading to the following distributional results (see Brillinger (1981))

$$\mathbf{d}^{(T)}_{\mathbf{X}}(\omega) \sim \begin{cases} N^C_r(\mathbf{0}, 2\pi T \mathbf{h}_{\mathbf{XX}}(\omega)) & \text{if } \omega \neq 0 \pmod{\pi} \\ N_r(\mathbf{0}, 2\pi T \mathbf{h}_{\mathbf{XX}}(\omega)) & \text{if } \omega = 0, \pm 2\pi, \ldots \\ N_r(\mathbf{0}, 2\pi \mathbf{h}_{\mathbf{XX}}(\omega)) & \text{if } \omega = \pm\pi, \pm, 3\pi, \ldots \end{cases}$$

where the normal distributions are multivariate of dimension r and the superscript C indicates a complex valued multivariate normal.

As in the univariate case, a natural estimate to consider is the following

$$\begin{aligned}\mathbf{I}^{(T)}_{\mathbf{XX}}(\omega) &= [I^{(T)}_{ij}(\omega)] \\ &= \left[\frac{1}{2\pi T} d^{(T)}_i(\omega)\overline{d^{(T)}_j(\omega)}\right], \quad i, j = 1, \ldots r.\end{aligned}$$

This estimator suffers the same draw backs as the univariate estimator, that is, the variance of our estimate does not go to zero as T goes to infinity. Therefore, in order to reduce the variance of our estimator we must smooth the multivariate periodogram.

Brillinger (1981) gives the following estimator and proves that it is asymptotically unbiased and that the ordinates (*i.e.* frequencies) are asymptotically independent

$$\mathbf{h}^{(T)}_{\mathbf{XX}}(\omega) = (2m+1)^{-1} \sum_{s=-m}^{m} \mathbf{I}^{(T)}_{XX}\left(\frac{2\pi[s(T)+s]}{T}\right) \text{ if } \omega \neq 0 \pmod{\pi}.$$

where $s(T)$ is an integer with $2\pi s(T)/T$ near $\omega \neq 0 \pmod{\pi}$. Again special considerations have to be paid for frequencies that are multiples of π and also the zero frequency, please refer to Brillinger (1981).

Although, the estimator considered here is asymptotically unbiased, Brillinger (1981) section 7.4 states that the above estimate is not generally consistent, that is, $\mathbf{h}^{(T)}_{\mathbf{XX}}(\omega)$ does not tend to $\mathbf{h}_{\mathbf{XX}}(\omega)$ in probability as $T \to \infty$. Brillinger (1981) considers classes of estimators which are asymptotically consistent which are of the following form.

$$\mathbf{h}^{(T)}_{\mathbf{XX}}(\omega) = \frac{2\pi}{T} \sum_{s=1}^{T-1} W^{(T)}\left(\omega - \frac{2\pi s}{T}\right) \mathbf{I}^{(T)}_{\mathbf{XX}}\left(\frac{2\pi s}{T}\right) \tag{4.14}$$

where

$$\mathbf{I}^{(T)}_{\mathbf{XX}}(\lambda) = (2\pi T)^{-1} \left[\sum_{t=0}^{T-1} X_t e^{-i\lambda t}\right] \overline{\left[\sum_{t=0}^{T-1} X_t e^{-i\lambda t}\right]}$$

and

$$W^{(T)}(\lambda) = \sum_{j=-\infty}^{\infty} W(B_T^{-1}[\lambda + 2\pi j])$$

with $W(\lambda)$ being concentrated near $\lambda = 0$ and B_T, $T = 1, 2, \ldots$, is a sequence of non–negative bandwidth parameters.

The inclusion of the spectral window $W^{(T)}(\lambda)$ gives us the required asymptotic consistency, provided $B_T T \to \infty$ as $T \to \infty$. This is similar to the univariate case.

It is well known that the quality of the cross–spectrum is greatly improved if an alignment is performed in advance of the spectral estimation (Brillinger, 1981; Priestley, 1981). We don't consider the alignment issue further, as we are dealing with cases where the maximal correlation occurs at lag 0. Tapering is also used to reduce the spectral leakage in multivariate spectral estimation.

4.8 Multivariate Test of Stationarity

Now that we have a basic introduction to multivariate evolutionary spectra, we consider how to generalize the test of Priestley and Subba Rao (1969) to multivariate processes,$\mathbf{X}(t)$. Proceeding along similar lines as the previous section we need an estimate of the spectral matrix $\mathbf{h}_{\mathbf{XX}}(t, \omega)$ for some collection of times $t_1, \ldots, t_I$ and frequencies $\omega_1, \ldots, \omega_J$. As in the univariate case, we have two–way layout with the factors being frequencies and times, however, rather than having a scalar valued response variable, we have a complex–valued matrix response variable, $\mathbf{h}_{\mathbf{XX}}(t, \omega)$. Brillinger (1981) shows that $\mathbf{h}^{(T)}_{\mathbf{XX}}(t, \omega)$ has an asymptotically complex–valued Wishart distribution. One could proceed to develop an analysis approach for this response variable, however, this is not the approach we shall pursue here.

Our approach is to consider a scalar function of the matrix $\mathbf{h}^{(T)}_{\mathbf{XX}}(t, \omega)$ and use this as our response variable in the stationarity test as described previously. Two natural choices are the determinant and trace of the spectral matrix $\mathbf{h}^{(T)}_{\mathbf{XX}}(t, \omega)$. We focus our attention on the determinant, as it has an interpretation as a generalized variance. However, we do present some simulation results on the trace.

We give some well known results for the determinant of a sample variance–covariance matrix and then proceed to investigate its properties for the spectral density matrix.

Note that we can write the determinant of a positive definite matrix as the product of its eigenvalues, that is,

$$\text{Det}(\mathbf{A}) = \prod_{i=1}^{r} \lambda_i$$

where λ_i are the eigenvalues of A.

Consider a sample, of size n, from a multivariate normal distribution with covariance matrix Σ with eigenvalues denoted by $\lambda_i,\ i = 1, \ldots, r$. Let the corresponding sample quantities be denoted by $\mathbf{S}$ and $\hat{\lambda}_i$ respectively. Anderson (1984) gives the following asymptotic results for the sample eigenvalues $\hat{\lambda}_i$

$$\hat{\lambda}_i \sim N(\lambda_i, 2\lambda_i^2/n)$$

in addition we also have

$$\log \hat{\lambda}_i \sim N(\log \lambda_i, 2/n),$$

which are asymptotically independent for each i.

As the generalized variance is the determinant of the sample covariance matrix, we have the following asymptotic sampling distribution

$$\log GV(S) = \log \text{det}(S) = \sum_{i=1}^{r} \log \hat{\lambda}_i \sim N(\sum_{i=1}^{r} \log \lambda_i, 2r/n).$$

Returning to the problem at hand, we have a spectral matrix $\mathbf{h}_{\mathbf{XX}}(\omega)$ and an associated estimate $\mathbf{h}_{\mathbf{XX}}^{(T)}(\omega)$.

Following Brillinger (1981) we decompose the spectral matrix $\mathbf{h}_{\mathbf{XX}}(t, \omega)$ into its eigenvalue and eigenvector decomposition.

Brillinger (1981) states the following theorem

Theorem 4.3. *Let*$\mathbf{X}(t), t = 0, \pm 1, \ldots$ *be an* $\mathbf{r}$ *vector–valued series satisfying assumption Assumption 2.6.1 of Brillinger (1981). Let* $\nu_j^{(T)}(\omega)$ $\mathbf{U}_j^{(T)}(\omega)$*,* $j = 1, \ldots, r$ *be the eigenvalues and eigenvectors of the matrix*

$$\int_0^{2\pi} W^{(T)}(\omega - \lambda)\mathbf{h}(\lambda) d\lambda$$

Let $\mathbf{h}_{\mathbf{XX}}^{(T)}(\omega)$ *be defined by (4.14) and assume that* $W(\beta)$ *satisfies*

$$\int_{-\infty}^{\infty} W(\beta) d\beta = 1$$

and

$$\int_{-\infty}^{\infty} |W(\beta)| d\beta < \infty.$$

Let $\mu_j^{(T)}(\omega)$*,* $V_j^{(T)}(\omega)$*,* $j = 1, \ldots, r$*, be the eigenvalues and eigenvectors of the matrix*

$\mathbf{h}_{\mathbf{XX}}^{(T)}(\omega)$. *If* $B_T T \to \infty$ *as* $T \to \infty$, *then*

$$E\mu_j^{(T)}(\omega) = \nu_j^{(T)}(\omega) + O(B^{-1/2}T^{1/2}).$$

If the eigenvalues of $\mathbf{h}_{\mathbf{XX}}(\omega)$ *are distinct then*

$$a\vec{ve}\,\mu_j^{(T)}(\omega) = \nu_j^{(T)}(\omega) + O(B_T^{-1}T^{-1})$$

and

$$a\vec{ve}\,V_j^{(T)}(\omega) = U_j^{(T)}(\omega) + O(B_T^{-1}T^{-1})$$

for $j = 1, \ldots, r$.

where $a\vec{ve}$ denotes an expected value derived in a term by term manner from a Taylor series expansion.

The key result, which is very similar in nature to the results of Anderson (1984), is the following theorem by Brillinger (1981)

Theorem 4.4. *Under the conditions of the previous theorem and if the eigenvalues of* $\mathbf{h}_{\mathbf{XX}}(\omega_m)$ *are distinct,* $m = 1, \ldots, M$, *the variates* $\mu_j^{(T)}(\omega_m)$, $\mathbf{V}_j^{(T)}(\omega_m)$, $j = 1, \ldots, r$, $m = 1, \ldots, M$ *are asymptotically jointly normal with asymptotic covariance structure*

$$\lim_{T\to\infty} B_T T co\vec{v}\,\{\mu_j^{(T)}(\omega_m), \mu_k^{(T)}(\omega_n)\}$$
$$= \begin{cases} 2\pi \int W(\alpha)^2 d\alpha[\eta\{\omega_m - \omega - n\} + \eta\{\omega_m + \omega_n\}]\mu_j(\omega_m)^2 & \textit{if } \; j = k \\ 0 & \textit{if } \; j \neq k \end{cases}$$

where

$$\eta\{\alpha\} = \begin{cases} 1 & \textit{if } \alpha = 0 \pmod{2\pi} \\ 0 & \textit{otherwise} \end{cases}$$

The theorem gives results for the asymptotic distribution of the eigenvectors, however, these are not of interest to us and will be omitted. Note, that this theorem implies the following very useful result

$$va\vec{r} \log \mu_j^{(T)}(\omega) \sim \begin{cases} B_T^{-1}T^{-1}2\pi \int W(\alpha)^2 d\alpha & \text{if } \omega \neq 0 \pmod{\pi} \\ B_T^{-1}T^{-1}4\pi \int W(\alpha)^2 d\alpha & \text{if } \omega = 0 \pmod{\pi} \end{cases}$$

Consider the generalized variance applied at each frequency to the spectral density matrix as follows

$$\log GV(\omega_m) = \sum_{i=1}^{r} \log \mu_j^{(T)} \sim N\left(\sum_{j=1}^{r} log\{\nu_j^{(T)}(\omega)\}, \frac{r2\pi \int W(\alpha)^2 d\alpha}{B_T T}\right).$$

Brillinger (1981) has the following key result

$$\begin{aligned} 2m+1 &\sim \frac{1}{\sum_s \left[\frac{2\pi}{T} W^{(T)}\left(\omega - \frac{2\pi s}{T}\right)\right]^2} \\ &\sim \frac{B_T T}{2\pi \int W(\alpha)^2 d\alpha} \end{aligned}$$

Thus, we can now express the sampling variability in the log generalized variance as follows

$$\log GV(\omega_m) = \sum_{i=1}^{r} \log \mu_j^{(T)} \sim N\left(\sum_{j=1}^{r} log\{\nu_j^{(T)}(\omega)\}, \frac{r}{2m+1}\right).$$

Using the log generalized variance allows us to form the multivariate test in an analogous manner as the univariate one in the previous section.

Given the asymptotic variance results, the form of the multivariate test will be equivalent to the univariate results with either the determinant or trace used in place of the univariate spectral estimate. Given the results of the univariate test we only consider the situation where we have multiple realizations of our multivariate process. We also only consider tapering and the non pre–whitened version of the test

4.9 Simulation Study

We now describe a simple simulation study to investigate the Neyman-Pearson properties of the test. Each multivariate process, $\mathbf{Z}(t) = \{Z_1(t), \ldots, Z_r(t)\}$, was generated in the following way. Firstly, $r = 5$ uncorrelated processes of the given type where generated (for a list of the processes used please see 4.5) for $t = 0, \ldots, 99, 999$ with $\sigma_W^2 = 1$. We then constructed a covariance matrix with variances given by $\sigma_Z^2 = (0.8, 0.9, 1.0, 1.1, 1.2)$ covariance elements were randomly selected by first choosing a correlation coefficient uniformly between $(-0.8, 0.8)$ for each pair for variables. The resulting covariance matrix was only kept if it was positive definite. Note a new covariance matrix was generated for each of the 10,000 trials, however, the variances were kept fixed. Finally, a multivariate series was constructed by the following relation

$$\mathbf{Z}(t) = \Sigma^{1/2}\mathbf{Z}(t)$$

where $A^{1/2}$ is the matrix square root.

The sampling result for eigenvalues given in the previous section assumes that the eigenvalues of the spectral matrix are distinct, this is the reason for having distinct variances for $\mathbf{Z}$.

Each of the eight processes were used to generate the univariate basis for the series, though, they were always of the same type. That is, we did not consider mixing the different processes in the same multivariate process, though this would certainly be an interesting sideline to pursue. The simulation consisted of generating 10,000 trials of length 100,000 with 2 replicates/realizations of the series.

Table 4.5 gives the multivariate performance for each of the 8 series types, with the effect of tapering and also the determinant and trace. The determinant outperforms the trace for almost every combination, the only exception being the white noise series. Remarkably, tapering moves the non-tapered version in the correct direction in every situation, making the test more conservative when the series are actually stationary and improving the power when the series are non–stationary.

4.10 Application to MCMC convergence

We have shown that the original test of Priestley and Subba Rao (1969) and the multivariate generalization considered here have desirable Neyman–Pearson properties. That is, they have the right level and power for the class of processes considered. Before proceeding with the modification to assess whether or not a MCMC sampler has reached its stationarity distribution, we need to consider a conceptual difficulty.

The processes considered previously in this chapter are either second order stationary or not for all t. That is, they do not go through periods of stationarity and non–stationarity as by definition a process like this would be non–stationary. By contrast, MCMC samplers potentially go through periods of non–stationarity as they are not necessarily initialized

		Determinant				Trace			
Series	Taper	Freq	Time	Int	T+I	Freq	Time	Int	T+I
WN	0	551	663	466	493	506	606	465	506
	10%	562	653	442	493	474	644	454	516
AR(1)	0	10000	641	498	521	10000	782	991	1144
$\phi = 0.3$	10%	10000	672	485	522	10000	765	947	1071
AR(1)	0	10000	680	459	534	10000	964	1564	1720
$\phi = 0.5$	10%	10000	643	421	490	10000	950	1574	1716
AR(1)	0	10000	590	471	520	10000	1053	1984	2100
$\phi = 0.7$	10%	10000	554	514	527	10000	1015	2020	2133
AR(1)	0	10000	848	705	801	10000	1161	2244	2362
$\phi = 0.9$	10%	10000	624	760	816	10000	1114	2179	2306
AR(2)	0	10000	671	442	503	10000	818	1179	1319
	10%	10000	627	442	491	10000	836	1194	1336
AR(2)	0	10000	10000	468	10000	10000	10000	10000	10000
UM	10%	10000	10000	500	10000	10000	10000	10000	10000
RW	0	10000	10000	0	1522	10000	1182	2112	2240
	10%	10000	8885	9310	9337	10000	5770	6138	6169

Table 4.5: The number of rejections in 10,000 trials of the multivariate test procedure with 2 replicates for each of the 8 processes described in the text. The effect of tapering and the use of determinant or trace are considered for each of the 4 factors of the basic model: frequency, time and their interaction, plus the combined effect of time and interaction.

in their stationary distribution. We additionally assume that MCMC samplers that have reached their stationary distribution will behave like second–order stationary processes.

The diagnostic we propose can be described, roughly speaking, as sequentially removing blocks of MCMC output until the resulting output appears to be stationary by the methods presented previously in this chapter.

Recall Y_{ijk} represents the log of the estimated spectral density or the log of the determinate of the estimated spectral density matrix for frequency i, time j and replicate k, that is, Y_{ijk}. The test statistic is then the sum of the time and frequency by time interaction effects, as we are not interested in the special case of uniformly modulated processes. We then compare this to the appropriate error term. Recall we have an asymptotic expression for the variance of log spectrum which allows us to test for the interaction plus time effect in the absence of replication.

The test statistic of interest is given by

$$F_i = \frac{(SS_T + SS_{T \times F})/((I-1)J)}{MSR}$$

which we index by i the starting block of the given test. We compute the test statistic for $i = 1, \ldots, I-1$ and the corresponding critical value, we suggest using $\alpha = 0.01$. The diagnostic plot would then depict the observed test statistics and their critical values. Note the same procedure would used for both the univariate and multivariate versions.

Alternatively, we could proceed in a similar fashion to Gelman and Rubin (1992) and derive a diagnostic on the basis of the expected mean squares, however, this was not pursued. We illustrate the diagnostic procedure using example 3.2 which illustrated the Metropolis–Hastings algorithm for generating a bivariate normal density with unit variances, and correlation 0.9 and mean vector of $(1, 2)$. To test the diagnostic procedure we used two different proposal distributions $[-\delta, \delta] \times [-\delta, \delta]$ for $\delta = 0.1, 2.0$. That is, one that will mix slowly and one close to the optimal scaling. We start the two replicate chains at (10,10) and (11,11) to illustrate the convergence to the stationary distribution.

Figures 4.2 and 4.3 show the results of applying the univariate and multivariate versions of our diagnostic procedure to the two Metropolis–Hastings examples. The procedures behave as expected, in that they indicate that it takes approximately 3000 iterations for the slowly mixing chain to forget its initial conditions while the optimally scaled chains move away from their initial conditions almost immediately.

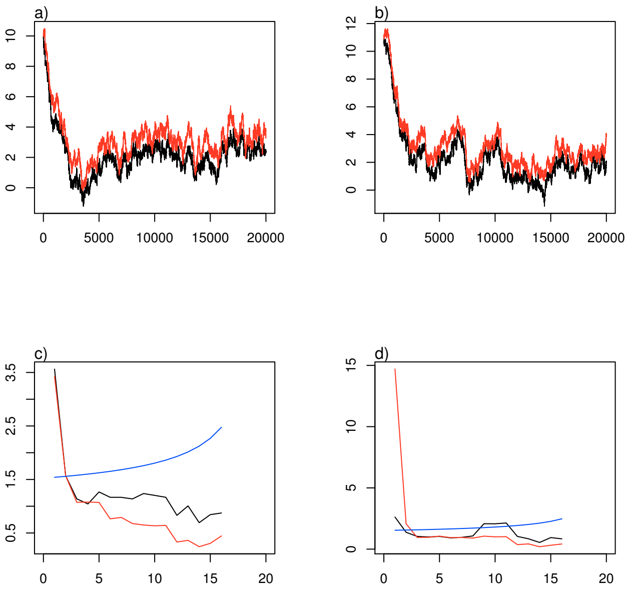

Figure 4.2: Trace plots (a-b) of the two replicates of the Metropolis–Hastings algorithm for $\delta = 0.1$ with starting values of (10,10) and (11,11) respectively. Panel c) gives the univariate diagnostic plot for each variable (black line= first variate, red line = second variate, blue=upper 95% of the appropriate F distribution). Panel d) gives the multivariate diagnostic plot (black line = determinant, red line = trace, blue=upper 95% of the appropriate F distribution).

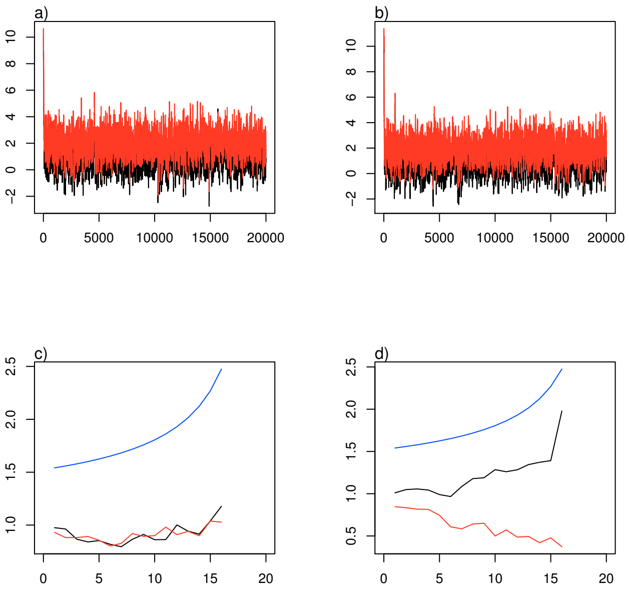

Figure 4.3: Trace plots (a-b) of the two replicates of the Metropolis–Hastings algorithm for $\delta = 2.0$ with starting values of (10,10) and (11,11) respectively. Panel c) gives the univariate diagnostic plot for each variable (black line= first variate, red line = second variate, blue=upper 95% of the appropriate F distribution). Panel d) gives the multivariate diagnostic plot (black line = determinant, red line = trace, blue=upper 95% of the appropriate F distribution).

CHAPTER 5

LINEAR MIXING MODELS

In this chapter, we introduce statistical models that have been used in the literature that solve similar problems to the diet problem which we study in more detail in the next chapter. We use the nomenclature of Wolbers and Stahel (2005) and refer to the general class of models under consideration as linear mixing models.

To fix ideas, consider the more commonly encountered mixture model, where we model the distribution of our, possibly vector valued, observations $\mathbf{y}_1, \ldots, \mathbf{y}_n$ as mixture of p components. That is, each observation, $\mathbf{y}_i$, comes from one of the p distributions, which we denote by $f_j(\mathbf{y}_i|\boldsymbol{\theta}_j), j = 1, \ldots, p$, which depend on a vector valued unknown parameter $\boldsymbol{\theta}_j$. Thus we denote the distribution of the $\mathbf{y}_i$ as follows

$$f(\mathbf{y}_i|\boldsymbol{\pi}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p) = \pi_1 f_1(\mathbf{y}_i|\theta_1) + \pi_2 f_1(\mathbf{y}_i|\theta_2) + \cdots + \pi_p f_p(\mathbf{y}_i|\theta_p) \;\; i = 1, \ldots, n$$

where $\sum_{i=1}^{p} \pi_i = 1$ are the proportions of the population that belong to each of the components. The parameters in this case are the mixing proportions $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_p)'$ and the unknown parameters of the distribution $\boldsymbol{\theta}_j$. There is a vast literature on mixture models see Titterington et al. (1985) and the references therein.

The linear constant mixing model can be written as follows

$$\mathbf{y}_i = \boldsymbol{\Theta}\boldsymbol{\alpha} + \boldsymbol{\epsilon}_i, \;\; i = 1, \ldots, n$$

where $\mathbf{y}_i$ is a column vector of length a, $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1| \ldots |\boldsymbol{\theta}_p]$ is a $a \times p$ dimensional matrix, $\boldsymbol{\theta}_j, j = 1, \ldots, p$ are column vectors of length a representing the components, sources or prey profiles, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)'$ is a column vector of length p of mixing coefficients and $\boldsymbol{\epsilon}_i$ is a column vector of length a.

The linear multilevel mixing model can then be written as follows:

$$\mathbf{y}_i = \boldsymbol{\Theta}\boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_i, \;\; i = 1, \ldots, n$$

where $\boldsymbol{\alpha}_i$ is the mixing vector for each observation.

Finite mixture models assert that each observation $\mathbf{y}_i$ belongs to one of the p distributions and the goal of the inference to determine the parameters of each of the component mixtures and the relative proportions of each that are present. Linear multilevel mixing models assert that each observation is a linear combination of the components plus noise, which is the key distinction between the two types of models.

Several disciplines have tackled the linear mixing problem in various guises and under different assumptions and terminology which makes deciphering the literature on this subject a difficult undertaking. We don't attempt a comprehensive review, but give some key references and a brief discussion of the issues relevant for the diet problem discussed in the next chapter. Perhaps the simplest linear mixing model is based on the concept of conservation of mass, typically known as chemical mass balance models. In our framework, we have

$$y_j = \sum_{i=1}^{p} \theta_{ij}\alpha_j + \epsilon_j$$

where $\mathbf{y} = (y_1, \ldots, y_a)$ is an a dimensional vector of chemical concentrations measured at a receptor site, $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{ia})'$ are a dimensional vectors of known sources, the mixing vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)$ and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_a)'$. Early chemical mass balance models solved the above equation for $\boldsymbol{\alpha}$ for each receptor site separately, usually by linear least squares of variants there of. Henry et al. (1984) gives an excellent review of the state of the art at that time, the methods include the tracer element method, the linear programming method, ordinary least squares, effective variance least squares and ridge regression to handle multicollinearity of the sources. This class of methods typically ignores the multivariate nature of the receptor measurements and also doesn't address the fact that the number of parameters increases with the sample size. This ignores the classical problem recognized by Neymann and Scott (1948) and further studied by Kiefer and Wolfowitz (1956) of having no asymptotic theory to show consistency of the resulting estimates. Henry et al. (1984) also considers multivariate models, where they deal with multiple measurements at the receptor site simultaneously. The goal of these models is to estimate both the sources $\boldsymbol{\theta}$'s and their contributions to the observed receptor $\boldsymbol{\alpha}_i$ in models like the following

$$\mathbf{y}_{ik} = \sum_{j=1}^{p} \theta_{jk}\alpha_{ij} + \boldsymbol{\epsilon}_j.$$

This model has received much recent attention. Specifically, estimating the number of sources, p, the source profiles θ_{jk} and the mixing vectors $\boldsymbol{\alpha}_{ij}$ all from receptor observations. Bandeen-Roche and Ruppert (1991); Bandeen-Roche (1994); Christensen and Sain (2002); Park et al. (2000, 2001); Billheimer (2001); Park et al. (2002); Wolbers and Stahel (2005) all deal with various aspects of these problems in the multivariate setting.

In the geological literature this problem is known as the end–member problem, see Renner (1993); Aitchison and Bacon-Shone (1999) and the references therein for an account of the linear mixing model in Geology.

An issue with the multivariate models, where the emphasis is typically on estimating both the sources and the mixing vectors is a lack of identifiability. To see this write the multivariate model in matrix notation as follows:

$$\mathbf{y}_i = \boldsymbol{\Theta}\boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_i$$

Let $\mathbf{T}$ be a $p \times p$ invertible matrix and define $\boldsymbol{\Theta}^* = \boldsymbol{\Theta}\mathbf{T}$ and $\boldsymbol{\alpha}^* = \mathbf{T}^{-1}\boldsymbol{\alpha}$ then

$$\mathbf{y}_i = \boldsymbol{\Theta}^*\boldsymbol{\alpha}_i^* + \boldsymbol{\epsilon}_i$$

gives an equivalent model and is hence non–identifiable, without some further restrictions. This is not of direct interest for our development, but it is a concern when there are no observations on the sources or when no prior information is available. Billheimer (2001) describes an approach to the problem using strong prior information on the source composition. We review the Bayesian approach to linear mixing models in some detail in this chapter before adapting them to the problem of predator diets in the next chapter. Also note, we refer to the models in the more generic sense of linear mixing, adapted from Wolbers and Stahel (2005).

5.1 Linear Mixing

In this section we introduce the two basic linear mixing models: the linear constant mixing model and the linear multilevel mixing model. We keep the description quite general at first and then introduce differing assumptions that characterize the many versions of the models.

The linear constant mixing model can be written as follows

$$\underset{(a\times 1)}{\mathbf{y}_i} = \underset{(a\times p)}{\boldsymbol{\Theta}}\underset{(p\times 1)}{\boldsymbol{\alpha}} + \underset{(a\times 1)}{\boldsymbol{\epsilon}_i} \quad i = 1, \ldots, n,$$

where $\mathbf{y}_i$ is a column vector of length a, $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1| \ldots |\boldsymbol{\theta}_p]$ is a $a \times p$ dimensional matrix,

$\boldsymbol{\theta}_j, j = 1, \ldots, p$ are column vectors of length a representing the components, sources or in the next chapter the prey fatty acid profiles, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)'$ is a column vector of length p of mixing coefficients and $\boldsymbol{\epsilon}_i$ is a column vector of length a. Note that we have written the model in its additive form which is appropriate when there are no restrictions on the spaces. However, the usual additive errors get replaced with component-wise multiplication, denoted by $\odot$, when the space is restricted to the positive quadrant or the perturbation operator, denoted by $\oplus$, when the space is restricted to the simplex (see 2.1).

The linear constant mixing model can be expressed in matrix notation as follows:

$$\begin{aligned} \underset{(a\times n)}{\mathbf{Y}} &= \underset{(a\times p)}{\boldsymbol{\Theta}}\underset{(p\times n)}{\Gamma} + \underset{(a\times n)}{\mathbf{E}} \\ \underset{(p\times n)}{\Gamma} &= \boldsymbol{\rho}_c^{-1}\left(\boldsymbol{\rho}\left(\underset{(p\times 1)}{\boldsymbol{\alpha}}\right)\underset{(1\times n)}{\mathbf{W}}\right) \end{aligned} \tag{5.1}$$

where $\mathbf{Y} = [\mathbf{y}_i|\ldots|\mathbf{y}_n]$ is an $a \times n$ matrix, $\mathbf{E} = [\boldsymbol{\epsilon}_1|\ldots|\boldsymbol{\epsilon}_n]$ is an $a \times n$ matrix of errors, $\boldsymbol{\rho}(.)$ represents a transformation, $\boldsymbol{\rho}^{-1}(.)$ its inverse, $\boldsymbol{\rho}_c^{-1}(.)$ its inverse applied to the columns of a matrix and $\mathbf{W}$ is a known $1 \times n$ design matrix of ones. Typical choices for the vector valued function, $\boldsymbol{\rho}(.)$, will depend on the restrictions imposed on the mixing vector. No restrictions would correspond to the identity transformation, $\boldsymbol{\rho}(\mathbf{x}) = \mathbf{x}$, while restricting the mixing vector to the positive quadrant would correspond to the usual log transformation, $\boldsymbol{\rho}(\mathbf{x}) = \log(\mathbf{x})$ and restricting the space to the simplex would correspond to the log–ratio transformation, $\boldsymbol{\rho}(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})$.

We can generalize this model to allow w different populations, by expanding the design matrix $\mathbf{W}$ to dimension $w \times n$. Thus our model becomes

$$\begin{aligned} \underset{(a\times n)}{\mathbf{Y}} &= \underset{(a\times p)}{\boldsymbol{\Theta}}\underset{(p\times n)}{\Gamma} + \underset{(a\times n)}{\mathbf{E}} \\ \underset{(a\times n)}{\Gamma} &= \boldsymbol{\rho}_c^{-1}\left(\boldsymbol{\rho}_c\left(\underset{(p\times w)}{\mathbf{A}}\right)\underset{(w\times n)}{\mathbf{W}}\right) \end{aligned}$$

where $\mathbf{A} = [\boldsymbol{\alpha}_1|\ldots|\boldsymbol{\alpha}_w]$ is a $p \times w$ dimensional matrix with the columns corresponding to the different population mixing vectors $\boldsymbol{\alpha}_i, i = 1, \ldots, w$. For example, assume we have two populations, one choice for the matrix $\mathbf{W}$ would be the following

$$\underset{(2\times(n_1+n_2))}{\mathbf{W}_1} = \begin{bmatrix} 1 & \ldots & 1 & 0 & \ldots & 0 \\ 0 & \ldots & 0 & 1 & \ldots & 1 \end{bmatrix}$$

where the first row has n_1 ones followed by n_2 zeros and the second row has n_1 zeros followed by n_2 ones. Or a second choice would be

$$\underset{(2\times(n_1+n_2))}{\mathbf{W}_2} = \begin{bmatrix} 1 & \dots & 1 & 1 & \dots & 1 \\ 0 & \dots & 0 & 1 & \dots & 1 \end{bmatrix},$$

where the first row is all ones and the second row has n_1 zeros followed by n_2 ones. Design matrix $\mathbf{W}_2$ corresponds to the usual dummy variable coding used in traditional analysis of variance models, while $\mathbf{W}_1$ would correspond to the effects modeling approach. With the $\mathbf{W}_2$ parametrization, $\boldsymbol{\alpha}_1$ represents the overall mixing vector in the two populations and $\boldsymbol{\alpha}_2$ is the difference between population two and population one. Using the $\mathbf{W}_1$ parametrization, $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ represent the mixing vectors in the two populations.

Consider the linear multilevel mixing model which can be seen as a generalization of the linear constant mixing model. Linear multilevel mixing models, typically, do not consider the case of multiple measurements per random effect or level. We develop the models in this slightly more general context. For example, assume we have air quality measurements taken in the morning and afternoon and we believe the air quality doesn't change substantially over the day, and we wish to apportion the air pollution to various sources on a daily basis.

The basic linear multilevel mixing model is given by the following

$$\underset{(a\times 1)}{\mathbf{y}_{ij}} = \underset{(a\times p)}{\boldsymbol{\Theta}}\underset{(p\times 1)}{\boldsymbol{\alpha}_i} + \underset{(a\times 1)}{\boldsymbol{\epsilon}_{ij}}\ ,\ \ i = 1, \dots, n,\ j = 1, \dots, n_i$$

where $\boldsymbol{\alpha}_i$ is a $p \times 1$ mixing vector for the ith level or random effect. For simplicity of presentation we assume that $n_i = r$, however, this is not required. If we collect the r replicates into a matrix as follows

$$\underset{(a\times r)}{\mathbf{Y}_i} = [\mathbf{y}_{i1}|\dots|\mathbf{y}_{ir}]$$

we can now write the above model in matrix notation as follows:

$$\underset{(a\times r)}{\mathbf{Y}_i} = \underset{(a\times p)}{\boldsymbol{\Theta}}\underset{(p\times 1)}{\boldsymbol{\alpha}_i}\underset{(1\times r)}{\mathbf{U}} + \underset{(a\times r)}{\mathbf{E}_i}\ ,\ \ i = 1, \dots, n$$

where, $\mathbf{E}_i$ is defined in an analogous fashion to $\mathbf{Y}_i$ and $\mathbf{U}$ is a $1 \times r$ matrix of ones.

Before presenting the linear multilevel mixing model in matrix form, we need the following definition of a Kronecker product:

Definition 5.1. *The Kronecker product, denoted by $\otimes$, of an $m \times n$ matrix $\mathbf{C}$ and $p \times q$ matrix $\mathbf{D}$ is an $(mp) \times (nq)$ matrix given by the following*

$$\mathbf{C} \otimes \mathbf{D} = \begin{bmatrix} c_{11}D & \dots & c_{1n}D \\ \vdots & \ddots & \vdots \\ c_{m1}D & \dots & c_{mn}D \end{bmatrix},$$

Finally we can join the matrices $\mathbf{Y}_i$ by concatenating the columns as follows

$$\underset{(a\times nr)}{\mathbf{Y}} = [\mathbf{Y}_1|\dots|\mathbf{Y}_n].$$

This allows us to write the model in matrix notation as follows

$$\underset{(a\times nr)}{\mathbf{Y}} = \underset{(a\times p)}{\mathbf{\Theta}} \underset{(p\times n)}{\mathbf{A}} \otimes \underset{(1\times r)}{\mathbf{U}} + \underset{(a\times nr)}{\mathbf{E}},$$

where $A = [\boldsymbol{\alpha}_1|\dots|\boldsymbol{\alpha}_n]$ an $p \times n$ matrix of mixing vectors, $\mathbf{E}$ is defined analogously to $\mathbf{Y}$ and $\mathbf{A} \otimes \mathbf{U}$ means replicate each mixing vector, $\boldsymbol{\alpha}_i$, r times column–wise.

We can extend the linear multilevel mixing model to allow for multiple populations in a similar fashion as for the linear constant mixing model. To make the models more explicit we adapt the structural equation model formulation (see Lee, 2007) which has connections to the state space formulation. The $\boldsymbol{\alpha}_i$ plays the role of the latent variable or the unobserved state in a state space model. We have the following equation describing the collection of mixing vectors $\mathbf{A}$:

$$\underset{(p\times n)}{\mathbf{A}} = \boldsymbol{\rho}_c^{-1}\left(\boldsymbol{\rho}_c\left(\underset{(p\times w)}{\mathbf{A}^f}\right)\underset{(w\times n)}{\mathbf{W}} + \underset{(p\times n)}{\mathbf{A}^*}\right)$$

where $\mathbf{W}$ is a known $w \times n$ design matrix for the linear multilevel mixing model. Note as in the linear constant mixing model we assume that the population means are additive on the $\boldsymbol{\rho}$ scale. We denote the population mean mixing vectors by $\boldsymbol{\mu}_{\boldsymbol{\alpha}_j}, j = 1, \dots, w$ and collect these into the matrix $\mathbf{A}^f$

$$\underset{(p\times w)}{\mathbf{A}^f} = [\boldsymbol{\mu}_{\boldsymbol{\alpha}_1}|\dots|\boldsymbol{\mu}_{\boldsymbol{\alpha}_w}].$$

Finally, $\mathbf{A}^*$ is a matrix of deviations from the populations means, the distribution of the columns of $\mathbf{A}^*$ will be discussed subsequently.

We are now in a position to write the generalization of the linear multilevel mixing model to multiple populations as follows

$$\begin{aligned}
\underset{(a\times nr)}{\mathbf{Y}} &= \underset{(a\times p)}{\boldsymbol{\Theta}}\underset{(p\times n)}{\mathbf{A}} \otimes \underset{(1\times r)}{\mathbf{U}} + \underset{(a\times nr)}{\mathbf{E}}, \\
\underset{(p\times n)}{\mathbf{A}} &= \boldsymbol{\rho}_c^{-1}\left(\boldsymbol{\rho}_c\left(\underset{(p\times w)}{\mathbf{A}^f}\right)\underset{(w\times n)}{\mathbf{V}} + \underset{(p\times n)}{\mathbf{A}^*}\right) \qquad (5.2)
\end{aligned}$$

For the present discussion we assume that the profiles/sources are known, that is, $\boldsymbol{\Theta}$ is a known matrix. To complete the model specifications we assign prior distributions to the unknown parameters and assign sampling distributions to the observables. Specifically for the linear constant mixing model for a single population we assign:

$$\begin{aligned}
\boldsymbol{\epsilon}_i|\Sigma_{\boldsymbol{\epsilon}} &\sim f(\boldsymbol{\epsilon}_i|\mathbf{0}, \Sigma_{\boldsymbol{\epsilon}}) \\
\boldsymbol{\alpha}|\boldsymbol{\mu}_{\boldsymbol{\alpha}}, \Sigma_{\boldsymbol{\alpha}} &\sim g(\boldsymbol{\alpha}|\boldsymbol{\mu}_{\boldsymbol{\alpha}}, \Sigma_{\boldsymbol{\alpha}}) \\
\Sigma_{\boldsymbol{\epsilon}}|\delta_{\boldsymbol{\epsilon}}, \Psi_{\boldsymbol{\epsilon}} &\sim h(\Sigma_{\boldsymbol{\epsilon}}|\delta_{\boldsymbol{\epsilon}}, \Psi_{\boldsymbol{\epsilon}}),
\end{aligned}$$

where f, g are multivariate density functions and h is a matrix valued density function. These assumptions induce the following joint conditional distribution on $\mathbf{y}_i$'s,

$$\mathbf{y}_1, \ldots, \mathbf{y}_n|\boldsymbol{\alpha}, \Sigma_{\boldsymbol{\epsilon}}, \mathcal{B} \sim \prod_{i=1}^{n} f(\mathbf{y}_i|\boldsymbol{\Theta}\boldsymbol{\alpha}, \Sigma_{\boldsymbol{\epsilon}})$$

where $\mathcal{B}$ is the collection of prior parameters and any other relevant background information, $\mathcal{B} = \{\boldsymbol{\mu}_{\boldsymbol{\alpha}}, \Sigma_{\boldsymbol{\alpha}}, \delta_{\boldsymbol{\epsilon}}, \Psi_{\boldsymbol{\epsilon}}, \boldsymbol{\Theta}\}$. The above equation implies that the $\mathbf{y}_i$'s are exchangeable or conditionally independent. The full posterior distribution is given by

$$p(\boldsymbol{\alpha}, \Sigma_{\boldsymbol{\epsilon}}, \boldsymbol{\Theta}|\mathcal{D}, \mathcal{B}) \propto h(\Sigma_{\boldsymbol{\epsilon}}|\delta, \Psi) \times g(\boldsymbol{\alpha}|\mu_{\boldsymbol{\alpha}}, \Sigma_{\boldsymbol{\alpha}}) \times \prod_{i=1}^{n} f(\mathbf{y}_i|\boldsymbol{\Theta}\boldsymbol{\alpha}, \Sigma_{\boldsymbol{\epsilon}}),$$

where $\mathcal{D} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ is the collection of all observed data. It is unlikely that the above full posterior distribution will be available in closed form and thus exact posterior inference will not be available to us. However, we can readily generate samples from the posterior distribution via MCMC. In addition, the hierarchical nature of the model will allow for relatively straightforward Gibbs or Metropolis-Hastings updates.

We assign the following sampling distributions and priors to complete the specification of the linear multilevel mixing problem for a single population as follows:

$$\begin{array}{rcl}
\boldsymbol{\epsilon}_i|\Sigma_{\boldsymbol{\epsilon}} & \sim & f_1(\boldsymbol{\epsilon}_i|\mathbf{0}, \Sigma_{\boldsymbol{\epsilon}}) \\
\boldsymbol{\alpha}_i|\boldsymbol{\mu}_{\boldsymbol{\alpha}}, \Sigma_{\boldsymbol{\alpha}} & \sim & f_2(\boldsymbol{\alpha}_i|\boldsymbol{\mu}_{\boldsymbol{\alpha}}, \Sigma_{\boldsymbol{\alpha}}) \\
\boldsymbol{\mu}_{\boldsymbol{\alpha}}|\tau, \Sigma_{\boldsymbol{\mu}_{\boldsymbol{\alpha}}} & \sim & g(\boldsymbol{\mu}_{\boldsymbol{\alpha}}|\boldsymbol{\tau}, \Sigma_{\boldsymbol{\mu}_{\boldsymbol{\alpha}}}) \\
\Sigma_{\boldsymbol{\epsilon}}|\delta_{\boldsymbol{\epsilon}}, \Psi_{\boldsymbol{\epsilon}} & \sim & h_1(\Sigma_{\boldsymbol{\epsilon}}|\delta_{\boldsymbol{\epsilon}}, \Psi_{\boldsymbol{\epsilon}}) \\
\Sigma_{\boldsymbol{\alpha}}|\delta_{\boldsymbol{\alpha}}, \Psi_{\boldsymbol{\alpha}} & \sim & h_2(\Sigma_{\boldsymbol{\alpha}}|\delta_{\boldsymbol{\alpha}}, \Psi_{\boldsymbol{\alpha}}),
\end{array}$$

where f_1, f_2, g are multivariate density functions and h_1, h_2 are matrix valued density functions. As with the linear constant mixing model, the above assumptions induce the following joint conditional distribution on $\mathbf{y}_{ij}$'s,

$$\mathbf{y}_{11}, \ldots, \mathbf{y}_{nr}|\boldsymbol{\alpha}_i, \boldsymbol{\mu}_{\boldsymbol{\alpha}}, \boldsymbol{\Theta}, \Sigma_{\boldsymbol{\alpha}}, \Sigma_{\boldsymbol{\epsilon}}, \mathcal{B} \sim \prod_{i=1}^{n}\prod_{j=1}^{r} f(\mathbf{y}_{ij}|\boldsymbol{\Theta}\boldsymbol{\alpha}_i, \Sigma_{\boldsymbol{\epsilon}})$$

where $\mathcal{B} = \{\boldsymbol{\tau}, \Sigma_{\boldsymbol{\mu}_{\boldsymbol{\alpha}}}, \delta_{\boldsymbol{\epsilon}}, \Psi_{\boldsymbol{\epsilon}}, \delta_{\boldsymbol{\alpha}}, \Psi_{\boldsymbol{\alpha}}\}$. That is, the $\mathbf{y}_{ij}$'s are exchangeable conditional on the parameters including α_i. The full posterior distribution is given by

$$\begin{aligned}
p(\boldsymbol{\alpha}_i, \boldsymbol{\mu}_{\boldsymbol{\alpha}}, \Sigma_{\boldsymbol{\alpha}}, \Sigma_{\boldsymbol{\epsilon}}|\boldsymbol{\Theta}, \mathcal{D}, \mathcal{B}) \propto\ & h_1(\Sigma_{\boldsymbol{\epsilon}}|\delta_{\boldsymbol{\epsilon}}, \Psi_{\boldsymbol{\epsilon}}) \times h_2(\Sigma_{\boldsymbol{\epsilon}}|\delta_{\boldsymbol{\alpha}}, \Psi_{\boldsymbol{\alpha}}) \times g(\boldsymbol{\mu}_{\boldsymbol{\alpha}}|\boldsymbol{\tau}, \Sigma_{\boldsymbol{\mu}_{\boldsymbol{\alpha}}}) \\
& \times \prod_{i=1}^{n} f_2(\boldsymbol{\alpha}_i|\boldsymbol{\mu}_{\boldsymbol{\alpha}}, \Sigma_{\boldsymbol{\alpha}}) \times \prod_{i=1}^{n}\prod_{j=1}^{n} f_1(\mathbf{y}_{ij}|\boldsymbol{\Theta}\boldsymbol{\alpha}_i, \Sigma_{\boldsymbol{\epsilon}}),
\end{aligned}$$

where $\mathcal{D} = \{\mathbf{y}_{11}, \ldots, \mathbf{y}_{nr}\}$. As with the linear constant mixing model, it is unlikely that the posterior distribution will available in closed form. Again we will use MCMC to generate samples form the posterior to do approximate posterior inference.

Figures 5.1 and 5.2 give Directed Acyclic Graphs (DAG's) for the generic linear constant mixing model and linear multilevel mixing model respectively. The methods from appendix A.5 enable us to write down the full conditional distributions directly from the DAG which form the basic building blocks of the Gibbs sampler and the Metropolis–Hastings within Gibbs sampler. Whether or not the sampling can be done using Gibbs or Metropolis–Hastings will depend on the functional form of the densities assigned to each component of the model.

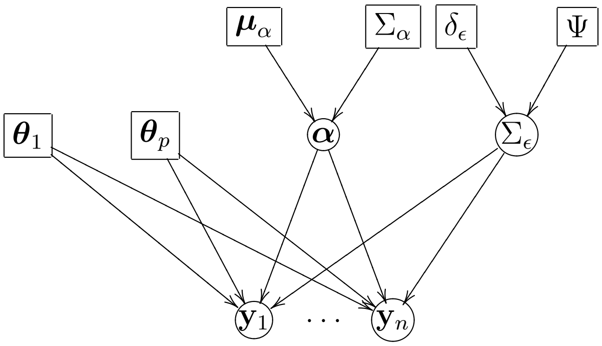

Figure 5.1: Directed Acyclic Graph (DAG) for the linear constant mixing model. Where $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p$ are the sources, $\boldsymbol{\alpha}$ is the mixing vector, $\Sigma_{\boldsymbol{\epsilon}}$ is the error covariance matrix, $\mathbf{y}_1, \ldots, \mathbf{y}_n$ are the observables, $\boldsymbol{\mu}_{\boldsymbol{\alpha}}$ and $\Sigma_{\boldsymbol{\alpha}}$ represent the prior for the mixing vector and $\delta_{\boldsymbol{\epsilon}}$ and $\Psi_{\boldsymbol{\epsilon}}$ represent the prior information for Σ_ϵ. The square nodes indicate parameters that are known a priori, while circular nodes represent unknown parameters that are updated when the data, $\mathbf{y}_i, i = 1, \ldots, n$, are observed.

The full conditional distributions for the linear constant mixing model are as follows:

$$
\begin{aligned}
p(\boldsymbol{\alpha}|\Sigma_{\boldsymbol{\epsilon}}, \boldsymbol{\Theta}, \mathcal{D}, \mathcal{B}) &\propto g(\boldsymbol{\alpha}|\boldsymbol{\mu}_{\boldsymbol{\alpha}}, \Sigma_{\boldsymbol{\alpha}}) \times \prod_{i=1}^{n} f(\mathbf{y}_i|\boldsymbol{\alpha}, \boldsymbol{\Theta}, \Sigma_{\boldsymbol{\epsilon}}) \\
p(\Sigma_{\boldsymbol{\epsilon}}|\boldsymbol{\alpha}, \boldsymbol{\Theta}, \mathcal{D}, \mathcal{B}) &\propto h(\Sigma_{\boldsymbol{\epsilon}}|\delta_{\boldsymbol{\epsilon}}, \Psi_{\boldsymbol{\epsilon}}) \times \prod_{i=1}^{n} f(\mathbf{y}_i|\boldsymbol{\alpha}, \boldsymbol{\Theta}, \Sigma_{\boldsymbol{\epsilon}})
\end{aligned}
$$

and for the linear multilevel mixing model

$$
\begin{aligned}
p(\boldsymbol{\alpha}_i|\boldsymbol{\alpha}_{-i}, \boldsymbol{\mu}_{\boldsymbol{\alpha}}, \Sigma_{\boldsymbol{\alpha}}, \Sigma_{\boldsymbol{\epsilon}}, \boldsymbol{\Theta}, \mathcal{D}, \mathcal{B}) &\propto f_2(\boldsymbol{\alpha}_i|\boldsymbol{\mu}_{\boldsymbol{\alpha}}, \Sigma_{\boldsymbol{\alpha}}) \times \prod_{j=1}^{r} f_1(\mathbf{y}_{ir}|\boldsymbol{\alpha}_i, \boldsymbol{\Theta}, \Sigma_{\boldsymbol{\epsilon}}) \\
p(\boldsymbol{\mu}_{\boldsymbol{\alpha}}|\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n, \Sigma_{\boldsymbol{\alpha}}, \Sigma_{\boldsymbol{\epsilon}}, \boldsymbol{\Theta}, \mathcal{D}, \mathcal{B}) &\propto g(\boldsymbol{\mu}_{\boldsymbol{\alpha}}|\boldsymbol{\tau}, \Sigma_{\boldsymbol{\mu}_{\boldsymbol{\alpha}}}) \times \prod_{i=1}^{n} f_2(\boldsymbol{\alpha}_i|\boldsymbol{\mu}_{\boldsymbol{\alpha}}, \Sigma_{\boldsymbol{\alpha}}) \\
p(\Sigma_{\boldsymbol{\alpha}}|\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n, \boldsymbol{\mu}_{\boldsymbol{\alpha}}, \Sigma_{\boldsymbol{\epsilon}}, \boldsymbol{\Theta}, \mathcal{D}, \mathcal{B}) &\propto h_2(\Sigma_{\boldsymbol{\alpha}}|\delta_{\boldsymbol{\alpha}}, \Psi_{\boldsymbol{\alpha}}) \times \prod_{i=1}^{n} f_2(\boldsymbol{\alpha}_i|\boldsymbol{\mu}_{\boldsymbol{\alpha}}, \Sigma_{\boldsymbol{\alpha}}) \\
p(\Sigma_{\boldsymbol{\epsilon}}|\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n, \boldsymbol{\mu}_{\boldsymbol{\alpha}}, \Sigma_{\boldsymbol{\alpha}}, \boldsymbol{\Theta}, \mathcal{D}, \mathcal{B}) &\propto h_1(\Sigma_{\boldsymbol{\epsilon}}|\delta_{\boldsymbol{\epsilon}}, \Psi_{\boldsymbol{\epsilon}}) \times \prod_{i=1}^{n} \prod_{j=1}^{r} f_1(\mathbf{y}_{ij}|\alpha_i^{'}\boldsymbol{\Theta}, \Sigma_{\boldsymbol{\epsilon}})
\end{aligned}
$$

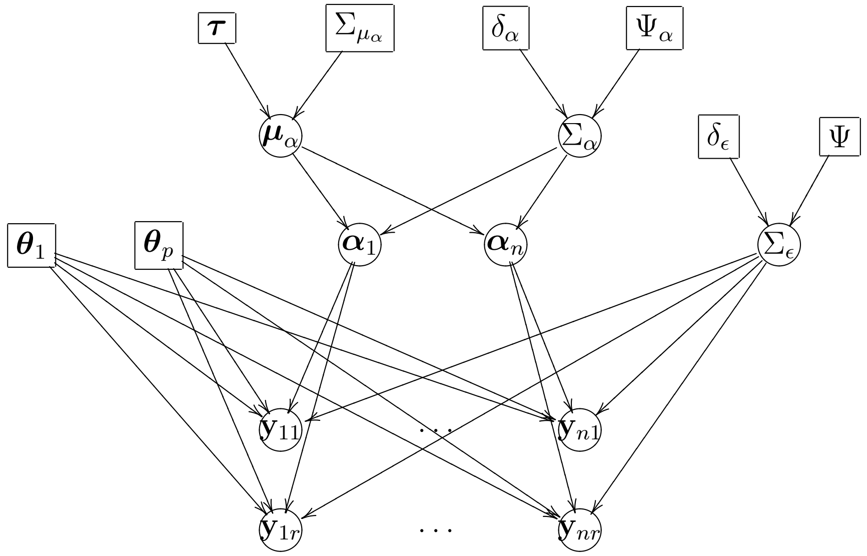

Figure 5.2: Directed Acyclic Graph (DAG) for the linear multilevel mixing model. Where $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p$ are the sources, $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n$ are the mixing vectors, $\boldsymbol{\mu}_{\boldsymbol{\alpha}}$ and $\Sigma_{\boldsymbol{\alpha}}$ describe the distribution of the $\boldsymbol{\alpha}_i$'s, $\Sigma_{\boldsymbol{\epsilon}}$ is the error covariance matrix, $\mathbf{y}_{11}, \ldots, \mathbf{y}_{nr}$ are the observables, $\boldsymbol{\tau}$ and $\Sigma_{\boldsymbol{\mu}_{\boldsymbol{\alpha}}}$ represent the prior information for mean of the mixing vector $\boldsymbol{\mu}_{\boldsymbol{\alpha}}$, $\delta_{\boldsymbol{\alpha}}$ and $\Psi_{\boldsymbol{\alpha}}$ represent the prior information for the variance of the mixing distribution $\Sigma_{\boldsymbol{\alpha}}$ and $\delta_{\boldsymbol{\epsilon}}$ and $\Psi_{\boldsymbol{\epsilon}}$ represent the prior information for Σ_ϵ. The square nodes indicate parameters that are known a priori, while circular nodes represent unknown parameters that are updated when the data, $\mathbf{y}_i, i = 1, \ldots, n$, are observed.

5.2 Special Cases of the Linear Mixing Model

Linear mixing models, as we have described them, have three distinct parts: the source matrix which we denote by $\boldsymbol{\Theta}$, the mixing vector denoted by $\boldsymbol{\alpha}$ or $\boldsymbol{\alpha}_i$ and the noise vector $\boldsymbol{\epsilon}_i$ or $\boldsymbol{\epsilon}_{ij}$. Differing assumptions on the components lead to different mixing models, which is the subject of this section.

As most of the literature on linear mixing models and their variants deal with physical systems, they typically restrict attention to cases where the source matrix $\boldsymbol{\Theta}$ and the errors $\boldsymbol{\epsilon}_i$ reside in the positive quadrant and the mixing vector(s) $(\boldsymbol{\alpha}, \boldsymbol{\alpha}_i)$ is(are) restricted to the positive quadrant. The most common applications are to receptor models in air pollution studies (see Billheimer, 2001; Wolbers and Stahel, 2005; Henry et al., 1984) and end member problems in Geology pollution (see Aitchison and Bacon-Shone, 1999; Renner, 1993) .

Wolbers and Stahel (2005) give an excellent description of the different types of models used in the air pollution literature and their basic assumptions. However, the assumptions given in Wolbers and Stahel (2005) are slightly more restrictive than necessary for our purposes as they consider the situation where the sources are not known and do not have observations on them.

The discussion, thus far, has not placed any restrictions on the mixing vectors, the sources or the errors, that is, they are all free to vary in the appropriate multi–dimensional real space. However, as pointed out by Wolbers and Stahel (2005), this form of the model is not physically realizable in air pollution studies, not to mention other areas of application. Table 5.1 lays out the various restrictions on $\mathbf{\Theta}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\epsilon}_i$ and gives potential candidate distributions to describe the various distributional assumptions over the reduced spaces. It also indicates which combinations of models are logically consistent. Note, it treats the triplet $\mathbf{\Theta}$, the choice of operator and $\boldsymbol{\epsilon}$ as one component of the model. It then places restrictions on the $\boldsymbol{\alpha}$ that are consistent with the choice of triplet. Note that the operators mentioned in the table are the usual addition, $\odot$ is elementwise multiplication and $\oplus$ is the compositional operator introduced in chapter 2.1.

Rather than discuss all six logically consistent models we give illustrative examples relying on the Markov Chain Monte Carlo machinery in all cases even in situations where the models are analytically tractable. Specifically we consider two cases of the linear mixing models: $\boldsymbol{\theta}_j \in \Re^a$, $\boldsymbol{\epsilon}_i \in \Re^a$, $\boldsymbol{\alpha} \in \Re^p$and $\boldsymbol{\rho}(\mathbf{x}) = \mathbf{x}$; and $\boldsymbol{\theta}_j \in \Re^a$, $\epsilon_{ij} \in \Re^a$, $\boldsymbol{\alpha} \in \mathcal{S}^p$ and $\boldsymbol{\rho}(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})$.

5.2.1 Case a: $\boldsymbol{\theta}_j \in \Re^a$ and $\boldsymbol{\alpha} \in \Re^p$

The linear constant mixing model for a single population given in equation (5.1) is described by three general distributions: a sampling distribution $f(\mathbf{y}_i|\mathbf{\Theta}\boldsymbol{\alpha}, \Sigma_{\boldsymbol{\epsilon}})$, a prior on $\boldsymbol{\alpha}$ denoted by $g(\boldsymbol{\alpha}|\boldsymbol{\mu}_{\boldsymbol{\alpha}}, \Sigma_{\boldsymbol{\alpha}})$ and a prior on Σ_ϵ denoted by $h(\Sigma_{\boldsymbol{\epsilon}}|\delta_{\boldsymbol{\epsilon}}, \Psi_{\boldsymbol{\epsilon}})$. For purposes of illustration we assign a multivariate normal distribution to the sampling distribution, with mean $\mathbf{\Theta}\boldsymbol{\alpha}$ and covariance matrix $\Sigma_{\boldsymbol{\epsilon}}$, a multivariate normal distribution to the prior distribution of $\boldsymbol{\alpha}$ with mean $\boldsymbol{\mu}_{\boldsymbol{\alpha}}$ and covariance matrix $\Sigma_{\boldsymbol{\alpha}}$ and finally a Inverse–Wishart, denoted by $\mathcal{IW}$, with degrees of freedom δ_ϵ and scale matrix Ψ_ϵ to $\Sigma_{\boldsymbol{\epsilon}}$. Note that, if $\mathbf{\Theta} = \mathbf{I}_a$ then the linear constant mixing model (5.1) reduces to the usual multivariate analysis of variance model.

Similarly for the linear multilevel mixing model, we assign multivariate normal distributions to the generic distributions f_1 , f_2 and g with appropriate meanings for the parameters and Inverse Wishart distributions to h_1 and h_2. If $\mathbf{\Theta} = \mathbf{I}_a$ then the linear multilevel mixing model 5.2) reduces to a multivariate random effects model.

When $\mathbf{\Theta}$ is known, it can be shown that (5.1) and (5.2) are identifiable provided $\mathbf{\Theta}$ is of

	$\boldsymbol{\theta}_j \in \Re^a$	$\boldsymbol{\theta}_j \in \Re^{a+}$	$\boldsymbol{\theta}_j \in \mathcal{S}^a$
Operator	$+$	$\odot$	$\oplus$
Linear Constant Mixing:			
$\boldsymbol{\alpha}$	$\boldsymbol{\epsilon}_i \sim \mathcal{N}^a$	$\boldsymbol{\epsilon}_i \sim \mathcal{LN}^a$	$\boldsymbol{\epsilon}_i \sim \mathcal{L}^{a-1}$
$\boldsymbol{\alpha} \in \Re^p$	$\surd$	$\times$	$\times$
$\boldsymbol{\alpha} \in \Re^{p+}$	$\surd$	$\surd$	$\times$
$\boldsymbol{\alpha} \in \mathcal{S}^p$	$\surd$	$\surd$	$\surd$
Linear Multilevel Mixing:			
$\boldsymbol{\alpha}_i$	$\boldsymbol{\epsilon}_{ij} \sim \mathcal{N}^a$	$\boldsymbol{\epsilon}_{ij} \sim \mathcal{LN}^a$	$\boldsymbol{\epsilon}_{ij} \sim \mathcal{L}^{a-1}$
$\boldsymbol{\alpha}_i \sim \mathcal{N}^p$	$\surd$	$\times$	$\times$
$\boldsymbol{\alpha}_i \sim \mathcal{LN}^p$	$\surd$	$\surd$	$\times$
$\boldsymbol{\alpha}_i \sim \mathcal{L}^{p-1}$	$\surd$	$\surd$	$\surd$

Table 5.1: Distributional assumptions for the linear constant mixing model and the linear multilevel mixing model. Where $\boldsymbol{\theta}_j$ is the jth column of the source matrix $\boldsymbol{\Theta}$, $\Re^a$ is a–dimensional real space, $\Re^{a+}$ is the positive quadrant of a–dimensional real space, $\mathcal{S}^a$ is the a– dimensional simplex, $\odot$ means elementwise multiplication and $\oplus$ is the perturbation operator introduced in chapter 2, $\mathcal{N}^a$ is the a–dimensional multivariate normal distribution, $\mathcal{LN}^a$ is multivariate log normal distribution defined on $\Re^{a+}$, $\mathcal{L}^{a-1}$ represents the logistic normal distribution on the $\mathcal{S}^a$. Appendix C gives the functional forms of the multivariate normal, the multivariate log–normal and the logistic normal distributions. Check marks indicate model combinations that are consistent.

rank p, (see Henry et al., 1984). Practically this means that the number of sources needs to be less than or equal to the dimension of the measurement vector $\mathbf{y}$. However, even if $\boldsymbol{\Theta}$ is of full rank, the model can still show signs of non-identifiability if the condition number of $\boldsymbol{\Theta}$ is large, which indicates multicollinearity among the columns $\boldsymbol{\Theta}$. This is well known in the chemical mass balance model literature; see Henry et al. (1984). It is of interest to see what effect this has on the approximate inference we perform on these models via MCMC sampling.

The condition number of an $n \times m$ matrix, with $n < m$ is defined as follows:

$$\text{cond}(A) = \frac{\max \text{eigen–value}(AA')}{\min \text{eigen–value}(AA')}.$$

To investigate the effects of an ill–conditioned source matrix $\boldsymbol{\Theta}$ we generated three source profiles of dimension $a = 5$ with different condition numbers, these are given in table 5.2. The profiles were generated from a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix:

$$\begin{bmatrix} 1 & \rho & 0 \\ \rho & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The three cases were generated by setting $\rho = 0, 0.99, 0.9999$ respectively, ranging from very little dependence to almost perfect dependence between the first two components of the source matrix. The resulting sources are given in table 5.2.

Example	Source	R1	R2	R3	R4	R5	Condition Number
a)	1	-1.301	-1.210	-0.547	0.038	-0.216	
	2	0.384	-0.970	-0.915	-0.027	0.734	
	3	-0.767	2.819	-0.004	0.686	1.669	7.063
b)	1	0.035	0.839	0.244	-0.046	-0.560	
	2	-0.065	0.986	0.117	-0.244	-0.677	
	3	2.052	-1.066	0.816	0.113	-0.414	282.97
c)	1	1.151	-0.879	-0.239	-0.602	0.950	
	2	1.152	-0.881	-0.242	-0.608	0.952	
	3	-0.571	0.087	-0.300	-1.480	0.163	5155448

Table 5.2: Source profiles used in generating synthetic data to demonstrate the effect of ill-conditioned matrices on the two versions of the linear mixing model given in equations (5.1) and (5.2). The rows of the table indicate the columns of $\boldsymbol{\Theta}$.

Using the linear constant mixing model (5.1) we generated 50 samples for three mixing vectors, $\boldsymbol{\alpha} = (1, 1, 1)', (3, -2, -1)', (-1, 3, 3)'$ and three covariance matrices $\Sigma_{\boldsymbol{\epsilon}} = \sigma^2_{\boldsymbol{\epsilon}} \mathbf{I}_d$ with $\sigma^2_{\boldsymbol{\epsilon}} = 0.5, 1.0, 1.5$. We used "vague" priors for both $\boldsymbol{\alpha}$ and $\Sigma_{\boldsymbol{\epsilon}}$ specifically

$$p(\boldsymbol{\alpha}) = \mathcal{N}^p(\mathbf{0}_p, 100\mathbf{I}_p)$$

and

$$p(\Sigma_{\boldsymbol{\epsilon}}) = \mathcal{IW}^a(p, 100\mathbf{I}_a).$$

We generated 100,000 posterior samples using a Metropolis–Hastings within Gibbs MCMC algorithm with a thinning factor of 10. The full conditional distribution for $\Sigma_{\boldsymbol{\epsilon}}$ is an inverse–Wishart distribution while the distribution for $\boldsymbol{\alpha}$ was updated using an adaptive Metropolis–Hastings algorithm. In this case, a full Gibbs sampler could be implemented, but we chose to use a Metropolis–Within–Gibbs algorithm to mimic what happens with the more complicated models to come. We implemented a simple adaptive algorithm, in that we didn't follow the state of the art as suggested in Andrieu and Thoms (2008) though the ideas were very much along the vein discussed there. A multivariate normal proposal distribution was used which was centered at the current state with an adaptive variance given by $\sigma^2_{\text{adapt}}\mathbf{I}_p$ where σ^2_{adapt} is adaptively updated or controlled by keeping the acceptance rate near the optimal value of 0.23. That is, we did not allow the algorithm to adaptively learn about the variance of the posterior. In essence our algorithm is an adaptively controlled algorithm, rather than a fully adaptive algorithm. For more details see the MCMC chapter 3. The starting values for $\boldsymbol{\alpha}$ were set to zero for each component.

Figure 5.3 shows trace plots of the MCMC output for the mixing vector $\boldsymbol{\alpha}$. Column one of the figure shows very rapidly mixing chains that quickly converge to the correct area of posterior space. Thus, the linear constant mixing model behaves very well when there is little or no dependence among the columns of the source matrix $\boldsymbol{\Theta}$. The picture for column three is not nearly as promising (the highest degree of dependence among the columns of the $\boldsymbol{\Theta}$ matrix). It shows all the tell-tale signs of a very slowly mixing chain and posterior inference without running the chain for much longer would be problematic. However, there is an interesting symmetry present. The total of the first two components matches quite well to the amount apportioned to the first two components. This is not surprising as the first two components are almost identical in this case.

Plotting the components in a pairwise fashion shows this strong dependence between the first two components and indicates that the posterior surface actually resides in lower dimensional subspace. This has implications for the diet problem discussed in the next chapter and will be much harder to diagnose as the dimension of the problem increases and the pattern of dependence becomes more complicated than presented in this simplified situation. The picture in the middle column is a comprise between the first and third columns of the figure. However, there is still a strong pattern of dependence between the first two components.

Table 5.3 gives the posterior mean of the mixing vector $\boldsymbol{\alpha}$ and the trace of $\Sigma_{\boldsymbol{\epsilon}}$ denoted by Σ^T_{ϵ} along with their associated component-wise 95% credible intervals. Results are given for the three $\boldsymbol{\Theta}$ matrices (see table 5.2), three error covariances denoted by $\sigma^2_{\boldsymbol{\epsilon}}$ and three different mixing vectors $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$ and $\boldsymbol{\alpha}_3$. The results for the nearly orthogonal source matrix (case a), indicate a well behaved posterior distribution with relatively short 95%

credible intervals, all of which contain the true mixing vector. Additionally the results for the trace of the error covariance indicate a good recovery of the covariance matrix. For moderate dependence (case b), the inference for the third component is unaffected by the dependence between the first two components, as the results are very comparable to case a). Not unexpectedly, the inference for the first two components is not correct, however, the total of the first two components is approximately correct. Another consequence of the strong dependence is the credible intervals for the first two components are substantially wider than in case a). A similar pattern emerges for the third case, but the credible intervals are much wider, indicating much more uncertainty in the first two components, but again the total is correct. The inference for the trace of the error covariance matrix are not adversely affected by the dependence in the sources, though some of the credible intervals are slightly wider.

We also generated synthetic data to study the effect of ill–conditioned source matrices on the linear multilevel mixing model (see equation 5.2) using the $\boldsymbol{\Theta}$ matrices given in table 5.2. Synthetic data for this case was generated in the following manner. Generate $n = 50$ mixing vectors $\boldsymbol{\alpha}_i$ from a multivariate normal distribution with mean $\boldsymbol{\mu}_{\boldsymbol{\alpha}}$ and covariance matrix $\Sigma_{\boldsymbol{\alpha}}$. Then for each i generate $r = 2$ samples $\mathbf{y}_{i1}$ and $\mathbf{y}_{i2}$ from a multivariate normal distribution with mean $\boldsymbol{\Theta}\boldsymbol{\alpha}_i$ and covariance matrix $\Sigma_{\boldsymbol{\epsilon}}$. We used the first sample to represent the $r = 1$ case and both for the $r = 1$ case. We used the following settings: $\Sigma_{\boldsymbol{\epsilon}} = \mathbf{I}_d$, $\boldsymbol{\mu}_{\boldsymbol{\alpha}} = (1, 1, 1)'$, $\Sigma_{\boldsymbol{\alpha}} = \sigma^2_{\boldsymbol{\alpha}}\mathbf{I}_p$ with $\sigma^2_{\boldsymbol{\alpha}} = 0.5, 1.0, 1.5$.

We used the following “vague” priors for $\boldsymbol{\mu}_{\boldsymbol{\alpha}}$, $\Sigma_{\boldsymbol{\epsilon}}$ and $\Sigma_{\boldsymbol{\alpha}}$:

$$p(\boldsymbol{\alpha}) = \mathcal{N}^p(\mathbf{0}_p, 100\mathbf{I}_p),$$

$$p(\Sigma_{\boldsymbol{\epsilon}}) = \mathcal{IW}^a(a, 100\mathbf{I}_a),$$

and

$$p(\Sigma_{\boldsymbol{\alpha}}) = \mathcal{IW}^p(p, 100\mathbf{I}_p).$$

We generated 100,000 posterior samples using a Metropolis–Hastings within Gibbs MCMC algorithm with a thinning factor of 10 to save storage space. To avoid the burn in period we started the algorithm at the generated values of $\boldsymbol{\alpha}_i$. The full conditional distributions for $\Sigma_{\boldsymbol{\epsilon}}$ and $\Sigma_{\boldsymbol{\alpha}}$ are both inverse–Wishart distribution, additionally the full conditional distribution for $\boldsymbol{\mu}_{\boldsymbol{\alpha}}$ is multivariate normal and the distribution for $\boldsymbol{\alpha}_i$ was updated by a similar adaptive

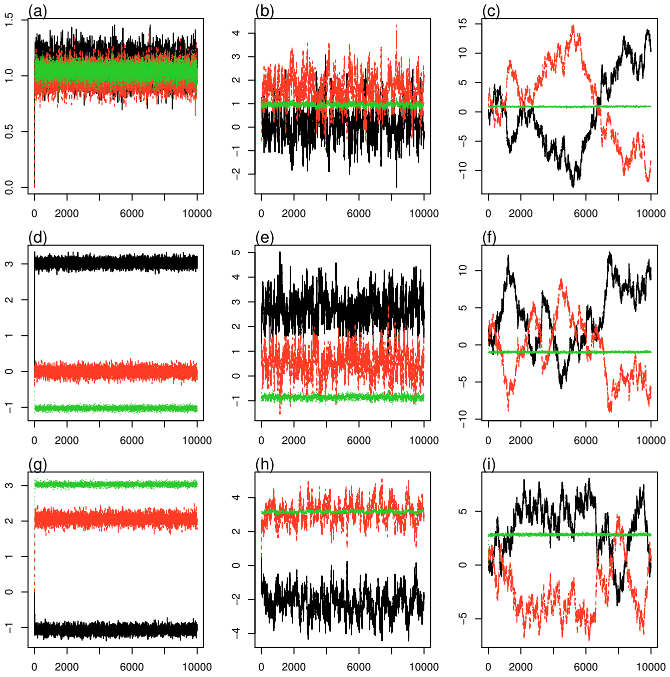

Figure 5.3: Trace plots of the linear constant mixing model for the mixing vector $\boldsymbol{\alpha}$ with $\Sigma_{\boldsymbol{\epsilon}} = \mathbf{I}_5$. The rows show the effect of changing the mixing vector: (a-c) $\boldsymbol{\alpha} = (1, 1, 1)'$, (d-f) $\boldsymbol{\alpha} = (3, 0, -1)'$ and (g-i) $\boldsymbol{\alpha} = (-1, 2, 3)'$. While the columns show the effect of dependence among the columns of $\boldsymbol{\Theta}$: (a,d,g) small condition number, (b,e,h) moderate condition number and (c,f,i) large condition number. For each panel, the black represents the first component, the red the second component and green the third component of the source matrix. See table 5.2 for the actual $\boldsymbol{\Theta}$'s used. Note the changing scale in all panels of the figure.

			$\boldsymbol{\alpha}_1$		$\boldsymbol{\alpha}_2$		$\boldsymbol{\alpha}_3$	
$\boldsymbol{\Theta}$	$\sigma^2_{\boldsymbol{\epsilon}}$	meas	$(1,1,1)^{'}$	Σ^T_{ϵ}	$(3,0,-1)^{'}$	Σ^T_{ϵ}	$(-1,2,3)^{'}$	Σ^T_{ϵ}
a	0.5	pm	(0.97,0.92,0.99)	2.77	(2.97,-0.04,-1.03)	2.86	(-1.04,2.05,3.01)	2.82
		lci	(0.85,0.79,0.92)	2.29	(2.86,-0.18,-1.09)	2.32	(-1.18,1.93,2.95)	2.27
		uci	(1.09,1.06,1.06)	3.36	(3.08,0.10,-0.96)	3.41	(-0.91,2.18,3.08)	3.40
	1.0	pm	(1.09,1.00,1.04)	5.47	(3.02,-0.01,-1.02)	5.16	(-1.07,2.06,3.03)	4.92
		lci	(0.89,0.84,0.93)	4.50	(2.86,-0.22,-1.12)	4.25	(-1.24,1.82,2.94)	4.00
		uci	(1.29,1.16,1.15)	6.64	(3.19,0.20,-0.93)	6.25	(-0.91,2.29,3.13)	5.93
	1.5	pm	(0.95,1.07,0.99)	7.65	(3.03,-0.06,-1.04)	6.89	(-1.20,2.06,3.00)	7.55
		lci	(0.76,0.82,0.89)	6.29	(2.82,-0.26,-1.15)	5.68	(-1.39,1.81,2.90)	6.18
		uci	(1.15,1.32,1.09)	9.32	(3.23,0.15,-0.94)	8.40	(-1.02,2.31,3.10)	9.19
b	0.5	pm	(0.49,1.42,1.03)	2.39	(3.30,-0.22,-1.03)	2.61	(-0.57,1.63,2.98)	2.44
		lci	(-0.32,0.71,0.95)	1.97	(1.89,-1.14,-1.12)	2.14	(-1.48,0.85,2.89)	2.01
		uci	(1.32,2.11,1.11)	2.94	(4.38,0.96,-0.93)	3.21	(0.34,2.41,3.08)	2.99
	1.0	pm	(0.18,1.56,0.95)	6.00	(2.68,0.55,-0.87)	5.05	(-2.19,3.12,3.16)	4.87
		lci	(-1.29,0.23,0.80)	4.93	(1.50,-0.60,-1.03)	4.13	(-3.67,1.95,3.03)	4.02
		uci	(1.64,2.91,1.10)	7.39	(3.96,1.61,-0.71)	6.17	(-0.81,4.35,3.29)	5.93
	1.5	pm	(0.45,1.29,0.97)	8.14	(1.89,0.81,-0.92)	7.84	(-2.85,3.70,3.13)	7.34
		lci	(-1.17,-0.35,0.80)	6.70	0.20,-0.60,-1.10)	6.50	(-4.45,2.24,2.96)	6.03
		uci	(2.27,2.74,1.15)	9.96	(3.51,2.29,-0.75)	9.49	(-1.21,5.10,3.29)	8.96
c	0.5	pm	(5.15,-3.07,1.01)	2.48	(0.08,2.93,-1.06)	2.16	(4.55,-3.51,3.12)	2.75
		lci	(-0.30,-8.56,0.89)	2.04	(-11.61,-7.17,-1.21)	1.78	(-2.09,-9.17,2.99)	2.27
		uci	(10.65,2.39,1.14)	3.00	(10.20,14.59,-0.91)	2.64	(10.23,3.13,3.26)	3.31
	1.0	pm	(0.07,1.98,0.89)	5.42	(3.94,-0.79,-0.98)	4.51	(3.50,-2.48,2.84)	5.15
		lci	(-10.88,-10.65,0.74)	4.45	(-4.15,-7.95,-1.16)	3.72	(-2.20,-6.06,2.68)	4.24
		uci	(12.59,12.78,1.04)	6.69	(11.11,7.27,-0.80)	5.48	(7.10,3.20,2.99)	6.34
	1.5	pm	(-3.49,5.55,0.85)	8.61	(1.78,1.13,-0.91)	7.74	(4.45,-3.44,3.10)	7.69
		lci	(-12.89,-3.07,0.59)	7.08	(-7.73,-9.97,-1.10)	6.38	(-5.83,-12.26,2.86)	6.34
		uci	(5.14,14.99,1.11)	10.55	(12.94,10.62,-0.72)	9.38	(13.28,6.80,3.33)	9.34

Table 5.3: Posterior summaries for the linear constant mixing model of 100,000 MCMC runs of a Metropolis–Hastings within Gibbs algorithm with a thinning factor of 10. The settings for $\boldsymbol{\Theta}$ are given in table 5.2 which correspond to increasing degrees of linear dependence between the first two rows of the $\boldsymbol{\Theta}$ matrix. σ^2 corresponds to the covariance matrix of the multivariate normal distribution of the errors, that is, $\Sigma_{\boldsymbol{\epsilon}} = \sigma^2_{\boldsymbol{\epsilon}}\mathbf{I}_d$, $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$ and $\boldsymbol{\alpha}_3$ are the three settings of the mixing vector and $\Sigma^T_{\boldsymbol{\epsilon}}$ is trace of the error covariance matrix. The posterior mean is denoted by pm, the upper and lower 95% element–wise credible intervals are denoted by lci, and uci respectively.

Metropolis–Hastings algorithm to the linear constant mixing model. Thus was done again for consistency with the more complicated models to come in the next chapter.

Before proceeding to the results, it is of interest to note the variance decomposition in the linear multilevel mixing model. Consider the covariance matrix of $\mathbf{y}_i|\Sigma_{\boldsymbol{\alpha}}, \Sigma_{\boldsymbol{\epsilon}}, \boldsymbol{\mu}_{\boldsymbol{\alpha}}$:

$$\begin{aligned} \text{var}(\mathbf{y}_i|\Sigma_{\boldsymbol{\alpha}}, \Sigma_{\boldsymbol{\epsilon}}, \boldsymbol{\mu}_{\boldsymbol{\alpha}},) &= \text{var}(\boldsymbol{\Theta}\boldsymbol{\alpha}_i) + \text{var}(\boldsymbol{\epsilon}_i) \\ \underset{(k\times p)}{\Sigma_{\mathbf{y}}} &= \underset{(k\times p)}{\boldsymbol{\Theta}}\underset{(p\times p)}{\Sigma_{\boldsymbol{\alpha}}}\underset{(p\times k)}{\boldsymbol{\Theta}^{'}} + \underset{(k\times k)}{\Sigma_{\epsilon}}, \end{aligned}$$

which was derived using the multivariate version of the law of total variance:

$$\text{var}(\mathbf{Y}) = E(\text{var}(\mathbf{Y}|\mathbf{X})) + \text{var}(E(\mathbf{Y}|\mathbf{X})).$$

The first term has rank at most p and the second term is of rank a. This can be seen as a classical factor analysis problem, in that, we are decomposing a rank a covariance matrix into two components one of potentially smaller rank. That is, we are hoping that the variability in the original data can be described by a smaller dimensional subset of latent variables.

Given the above breakdown in terms of covariance matrices we summarize the variability due to the first and second terms respectively by their traces. We label the first one $(\boldsymbol{\Theta}\Sigma_{\boldsymbol{\alpha}}\boldsymbol{\Theta}^{'})^T$ and the second one $\Sigma_{\boldsymbol{\epsilon}}^T$. As synthetic data was generated with $\Sigma_{\boldsymbol{\epsilon}} = \mathbf{I}_5$ the trace is just 5, the trace for the first term is given by

$$(\boldsymbol{\Theta}\Sigma_{\boldsymbol{\alpha}}\boldsymbol{\Theta}^{'})^T = \sigma_{\alpha}^2\text{trace}(\boldsymbol{\Theta}\boldsymbol{\Theta}^{'})$$

since $\Sigma_{\boldsymbol{\alpha}} = \sigma_{\alpha}^2\mathbf{I}_3$. The traces for $\boldsymbol{\Theta}_a$, $\boldsymbol{\Theta}_b$ and $\boldsymbol{\Theta}_c$ given in table 5.2 are 17.76, 8.78 and 9.49 respectively.

Tables 5.4 and 5.5 give the posterior mean of the mean of the mixing distribution $\boldsymbol{\mu}_{\boldsymbol{\alpha}}$, the diagonal of covariance of the mixing distribution $\Sigma_{\boldsymbol{\alpha}}$, the trace of $\Theta\Sigma_{\boldsymbol{\alpha}}\Theta^{'}$, the trace of the error covariance matrix Σ_{ϵ} and an individual mixing vector $\boldsymbol{\alpha}_1$ for $r = 1$ and $r = 2$ respectively. We consider each parameter in turn. The results for $\boldsymbol{\mu}_{\boldsymbol{\alpha}}$ are consistent with the linear constant mixing model, with good performance for case a) but the performance doesn't suffer as much when the dependence among the columns of the source profile matrix increases. The multilevel model seems to control the wild fluctuations in posterior space, but components one and two are still not well determined. Generally, with one observation, the model attributes more variability to the mixing distribution at the expense of the residual variability. By contrast, with two replicates, the partitioning of the covariance to the two components is more consistent with the generating mechanism. Of course there are exceptions to this rule. Not surprisingly, both models have a difficult time with the inference for an single individual, as suggested by the credible intervals being quite wide in all cases.

It is surprising that we are able to do so well in the $r = 1$ case, which contradicts the usual intuition one has in the usual one–way random effects model. This suggests that there is enough structure in the linear multilevel mixing model to partition the covariances properly.

$\boldsymbol{\Theta}$	$\sigma^2_{\boldsymbol{\alpha}}$	meas	$\boldsymbol{\mu}_{\boldsymbol{\alpha}} = (1,1,1)^{'}$	diag$(\Sigma_{\boldsymbol{\alpha}})$	$(\boldsymbol{\Theta}\Sigma_{\boldsymbol{\alpha}}\boldsymbol{\Theta}^{'})^T$	Σ^T_{ϵ}	$\boldsymbol{\alpha}_1$
a	0.5	pm	(0.97,0.85,0.84)	(0.81,0.61,0.45)	8.78	5.08	(1.05,0.85,0.67)
		lci	(0.70,0.55,0.64)	(0.46,0.25,0.21)	6.37	3.83	(0.65,0.51,0.50)
		uci	(1.25,1.13,1.06)	(1.31,1.12,0.76)	12.10	6.79	(1.55,1.22,0.88)
		gen	(0.90,0.80,0.82)	(0.52,0.52,0.42)	7.81	5.13	(1.21,1.61,0.71)
	1.0	pm	(1.13,1.10,1.13)	(1.29,1.49,1.32)	22.86	3.20	(1.06,-1.20,0.63)
		lci	(0.80,0.76,0.80)	(0.85,0.97,0.87)	17.02	2.12	(0.61,-1.57,0.27)
		uci	(1.45,1.45,1.46)	(1.94,2.21,1.99)	31.15	5.64	(1.35,-0.45,1.26)
		gen	(1.16,1.01,1.12)	(1.02,0.90,1.22)	19.55	4.10	(0.79,-0.48,0.72)
	1.5	pm	(1.04,1.11,1.37)	(2.17,2.60,1.70)	32.94	3.09	(0.47,1.65,2.42)
		lci	(0.62,0.65,1.01)	(1.45,1.60,1.12)	24.88	1.78	(0.26,0.95,2.29)
		uci	(1.46,1.56,1.74)	(3.26,3.93,2.56)	44.27	6.31	(0.68,2.00,2.64)
		gen	(1.12,1.07,1.40)	(1.84,1.67,1.46)	26.73	5.23	(0.64,1.99,2.22)
b	0.5	pm	(0.86,1.45,1.03)	(0.15,0.64,0.54)	5.02	4.76	(0.88,2.50,0.07)
		lci	(0.29,0.91,0.79)	(0.00,0.02,0.24)	3.16	2.74	(0.05,1.33,-0.34)
		uci	(1.50,2.08,1.24)	(0.51,1.76,0.87)	7.98	6.45	(2.11,3.99,0.39)
		gen	(1.17,1.08,1.06)	(0.54,0.59,0.44)	4.14	4.97	(0.29,2.22,0.84)
	1.0	pm	(0.99,0.60,1.24)	(0.65,3.98,1.01)	8.69	4.23	(0.38,2.44,1.99)
		lci	(0.54,-0.03,0.96)	(0.05,1.97,0.67)	6.40	3.29	(-0.48,1.6,1.71)
		uci	(1.45,1.26,1.54)	(1.77,7.22,1.50)	11.83	5.47	(1.20,3.29,2.30)
		gen	(0.92,0.99,1.25)	(1.01,0.92,0.80)	7.56	5.76	(0.82,2.17,2.25)
	1.5	pm	(0.24,1.68,1.41)	(3.07,2.18,1.61)	13.34	4.08	(-1.86,-1.29,0.63)
		lci	(-0.48,0.88,1.05)	(1.67,0.63,1.07)	9.75	3.07	(-3.55,-2.72,0.25)
		uci	(1.25,2.35,1.77)	(5.64,4.46,2.41)	18.33	5.51	(-0.44,0.46,0.92)
		gen	(1.03,0.94,1.30)	(1.12,1.26,1.31)	11.26	5.09	(-0.07,-2.41,1.09)
c	0.5	pm	(1.25,0.67,1.25)	(1.20,0.31,0.97)	8.98	4.00	(1.44,0.87,1.66)
		lci	(0.75,0.27,0.97)	(0.41,0.04,0.64)	6.53	3.09	(1.03,0.43,1.31)
		uci	(1.69,1.00,1.52)	(2.78,0.86,1.46)	12.39	5.23	(1.96,1.27,2.19)
		gen	(0.82,1.18,1.17)	(0.58,0.55,0.54)	6.79	5.67	(0.65,2.21,1.04)
	1.0	pm	(1.53,0.52,1.01)	(3.60,3.07,1.76)	13.21	4.22	(-2.03,2.67,-1.22)
		lci	(0.74,-0.21,0.64)	(1.77,1.12,0.99)	9.56	2.92	(-2.90,1.43,-1.69)
		uci	(2.28,1.33,1.39)	(5.87,5.00,2.68)	17.99	6.17	(-0.43,3.69,-0.37)
		gen	(1.07,0.89,1.05)	(1.02,1.18,1.18)	9.88	5.13	(-0.15,1.67,-0.23)
	1.5	pm	(0.36,1.34,0.65)	(0.88,1.77,1.46)	17.30	3.70	(0.21,0.18,-0.50)
		lci	(-0.10,0.88,0.32)	(,0.16,0.76,0.97)	12.48	2.65	(-0.62,-0.25,-1.08)
		uci	(0.71,1.84,1.00)	(1.89,3.28,2.18)	24.33	5.55	(0.79,0.67,-0.03)
		gen	(0.77,0.97,0.70)	(2.23,1.73,1.08)	16.25	4.71	(-0.39,0.48,-0.05)

Table 5.4: Posterior summaries for the linear multilevel mixing model with $r = 1$ of 100,000 MCMC runs of a Metropolis–Hastings within Gibbs algorithm. The settings for $\boldsymbol{\Theta}$ are given in table 5.2 which correspond to increasing degrees of linear dependence between the first two rows of the $\boldsymbol{\Theta}$ matrix. $\sigma^2_{\boldsymbol{\alpha}}$ corresponds to the covariance matrix of the multivariate normal distribution of the mixing distribution, that is, $\Sigma_{\boldsymbol{\alpha}} = \sigma^2_{\boldsymbol{\alpha}}\mathbf{I}_3$. $\boldsymbol{\mu}_{\boldsymbol{\alpha}}$ is the mean of the mixing distribution, diag$(\Sigma_{\boldsymbol{\alpha}})$ is diagonal of the covariance matrix of the mixing distribution, $(\boldsymbol{\Theta}\Sigma_{\boldsymbol{\alpha}}\boldsymbol{\Theta}^{'})^T$ is the trace of the covariance of the mixing term, $\Sigma^T_{\boldsymbol{\epsilon}}$ is trace of the error covariance matrix and $\boldsymbol{\alpha}_1$ is the mixing vector for one level. The posterior mean is denoted by pm, the upper and lower 95% element–wise credible intervals are denoted by lci, and uci respectively. The lines indicated by gen, give the means of the 50 $\boldsymbol{\alpha}_i$'s generated from $\mathcal{N}^3(\boldsymbol{\mu}_{\boldsymbol{\alpha}}, \sigma^2_{\boldsymbol{\alpha}}\mathbf{I}_3)$ in column 4, the variances of the generated $\boldsymbol{\alpha}_i$'s in column 5, the trace of $\boldsymbol{\Theta}\Sigma_{\boldsymbol{\alpha}}\boldsymbol{\Theta}^{'}$ using the variance of the generated $\boldsymbol{\alpha}$'s in column 6, the trace of $\Sigma_{\boldsymbol{\epsilon}}$ the variance of the actual errors in column 7 and the individual $\boldsymbol{\alpha}_1$ in column 8.

$\mathbf{\Theta}$	$\sigma^2_{\boldsymbol{\alpha}}$	meas	$\boldsymbol{\mu}_{\boldsymbol{\alpha}} = (1,1,1)^{'}$	diag$(\Sigma_{\boldsymbol{\alpha}})$	$(\mathbf{\Theta}\Sigma_{\boldsymbol{\alpha}}\mathbf{\Theta}^{'})^T$	Σ^T_{ϵ}	$\boldsymbol{\alpha}_1$
a	0.5	pm	(0.88,0.89,0.82)	(0.28,0.37,0.40)	6.17	6.68	(0.92,0.98,0.55)
		lci	(0.66,0.67,0.63)	(0.06,0.14,0.24)	4.18	5.70	(0.36,0.36,0.18)
		uci	(1.09,1.11,1.01)	(0.61,0.70,0.63)	9.11	7.81	(1.57,1.65,0.93)
		gen	(0.90,0.80,0.82)	(0.52,0.52,0.42)	7.81	5.34	(1.21,1.61,0.71)
	1.0	pm	(1.15,1.05,1.09)	(1.04,0.78,1.12)	18.59	5.66	(0.98,0.07,1.03)
		lci	(0.84,0.76,0.79)	(0.64,0.13,0.73)	13.37	4.72	(0.25,-0.87,0.53)
		uci	(1.45,1.33,1.39)	(1.63,1.35,1.71)	26.17	7.21	(1.72,1.13,1.56)
		gen	(1.16,1.01,1.12)	(1.02,0.90,1.22)	19.55	5.26	(0.79,-0.48,0.72)
	1.5	pm	(1.07,1.13,1.37)	(1.71,1.84,1.42)	26.56	5.48	(0.42,1.62,2.38)
		lci	(0.69,0.74,1.04)	(1.08,1.13,0.93)	19.70	4.60	(-0.36,0.74,1.88)
		uci	(1.45,1.55,1.72)	(2.62,2.91,2.13)	36.07	6.56	(1.23,2.48,2.88)
		gen	(1.12,1.07,1.40)	(1.84,1.67,1.46)	26.73	5.42	(0.64,1.99,2.22)
b	0.5	pm	(1.33,0.85,0.99)	(0.67,0.32,0.45)	4.03	5.10	(2.12,1.14,0.47)
		lci	(0.45,0.22,0.79)	(0.04,0.03,0.27)	2.60	4.37	(0.71,-0.19,-0.04)
		uci	(2.05,1.61,1.19)	(2.61,0.91,0.72)	5.95	5.97	(3.89,2.53,0.96)
		gen	(1.17,1.08,1.06)	(0.54,0.59,0.44)	4.15	5.04	(0.29,2.22,0.84)
	1.0	pm	(1.64,0.33,1.20)	(1.26,0.42,0.86)	8.03	5.28	(1.90,0.34,1.96)
		lci	(0.64,-0.31,0.93)	(0.04,0.01,0.55)	5.70	4.51	(0.62,-0.47,1.50)
		uci	(2.40,1.26,1.47)	(3.28,1.79,1.30)	11.14	6.23	(3.19,1.46,2.44)
		gen	(0.92,0.99,1.25)	(1.01,0.92,0.80)	7.56	5.54	(0.82,2.17,2.25)
	1.5	pm	(0.40,1.54,1.39)	(0.33,2.00,1.33)	11.22	5.68	(0.35,-2.17,0.92)
		lci	(-0.28,0.94,1.06)	(0.01,0.57,0.87)	7.95	4.82	(-1.58,-5.01,0.31)
		uci	(0.96,2.22,1.73)	(1.55,5.58,2.02)	15.76	6.74	(3.67,-0.19,1.55)
		gen	(1.03,0.94,1.30)	(1.12,1.26,1.31)	11.26	5.28	(-0.07,-2.41,1.09)
c	0.5	pm	(1.89,0.05,1.14)	(0.86,0.35,0.52)	6.10	5.86	(2.46,0.28,1.03)
		lci	(0.50,-0.71,0.90)	(0.16,0.02,0.26)	4.11	5.02	(0.97,-0.64,0.27)
		uci	(2.67,1.42,1.38)	(2.96,1.13,0.89)	8.91	6.85	(3.69,1.76,1.79)
		gen	(0.82,1.18,1.17)	(0.58,0.55,0.54)	6.79	5.54	(0.65,2.21,1.04)
	1.0	pm	(1.08,0.89,0.99)	(1.40,0.67,1.41)	10.88	5.08	(0.53,0.52,-0.10)
		lci	(0.50,0.19,0.64)	(0.02,0.01,0.87)	7.83	4.37	(-0.43,-0.24,-0.89)
		uci	(1.87,1.49,1.35)	(5.65,2.18,2.20)	15.03	5.93	(1.31,1.38,0.72)
		gen	(1.07,0.89,1.05)	(1.02,1.18,1.18)	9.88	5.03	(-0.15,1.67,-0.23)
	1.5	pm	(-0.28,1.93,0.64)	(1.22,2.58,1.26)	16.61	4.62	(-0.89,0.94,-0.08)
		lci	(-1.77,0.78,0.31)	(0.13,0.34,0.78)	11.88	3.97	(-2.08,-0.26,-0.81)
		uci	(0.85,3.63,0.97)	(2.77,8.65,1.97)	23.33	5.40	(0.29,2.33,0.65)
		gen	(0.77,0.97,0.70)	(2.23,1.73,1.08)	16.25	4.71	(-0.39,0.48,-0.05)

Table 5.5: Posterior summaries for the linear multilevel mixing model with $r = 2$ of 100,000 MCMC runs of a Metropolis–Hastings within Gibbs algorithm. The settings for $\mathbf{\Theta}$ are given in table 5.2 which correspond to increasing degrees of linear dependence between the first two rows of the $\mathbf{\Theta}$ matrix. $\sigma^2_{\boldsymbol{\alpha}}$ corresponds to the covariance matrix of the multivariate normal distribution of the mixing distribution, that is, $\Sigma_{\boldsymbol{\alpha}} = \sigma^2_{\boldsymbol{\alpha}}\mathbf{I}_3$. $\boldsymbol{\mu}_{\boldsymbol{\alpha}}$ is the mean of the mixing distribution, diag$(\Sigma_{\boldsymbol{\alpha}})$ is diagonal of the covariance matrix of the mixing distribution, $(\mathbf{\Theta}\Sigma_{\boldsymbol{\alpha}}\mathbf{\Theta}^{'})^T$ is the trace of the covariance of the mixing term, $\Sigma^T_{\boldsymbol{\epsilon}}$ is trace of the error covariance matrix and $\boldsymbol{\alpha}_1$ is the mixing vector for one level. The posterior mean is denoted by pm, the upper and lower 95% element–wise credible intervals are denoted by lci, and uci respectively. The lines indicated by gen, give the means of the 50 $\boldsymbol{\alpha}_i$'s generated from $\mathcal{N}^3(\boldsymbol{\mu}_{\boldsymbol{\alpha}}, \sigma^2_{\boldsymbol{\alpha}}\mathbf{I}_3)$ in column 4, the variances of the generated $\boldsymbol{\alpha}_i$'s in column 5, the trace of $\mathbf{\Theta}\Sigma_{\boldsymbol{\alpha}}\mathbf{\Theta}^{'}$ using the variance of the generated $\boldsymbol{\alpha}$'s in column 6, the trace of $\Sigma_{\boldsymbol{\epsilon}}$ the variance of the actual errors in column 7 and the individual $\boldsymbol{\alpha}_1$ in column 8.

5.2.2 Case b: $\boldsymbol{\theta}_j \in \Re^a$ and $\boldsymbol{\alpha} \in \mathcal{S}^p$

In this section, we examine the case where $\boldsymbol{\alpha}$ is constrained to lie in the simplex for the linear constant mixing model and $\boldsymbol{\alpha}_i$ is sampled from a logistic normal distribution for the linear multilevel mixing model. The distributional assumptions for the linear constant mixing model are as follows:

$$\begin{aligned}
\mathbf{y}_i|\boldsymbol{\Theta}, \boldsymbol{\alpha}, \Sigma_{\boldsymbol{\epsilon}} &\sim \mathcal{N}^a(\mathbf{y}_i|\boldsymbol{\Theta}\boldsymbol{\alpha}, \Sigma_{\boldsymbol{\epsilon}}) \\
\boldsymbol{\alpha}|\boldsymbol{\mu}_{\boldsymbol{\alpha}}, \Sigma_{\boldsymbol{\alpha}} &\sim \mathcal{L}^p(\boldsymbol{\alpha}|\boldsymbol{\mu}_{\boldsymbol{\alpha}}, \Sigma_{\boldsymbol{\alpha}}) \\
\Sigma_{\boldsymbol{\epsilon}}|\delta, \Psi &\sim \mathcal{IW}^a(\Sigma_{\boldsymbol{\epsilon}}|\delta, \Psi).
\end{aligned}$$

We chose the logistic normal distribution over the Dirichlet distribution for the reasons summarized in chapter 2 and expanded upon in Aitchison (2003). And for the linear multilevel mixing model

$$\begin{aligned}
\boldsymbol{\epsilon}_i|\Sigma_{\boldsymbol{\epsilon}} &\sim \mathcal{N}^a(\boldsymbol{\epsilon}_i|\mathbf{0}, \Sigma_{\boldsymbol{\epsilon}}) \\
\boldsymbol{\alpha}_i|\boldsymbol{\mu}_{\boldsymbol{\alpha}}, \Sigma_{\boldsymbol{\alpha}} &\sim \mathcal{L}^p(\boldsymbol{\alpha}_i|\boldsymbol{\mu}_{\boldsymbol{\alpha}}, \Sigma_{\boldsymbol{\alpha}}) \\
\boldsymbol{\mu}_{\boldsymbol{\alpha}}|\tau, \Sigma_{\boldsymbol{\mu}_{\boldsymbol{\alpha}}} &\sim \mathcal{N}^{p-1}(\boldsymbol{\mu}_{\boldsymbol{\alpha}}|\boldsymbol{\tau}, \Sigma_{\boldsymbol{\mu}_{\boldsymbol{\alpha}}}) \\
\Sigma_{\boldsymbol{\epsilon}}|\delta_{\boldsymbol{\epsilon}}, \Psi_{\boldsymbol{\epsilon}} &\sim \mathcal{IW}^a(\Sigma_{\boldsymbol{\epsilon}}|\delta_{\boldsymbol{\epsilon}}, \Psi_{\boldsymbol{\epsilon}}) \\
\Sigma_{\boldsymbol{\alpha}}|\delta_{\boldsymbol{\alpha}}, \Psi_{\boldsymbol{\alpha}} &\sim \mathcal{IW}^{p-1}(\Sigma_{\boldsymbol{\alpha}}|\delta_{\boldsymbol{\alpha}}, \Psi_{\boldsymbol{\alpha}}),
\end{aligned}$$

To make the connection to the assumptions more clear we refer to the models as convex constant normal mixing and convex multilevel normal mixing.

Before considering the implementation of the convex mixing models we discuss the assignment of parameters to the logistic normal prior distributions for the convex mixing models. Recall from chapter 2.1 that we can construct the logistic normal distribution from the closure of a multivariate log–normal distribution. That is, let $\mathbf{z}$ be a vector of dimension p and further assume that

$$\mathbf{z} \sim \mathcal{LN}^p(\mathbf{0}_p, \sigma^2\mathbf{I}_p)$$

that is, a multivariate log–normal distribution with zero means and independent components with a common variance σ^2. If we then apply the closure operator $\mathcal{C}$ we have the following

$$\mathcal{C}(\mathbf{z}) \sim \mathcal{L}^{p-1}(\mathbf{0}_{p-1}, \sigma^2(\mathbf{I}_{p-1} + J_{p-1}))$$

where J_p is a matrix of all ones. We use distributions of this form to assign prior distributions to the parameters of the convex models. Before turning to the assignment of σ^2 it is instructive to consider what ranges of values are practically reasonable. Due to the constrained nature of the space on the simplex and the way the mixing vector enters into the convex linear mixing model $\Theta\boldsymbol{\alpha}$ large changes in $\boldsymbol{\alpha}$ on the logistic normal scale translate into relatively small changes in $\boldsymbol{\Theta\alpha}$. To see this consider, $p = 3$ and consider some possible configurations of $\phi(\boldsymbol{\alpha})$, where ϕ is the log–ratio transformation, and how they transform back to the original scale via ϕ^{-1}, see the following table:

$\phi(\boldsymbol{\alpha})^{'}$	$\boldsymbol{\alpha}^{'}$	$\phi(\boldsymbol{\alpha})^{'}$	$\boldsymbol{\alpha}^{'}$
(0,0)	(0.333,0.333,0.333)		
(1,-1)	(0.665,0.090,0.245)	(5,-5)	(0.99326,0.00005,0.00669)
(-1,-1)	(0.212,0.212,0.576)	(-5,-5)	(0.00665,0.00665,0.98670)
(1,1)	(0.422,0.422,0.155)	(5,5)	(0.49832,0.49832,0.00336)
(2,-2)	(0.867,0.016,0.117)	(10,-10)	(0.999955,2.06e-09,0.2,0.000045)
(-2,-2)	(0.107,0.107, 0.787)	(-10,-10)	(0.00005,0.00005,0.99991)
(2,2)	(0.468,0.468,0.063)	(10,10)	(0.49999,0.49999,0.00002)

Thus for almost all practical situations, priors that have most of their mass within, -10, 10 on the log–ratio scale will be sufficient. To err on the side of being non-informative, we chose $\sigma^2 = 10$. Also, note that we have chosen a prior distribution that is in the region of multi–modality (see chapter 2.5.3).

The synthetic data for convex constant normal mixing model were generated using the same source matrices $\boldsymbol{\Theta}$ as given in table 5.2 and three settings for $\Sigma_\epsilon = \sigma_\epsilon^2\mathbf{I}_5$, $\sigma_{\boldsymbol{\epsilon}}^2 = 0.5, 1.0, 1.5$. The following three settings were used for $\boldsymbol{\alpha}$: $\phi^{-1}((0,0)^{'}) = (0.33, 0.33, 0.33)^{'}$, $\phi^{-1}((1,-1)^{'}) = (0.67, 0.09, 0.24)^{'}$,$\phi^{-1}((-2,2)^{'}) = (0.02, 0.87, 0.12)^{'}$, where ϕ^{-1} is the inverse logistic transform described in 2. The MCMC algorithm was started at the parameter settings used to generate the data, to avoid burn in effects.

Figure 5.4 gives two-dimensional density plots of $\phi(\alpha)$ for $\sigma_{\boldsymbol{\epsilon}}^2 = 1.0$ and each of the three settings of $\boldsymbol{\alpha}$ across the columns of the figure and the three settings of $\boldsymbol{\Theta}$ given in table 5.2 across the rows. The results for the first column, corresponding to the case where the source matrix is well conditioned, show very well behaved posterior surfaces, suggesting a uni-modal posterior. The picture for increasing level of dependence is not so straightforward. They all have a boomerang shape, which suggests the switching of components between the two highly dependent sources. The pattern deviates slightly depending on the particular setting for $\boldsymbol{\alpha}$. These plots provide very good indicators for the component switching, however, it isn't clear how to deal with higher dimensional situations which are explored in the next chapter.

Another way to examine the posterior surfaces with three components is to use ternary diagrams which were introduced in chapter 2. Figure 5.5 gives ternary diagrams of $\phi(\alpha)$ for $\sigma_{\boldsymbol{\epsilon}}^2 = 1.0$ and each of the three settings of $\boldsymbol{\alpha}$ across the columns of the figure and the three settings of $\boldsymbol{\Theta}$ given in table 5.2 across the rows. As seen in figure 5.4, the posterior surface is well behaved when there is little dependence between the components of the source matrix. Similarly the ternary diagrams show the characteristic switching between components one and two. Again, this is hard to generalize to higher dimensional mixing problems where ill–conditioned source matrices are more likely to be an issue.

Table 5.6 gives posterior summaries for the convex constant normal mixing model. We present the mixing vector results on the original scale for ease of interpretation, however, we equally could have presented the results on the $\phi(\boldsymbol{\alpha})$ scale. Focusing on the equally weighted mixing vector $\boldsymbol{\alpha} = (0.33, 0.33, 0.33)'$, we do reasonably well when the sources are nearly orthogonal, and as before the performance of final component which is uncorrelated with the first two. The partitioning of the first two components breaks down as the source matrix becomes more ill–conditioned as was seen before. Rare components are very difficult to apportion properly and not surprisingly deteriorates as residual variance increases as evidenced in part a) of the table for $\boldsymbol{\alpha}_2$ and $\boldsymbol{\alpha}_3$. The effect of the high degree of correlation results in the inability to separate out those two components regardless of how rare they are. Also note that there is substantial uncertainty in the posterior with high degree of dependence as evidenced by most of the credible intervals spanning the entire 0-1 range to 2 decimal places of course. The residual variance is well recovered, however, the posterior mean estimates are on the high side for case c) with residual variance $\sigma_{\boldsymbol{\epsilon}} = 1.5$.

Consider the convex multilevel normal model and recall the that the covariance breakdown is given by:

$$\begin{aligned}
\text{var}(\mathbf{y}_i) &= \text{var}(\boldsymbol{\Theta}\boldsymbol{\alpha}_i) + \text{var}(\boldsymbol{\epsilon}_i) \\
\underset{(k\times p)}{\Sigma_{\mathbf{y}}} &= \underset{(k\times p)}{\boldsymbol{\Theta}}\underset{(p\times p)}{\Sigma^*_{\boldsymbol{\alpha}}}\underset{(p\times k)}{\boldsymbol{\Theta}'} + \underset{(k\times k)}{\Sigma_{\epsilon}} .
\end{aligned}$$

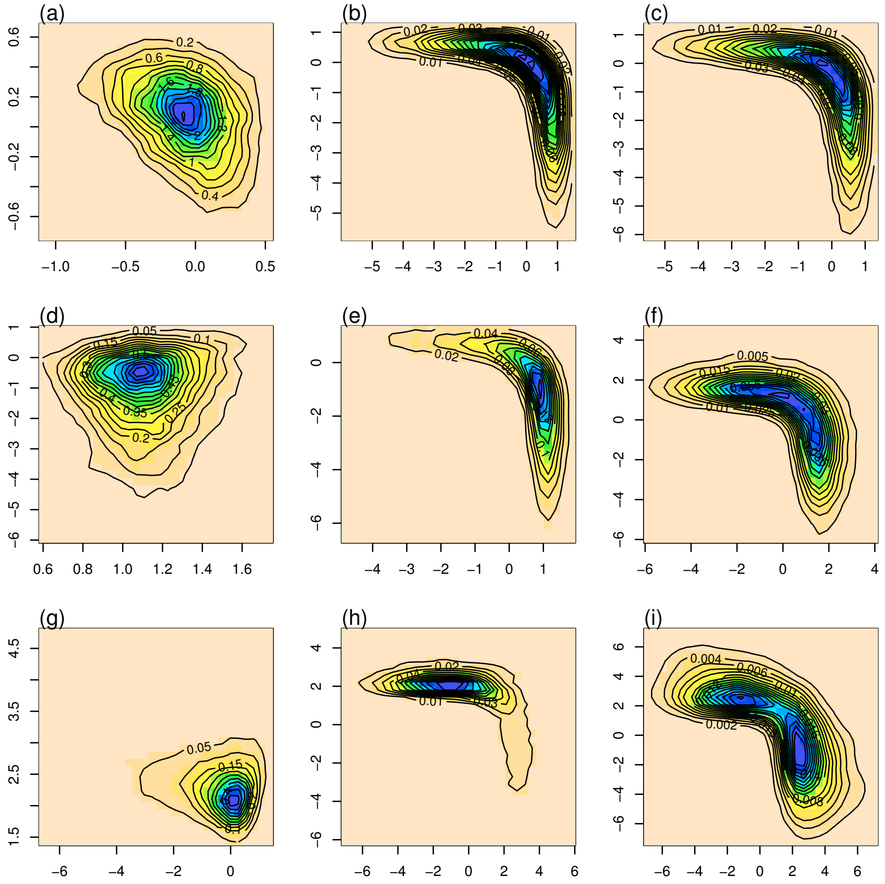

Figure 5.4: Posterior density plots of the convex constant mixing model for the log–ratio transformation of the mixing vector denoted by $\phi(\boldsymbol{\alpha})$. The rows show the effect of changing the mixing vector: (a-c) $\phi(\boldsymbol{\alpha}) = (0, 0)'$, (d-f) $\phi(\boldsymbol{\alpha}) = (1, -1)'$ and (g-i) $\phi(\boldsymbol{\alpha}) = (-2, 2)'$. While the columns show the effect of dependence among the columns of $\boldsymbol{\Theta}$: (a,d,g) small condition number, (b,e,h) moderate condition number and (c,f,i) large condition number. See table 5.2 for the actual $\boldsymbol{\Theta}$'s used. Note the changing scale across the rows of the panel plot. The plots were constructed using the kde2d function from the MASS package of R. The ranges were restricted to the 0.5% and 99.5% quantiles for each of the components of $\phi(\boldsymbol{\alpha})$ to avoid large regions with very little density variation.

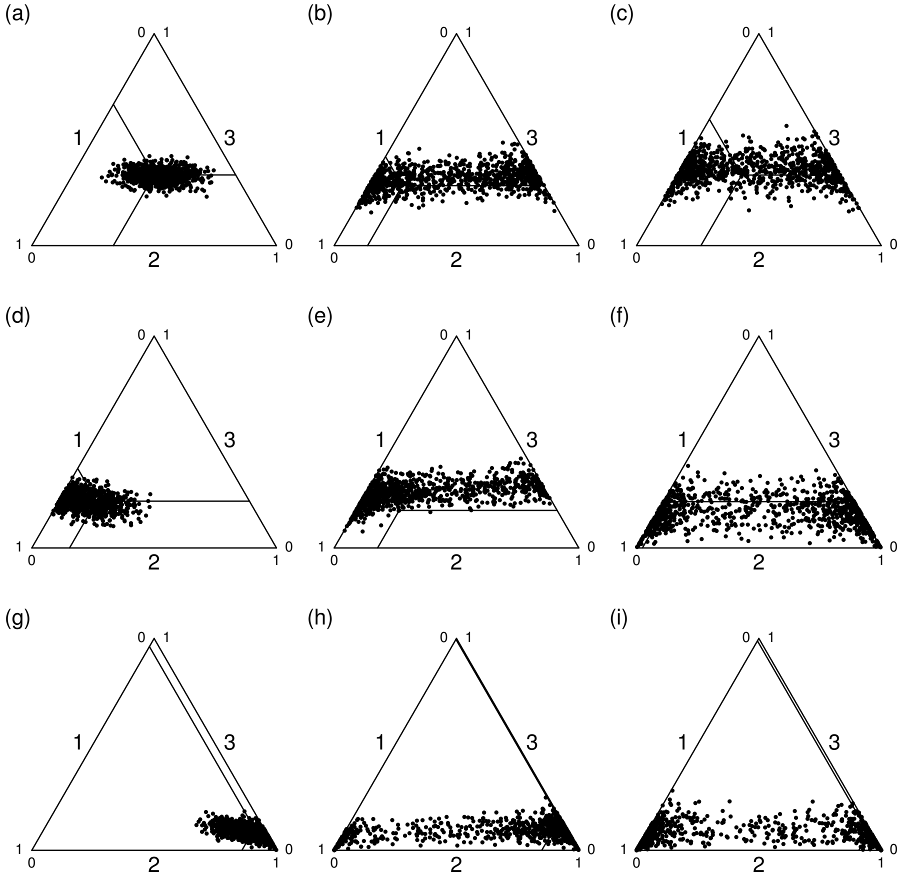

Figure 5.5: Posterior density ternary diagrams of the convex constant mixing model for $\phi(\boldsymbol{\alpha})$. The rows show the effect of changing the mixing vector: (a-c) $\phi(\boldsymbol{\alpha}) = (0, 0)'$, (d-f) $\phi(\boldsymbol{\alpha}) = (1, -1)'$ and (g-i) $\phi(\boldsymbol{\alpha}) = (-2, 2)'$. While the columns show the effect of dependence among the columns of $\boldsymbol{\Theta}$: (a,d,g) small condition number, (b,e,h) moderate condition number and (c,f,i) large condition number. See table 5.2 for the actual $\boldsymbol{\Theta}$'s used. Note that only every 10 posterior sample was used in the construction of the ternary diagrams.

$\mathbf{\Theta}$	$\sigma^2_{\boldsymbol{\epsilon}}$	meas	$\boldsymbol{\alpha}_1$		$\boldsymbol{\alpha}_2$		$\boldsymbol{\alpha}_3$	
			$(0.33, 0.33, 0.33)'$	Σ^T_ϵ	$(0.67, 0.09, 0.24)'$	Σ^T_ϵ	$(0.02, 0.87, 0.12)'$	Σ^T_ϵ
a	0.5	pm	(0.34,0.28,0.37)	2.86	(0.64,0.12,0.24)	2.89	(0.01,0.85,0.13)	1.87
		lci	(0.24,0.17,0.32)	2.35	(0.53,0.01,0.19)	2.37	(0.00,0.80,0.10)	1.54
		uci	(0.45,0.39,0.43)	3.48	(0.74,0.23,0.30)	3.54	(0.05,0.90,0.17)	2.27
	1.0	pm	(0.31,0.36,0.33)	5.46	(0.67,0.11,0.22)	5.60	(0.08,0.84,0.09)	4.55
		lci	(0.17,0.21,0.27)	4.51	(0.51,0.00,0.15)	4.62	(0.00,0.68,0.03)	3.74
		uci	(0.45,0.51,0.39)	6.64	(0.79,0.30,0.29)	6.87	(0.20,0.96,0.14)	5.54
	1.5	pm	(0.25,0.42,0.33)	6.85	(0.77,0.03,0.20)	6.98	(0.02,0.89,0.10)	8.64
		lci	(0.08,0.26,0.26)	5.66	(0.67,0.00,0.13)	5.76	(0.00,0.77,0.01)	7.14
		uci	(0.41,0.58,0.41)	8.35	(0.85,0.12,0.28)	8.45	(0.09,0.99,0.19)	10.47
b	0.5	pm	(0.49,0.17,0.34)	3.04	(0.51,0.22,0.27)	2.41	(0.11,0.81,0.07)	2.82
		lci	(0.02,0.00,0.26)	2.52	(0.02,0.00,0.20)	1.98	(0.00,0.00,0.01)	2.33
		uci	(0.72,0.61,0.42)	3.68	(0.78,0.67,0.35)	2.95	(0.98,0.98,0.14)	3.43
	1.0	pm	(0.37,0.32,0.31)	5.33	(0.52,0.22,0.26)	5.08	(0.25,0.66,0.09)	4.69
		lci	(0.01,0.00,0.21)	4.41	(0.01,0.00,0.16)	4.18	(0.00,0.00,0.01)	3.89
		uci	(0.75,0.70,0.41)	6.47	(0.81,0.71,0.35)	6.21	(0.98,0.96,0.17)	5.72
	1.5	pm	(0.44,0.21,0.34)	6.07	(0.54,0.26,0.20)	7.99	(0.24,0.60,0.17)	6.35
		lci	(0.01,0.00,0.23)	4.99	(0.01,0.00,0.06)	6.63	(0.00,0.00,0.06)	5.24
		uci	(0.74,0.63,0.45)	7.45	(0.92,0.79,0.34)	9.66	(0.92,0.88,0.26)	7.75
c	0.5	pm	(0.36,0.31,0.32)	2.44	(0.40,0.38,0.22)	2.32	(0.47,0.45,0.07)	2.56
		lci	(0.00,0.00,0.24)	2.02	(0.00,0.00,0.16)	1.90	(0.00,0.00,0.01)	2.10
		uci	(0.71,0.71,0.40)	2.96	(0.80,0.80,0.29)	2.85	(0.98,0.98,0.15)	3.14
	1.0	pm	(0.34,0.31,0.36)	4.99	(0.40,0.44,0.16)	5.73	(0.49,0.45,0.07)	4.96
		lci	(0.00,0.00,0.23)	4.11	(0.00,0.00,0.03)	4.73	(0.00,0.00,0.00)	4.08
		uci	(0.71,0.70,0.48)	6.05	(0.93,0.94,0.30)	6.95	(0.99,0.99,0.17)	6.08
	1.5	pm	(0.24,0.24,0.52)	8.47	(0.45,0.46,0.09)	8.53	(0.45,0.46,0.09)	9.06
		lci	(0.01,0.01,0.38)	6.94	(0.00,0.00,0.00)	7.05	(0.00,0.00,0.00)	7.45
		uci	(0.54,0.55,0.65)	10.39	(0.99,0.99,0.24)	10.38	(0.99,0.99,0.24)	11.03

Table 5.6: Posterior summaries for the convex linear constant mixing model of 100,000 MCMC runs of a Metropolis–Hastings within Gibbs algorithm with a thinning factor of 10. The settings for $\mathbf{\Theta}$ are given in table 5.2 which correspond to increasing degrees of linear dependence between the first two rows of the $\mathbf{\Theta}$ matrix. σ^2 corresponds to the covariance matrix of the multivariate normal distribution of the errors, that is, $\Sigma_{\boldsymbol{\epsilon}} = \sigma^2_{\boldsymbol{\epsilon}}\mathbf{I}_d$, $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$ and $\boldsymbol{\alpha}_3$ are the three settings of the mixing vector and $\Sigma^T_{\boldsymbol{\epsilon}}$ is trace of the error covariance matrix. The posterior mean is denoted by pm, the upper and lower 95% element–wise credible intervals are denoted by lci, and uci respectively. Note that the posterior is summarized for the original scale rather than the ϕ scale for ease of interpretation.

The mixing occurs on the simplex scale, however, the $(p-1) \times (p-1)$ covariance matrix $\Sigma_{\boldsymbol{\alpha}}$ is expressed on the log–ratio scale (see chapter 2). In order to apply the above equation we need an expression for the variance on the simplex scale, which we denote by $\Sigma^*_{\boldsymbol{\alpha}}$ which is $p \times p$ matrix. Note that this is a singular matrix, that is, one of the eigenvalues is zero due to the unit sum constraint. Rather than work out an analytical approximation, we use the generated α_i's to compute the covariance matrix. Aitchison and Bacon-Shone (1999) give an approximation to this distribution using the delta method for the compositional case. By trial and error, we deduced that scaling the $\mathbf{\Theta}$ by a scale factor of 5, gives approximately the correct signal to noise ratio with $\Sigma_{\boldsymbol{\epsilon}} = \mathbf{I}_5$.

Synthetic data was generated using the convex normal mixing model using $\boldsymbol{\mu}_{\boldsymbol{\alpha}} = (0.33, 0.33, 0.33)'$, $\Sigma_{\boldsymbol{\alpha}} = \sigma^2_{\boldsymbol{\alpha}}\mathbf{I}_3$, with $\sigma^2_{\boldsymbol{\alpha}} = 0.5, 1.0, 1.5$ and $\Sigma_{\epsilon} = \mathbf{I}_5$. The source matrix is multiplied by 5 to get a reasonable signal to noise ratio. The results of generating $n = 50$ and $r = 1$ and $r = 2$ replicates with first one serving as one of the replicates for the $r = 2$ case.

The results are given tables in 5.7 and 5.8 for $r = 1$ and $r = 2$ respectively. The results for case a) indicate that we do remarkably well in reconstructing the mean mixing vector, $\boldsymbol{\mu}_{\boldsymbol{\alpha}}$ for both one and two replicate observations and also in the recovery of the individual level result given by $\boldsymbol{\alpha}_1$. The partitioning of the covariance between the mixing process and the error process appears to be well recovered as well, though there is a tendency to attribute more variability to the mixing process. A similar pattern emerges for the other two cases, an inability to correctly partition the first two components, however, the final component appears to be recovered well. There is no obvious improvement in the partitioning of the covariance between the two model components.

$\boldsymbol{\Theta}$	$\sigma^2_{\boldsymbol{\alpha}}$	meas	$\boldsymbol{\mu}_{\boldsymbol{\alpha}} = (0.33, 0.33, 0.33)^{'}$	diag$(\Sigma_{\boldsymbol{\alpha}})$	$(\boldsymbol{\Theta}\Sigma_{\boldsymbol{\alpha}}\boldsymbol{\Theta}^{'})^T$	σ_T	$\boldsymbol{\alpha}_1$
a	0.5	pm	(0.33,0.34,0.33)	(0.82,0.78)	9.67	4.73	(0.72,0.15,0.14)
		lci	(0.27,0.28,0.29)	(0.54,0.49)	9.15	3.50	(0.70,0.12,0.12)
		uci	(0.40,0.40,0.37)	(1.24,1.21)	10.48	6.49	(0.73,0.17,0.16)
		gen	(0.34,0.36,0.30)	(0.60,0.50)	7.67	5.15	(0.63,0.24,0.13)
	1.0	pm	(0.31,0.41,0.27)	(1.39,0.32)	9.28	8.52	(0.24,0.48,0.28)
		lci	(0.23,0.30,0.22)	(0.82,0.02)	7.70	4.77	(0.10,0.31,0.20)
		uci	(0.40,0.52,0.34)	(2.29,1.68)	13.3	12.65	(0.37,0.70,0.34)
		gen	(0.35,0.38,0.27)	(0.98,1.00)	11.76	4.57	(0.17,0.57,0.26)
	1.5	pm	(0.32,0.38,0.30)	(2.23,1.50)	22.07	6.37	(0.13,0.17,0.71)
		lci	(0.23,0.28,0.24)	(1.33,0.69)	18.7	4.54	(0.03,0.04,0.61)
		uci	(0.41,0.47,0.38)	(3.64,2.81)	25.73	8.94	(0.23,0.34,0.79)
		gen	(0.36,0.34,0.30)	(1.83,1.87)	19.08	5.09	(0.3,0.14,0.55)
b	0.5	pm	(0.53,0.17,0.30)	(0.41,0.54)	3.76	4.38	(0.49,0.15,0.36)
		lci	(0.30,0.08,0.26)	(0.13,0.11)	3.48	3.45	(0.30,0.06,0.27)
		uci	(0.64,0.38,0.34)	(0.73,1.31)	4.13	5.62	(0.60,0.38,0.42)
		gen	(0.32,0.37,0.31)	(0.45,0.48)	2.79	4.90	(0.26,0.41,0.33)
	1.0	pm	(0.51,0.16,0.33)	(2.00,0.54)	9.32	5.76	(0.45,0.16,0.39)
		lci	(0.31,0.06,0.26)	(1.14,0.04)	8.62	4.37	(0.27,0.06,0.35)
		uci	(0.65,0.34,0.42)	(3.41,1.77)	10.42	7.92	(0.58,0.33,0.43)
		gen	(0.29,0.34,0.36)	(0.90,1.09)	5.65	5.26	(0.24,0.13,0.63)
	1.5	pm	(0.23,0.41,0.36)	(1.60,1.01)	7.11	4.48	(0.25,0.64,0.11)
		lci	(0.11,0.28,0.27)	(0.24,0.40)	6.70	3.59	(0.00,0.19,0.04)
		uci	(0.42,0.55,0.47)	(3.88,2.11)	7.57	5.60	(0.76,0.89,0.16)
		gen	(0.33,0.35,0.32)	(1.36,1.43)	6.99	4.60	(0.09,0.74,0.16)
c	0.5	pm	(0.42,0.25,0.33)	(0.61,4.75)	3.36	4.38	(0.46,0.13,0.41)
		lci	(0.08,0.08,0.25)	(0.18,2.14)	3.10	3.51	(0.08,0.03,0.37)
		uci	(0.61,0.59,0.44)	(2.05,8.22)	3.72	5.49	(0.57,0.50,0.44)
		gen	(0.35,0.34,0.31)	(0.54,0.54)	1.46	5.15	(0.51,0.17,0.32)
	1.0	pm	(0.29,0.43,0.28)	(2.77,0.31)	3.16	4.57	(0.65,0.26,0.09)
		lci	(0.18,0.25,0.22)	(1.41,0.11)	2.98	3.69	(0.27,0.03,0.01)
		uci	(0.45,0.53,0.35)	(4.87,0.69)	3.48	5.69	(0.94,0.54,0.19)
		gen	(0.39,0.32,0.29)	(1.04,0.92)	2.55	5.20	(0.59,0.26,0.15)
	1.5	pm	(0.61,0.14,0.25)	(1.38,0.84)	4.71	4.30	(0.69,0.12,0.19)
		lci	(0.36,0.07,0.19)	(0.83,0.09)	4.41	3.45	(0.54,0.03,0.14)
		uci	(0.72,0.33,0.33)	(2.21,2.92)	5.26	5.41	(0.78,0.27,0.26)
		gen	(0.37,0.35,0.29)	(1.58,1.28)	3.32	4.65	(0.14,0.67,0.20)

Table 5.7: Posterior summaries for the convex multilevel normal mixing model with $r = 1$ of 100,000 MCMC runs of a Metropolis–Hastings within Gibbs algorithm. The settings for $\boldsymbol{\Theta}$ are given in table 5.2 which correspond to increasing degrees of linear dependence between the first two rows of the $\boldsymbol{\Theta}$ matrix. $\sigma^2_{\boldsymbol{\alpha}}$ corresponds to the covariance matrix of the multivariate normal distribution of the mixing distribution, that is, $\Sigma_{\boldsymbol{\alpha}} = \sigma^2_{\boldsymbol{\alpha}}\mathbf{I}_3$. $\boldsymbol{\mu}_{\boldsymbol{\alpha}}$ is the mean of the mixing distribution, diag$(\Sigma_{\boldsymbol{\alpha}})$ is diagonal of the covariance matrix of the mixing distribution, $(\boldsymbol{\Theta}\Sigma_{\boldsymbol{\alpha}}\boldsymbol{\Theta}^{'})^T$ is the trace of the covariance of the mixing term, $\Sigma^T_{\boldsymbol{\epsilon}}$ is trace of the error covariance matrix and $\boldsymbol{\alpha}_1$ is the mixing vector for one level. The posterior mean is denoted by pm, the upper and lower 95% element–wise credible intervals are denoted by lci, and uci respectively. The lines indicated by gen, give the means of the 50 $\boldsymbol{\alpha}_i$'s generated from $\mathcal{L}^2(\boldsymbol{\mu}_{\boldsymbol{\alpha}}, \sigma^2_{\boldsymbol{\alpha}}\mathbf{I}_2)$ in column 4, the variances of the generated $\boldsymbol{\alpha}_i$'s in column 5, the trace of $\boldsymbol{\Theta}\Sigma_{\boldsymbol{\alpha}}\boldsymbol{\Theta}^{'}$ using the variance of the generated $\boldsymbol{\alpha}$'s in column 6, the trace of $\Sigma_{\boldsymbol{\epsilon}}$ the variance of the actual errors in column 7 and the individual $\boldsymbol{\alpha}_1$ in column 8.

$\boldsymbol{\Theta}$	$\sigma^2_{\boldsymbol{\alpha}}$	meas	$\boldsymbol{\mu}_{\boldsymbol{\alpha}} = (0.33, 0.33, 0.33)^{'}$	diag$(\Sigma_{\boldsymbol{\alpha}})$	$(\boldsymbol{\Theta}\Sigma_{\boldsymbol{\alpha}}\boldsymbol{\Theta}^{'})^T$	σ_T	$\boldsymbol{\alpha}_1$
a	0.5	pm	(0.34,0.34,0.32)	(0.55,0.53)	7.71	4.79	(0.63,0.22,0.16)
		lci	(0.29,0.29,0.29)	(0.34,0.31)	6.75	4.13	(0.52,0.11,0.11)
		uci	(0.39,0.39,0.36)	(0.87,0.87)	8.67	5.57	(0.73,0.33,0.21)
		gen	(0.34,0.36,0.30)	(0.60,0.50)	7.67	4.85	(0.63,0.24,0.13)
	1.0	pm	(0.31,0.40,0.29)	(1.55,1.23)	13.11	4.63	(0.20,0.55,0.24)
		lci	(0.23,0.31,0.24)	(0.94,0.72)	11.93	3.98	(0.08,0.40,0.19)
		uci	(0.41,0.49,0.34)	(2.53,2.01)	14.29	5.39	(0.34,0.70,0.29)
		gen	(0.35,0.38,0.27)	(0.98,1.00)	11.76	4.77	(0.17,0.57,0.26)
	1.5	pm	(0.34,0.35,0.32)	(2.24,1.82)	19.71	5.05	(0.24,0.14,0.62)
		lci	(0.24,0.26,0.25)	(1.34,1.03)	18.03	4.32	(0.13,0.03,0.55)
		uci	(0.44,0.45,0.39)	(3.67,3.03)	21.35	5.93	(0.33,0.27,0.69)
		gen	(0.36,0.34,0.30)	(1.83,1.87)	19.08	5.01	(0.30,0.14,0.55)
b	0.5	pm	(0.59,0.11,0.30)	(0.42,0.35)	3.13	4.76	(0.57,0.11,0.32)
		lci	(0.42,0.01,0.26)	(0.21,0.01)	2.43	4.14	(0.39,0.01,0.22)
		uci	(0.71,0.27,0.34)	(0.76,1.45)	3.85	5.49	(0.73,0.27,0.41)
		gen	(0.32,0.37,0.31)	(0.45,0.48)	2.79	4.81	(0.26,0.41,0.33)
	1.0	pm	(0.39,0.25,0.37)	(0.49,1.32)	5.90	5.38	(0.30,0.13,0.56)
		lci	(0.22,0.07,0.31)	(0.12,0.36)	4.90	4.68	(0.14,0.02,0.46)
		uci	(0.57,0.41,0.43)	(1.41,3.33)	6.93	6.21	(0.45,0.29,0.67)
		gen	(0.29,0.34,0.36)	(0.90,1.09)	5.65	5.22	(0.24,0.13,0.63)
	1.5	pm	(0.22,0.46,0.32)	(2.10,0.69)	7.19	5.12	(0.49,0.45,0.06)
		lci	(0.08,0.31,0.26)	(0.27,0.31)	6.41	4.45	(0.07,0.09,0.01)
		uci	(0.39,0.58,0.40)	(6.49,1.29)	8.00	5.90	(0.89,0.85,0.13)
		gen	(0.33,0.35,0.32)	(1.36,1.43)	6.99	5.27	(0.09,0.74,0.16)
c	0.5	pm	(0.14,0.54,0.31)	(0.46,0.39)	1.83	4.98	(0.14,0.51,0.35)
		lci	(0.03,0.24,0.27)	(0.01,0.05)	1.15	4.30	(0.03,0.24,0.23)
		uci	(0.45,0.67,0.36)	(2.22,0.92)	2.53	5.80	(0.39,0.68,0.47)
		gen	(0.35,0.34,0.31)	(0.54,0.54)	1.46	4.88	(0.51,0.17,0.32)
	1.0	pm	(0.50,0.23,0.28)	(0.46,0.41)	1.98	5.60	(0.58,0.26,0.16)
		lci	(0.16,0.02,0.23)	(0.06,0.01)	1.25	4.85	(0.12,0.01,0.06)
		uci	(0.72,0.57,0.32)	(1.31,1.46)	2.72	6.50	(0.88,0.75,0.27)
		gen	(0.39,0.32,0.29)	(1.04,0.92)	2.55	5.54	(0.59,0.26,0.15)
	1.5	pm	(0.13,0.62,0.25)	(0.89,0.93)	3.46	4.84	(0.13,0.66,0.21)
		lci	(0.00,0.35,0.20)	(0.02,0.45)	2.86	4.21	(0.00,0.38,0.12)
		uci	(0.38,0.78,0.31)	(4.01,2.01)	4.09	5.54	(0.39,0.86,0.31)
		gen	(0.37,0.35,0.29)	(1.58,1.28)	3.32	4.71	(0.14,0.67,0.20)

Table 5.8: Posterior summaries for the convex multilevel normal mixing model with $r = 2$ of 100,000 MCMC runs of a Metropolis–Hastings within Gibbs algorithm. The settings for $\boldsymbol{\Theta}$ are given in table 5.2 which correspond to increasing degrees of linear dependence between the first two rows of the $\boldsymbol{\Theta}$ matrix. $\sigma^2_{\boldsymbol{\alpha}}$ corresponds to the covariance matrix of the multivariate normal distribution of the mixing distribution, that is, $\Sigma_{\boldsymbol{\alpha}} = \sigma^2_{\boldsymbol{\alpha}}\mathbf{I}_3$. $\boldsymbol{\mu}_{\boldsymbol{\alpha}}$ is the mean of the mixing distribution, diag$(\Sigma_{\boldsymbol{\alpha}})$ is diagonal of the covariance matrix of the mixing distribution, $(\boldsymbol{\Theta}\Sigma_{\boldsymbol{\alpha}}\boldsymbol{\Theta}^{'})^T$ is the trace of the covariance of the mixing term, $\Sigma^T_{\boldsymbol{\epsilon}}$ is trace of the error covariance matrix and $\boldsymbol{\alpha}_1$ is the mixing vector for one level. The posterior mean is denoted by pm, the upper and lower 95% element–wise credible intervals are denoted by lci, and uci respectively. The lines indicated by gen, give the means of the 50 $\boldsymbol{\alpha}_i$'s generated from $\mathcal{L}^2(\boldsymbol{\mu}_{\boldsymbol{\alpha}}, \sigma^2_{\boldsymbol{\alpha}}\mathbf{I}_2)$ in column 4, the variances of the generated $\boldsymbol{\alpha}_i$'s in column 5, the trace of $\boldsymbol{\Theta}\Sigma_{\boldsymbol{\alpha}}\boldsymbol{\Theta}^{'}$ using the variance of the generated $\boldsymbol{\alpha}$'s in column 6, the trace of $\Sigma_{\boldsymbol{\epsilon}}$ the variance of the actual errors in column 7 and the individual $\boldsymbol{\alpha}_1$ in column 8.

5.3 Observations on the Sources

Throughout this chapter we assumed that the source matrix $\boldsymbol{\Theta}$ was known, however, this is rarely the case in practice. Several attempts have been made in the literature to address this, for example Bandeen-Roche and Ruppert (1991); Bandeen-Roche (1994); Billheimer (2001); Park et al. (2000, 2001, 2002); Wolbers and Stahel (2005) to name but a few.

The studies and methods considered above did not have access to measurements on the sources, we briefly introduce a model that allows for this and we take this up in much more detail in the next chapter where we analyze the diet of marine predators. The model under consideration has some similarities to the work of Aitchison and Bacon-Shone (1999); Billheimer (2001).

To keep things simple, we consider the most basic version of the mixing model, the linear constant mixing model without any restrictions and assume that $\boldsymbol{\epsilon}_i \sim \mathcal{N}^a(\mathbf{0}, \Sigma_{\boldsymbol{\epsilon}})$. The model is again

$$\mathbf{y}_i = \boldsymbol{\Theta}\boldsymbol{\alpha} + \boldsymbol{\epsilon}_i \;\; i = 1, \dots, n.$$

Let the observations on the sources be denoted by $\mathbf{x}_{jk}$, $j = 1, \dots, p$, $k = 1, \dots, n_j$, which for simplicity we assume follow the model

$$\mathbf{x}_{jk} = \boldsymbol{\theta}_j + \boldsymbol{\epsilon}^{\mathbf{x}}_{jk} \;\; j = 1, \dots, p, \;\; k = 1, \dots, n_j.$$

where $\boldsymbol{\epsilon}^{\mathbf{x}}_{jk} \sim \mathcal{N}^a(\mathbf{0}, \Sigma_{\mathbf{x}_j})$. Or writing it in the structural equation framework (see Lee, 2007) in matrix notation

$$\begin{aligned}
\underset{(a\times n)}{\mathbf{Y}} &= \underset{(a\times p)}{\boldsymbol{\Theta}}\underset{(p\times 1)}{\boldsymbol{\alpha}}\underset{(1\times n)}{\mathbf{W}} + \underset{(a\times n)}{\mathbf{E}} \\
\underset{(a\times n_j)}{\mathbf{X}_j} &= \underset{(k\times 1)}{\boldsymbol{\theta}_j}\underset{(1\times n_j)}{\mathbf{W}_j} + \underset{(a\times n_j)}{\mathbf{E}^{\mathbf{X}}_j} \quad j = 1, \dots, p
\end{aligned}$$

where $\underset{(a\times n_j)}{\mathbf{X}_j} = [\mathbf{x}_{j1}|\dots|\mathbf{x}_{jn_j}]$ is an $a \times n_j$ matrix of observations on the jth source, $\underset{(1\times n_j)}{\mathbf{W}_j}$ is a matrix of ones, and finally $\mathbf{E}^{\mathbf{X}}_j$ is defined analogously to $\mathbf{X}_j$ with the errors $\boldsymbol{\epsilon}^{\mathbf{x}}_{jk}, k = 1, \dots, n_j$ forming the columns of the matrix. This model is highly non–linear as two parameters $\boldsymbol{\Theta}$ and $\boldsymbol{\alpha}$ appear as a product which makes the analysis of this model much more complicated compared to when $\boldsymbol{\Theta}$ is known. However, it turns out that the MCMC algorithm for this model is relatively straightforward due to the conditional nature of the updates. The DAG for constant mixing version is given in figure 5.6 and the multilevel version in figure 5.7.

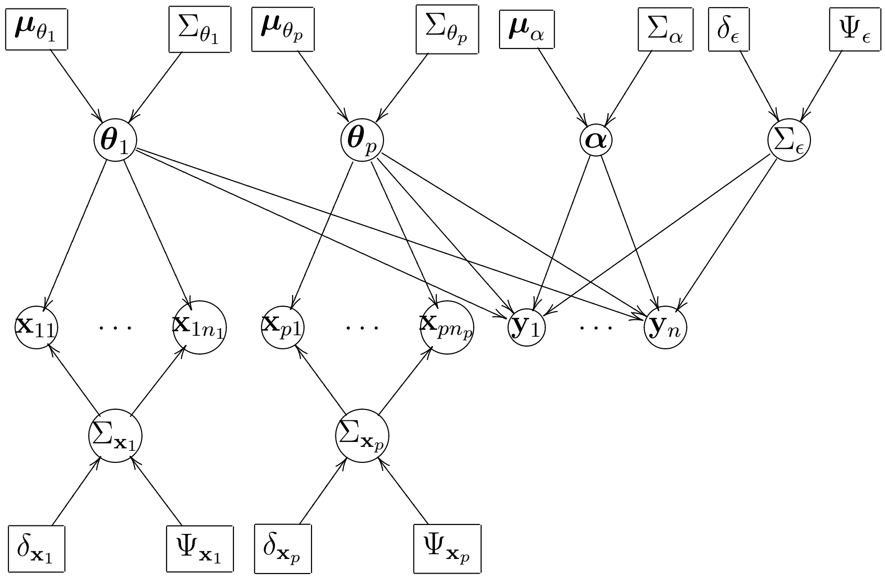

Figure 5.6: Directed Acyclic Graph (DAG) linear constant mixing model with observations on the sources. The square nodes indicate parameters that are known a priori, while circular nodes represent unknown parameters that are updated when the data, $\mathbf{y}_i, i = 1, \ldots, n$ and $\mathbf{x}_{jk}, j = 1, \ldots, p, k = 1, \ldots, n_j$ are observed.

This model can be see as a halfway house between the chemical mass balance models where the source matrix $\boldsymbol{\Theta}$ is assumed to be known and the models that assume that it is not known and can be inferred from receptor observations see Bandeen-Roche and Ruppert (1991); Bandeen-Roche (1994); Billheimer (2001); Park et al. (2000, 2001, 2002); Wolbers and Stahel (2005). The identifiability conditions are less stringent than usual as the observations on the sources pin down the sources, however, we still require $\boldsymbol{\Theta}$ to be of full rank. Seen in this way this is a natural extension of the Billheimer (2001) model, however, we defer discussion of this until the next chapter.

The full posterior distribution of the linear constant mixing model with observations on the sources is given by:

$$p(\boldsymbol{\alpha}, \Sigma_{\boldsymbol{\epsilon}}, \boldsymbol{\Theta}, \Sigma_{\mathbf{x}} | \mathbf{y}_1, \ldots, \mathbf{y}_n, \mathbf{X}_1, \ldots, \mathbf{X}_p) \propto h(\Sigma_{\boldsymbol{\epsilon}} | \delta, \Psi) \times g(\boldsymbol{\alpha} | \mu_{\boldsymbol{\alpha}}, \Sigma_{\boldsymbol{\alpha}}) \times \prod_{i=1}^{n} f(\mathbf{y}_i | \boldsymbol{\Theta}\boldsymbol{\alpha}, \Sigma_{\boldsymbol{\epsilon}})$$

$$\times h(\Sigma_{\mathbf{x}} | \delta_{\mathbf{x}}, \Psi_{\mathbf{x}}) \times \prod_{j=1}^{p} g(\boldsymbol{\theta}_j | \mu_{\boldsymbol{\theta}_j}, \Sigma_{\boldsymbol{\theta}_j}) \times \prod_{j=1}^{p} \prod_{k=1}^{n_j} f(\mathbf{x}_{jk} | \boldsymbol{\theta}_j, \Sigma_{\mathbf{x}}).$$

where we have assumed the functional form of the prior for $\boldsymbol{\theta}$, $\Sigma_{\mathbf{x}_1}$ and sampling distribution

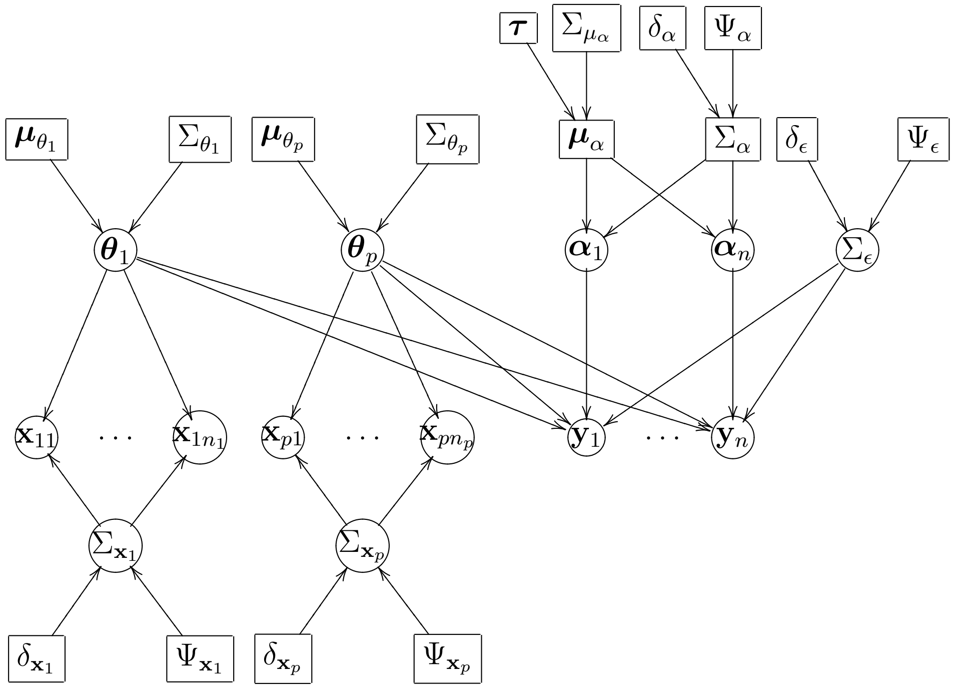

Figure 5.7: Directed Acyclic Graph (DAG) linear multilevel mixing model with observations on the sources. The square nodes indicate parameters that are known a priori, while circular nodes represent unknown parameters that are updated when the data, $\mathbf{y}_i, i = 1, \ldots, n$ and $\mathbf{x}_{jk}, j = 1, \ldots, p, k = 1, \ldots, n_j$ are observed.

for $\mathbf{x}_{jk}$ are g, h and f respectively, similar to distributional assumptions for $\boldsymbol{\alpha}$, Σ_ϵ and $\mathbf{y}_i$. This posterior distribution has more complicated structure than previously as the term $\boldsymbol{\Theta\alpha}$ now involves two unknown parameters. However, using the properties of conditional independence which are graphically depicted in the DAG (see figure 5.6) of the linear constant mixing model with observations on the sources, we can write down the full conditional distributions as follows:

$$
\begin{aligned}
p(\boldsymbol{\alpha}|\Sigma_{\boldsymbol{\epsilon}},) &\propto g(\boldsymbol{\alpha}|\boldsymbol{\mu}_{\boldsymbol{\alpha}}, \Sigma_{\boldsymbol{\alpha}}) \times \prod_{i=1}^{n} f(\mathbf{y}_i|\boldsymbol{\alpha}, \boldsymbol{\Theta}, \Sigma_{\boldsymbol{\epsilon}}) \\
p(\Sigma_{\boldsymbol{\epsilon}}|\boldsymbol{\alpha},) &\propto h(\Sigma_{\boldsymbol{\epsilon}}|\delta_{\boldsymbol{\epsilon}}, \Psi_{\boldsymbol{\epsilon}}) \times \prod_{i=1}^{n} f(\mathbf{y}_i|\boldsymbol{\alpha}, \boldsymbol{\Theta}, \Sigma_{\boldsymbol{\epsilon}}) \\
p(\Sigma_{\mathbf{x}_j}|\boldsymbol{\theta}_j,) &\propto h(\Sigma_{\mathbf{x}_j}|\delta_{\mathbf{x}_j}, \Psi_{\mathbf{x}_j}) \times \prod_{k=1}^{n_j} f(\mathbf{x}_{jk}|\boldsymbol{\theta}_j, \Sigma_{\mathbf{x}_j}) \\
p(\boldsymbol{\theta}_j|\Sigma_{\boldsymbol{\epsilon}},) &\propto g(\boldsymbol{\theta}_j|\boldsymbol{\mu}_{\boldsymbol{\theta}_j}, \Sigma_{\boldsymbol{\theta}_j}) \times \prod_{k=1}^{n_j} f(\mathbf{x}_{jk}|\boldsymbol{\theta}_j, \Sigma_{\mathbf{x}_j}) \times \prod_{i=1}^{n} f(\mathbf{y}_i|\boldsymbol{\alpha}, \boldsymbol{\Theta}, \Sigma_{\boldsymbol{\epsilon}})
\end{aligned}
$$

The full conditional distribution for $\boldsymbol{\theta}_j$ has information from three sources: the prior for $\boldsymbol{\theta}_j$, the observations on each source (prey) $\mathbf{x}_{jk}, k = 1, \ldots, n_j$ and also on the observations receptor (predator), $\mathbf{y}_i, i = 1, \ldots, n$. This is consistent with the approach taken by Billheimer (2001) where he used informative priors on the sources to help make the problem identifiable, though he didn't have observations on the sources. For reasons that will become clear in the next chapter we also consider modifying the fully Bayesian MCMC approach and breaking the information flow between the predator and prey, that is, we consider having the updates for $\boldsymbol{\theta}_j$ not depend on the $\mathbf{y}_i$.

To illustrate the model, we consider the linear constant mixing model with observations on the sources. We use identical settings as in the previous sections, that is, we consider the three source matrices, $\boldsymbol{\Theta}$ given in table 5.2, $\boldsymbol{\alpha} = (1, 1, 1)'$ and $\Sigma_\epsilon = \mathbf{I}_5$. Additionally, we assume we have $n_j = 50$ observations on the sources generated according to $\mathbf{x}_{jk} \sim \mathcal{N}^a(\boldsymbol{\theta}_j, \sigma^2_{\mathbf{x}}\mathbf{I}_5)$, with $\sigma^2_{\mathbf{x}} = 0.5, 1.0, 1.5$. We assume that each source has the same covariance matrix, though this assumption will be relaxed in the next chapter.

We used the same observations on the receptors/predators as were used to generate table 5.3. We focus on the three parameters of the model, namely the mixing vector $\boldsymbol{\alpha}$, the source matrix $\boldsymbol{\Theta}$, and the trace of the error covariance matrix, denoted by $\Sigma^T_{\boldsymbol{\epsilon}}$.

The results of 100,000 MCMC runs with a thinning of 10 are given in table 5.9 which can be compared to the inference given in table 5.3 where the source matrix, $\boldsymbol{\Theta}$, was known. The results for the essentially orthogonal case (a), indicate that the inference for the mixing vector, $\boldsymbol{\alpha}$, is more uncertain as indicated by the wider credible intervals, however, the posterior expectation is approximately the same in both cases. The inference for the trace of the error covariance matrix is not affected by the increased uncertainty of not knowing the sources exactly, the credible intervals are essentially the same length and the posterior mean is essential unchanged. Also, the variability in the sources themselves does not appear to have much effect on the mixing vector or the trace of the error covariance matrix, however, this is probably due to the small range of variances considered. The inference for the source profiles $\boldsymbol{\theta}$ is affected by the variability in the sources, not surprisingly, with the credible intervals getting slightly wider as the $\sigma^2_{\mathbf{x}}$ increases. The results for cases b) and c) behave in a similar way for the inference of the mixing vector $\boldsymbol{\alpha}$, with some interesting differences. Specifically, the third component has the same posterior mean, but increased uncertainty as indicated by widening credible intervals. However, the first two components behave in a similar way for case b), that is, increased variability compared to the known source profile case. For the last case, the increased uncertainty in not knowing the true source profile matrix, seemingly helps, in that the posterior uncertainty is reduced. This can be explained by the fact that the uncertainty actually helps by reducing the dependence between the sources. The other components namely the sources and the trace of the covariance matrix have similar properties as in case a.

We now discuss the impact of breaking the flow of information from the receptors to the sources by changing the full conditional distribution for $\boldsymbol{\theta}_j$ from the following:

$$p(\boldsymbol{\theta}_j|\Sigma_{\boldsymbol{\epsilon}},) \propto g(\boldsymbol{\theta}_j|\boldsymbol{\mu}_{\boldsymbol{\theta}_j}, \Sigma_{\boldsymbol{\theta}_j}) \times \prod_{k=1}^{n_j} f(\mathbf{x}_{jk}|\boldsymbol{\theta}_j, \Sigma_{\mathbf{x}_j}) \times \prod_{i=1}^{n} f(\mathbf{y}_i|\boldsymbol{\alpha}, \boldsymbol{\Theta}, \Sigma_{\boldsymbol{\epsilon}})$$

to

$$p(\boldsymbol{\theta}_j|\Sigma_{\boldsymbol{\epsilon}},) \propto g(\boldsymbol{\theta}_j|\boldsymbol{\mu}_{\boldsymbol{\theta}_j}, \Sigma_{\boldsymbol{\theta}_j}) \times \prod_{k=1}^{n_j} f(\mathbf{x}_{jk}|\boldsymbol{\theta}_j, \Sigma_{\mathbf{x}_j})$$

The physical reason for potentially considering this, is that the information on the sources should trump the information contained in the receptors.

The results of 100,000 MCMC runs with a thinning of 10 are given in table 5.10 and we compare them to the inference given in table 5.9. For case a), the inference for the various parameters doesn't appear to be affected by the elimination of the information flow from the receptor to the source, with one exception. The posterior distribution of Σ_ϵ^T has shifted to larger values, indicating a poorer fit to the receptors. In fact, this result is consistent across the other two cases as well. Interestingly, the posterior for the mixing process for cases b) and c) has reduced uncertainty in the posterior samples as indicated by shorter credible intervals, additionally, the posterior means have changed.

There are some subtle changes in the posterior means of the $\boldsymbol{\theta}$ parameters, but no obvious move away from the original parameter values. One potential reason for this is that the sample sizes and hence information content is the same for both the sources and receptors in the synthetic data.

5.4 Conclusion

This chapter has shown that Bayesian inference applied to linear mixing models can help diagnose ill–conditioned source matrices and has also shown that we can extend the model

$\boldsymbol{\Theta}$	$\sigma^2_{\mathbf{x}}$	meas	$\boldsymbol{\alpha} = (1,1,1)'$	$\boldsymbol{\theta}_1$	$\boldsymbol{\theta}_2$	$\boldsymbol{\theta}_3$	Σ^T_ϵ
				(-1.30,-1.21,-0.55,0.04,-0.22)	**(0.38,-0.97,-0.92,-0.03,0.73)**	**(-0.77,2.81,0.00,0.69,1.67)**	
a	0.5	pm	(0.96,1.14,0.96)	(-1.45,-1.06,-0.56,0.06,-0.21)	(0.34,-0.90,-1.04,-0.05,0.81)	(-0.84,2.86,0.09,0.76,1.48)	5.49
		lci	(0.55,0.70,0.59)	(-1.80,-1.80,-0.80,-0.17,-0.54)	(-0.01,-1.65,-1.29,-0.28,0.49)	(-1.20,2.16,-0.15,0.55,1.16)	4.51
		uci	(1.38,1.60,1.34)	(-1.11,-0.36,-0.30,0.28,0.11)	(0.71,-0.15,-0.80,0.17,1.15)	(-0.49,3.61,0.34,0.99,1.82)	6.66
	1.0	pm	(1.19,1.25,0.94)	(-1.18,-1.13,-0.51,0.12,-0.41)	(0.31,-0.83,-0.87,0.14,0.59)	(-0.83,3.28,-0.01,0.46,1.96)	5.51
		lci	(0.67,0.64,0.47)	(-1.55,-2.02,-0.82,-0.13,-0.93)	(-0.05,-1.72,-1.20,-0.10,0.09)	(-1.20,2.45,-0.31,0.22,1.46)	4.55
		uci	(1.79,1.89,1.49)	(-0.82,-0.25,-0.21,0.35,0.11)	(0.68,0.00,-0.58,0.37,1.06)	(-0.46,4.22,0.30,0.71,2.50)	6.70
	1.5	pm	(0.98,0.68,0.89)	(-1.32,-1.34,-0.69,0.23,-0.12)	(0.33,-0.95,-1.08,-0.10,0.95)	(-0.87,2.98,-0.26,0.68,1.81)	5.50
		lci	(0.56,0.23,0.57)	(-1.75,-2.21,-1.02,-0.09,-0.59)	(-0.09,-1.80,-1.43,-0.46,0.52)	(-1.29,2.19,-0.58,0.36,1.39)	4.53
		uci	(1.47,1.17,1.23)	(-0.90,-0.48,-0.37,0.55,0.29)	(0.76,-0.12,-0.75,0.25,1.40)	(-0.47,3.86,0.07,1.02,2.25)	6.72
				(0.04,0.84,0.24,-0.05,-0.57)	**(-0.07,0.99,0.12,-0.24,-0.68)**	**(2.05,-1.07,0.82,0.11,-0.41)**	
b	0.5	pm	(1.20,0.74,1.01)	(-0.06,0.89,0.07,-0.07,-0.58)	(-0.14,0.93,0.18,-0.11,-0.46)	(1.97,-1.00,0.70,0.16,-0.44)	6.01
		lci	(-2.11,-1.44,0.44)	(-0.48,0.50,-0.14,-0.25,-0.76)	(-0.57,0.49,-0.04,-0.31,-0.64)	(1.55,-1.43,0.47,-0.04,-0.63)	4.93
		uci	(3.27,4.39,1.54)	(0.34,1.32,0.27,0.14,-0.41)	(0.34,1.38,0.41,0.09,-0.26)	(2.41,-0.58,0.92,0.36,-0.24)	7.40
	1.0	pm	(1.66,0.24,0.93)	(0.04,0.95,0.24,-0.11,-0.65)	(-0.20,1.08,0.24,-0.17,-0.61)	(2.16,-1.06,0.65,0.13,-0.05)	6.02
		lci	(-3.10,-3.66,0.03)	(-0.46,0.53,-0.01,-0.30,-0.90)	(-0.69,0.60,-0.05,-0.40,-0.90)	(1.64,-1.55,0.36,-0.11,-0.34)	4.93
		uci	(5.98,4.70,2.01)	(0.49,1.44,0.54,0.12,-0.37)	(0.29,1.57,0.56,0.06,-0.29)	(2.71,-0.57,0.95,0.38,0.24)	7.40
	1.5	pm	(0.86,0.79,0.81)	(0.08,1.20,0.32,0.05,-0.63)	(0.01,0.81,0.04,-0.53,-0.55)	(2.29,-1.18,0.68,0.23,-0.68)	6.01
		lci	(-0.04,-0.31,0.46)	(-0.44,0.71,0.00,-0.32,-0.95)	(-0.50,0.32,-0.32,-0.93,-0.88)	(1.79,-1.71,0.35,-0.15,-1.01)	4.91
		uci	(1.71,2.03,1.20)	(0.58,1.72,0.65,0.44,-0.31)	(0.53,1.32,0.35,-0.15,-0.22)	(2.82,-0.69,1.02,0.62,-0.35)	7.35
				(1.15,-0.88,-0.24,-0.60,0.95)	**(1.15,-0.88,-0.24,0.61,0.95)**	**(-0.57,0.09,-0.30,-1.48,0.16)**	
c	0.5	pm	(1.21,0.61,0.92)	(1.16,-0.87,-0.06,-0.51,1.03)	(1.28,-0.97,-0.32,-0.69,1.01)	(-0.48,0.09,-0.27,-1.64,0.17)	5.41
		lci	(-1.27,-2.48,0.43)	(0.79,-1.12,-0.25,-0.79,0.80)	(0.93,-1.21,-0.53,-0.96,0.79)	(-0.85,-0.15,-0.47,-1.92,-0.06)	4.42
		uci	(4.46,2.81,1.66)	(1.56,-0.61,0.14,-0.23,1.26)	(1.66,-0.73,-0.14,-0.41,1.24)	(-0.13,0.34,-0.07,-1.37,0.39)	6.65
	1.0	pm	(1.72,-0.25,1.22)	(1.31,-1.05,-0.22,-0.66,1.04)	(1.10,-0.80,-0.36,-0.84,0.89)	(-0.36,0.01,-0.20,-1.35,0.32)	5.41
		lci	(0.09,-5.40,0.47)	(0.95,-1.34,-0.46,-0.95,0.80)	(0.68,-1.13,-0.65,-1.13,0.62)	(-0.77,-0.29,-0.48,-1.65,0.04)	4.45
		uci	(5.02,1.74,3.36)	(1.72,-0.77,0.02,-0.36,1.29)	(1.52,-0.50,-0.10,-0.56,1.17)	(0.05,0.34,0.06,-1.06,0.58)	6.66
	1.5	pm	(-0.06,1.97,1.04)	(1.19,-0.75,-0.13,-0.78,0.89)	(1.30,-0.83,-0.09,-0.45,0.96)	(-0.86,0.17,-0.57,-1.45,0.30)	5.43
		lci	(-5.59,-1.58,-0.04)	(0.75,-1.12,-0.44,-1.13,0.53)	(0.83,-1.16,-0.41,-0.76,0.66)	(-1.34,-0.20,-0.90,-1.80,-0.06)	4.43
		uci	(3.63,7.55,2.27)	(1.67,-0.39,0.17,-0.45,1.22)	(1.76,-0.47,0.18,-0.12,1.29)	(-0.38,0.55,-0.25,-1.11,0.65)	6.69

Table 5.9: Posterior summaries for the linear constant mixing model with observations on the sources of 100,000 MCMC runs of a Metropolis–Hastings within Gibbs algorithm with a thinning factor of 10. The settings for $\boldsymbol{\Theta}$ are given in table 5.2 which correspond to increasing degrees of linear dependence between the first two rows of the $\boldsymbol{\Theta}$ matrix. $\sigma^2_{\mathbf{x}}$ corresponds to the covariance matrix of the multivariate normal distribution of observations on the sources that is, $\Sigma_{\mathbf{x}} = \sigma^2_{\mathbf{x}}\mathbf{I}_d$, $\boldsymbol{\alpha}$ is the mixing vector, $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_3$ are the source profiles and $\Sigma^T_{\boldsymbol{\epsilon}}$ is trace of the error covariance matrix. The observations on receptors were generated as in table 5.3 with $\Sigma^2_{\boldsymbol{\epsilon}} = \mathbf{I}_5$. The posterior mean is denoted by pm, the upper and lower 95% element–wise credible intervals are denoted by lci, and uci respectively. The actual source profile is given in bold for each setting.

Θ	$\sigma^2_{\mathbf{x}}$	meas	$\boldsymbol{\alpha}=(1,1,1)^{'}$	$\boldsymbol{\theta}_1$	$\boldsymbol{\theta}_2$	$\boldsymbol{\theta}_3$	Σ^T_ϵ
				(-1.30,-1.21,-0.55,0.04,-0.22)	**(0.38,-0.97,-0.92,-0.03,0.73)**	**(-0.77,2.81,0.00,0.69,1.67)**	
a	0.5	pm	(0.92,1.13,0.95)	(-1.46,-1.11,-0.58,0.01,-0.24)	(0.33,-0.91,-1.06,-0.09,0.79)	(-0.85,2.90,0.08,0.74,1.48)	5.77
		lci	(0.50,0.75,0.59)	(-1.82,-1.87,-0.83,-0.23,-0.58)	(-0.01,-1.66,-1.32,-0.33,0.46)	(-1.20,2.14,-0.17,0.50,1.13)	4.68
		uci	(1.35,1.55,1.33)	(-1.11,-0.39,-0.32,0.25,0.09)	(0.69,-0.22,-0.81,0.15,1.13)	(-0.50,3.67,0.34,0.99,1.82)	7.29
	1.0	pm	(1.10,1.15,0.90)	(-1.17,-1.14,-0.61,0.06,-0.49)	(0.34,-0.80,-0.99,0.10,0.53)	(-0.83,3.33,-0.08,0.43,1.91)	6.37
		lci	(0.63,0.69,0.50)	(-1.55,-2.01,-0.94,-0.21,-1)	(-0.02,-1.67,-1.33,-0.18,0.01)	(-1.19,2.51,-0.41,0.16,1.41)	4.95
		uci	(1.61,1.66,1.32)	(-0.80,-0.29,-0.28,0.34,0.01)	(0.72,0.04,-0.65,0.37,1.04)	(-0.45,4.18,0.26,0.71,2.43)	8.69
	1.5	pm	(0.94,0.69,0.84)	(-1.34,-1.25,-0.73,0.17,-0.16)	(0.35,-0.90,-1.12,-0.16,0.94)	(-0.87,3.01,-0.30,0.62,1.77)	6.08
		lci	(0.54,0.32,0.52)	(-1.76,-2.15,-1.09,-0.22,-0.61)	(-0.06,-1.80,-1.47,-0.54,0.49)	(-1.30,2.14,-0.65,0.23,1.31)	4.82
		uci	(1.36,1.06,1.19)	(-0.91,-0.38,-0.38,0.54,0.28)	(0.78,-0.05,-0.77,0.22,1.40)	(-0.44,3.92,0.04,1.00,2.23)	8.01
				(0.04,0.84,0.24,-0.05,-0.57)	**(-0.07,0.99,0.12,-0.24,-0.68)**	**(2.05,-1.07,0.82,0.11,-0.41)**	
b	0.5	pm	(1.20,0.60,1.01)	(-0.09,0.91,0.08,-0.01,-0.59)	(-0.16,0.95,0.22,-0.09,-0.44)	(1.98,-1.02,0.73,0.20,-0.43)	6.40
		lci	(-0.29,-0.98,0.60)	(-0.54,0.46,-0.16,-0.22,-0.79)	(-0.60,0.54,-0.01,-0.30,-0.64)	(1.55,-1.48,0.49,-0.01,-0.63)	5.13
		uci	(2.75,2.09,1.45)	(0.35,1.35,0.31,0.20,-0.39)	(0.25,1.39,0.45,0.11,-0.25)	(2.44,-0.59,0.97,0.41,-0.23)	8.17
	1.0	pm	(0.92,0.73,0.88)	(0.06,1.02,0.30,-0.06,-0.58)	(-0.21,1.12,0.28,-0.18,-0.59)	(2.19,-1.04,0.69,0.15,0)	6.77
		lci	(-0.39,-0.62,0.46)	(-0.47,0.55,0.00,-0.33,-0.86)	(-0.75,0.64,-0.02,-0.45,-0.89)	(1.67,-1.54,0.39,-0.12,-0.29)	5.33
		uci	(2.27,2.08,1.33)	(0.57,1.50,0.61,0.20,-0.29)	(0.31,1.61,0.59,0.08,-0.29)	(2.73,-0.55,1.00,0.41,0.30)	8.82
	1.5	pm	(0.89,0.68,0.80)	(0.07,1.19,0.32,0.10,-0.63)	(-0.01,0.81,0.02,-0.57,-0.56)	(2.31,-1.21,0.67,0.25,-0.68)	6.44
		lci	(0.28,-0.12,0.47)	(-0.49,0.67,-0.05,-0.32,-0.99)	(-0.54,0.29,-0.35,-0.98,-0.92)	(1.77,-1.74,0.31,-0.15,-1.04)	5.16
		uci	(1.54,1.49,1.18)	(0.60,1.72,0.68,0.50,-0.27)	(0.53,1.33,0.38,-0.17,-0.21)	(2.85,-0.70,1.03,0.66,-0.32)	8.22
				(1.15,-0.88,-0.24,-0.60,0.95)	**(1.15,-0.88,-0.24,0.61,0.95)**	**(-0.57,0.09,-0.30,-1.48,0.16)**	
c	0.5	pm	(1.02,0.79,0.87)	(1.17,-0.83,-0.04,-0.52,1.05)	(1.31,-0.96,-0.37,-0.69,1)	(-0.47,0.11,-0.28,-1.66,0.17)	5.60
		lci	(0.09,-0.09,0.50)	(0.80,-1.09,-0.24,-0.80,0.82)	(0.95,-1.22,-0.57,-0.97,0.76)	(-0.85,-0.14,-0.47,-1.94,-0.06)	4.53
		uci	(1.99,1.64,1.27)	(1.54,-0.58,0.17,-0.23,1.29)	(1.69,-0.71,-0.17,-0.42,1.23)	(-0.11,0.38,-0.07,-1.38,0.40)	6.98
	1.0	pm	(1.16,0.51,0.94)	(1.30,-1.02,-0.25,-0.65,1.03)	(1.14,-0.80,-0.41,-0.85,0.89)	(-0.39,0.05,-0.23,-1.37,0.30)	5.82
		lci	(0.18,-0.66,0.45)	(0.91,-1.37,-0.54,-0.96,0.73)	(0.74,-1.13,-0.70,-1.14,0.59)	(-0.80,-0.29,-0.53,-1.67,0.01)	4.64
		uci	(2.20,1.58,1.49)	(1.70,-0.68,0.04,-0.35,1.33)	(1.54,-0.45,-0.12,-0.55,1.19)	(0,0.39,0.05,-1.08,0.60)	7.53
	1.5	pm	(1.15,0.82,0.72)	(1.22,-0.70,-0.14,-0.82,0.85)	(1.28,-0.78,-0.17,-0.39,1.03)	(-0.87,0.23,-0.62,-1.45,0.32)	6.19
		lci	(0.17,-0.15,0.28)	(0.74,-1.10,-0.48,-1.19,0.47)	(0.81,-1.16,-0.50,-0.74,0.66)	(-1.36,-0.16,-0.95,-1.80,-0.06)	4.83
		uci	(2.15,1.78,1.16)	(1.69,-0.31,0.20,-0.47,1.22)	(1.74,-0.40,0.17,-0.04,1.41)	(-0.41,0.61,-0.29,-1.08,0.70)	8.30

Table 5.10: Posterior summaries for the linear constant mixing model with observations on the sources with no information flow from the receptors to the sources of 100,000 MCMC runs of a Metropolis–Hastings within Gibbs algorithm with a thinning factor of 10. The settings for $\boldsymbol{\Theta}$ are given in table 5.2 which correspond to increasing degrees of linear dependence between the first two rows of the $\boldsymbol{\Theta}$ matrix. $\sigma^2_{\mathbf{x}}$ corresponds to the covariance matrix of the multivariate normal distribution of observations on the sources that is, $\Sigma_{\boldsymbol{\epsilon}} = \sigma^2_{\boldsymbol{\epsilon}}\mathbf{I}_d$, $\boldsymbol{\alpha}$ is the mixing vector, $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_3$ are the source profiles and $\Sigma^T_{\boldsymbol{\epsilon}}$ is trace of the error covariance matrix. The observations on receptors were generated as in table 5.3 with $\Sigma^2_{\boldsymbol{\epsilon}} = \mathbf{I}_5$. The posterior mean is denoted by pm, the upper and lower 95% element–wise credible intervals are denoted by lci, and uci respectively. The actual source profile is given in bold for each setting.

to cases where we don't have exact knowledge of the source profiles.

It is widely known that multicollinearity is a serious problem and given the nature of the sources considered here and in the next chapter ways of dealing with it are desperately needed. Some potential ways are:

- Examine the source matrix $\boldsymbol{\Theta}$ and determine whether the offending columns can be removed or combined. This can be difficult in large dimensional problems where the relationships between the columns can be quite complex.

- Perform a singular value decomposition of the columns and form principal components of the columns. There is a practical difficulty here that the components lose their physical or biological interpretation. This approach can work well when there are no restrictions on where the sources lie, however, this is problematic when the sources lie in the positive quadrant or in the simplex. Principal components for compositional data (see Aitchison, 2003) offers some promise, though, there is still no guarantee of physical interpretability.

- If prior information on the relative contributions of each component to the mixing vector is available then this information can be used to help locate the proper range of parameter space and hence deal with the non–identifiability. This is theoretically appealing, however, such information is not typically available in practice.

- Regression approaches to the linear mixing models have used variants of ridge regression (see Henry et al., 1984). However, it is not obvious how one would implement this approach in the more complicated situation considered in the next chapter, where we deal with the case where the sources are themselves compositions.

- Our preferred method of handling this issue is a model selection method, however, due to time constraints this approach was not possible. The ideas contained in Green (1995) would be essential here.

CHAPTER 6

INFERENCE ON PREDATOR DIET COMPOSITIONS: AN APPLICATION OF LINEAR MIXING MODELS

6.1 Introduction

The overall aim of this chapter is to develop Bayesian inference for the convex linear mixing model applied to fatty acid signatures. Specifically, the Bayesian approach allows for incorporation of prior information on diet compositions when available and perhaps more importantly gives an accurate account of uncertainty. We present some details on the chemical mass balance approach to fatty acids. The bulk of the chapter develops the convex linear mixing model and several variants and presents results using synthetic data. Finally, we give two applications: a captive sea bird study (see Iverson et al., 2007) and a wild harbour seal study for which the individuals animals had an attached critter cam that recorded prey encounters (see Iverson et al., 2004; Bowen et al., 2002).

6.2 Biological Basis of the Model

The biological basis of the model follows a very similar line of reasoning as the chemical mass balance models discussed in the previous chapter with some modifications. Firstly, the fatty acids are compositional in nature, that is, we only know the relative amounts of the various fatty acids due to the sampling method used. Specifically, a small blubber biopsy (or adipose tissue) is taken from the predator rather a complete analysis of the animal's blubber. In fact, even the total amount of blubber of a given animal is not known as estimating the fat content of each individual animal is not practically feasible.

Therefore we assume that the fatty acid profiles are mostly deposited in the adipose tissue without modification. However, as Iverson et al. (2004) noted, this assumption isn't strictly speaking true, that is, predators do modify some fatty acids, in other words, some biosynthesi , modification and differing deposition characteristics occur. That is, the fatty acid profile of the fat deposited in the adipose tissue is not the same as the fatty acid profile of the ingested prey. Iverson et al. (2004) dealt with this issue via the idea of calibration coefficients. These were calculated by a long term feeding study, where a group of predators was fed the same diet for an extended period, assumed to represent complete turnover of all fatty acids. The fatty acid composition of the diet was then compared to the deposited fatty acid signature and a calibration coefficient was computed to mimic predator metabolism.

For illustration consider the following simplified ecosystem. Assume the predator has access to p prey types and let $a_1, \ldots, a_p$ be the amounts in arbitrary units that the predator consumed of each prey type and denote the total amount consumed by $a_T = \sum_{i=1}^{p} a_i$. As mentioned previously, we can't reconstruct the actual amounts due to the nature of the sampling. However, expressing this in proportional terms we have $\tau_i = a_i/a_T$.

Now consider the amount of fat that was actually deposited in the predator, each prey has an associated percentage fat, which we denote by f_i. Thus, the total amount of fat deposited is given by

$$f_T = \sum_{i=1}^{p} a_i f_i$$

however, we can substitute $a_i = \tau_i a_T$ giving the total fat deposited in terms of the proportions of prey consumed, their fat contents and the total amount consumed:

$$f_T = \left(\sum_{i=1}^{p} \tau_i f_i \right) a_T.$$

If we consider the proportions of fat deposited, denoted by α_i we have

$$\alpha_i = \frac{\tau_i f_i a_T}{\sum_{i=1}^{p} \tau_i f_i a_T} = \frac{\tau_i f_i}{\sum_{i=1}^{p} \tau_i f_i}$$

Thus, if we can estimate the composition deposited in fat, α_i's, we can then invert the last equation to estimate the true proportions τ_i's which are of interest. That is,

$$\tau_i = \frac{\alpha_i / f_i}{\sum_{i=1}^{p} \alpha_i / f_i}.$$

The QFASA method introduced by Iverson et al. (2004) deals directly with deposited fat, therefore we can estimate the α_i's and hence the actual proportions of each prey item consumed.

6.3 Constant Convex Mixing Model With Sources Known

In this section we consider the constant convex mixing model which we study in detail, as it illustrates the basis for all subsequent models considered in the sequel. Biologically this model is unrealistic as it assumes that there is no predator metabolism effects, that is, the fat the predator eats is deposited in its adipose tissue without modification. In addition, it assumes that each potential prey source has an equal amount of fat. That is, the same amount of fat is deposited from herring which is high in fat as a cod which is lower in fat. We gradually introduce more complicated models which address these biological issues.

The basic model can be stated as follows:

$$\mathbf{y}_i = \sum_{j=1}^{p} \alpha_j \boldsymbol{\theta}_j \oplus \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n$$

where $\mathbf{y}_i$ is an a–dimensional compositional vector also referred to as the predator fatty acid profile, $\boldsymbol{\theta}_j, j = 1, \ldots, p$ is an a–dimensional vector which represent the mixing components or the jth prey fatty acid profile, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)'$ is a p–dimensional compositional vector of mixing coefficients also called the diet composition vector, $\oplus$ indicates the perturbation operator introduced in chapter 2 and $\boldsymbol{\epsilon}_i$ is an a–dimensional error compositional vector. Note that all vectors belong to a simplex of appropriate dimension, that is, all components must lie strictly between zero and one and must obey the unit sum constraint. All vectors are assumed to be column vectors and we denote matrix transposition by $'$. Note this model is a special case of the Billheimer (2001) model where he had a separate mixing vector for each observation $\mathbf{y}_i$.

The basic model can be expressed more compactly as follows:

$$\underset{(a\times 1)}{\mathbf{y}_i} = \underset{(a\times p)}{\boldsymbol{\Theta}}\underset{(p\times 1)}{\boldsymbol{\alpha}} \oplus \underset{(a\times 1)}{\boldsymbol{\epsilon}_i}, \quad i = 1, \ldots, n$$

where $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1 | \ldots | \boldsymbol{\theta}_p]$ is an $a \times p$ dimensional matrix with each prey fatty acid profile forming the columns and | indicates column concatenation.

We collect the sample of n predators $\mathbf{y}_i, i = 1, \ldots, n$ into a matrix $\mathbf{Y}$ by concatenating them column–wise as given below

$$\underset{(a\times n)}{\mathbf{Y}} = [\mathbf{y}_1 | \ldots | \mathbf{y}_n].$$

The matrix version of the basic model written in terms of a structural equation is given by:

$$\begin{aligned} \underset{(a\times n)}{\mathbf{Y}} &= \underset{(a\times p)}{\boldsymbol{\Theta}}\underset{(p\times n)}{\Gamma} \oplus_c \underset{(a\times n)}{\mathbf{E}} \\ \underset{(p\times n)}{\Gamma} &= \boldsymbol{\phi}_c^{-1}\left(\underset{((p-1)\times 1)}{\phi(\boldsymbol{\alpha})}\underset{(1\times n)}{\mathbf{W}} \right) \end{aligned}$$

where $\oplus_c$ is the column-wise perturbation operator, $\mathbf{E} = [\boldsymbol{\epsilon}_1|\ldots|\boldsymbol{\epsilon}_n]$ is an $a \times n$ matrix of compositional errors, $\boldsymbol{\phi}_c^{-1}$ is the inverse log–ratio function applied to the columns of its matrix argument, ϕ_c is the log–ratio transformation applied to the columns of a matrix, $\mathbf{W}$ is a known $1 \times n$ design matrix of ones and Γ is a $p \times n$ matrix. The usual order of precedence applies, that is, the matrix multiplications are done before the compositional addition $\oplus_c$.

To complete the model specification we need to assign a distribution to the compositional errors $\boldsymbol{\epsilon}_i$. We assume they follow a logistic normal distribution

$$\pi(\boldsymbol{\epsilon}_i|\mathbf{0}_{a-1}, \Sigma_{\boldsymbol{\epsilon}})$$

which we denote by $\mathcal{L}^a(\mathbf{0}_{a-1}, \Sigma_{\boldsymbol{\epsilon}})$, where $\mathbf{0}_{a-1}$ is a vector of zeros of dimension $a-1$ and covariance matrix $\Sigma_{\boldsymbol{\epsilon}}$. The density function is (see chapter 2 for more details)

$$\begin{aligned}\pi(\boldsymbol{\epsilon}_i|\mathbf{0}_{a-1}, \Sigma_{\boldsymbol{\epsilon}}) &= \left(\frac{1}{2\pi}\right)^{(a-1)/2} |\Sigma_{\boldsymbol{\epsilon}}|^{-1/2} \left(\frac{1}{\prod_{t=1}^{a}[\boldsymbol{\epsilon}_i]_j}\right) \\ &\times \; exp\left\{-\frac{1}{2}(\phi(\mathbf{x}_i) - \mathbf{0}_{a-1})' \Sigma_{\boldsymbol{\epsilon}}^{-1}(\phi(\mathbf{x}_i) - \mathbf{0}_{a-1})\right\},\end{aligned}$$

where $[\boldsymbol{\epsilon}_i]_t$ denotes the t element of the vector $\boldsymbol{\epsilon}_i$.

We assign a logistic normal prior distribution to the diet composition vector $\boldsymbol{\alpha}$, denoted by $\mathcal{L}^p(\boldsymbol{\mu}_{\boldsymbol{\alpha}}, \Sigma_{\boldsymbol{\alpha}})$, where $\boldsymbol{\mu}_\alpha$ and $\Sigma_{\boldsymbol{\alpha}}$ are parameters representing our uncertainty/knowledge about the diet. Similarly, we assign a logistic normal prior distribution for each of the p prey types denoted by $\mathcal{L}^a(\boldsymbol{\mu}_{\boldsymbol{\theta}_j}, \Sigma_{\boldsymbol{\theta}_j})$. Finally, we assign an inverse Wishart distribution to $\Sigma_{\boldsymbol{\epsilon}}$, denoted by $\mathcal{IW}^{a-1}(\delta_{\boldsymbol{\epsilon}}, \Psi_{\boldsymbol{\epsilon}})$ (see appendix C). This induces the following conditional distribution on $\mathbf{y}_i$

$$\pi(\mathbf{y}_i|\boldsymbol{\alpha}, \boldsymbol{\Theta}, \Sigma_{\boldsymbol{\epsilon}}, B) = \mathcal{L}^a(\phi(\boldsymbol{\Theta}\boldsymbol{\alpha}), \Sigma_{\boldsymbol{\epsilon}})$$

where B indicates the additional background information, typically the parameters of the prior distributions. However, the presence of the background information is assumed and will be dropped.

The joint distribution of a single observation $\mathbf{y}_i$ and the model parameters is,

$$(\mathbf{y}_i, \boldsymbol{\Theta}, \boldsymbol{\alpha}, \Sigma_{\boldsymbol{\epsilon}}) = \pi(\boldsymbol{\alpha}|\boldsymbol{\mu}_{\boldsymbol{\alpha}}, \Sigma_{\boldsymbol{\alpha}}) \left\{\prod_{j=1}^{p} \pi(\boldsymbol{\theta}_j|\boldsymbol{\mu}_{\boldsymbol{\theta}_j}, \Sigma_{\boldsymbol{\theta}_j})\right\} \pi(\Sigma_{\boldsymbol{\epsilon}}^{-1}|\Psi_{\boldsymbol{\epsilon}}, \delta_{\boldsymbol{\epsilon}}) \pi(\mathbf{Y}_i|\boldsymbol{\alpha}, \boldsymbol{\Theta}, \Sigma_{\boldsymbol{\epsilon}}).$$

The joint distribution is then the product over the n observations, as they are conditionally independent given the parameters. The posterior distribution is given by

$$\pi(\boldsymbol{\Theta}, \boldsymbol{\alpha}, \Sigma_{\boldsymbol{\epsilon}} | \mathbf{y}_1, \ldots, \mathbf{y}_n, B) = \frac{\prod_{i=1}^{n} \pi(\mathbf{y}_i, \boldsymbol{\Theta}, \boldsymbol{\alpha}, \Sigma_{\boldsymbol{\epsilon}} | B)}{\int \prod_{i=1}^{n} \pi(\mathbf{y}_i, \boldsymbol{\Theta}, \boldsymbol{\alpha}, \Sigma_{\boldsymbol{\epsilon}} | B) d\boldsymbol{\Theta} d\boldsymbol{\alpha} d\Sigma_{\boldsymbol{\epsilon}}}.$$

Aitchison and Bacon-Shone (1999) approached a similar problem from a likelihood perspective, however, they were unable to derive an exact likelihood and developed several approximations using Taylor series methods. The complication is the product term, $\boldsymbol{\Theta\alpha}$, which also makes the posterior non–analytically tractable. However, we can use Markov Chain Monte Carlo methods to sample from the posterior, provided it is a proper distribution.

It is instructive to look at the Directed Acyclic Graph (DAG) for this model. The main advantage of graphical representation is that enables one to write down the conditional distributions required for MCMC in a straightforward way (see chapter A). Figure 6.1 gives the DAG for the base model. Square nodes indicate parameters known a priori, while circular nodes represent unknown parameters that are updated when the data, $\mathbf{y}_i, i = 1, \ldots, n$, is observed. The joint distribution of the data and the unknown parameters can be written using the rules given in A, but the more important use of DAG's is the fact that the full conditional distributions can be read directly off the graph. The full conditional distributions are key ingredients to Gibbs samplers, more specifically, Metropolis–within–Gibbs algorithms which we employ.

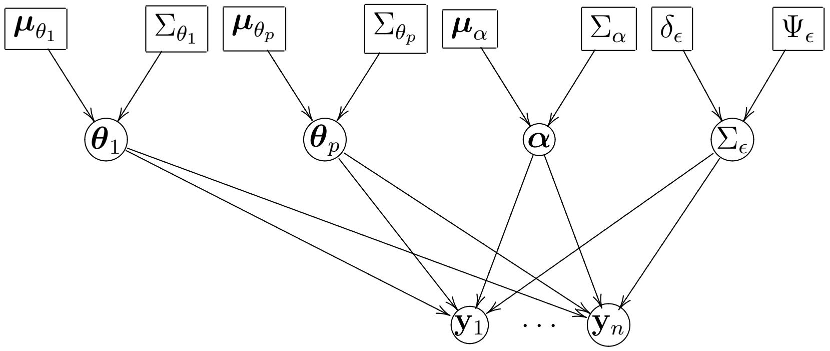

Figure 6.1: Directed Acyclic Graph (DAG) for the base model. The square nodes indicate parameters that are known a priori, while circular nodes represent unknown parameters that are updated when the data, $\mathbf{y}_i, i = 1, \ldots, n$, are observed.

DAG theory states that the full conditional distributions for any node depend only on its immediate parents and children. For more details see the section of on Graphical models in Appendix A. The conditional distributions are as follows:

$$\begin{aligned}
\pi(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{-j},\boldsymbol{\alpha},\Sigma_{\boldsymbol{\epsilon}},\mathbf{Y}) &= \pi(\boldsymbol{\theta}_j|\boldsymbol{\mu}_{\boldsymbol{\theta}_j},\Sigma_{\boldsymbol{\theta}_j})\prod_{i=1}^{n}\pi(\mathbf{y}_i|\boldsymbol{\alpha},\boldsymbol{\Theta},\Sigma_{\boldsymbol{\epsilon}}),\quad j=1,\ldots,p,\\
\pi(\boldsymbol{\alpha}|\boldsymbol{\Theta},\Sigma_{\boldsymbol{\epsilon}},\mathbf{Y}) &= \pi(\boldsymbol{\alpha}|\boldsymbol{\mu}_{\boldsymbol{\alpha}},\Sigma_{\boldsymbol{\alpha}})\prod_{i=1}^{n}\pi(\mathbf{y}_i|\boldsymbol{\alpha},\boldsymbol{\Theta},\Sigma_{\boldsymbol{\epsilon}}),\\
\pi(\Sigma_{\boldsymbol{\epsilon}}|\boldsymbol{\alpha},\boldsymbol{\Theta},\mathbf{Y}) &= \pi(\Sigma_{\boldsymbol{\epsilon}}|\Psi_{\boldsymbol{\epsilon}},\delta_{\boldsymbol{\epsilon}})\prod_{i=1}^{n}\pi(\mathbf{y}_i|\boldsymbol{\alpha},\boldsymbol{\Theta},\Sigma_{\boldsymbol{\epsilon}}),
\end{aligned}$$

where $\boldsymbol{\theta}_{-j}$ is the collection of all other $\boldsymbol{\theta}$'s other than $\boldsymbol{\theta}_j$.

The term $\phi(\boldsymbol{\Theta}\boldsymbol{\alpha})$ in the sampling distribution for $\mathbf{y}_i$ means that the full conditional distributions for $\boldsymbol{\alpha}$ and each of the $\boldsymbol{\theta}_j$'s do not follow standard distributions. Metropolis–Hastings updates will be needed for these parameters and hence we employ a Metropolis–within–Gibbs algorithm. We can use the Gibbs sampler for updating $\Sigma_{\boldsymbol{\epsilon}}$, as our prior distribution for $\Sigma_{\boldsymbol{\epsilon}}$ is inverse–Wishart and this is conjugate for $\Sigma_{\boldsymbol{\epsilon}}$ in the logistic normal distribution. Our MCMC algorithm, the reversible systematic scan Metropolis–within–Gibbs, for the base model consists of the following:

0. Choose starting values for $\boldsymbol{\alpha}_0$ and $\boldsymbol{\theta}_{j,0}, j=1,\ldots,p$.

1. Sample $\Sigma^*_{\boldsymbol{\epsilon}}$ from
$$\mathcal{IW}^{a-1}(\delta_{\boldsymbol{\epsilon}}+n,\Psi_{\boldsymbol{\epsilon}}+S(\mathbf{y}_i,\boldsymbol{\Theta}_{t-1}\boldsymbol{\alpha}_{t-1}))$$
   where
$$S(\mathbf{y}_i,\boldsymbol{\Theta}_{t-1}\boldsymbol{\alpha}_{t-1})=\sum_{i=1}^{n}\left(\phi(\mathbf{y}_i)-\phi(\boldsymbol{\Theta}_{t-1}\boldsymbol{\alpha}_{t-1})\right)\left(\phi(\mathbf{y}_i)-\phi(\boldsymbol{\Theta}_{t-1}\boldsymbol{\alpha}_{t-1})\right)'$$
   and $\boldsymbol{\Theta}_{t-1}=[\boldsymbol{\theta}_{1,t-1}|\ldots|\boldsymbol{\theta}_{p,t-1}]$.

2. For $j=1,\ldots,p$ generate $\boldsymbol{\theta}^*_j$ from the following Metropolis–Hastings algorithm with proposal distribution $q_{\boldsymbol{\theta}_j}(.|\boldsymbol{\theta}_{j,t-1})$

   (a) Generate $\boldsymbol{\nu}\sim q_{\boldsymbol{\theta}_j}(\boldsymbol{\nu}|\boldsymbol{\theta}_{j,t-1})$

(b)

$$\boldsymbol{\theta}_j^* = \begin{cases} \boldsymbol{\nu} & \text{with probability} \rho_{\boldsymbol{\theta}_j}(\boldsymbol{\theta}_{j,t-1}, \boldsymbol{\nu}) \\ \boldsymbol{\theta}_{j,t-1} & \text{with probability} 1 - \rho_{\boldsymbol{\theta}_j}(\boldsymbol{\theta}_{j,t-1}, \boldsymbol{\nu}) \end{cases}$$

where

$$\rho_{\boldsymbol{\theta}_j}(\boldsymbol{\nu}^o, \boldsymbol{\nu}^n) = \min\left\{ \frac{f_{\boldsymbol{\theta}_j}(\boldsymbol{\nu}^n) q_{\boldsymbol{\theta}_j}(\boldsymbol{\nu}^o|\boldsymbol{\nu}^n)}{f_{\boldsymbol{\theta}_j}(\boldsymbol{\nu}^o) q_{\boldsymbol{\theta}_j}(\boldsymbol{\nu}^n|\boldsymbol{\nu}^o)}, 1 \right\}$$

3. Generate $\boldsymbol{\alpha}_t$ from the following Metropolis–Hastings algorithm with proposal distribution $q_{\boldsymbol{\alpha}}(.|\boldsymbol{\alpha}_{t-1})$

   (a) Generate $\boldsymbol{\nu} \sim q_{\boldsymbol{\alpha}}(\boldsymbol{\nu}|\boldsymbol{\alpha}_{t-1})$

   (b)

$$\boldsymbol{\alpha}_t = \begin{cases} \boldsymbol{\nu} & \text{with probability} \rho_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}_{t-1}, \boldsymbol{\nu}) \\ \boldsymbol{\alpha}_{t-1} & \text{with probability} 1 - \rho_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}_{t-1}, \boldsymbol{\nu}) \end{cases}$$

   where

$$\rho_{\boldsymbol{\alpha}}(\boldsymbol{\nu}^o, \boldsymbol{\nu}^n) = \min\left\{ \frac{f_{\boldsymbol{\alpha}}(\boldsymbol{\nu}^n) q_{\boldsymbol{\alpha}}(\boldsymbol{\nu}^o|\boldsymbol{\nu}^n)}{f_{\boldsymbol{\alpha}}(\boldsymbol{\nu}^o) q_{\boldsymbol{\alpha}}(\boldsymbol{\nu}^n|\boldsymbol{\nu}^o)}, 1 \right\}$$

4. For $j = p, \dots, 1$ generate $\boldsymbol{\theta}_{j,t}$ from the following Metropolis–Hastings algorithm with proposal distribution $q_{\boldsymbol{\theta}_j}(.|\boldsymbol{\theta}_j^*)$

   (a) Generate $\boldsymbol{\nu} \sim q_{\boldsymbol{\theta}_j}(\boldsymbol{\nu}|\boldsymbol{\theta}_j^*)$

   (b)

$$\boldsymbol{\theta}_{j,t} = \begin{cases} \boldsymbol{\nu} & \text{with probability} \rho_{\boldsymbol{\theta}_j}(\boldsymbol{\theta}_j^*, \boldsymbol{\nu}) \\ \boldsymbol{\theta}_j^* & \text{with probability} 1 - \rho_{\boldsymbol{\theta}_j}(\boldsymbol{\theta}_j^*, \boldsymbol{\nu}) \end{cases}$$

5. Sample $\Sigma_{\boldsymbol{\epsilon},t}$ from

$$\mathcal{IW}^{a-1}(\delta_{\boldsymbol{\epsilon}} + n, \Psi_{\boldsymbol{\epsilon}} + S(\mathbf{y}_i, \boldsymbol{\Theta}_t \boldsymbol{\alpha}_t))$$

   where

$$S(\mathbf{y}_i, \boldsymbol{\alpha}_t^{'} \Theta_t) = \sum_{i=1}^{n} (\phi(\mathbf{y}_i) - \phi(\boldsymbol{\Theta}_t \boldsymbol{\alpha}_t))(\phi(\mathbf{y}_i) - \phi(\boldsymbol{\Theta}_t \boldsymbol{\alpha}_t))^{'}$$

   and $\boldsymbol{\Theta} = [\boldsymbol{\theta}_{1,t}| \dots |\boldsymbol{\theta}_{p,t}]$.

6. Repeat steps 1-5 and increment t till “convergence”.

The target densities $f_{\boldsymbol{\alpha}}(\boldsymbol{\nu})$ and $f_{\boldsymbol{\theta}_j}(\boldsymbol{\nu})$ are given by the following

$$\begin{aligned} f_{\boldsymbol{\alpha}}(\boldsymbol{\nu}) &= \pi(\boldsymbol{\nu}|\boldsymbol{\mu}_{\boldsymbol{\alpha}}, \Psi_{\boldsymbol{\alpha}}) \prod_{i=1}^{n} \pi(\mathbf{y}_i|\boldsymbol{\alpha} = \boldsymbol{\nu}, \boldsymbol{\Theta}, \Psi_{\boldsymbol{\epsilon}}), \\ f_{\boldsymbol{\theta}_j}(\boldsymbol{\nu}) &= \pi(\boldsymbol{\nu}|\boldsymbol{\mu}_{\boldsymbol{\theta}_j}, \Psi_{\boldsymbol{\theta}_j}) \prod_{i=1}^{n} \pi(\mathbf{y}_i|\boldsymbol{\alpha}, \boldsymbol{\theta}_j = \boldsymbol{\nu}, \boldsymbol{\theta}_{-j}, \Psi_{\boldsymbol{\epsilon}}). \end{aligned}$$

We should explicitly include everything we are conditioning on in the form of the distributions given above, however, that notation gets cumbersome and clumsy very quickly. The notation we have adopted means that when implementing the Metropolis–within–Gibbs steps, it isn't clear exactly what is being conditioned on. To clarify, assume $p = 2$ and that we are updating $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ in step 2 of the algorithm. The full version of $f_{\boldsymbol{\theta}_1}(\boldsymbol{\nu})$ and $f_{\boldsymbol{\theta}_1}(\boldsymbol{\nu})$ are as follows

$$\begin{aligned} & f_{\boldsymbol{\theta}_1}(\boldsymbol{\nu}|\boldsymbol{\alpha} = \boldsymbol{\alpha}_{t-1}, \boldsymbol{\theta}_2 = \boldsymbol{\theta}_{2,t-1}, \Psi_{\boldsymbol{\epsilon}} = \Psi^*_{\boldsymbol{\epsilon}}) = \\ & \pi(\boldsymbol{\theta}_1 = \boldsymbol{\nu}|\boldsymbol{\mu}_{\boldsymbol{\theta}_1}, \Psi_{\boldsymbol{\theta}_1}) \prod_{i=1}^{n} \pi(\mathbf{y}_i|\boldsymbol{\alpha} = \boldsymbol{\alpha}_{t-1}, \boldsymbol{\theta}_1 = \boldsymbol{\nu}, \boldsymbol{\theta}_2 = \boldsymbol{\theta}_{2,t-1}, \Psi_{\boldsymbol{\epsilon}} = \Psi^*_{\boldsymbol{\epsilon}}) \\ & f_{\boldsymbol{\theta}_2}(\boldsymbol{\nu}|\boldsymbol{\alpha} = \boldsymbol{\alpha}_{t-1}, \boldsymbol{\theta}_1 = \boldsymbol{\theta}^*_1, \Psi_{\boldsymbol{\epsilon}} = \Psi_{\boldsymbol{\epsilon}_{t-1}}) = \\ & \pi(\boldsymbol{\theta}_2 = \boldsymbol{\nu}|\boldsymbol{\mu}_{\boldsymbol{\theta}_2}, \Psi_{\boldsymbol{\theta}_2}) \prod_{i=1}^{n} \pi(\mathbf{y}_i|\boldsymbol{\alpha} = \alpha_{t-1}, \boldsymbol{\theta}_1 = \theta^*_1, \boldsymbol{\theta}_2 = \nu, \Psi_{\boldsymbol{\epsilon}} = \Psi^*_{\boldsymbol{\epsilon}}). \end{aligned}$$

We will use the shorthand notation in the sequel but implicitly mean the full version in the example above.

The proposal distributions are given by

$$\begin{aligned} q_{\boldsymbol{\alpha}}(\boldsymbol{\nu}^n|\boldsymbol{\nu}^o) &= \mathcal{L}^p\left(\phi(\boldsymbol{\nu}^o), \beta_{\boldsymbol{\alpha}}(\mathbf{I} + \mathbf{J})\right) \\ q_{\boldsymbol{\theta}_j}(\boldsymbol{\nu}^n|\boldsymbol{\nu}^o) &= \mathcal{L}^a\left(\phi(\boldsymbol{\nu}^o), \beta_{\boldsymbol{\theta}_j}(\mathbf{I} + \mathbf{J})\right), \end{aligned}$$

where $\beta_{\boldsymbol{\alpha}}$ and $\beta_{\boldsymbol{\theta}_j}$ are scale factors that control the acceptance rates for the Metropolis–Hastings algorithm and the identity matrix $\mathbf{I}$ and the matrix $\mathbf{J}$ are of the appropriate size (see 2.6) . The terms $q_{\boldsymbol{\alpha}}(\boldsymbol{\nu}^o|\boldsymbol{\nu}^n)/q_{\boldsymbol{\alpha}}(\boldsymbol{\nu}^n|\boldsymbol{\nu}^o)$ and $q_{\boldsymbol{\theta}_j}(\boldsymbol{\nu}^o|\boldsymbol{\nu}^n)/q_{\boldsymbol{\theta}_j}(\boldsymbol{\nu}^n|\boldsymbol{\nu}^o)$ can be expressed as follows,

$$\frac{q_{\boldsymbol{\alpha}}(\boldsymbol{\nu}^o|\boldsymbol{\nu}^n)}{q_{\boldsymbol{\alpha}}(\boldsymbol{\nu}^n|\boldsymbol{\nu}^o)} = \frac{\prod_{j=1}^{p}[\boldsymbol{\nu}^n]_j}{\prod_{j=1}^{p}[\boldsymbol{\nu}^o]_j}$$

and

$$\frac{q_{\boldsymbol{\theta}_j}(\boldsymbol{\nu}^o|\boldsymbol{\nu}^n)}{q_{\boldsymbol{\theta}_j}(\boldsymbol{\nu}^n|\boldsymbol{\nu}^o)} = \frac{\prod_{j=1}^{a}[\boldsymbol{\nu}^n]_j}{\prod_{j=1}^{a}[\boldsymbol{\nu}^o]_j}.$$

The acceptance probabilities $\rho_{\boldsymbol{\alpha}}(\boldsymbol{\nu}^o, \boldsymbol{\nu}^n)$ and $\rho_{\boldsymbol{\theta}_j}(\boldsymbol{\nu}^o, \boldsymbol{\nu}^n)$ are given by

$$\rho_{\boldsymbol{\alpha}}(\boldsymbol{\nu}^o, \boldsymbol{\nu}^n) = \min\left\{\frac{f_{\boldsymbol{\alpha}}(\boldsymbol{\nu}^n)\prod_{j=1}^{p}[\boldsymbol{\nu}^n]_j}{f_{\boldsymbol{\alpha}}(\boldsymbol{\nu}^o)\prod_{j=1}^{p}[\boldsymbol{\nu}^o]_j}, 1\right\}$$

and

$$\rho_{\boldsymbol{\theta}_j}(\boldsymbol{\nu}^o, \boldsymbol{\nu}^n) = \min\left\{\frac{f_{\boldsymbol{\theta}_j}(\boldsymbol{\nu}^n)\prod_{j=1}^{a}[\boldsymbol{\nu}^n]_j}{f_{\boldsymbol{\theta}_j}(\boldsymbol{\nu}^o)\prod_{j=1}^{a}[\boldsymbol{\nu}^o]_j}, 1\right\}$$

To tune the reversible systematic scan Metropolis–within–Gibbs algorithm we have $p+1$ scale parameters to vary ($\beta_{\boldsymbol{\alpha}}$ and $\beta_{\boldsymbol{\theta}_j}, j = 1, \ldots, p$). Ideally, these parameters are chosen to give acceptance rates in the $20 - 40\%$ range. In practice this can be difficult as the number of parameters requiring Metropolis–Hastings updates increases. As we have seen in chapter 3, adaptive algorithms offer a solution to this tuning problem. We discuss this further for the more complicated models to come.

6.3.1 Observations on the Sources

The data set that Billheimer (2001) used does not have any observations on the chemical sources, that is, woodsmoke and automobile emissions. Rather, Billheimer (2001) uses informative priors on the source profiles. In the diet estimation problem observations on the fatty acid profiles of the potential prey items is available. We develop a model to allow for this new source of information.

The notation for the p prey types is slightly more complicated. We let $\mathbf{x}_{jk}$ represent the fatty acid signature for the kth sample of the jth prey type and note that for each prey type j $\mathbf{x}_{jk} \in \mathcal{S}^a$. Let $\mathbf{X}_j$ denote the matrix of n_j samples of the jth prey type.

Consider the following model

$$\begin{aligned} \mathbf{y}_i &= \boldsymbol{\Theta}\boldsymbol{\alpha} \oplus \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n, \\ \mathbf{x}_{jk} &= \boldsymbol{\theta}_j \oplus \boldsymbol{\epsilon}_{jk}^{\mathbf{x}}, \quad j = 1, \ldots, p, \; k = 1, \ldots, n_j. \end{aligned}$$

where $\boldsymbol{\theta}_j$ is the fatty acid profile for the jth prey type, $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1| \ldots |\boldsymbol{\theta}_p]$ is an $a \times p$ matrix and $\boldsymbol{\epsilon}_{jk}^{\mathbf{x}}$ is an a–dimensional compositional vector of errors. We can also write this model in structural equation or matrix notation as follows:

$$
\begin{aligned}
\underset{(a\times n)}{\mathbf{Y}} &= \underset{(a\times p)}{\boldsymbol{\Theta}}\underset{(p\times n)}{\Gamma} \oplus_c \underset{(a\times n)}{\mathbf{E}}, \\
\underset{(p\times n)}{\Gamma} &= \boldsymbol{\phi}_c^{-1}\left(\underset{((p-1)\times 1)}{\phi(\boldsymbol{\alpha})}\underset{(1\times n)}{\mathbf{W}}\right), \\
\underset{(a\times n_j)}{\mathbf{X}_j} &= \underset{(a\times 1)}{\boldsymbol{\theta}_j}\underset{(1\times n_j)}{\mathbf{W}_{\mathbf{X}_j}} \oplus_c \underset{(a\times n_j)}{\mathbf{E}_{\mathbf{x}_j}}, \quad j = 1, \ldots, p,
\end{aligned}
$$

where $\mathbf{X}_j = [\mathbf{x}_{j1}|\ldots|\mathbf{x}_{jn_j}]$ is an $a \times n_j$ matrix with the individual samples of the jth source forming the columns, $\mathbf{W}_j$ is an $1 \times n_j$ matrix of ones and $\mathbf{E}_{\mathbf{x}_j}$ is defined analogously to $\mathbf{X}_j$. Note we do not allow general design matrices for the prey sources in our formulation but they could easily be accommodated. However, we do not pursue this avenue.

We assign the following prior distributions

$$
\begin{aligned}
\pi(\boldsymbol{\theta}_j|\boldsymbol{\mu}_{\theta_j}, \Sigma_{\theta_j}) &\sim \mathcal{L}^a(\boldsymbol{\theta}_j|\boldsymbol{\mu}_{\theta_j}, \Sigma_{\theta_j}), \;\; j = 1, \ldots, p \\
\pi(\boldsymbol{\alpha}|\boldsymbol{\mu}_{\boldsymbol{\alpha}}, \Sigma_{\boldsymbol{\alpha}}) &\sim \mathcal{L}^p(\boldsymbol{\alpha}|\boldsymbol{\mu}_{\boldsymbol{\alpha}}, \Sigma_{\boldsymbol{\alpha}}) \\
\pi(\Sigma_{\mathbf{x}_j}|\delta_{\mathbf{x}_j}, \Psi_{\mathbf{x}_j}) &\sim \mathcal{IW}^{a-1}(\Sigma_{\mathbf{x}_j}|\delta_{\mathbf{x}_j}, \Psi_{\mathbf{x}_j}), \;\; j = 1, \ldots, p \\
\pi(\Sigma_{\boldsymbol{\epsilon}}|\delta_{\boldsymbol{\epsilon}}, \Psi_{\boldsymbol{\epsilon}}) &\sim \mathcal{IW}^{a-1}(\Sigma_{\boldsymbol{\epsilon}}|\delta_{\boldsymbol{\epsilon}}, \Psi_{\boldsymbol{\epsilon}})
\end{aligned}
$$

and the following sampling distributions

$$
\begin{aligned}
\pi(\mathbf{x}_{jk}|\boldsymbol{\theta}_j, \Sigma_{\mathbf{x}_j}) &\sim \mathcal{L}^a(\phi(\boldsymbol{\theta}_j), \Sigma_{\mathbf{x}_j}), \;\; j = 1, \ldots, p; \;\; k = 1, \ldots, n_j \\
\pi(\mathbf{y}_i|\boldsymbol{\Theta}, \boldsymbol{\alpha}, \Sigma_{\boldsymbol{\epsilon}}) &\sim \mathcal{L}^a(\phi(\boldsymbol{\Theta}\boldsymbol{\alpha}), \Sigma_{\boldsymbol{\epsilon}}), \;\; i = 1, \ldots, n.
\end{aligned}
$$

The joint prior distribution for $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p, \boldsymbol{\alpha}, \Sigma_{\boldsymbol{\epsilon}}, \Sigma_{\mathbf{x}_1}, \ldots, \Sigma_{\mathbf{x}_p}$ has the following form

$$
\pi(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p, \boldsymbol{\alpha}, \Sigma_{\boldsymbol{\epsilon}}, \Sigma_{\mathbf{x}_1}, \ldots, \Sigma_{\mathbf{x}_p}) = \prod_{j=1}^{p} \pi(\boldsymbol{\theta}_j) \times \pi(\boldsymbol{\alpha}) \times \pi(\Sigma_{\boldsymbol{\epsilon}}) \times \prod_{j=1}^{p} \pi(\Sigma_{\mathbf{x}_j}).
$$

Each component of the model is logically independent of the other components a priori.

The DAG for this model is given in Figure 6.2. The local character of the DAG is very similar in nature to the previous model and most of the Gibbs updates will be essentially the same as the previous section, with the exception of the $\boldsymbol{\theta}_j$'s and of course the additional covariance matrices $\Sigma_{\mathbf{x}_j}$'s. We omit the details of the Metropolis–within–Gibbs sampler for this model as it is a special case of the more complicated models to come and isn't particularly illuminating.

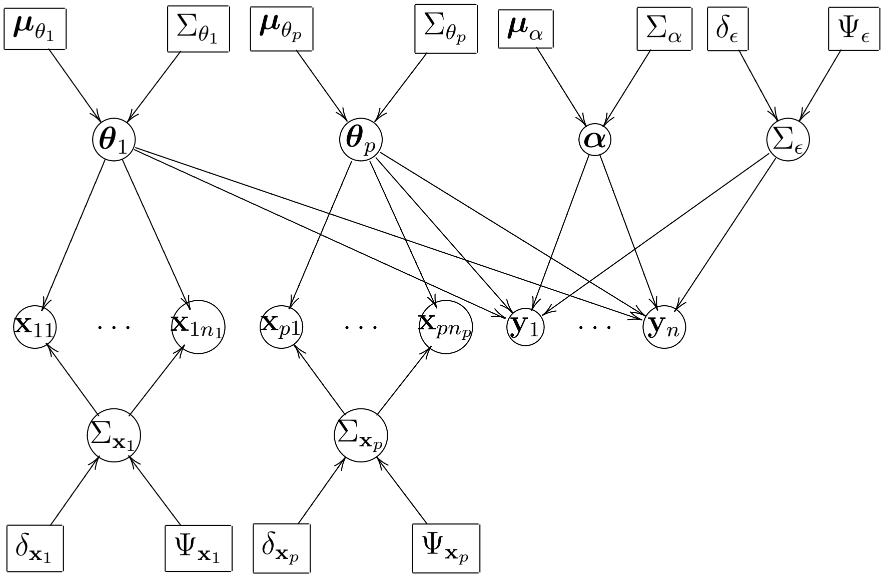

Figure 6.2: Directed Acyclic Graph (DAG) for the base model with observations on the prey types. The square nodes indicate parameters that are known a priori, while circular nodes represent unknown parameters that are updated when the data, $\mathbf{y}_i, i = 1, \ldots, n$, are observed.

6.3.2 Fat Content and Predator Biosynthesis

Long term diet studies (see Iverson et al. (2004) and the references therein) cast doubt on a key biological assumption the previous models have made. The previous models assumed, essentially that you are what you eat, that is, they assumed that the fat a predator consumed from its diet was deposited directly in its adipose tissue (fat stores). The studies reported in Iverson et al. (2004) consisted of long term feeding experiments where groups of predators were fed a known diet for extended periods (months or representing complete fattening). Comparing the fatty acid profiles of the predator to that of the known diet showed that, the fatty acid profile in the predator differed predictably from the fatty acid profile of consumed diet due to some effects of metabolism. Iverson et al. (2004) developed so called "calibration coefficients" to mimic predator biosynthesis. The model we develop in this section includes predator metabolism and its associated uncertainty.

The previous model also assumed that kilogram of herring has the same amount of fat as a kilogram of cod, however, this is not the case as was seen in section 6.2. Assume we know the true diet composition, denoted by $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_p)$ and let $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)$ be the corresponding percentage fat. The amount of fat deposited in the predator from the jth prey type is $\tau_j \lambda_j$ $j = 1, \ldots, p$ or more compactly $\boldsymbol{\tau} \oplus \boldsymbol{\xi}$. That is, if a predator consumes 1 kilogram of a high fat species, they have consumed proportionally more of those fatty acids compared to a kilogram of a low fat species.

The $\boldsymbol{\alpha}$ considered in the previous models corresponds to the diet based on fatty acid signatures and not the actual diet that was consumed by the predator. This can be related to the true diet by way of the compositional closure operator as follows

$$\boldsymbol{\alpha} = \boldsymbol{\tau} \oplus \boldsymbol{\lambda} = \frac{\tau_j \lambda_j}{\sum_{j=1}^{p} \tau_j \lambda_j}.$$

We can relate the actual diet to the signature diet by inverting the above equation as follows

$$\boldsymbol{\tau} = \boldsymbol{\alpha} \ominus \boldsymbol{\lambda} = \frac{\alpha_j / \lambda_j}{\sum_{j=1}^{p} \alpha_j / \lambda_j}.$$

The compositional, $\boldsymbol{\alpha}$, considered in the previous model, can now be seen to represent the mixing vector for the fatty acid signatures. That is, to reconstruct the diet composition vector, $\boldsymbol{\tau}$, we need to consider the percentage fat of the consumed prey. The approach we take models the true diet $\boldsymbol{\tau}$ and the fat content and its inherent uncertainty.

The model developed in this section takes both fat content and predator biosynthesis into account.

Before proceeding with the model we need some additional notation, let $\mathbf{u}_l, l = 1, \ldots, L$ be the fatty acid profiles of the prey used in the calibration experiment and similarly let $\mathbf{v}_m, m = 1, \ldots, M$ be the fatty acid profiles of the predators used in the calibration experiment. Let $\mathbf{z}_{jk}$ be the fat composition vector for the kth sample of the jth predator. The fat composition vector is of length 2, as we classify the prey as percentage fat and non–fat.

Consider the following model

$$\begin{aligned}
\mathbf{y}_i &= (\Theta\boldsymbol{\gamma}) \oplus (\boldsymbol{\theta}_{\mathbf{v}} \ominus \boldsymbol{\theta}_{\mathbf{u}}) \oplus \boldsymbol{\epsilon}_i, \;\; i = 1, \ldots, n, \\
\boldsymbol{\gamma} &= \boldsymbol{\tau} \oplus \boldsymbol{\lambda} \\
\mathbf{x}_{jk} &= \boldsymbol{\theta}_j \oplus \boldsymbol{\epsilon}^{\mathbf{x}}_{jk}, \;\; j = 1, \ldots, p, \; k = 1, \ldots, n_j, \\
\mathbf{z}_{jk} &= \boldsymbol{\lambda}^v_j \oplus \boldsymbol{\epsilon}^{\mathbf{z}}_{jk}, \;\; j = 1, \ldots, p, \; k = 1, \ldots, n_j, \\
\mathbf{u}_l &= \boldsymbol{\theta}_{\mathbf{u}} \oplus \boldsymbol{\epsilon}^{\mathbf{u}}_l, \;\; l = 1, \ldots, L \\
\mathbf{v}_m &= \boldsymbol{\theta}_{\mathbf{v}} \oplus \boldsymbol{\epsilon}^{\mathbf{v}}_m, \;\; m = 1, \ldots, M
\end{aligned}$$

where $\boldsymbol{\theta}_j$ is the fatty acid profile of the jth prey type, $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1| \ldots |\boldsymbol{\theta}_p]$ is an $a \times p$ matrix, $\boldsymbol{\theta}_{\mathbf{u}}$ is the fatty acid profile of the calibration prey, $\boldsymbol{\theta}_{\mathbf{v}}$ is the fatty acid profile of the calibration predator, $\boldsymbol{\lambda}^v_j = (\lambda_j, 1 - \lambda_j)'$ is the fat content profile consisting of the proportion of fat for the jth prey type and the the proportion of non–fat and $\boldsymbol{\lambda}$ is the vector of percentage fats, that is $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)'$. Note $\boldsymbol{\lambda}$ is not a composition as its components don't necessarily sum to 1, however, it is a positive vector and can be used to perturb compositions (see 2).

The matrix version of the model is given as follows:

$$\begin{array}{rcl}
\underset{(a\times n)}{\mathbf{Y}} & = & \left[\left(\underset{(a\times p)}{\Theta}\underset{(p\times n)}{\Gamma}\right)\oplus_c\left(\underset{(a\times 1)}{\boldsymbol{\theta}_{\mathbf{v}}}\ominus\underset{(a\times 1)}{\boldsymbol{\theta}_{\mathbf{u}}}\right)\right]\oplus_c\underset{(a\times n)}{\mathbf{E}} \\
\underset{(p\times n)}{\Gamma} & = & \boldsymbol{\phi}_c^{-1}\left(\boldsymbol{\phi}(\underset{(p\times 1)}{\boldsymbol{\tau}}\oplus\underset{(p\times 1)}{\boldsymbol{\lambda}})\underset{(1\times n)}{\mathbf{W}}\right) \\
\underset{(a\times n_j)}{\mathbf{X}_j} & = & \underset{(a\times 1)}{\boldsymbol{\theta}_j}\underset{(1\times n_j)}{\mathbf{W}_{\mathbf{X}_j}}\oplus_c\underset{(a\times n_j)}{\mathbf{E}_{\mathbf{X}_j}},\ \ j=1,\ldots,p, \\
\underset{(2\times n_j)}{\mathbf{Z}_j} & = & \underset{(2\times 1)}{\boldsymbol{\lambda}_j^v}\underset{(1\times n_j)}{\mathbf{W}_{\mathbf{Z}_j}}\oplus_c\underset{(a\times n_j)}{\mathbf{E}_{\mathbf{Z}_j}},\ \ j=1,\ldots,p, \\
\underset{(a\times L)}{\mathbf{U}} & = & \underset{(a\times 1)}{\boldsymbol{\theta}_{\mathbf{u}}}\underset{(1\times L)}{\mathbf{W}_{\mathbf{U}}}\oplus_c\underset{(a\times L)}{\mathbf{E}_{\mathbf{U}}} \\
\underset{(a\times M)}{\mathbf{V}} & = & \underset{(a\times 1)}{\boldsymbol{\theta}_{\mathbf{v}}}\underset{(1\times M)}{\mathbf{W}_{\mathbf{V}}}\oplus_c\underset{(a\times M)}{\mathbf{E}_{\mathbf{V}}}
\end{array}$$

where $\oplus_c$ is the column–wise perturbation operator and the matrices are formed as in previous sections, that is, they are formed by column–concatenation.

We assign the following prior distributions for the location parameters

$$\begin{array}{rcl}
\pi(\boldsymbol{\tau}_r|\boldsymbol{\mu}_{\boldsymbol{\tau}},\Sigma_{\boldsymbol{\tau}}) & \sim & \mathcal{L}^p(\boldsymbol{\mu}_{\boldsymbol{\tau}},\Sigma_{\boldsymbol{\tau}}),\ \ r=1,\ldots,p, \\
\pi(\boldsymbol{\theta}_j|\boldsymbol{\mu}_{\theta_j},\Sigma_{\theta_j}) & \sim & \mathcal{L}^a(\boldsymbol{\mu}_{\theta_j},\Sigma_{\theta_j}),\ \ j=1,\ldots,p, \\
\pi(\boldsymbol{\lambda}_j^v|\mu_{\lambda_j},\Sigma_{\lambda_j}) & \sim & \mathcal{L}^2(\boldsymbol{\mu}_{\lambda_j},\Sigma_{\lambda_j}),\ \ j=1,\ldots,p, \\
\pi(\boldsymbol{\theta}_{\mathbf{u}}|\boldsymbol{\mu}_{\theta_{\mathbf{u}}},\Sigma_{\theta_{\mathbf{u}}}) & \sim & \mathcal{L}^a(\boldsymbol{\mu}_{\theta_{\mathbf{u}}},\Sigma_{\theta_{\mathbf{u}}}), \\
\pi(\boldsymbol{\theta}_{\mathbf{v}}|\boldsymbol{\mu}_{\theta_{\mathbf{v}}},\Sigma_{\theta_{\mathbf{v}}}) & \sim & \mathcal{L}^a(\boldsymbol{\mu}_{\theta_{\mathbf{v}}},\Sigma_{\theta_{\mathbf{v}}}),
\end{array}$$

and for the covariance matrices

$$\begin{array}{rcl}
\pi(\Sigma_{\boldsymbol{\epsilon}}|\delta_{\boldsymbol{\epsilon}},\Psi_{\boldsymbol{\epsilon}}) & \sim & \mathcal{IW}^{a-1}(\delta_{\boldsymbol{\epsilon}},\Psi_{\boldsymbol{\epsilon}}), \\
\pi(\Sigma_{\mathbf{x}_j}|\delta_{\mathbf{x}_j},\Psi_{\mathbf{x}_j}) & \sim & \mathcal{IW}^{a-1}(\delta_{\mathbf{x}_j},\Psi_{\mathbf{x}_j})\ \ j=1,\ldots,p, \\
\pi(\Sigma_{\mathbf{z}_j}|\delta_{\mathbf{z}_j},\Psi_{\mathbf{z}_j}) & \sim & \mathcal{IW}^{1}(\delta_{\mathbf{z}_j},\Psi_{\mathbf{z}_j}),\ \ j=1,\ldots,p, \\
\pi(\Sigma_{\mathbf{u}}|\delta_{\mathbf{u}},\Psi_{\mathbf{u}}) & \sim & \mathcal{IW}^{a-1}(\delta_{\mathbf{u}},\Psi_{\mathbf{u}}), \\
\pi(\Sigma_{\mathbf{v}}|\delta_{\mathbf{v}},\Psi_{\mathbf{v}}) & \sim & \mathcal{IW}^{a-1}(\delta_{\mathbf{v}},\Psi_{\mathbf{v}}).
\end{array}$$

The sampling distributions are given by

$$\begin{aligned}
\pi(\mathbf{y}_1, \ldots, \mathbf{y}_n | \boldsymbol{\Theta}, \boldsymbol{\tau}, \Sigma_{\boldsymbol{\epsilon}}, \boldsymbol{\lambda}, \boldsymbol{\theta}_{\mathbf{u}}, \boldsymbol{\theta}_{\mathbf{v}}) &\sim \mathcal{L}^a(\phi\left(\Theta\Gamma^i \oplus_c (\boldsymbol{\theta}_{\mathbf{v}} \ominus \boldsymbol{\theta}_{\mathbf{u}})\right), \Sigma_{\boldsymbol{\epsilon}}), \quad i = 1, \ldots, n \\
\pi(\mathbf{x}_{jk} | \boldsymbol{\theta}_j, \Sigma_{\mathbf{x}_j}) &\sim \mathcal{L}^a(\boldsymbol{\phi}(\boldsymbol{\theta}_j), \Sigma_{\mathbf{x}_j}), \quad j = 1, \ldots, p, \quad k = 1, \ldots, n_j, \\
\pi(\mathbf{z}_{jk} | \boldsymbol{\lambda}_j^v, \Sigma_{\mathbf{z}_j}) &\sim \mathcal{L}^2(\boldsymbol{\phi}(\lambda_j^v), \Sigma_{\mathbf{z}_j}), \quad j = 1, \ldots, p, \quad k = 1, \ldots, n_j, \\
\pi(\mathbf{u}_l | \boldsymbol{\theta}_{\mathbf{u}}, \Sigma_{\mathbf{u}}) &\sim \mathcal{L}^a(\boldsymbol{\phi}(\boldsymbol{\theta}_{\mathbf{u}}), \Sigma_{\mathbf{u}}), \quad l = 1, \ldots, L, \\
\pi(\mathbf{v}_m | \boldsymbol{\theta}_{\mathbf{v}}, \Sigma_{\mathbf{v}}) &\sim \mathcal{L}^a(\boldsymbol{\phi}(\boldsymbol{\theta}_{\mathbf{v}}), \Sigma_{\mathbf{v}}), \quad m = 1, \ldots, M,
\end{aligned}$$

where the notation $\mathbf{A}^i$ means the ith column of the matrix $\mathbf{A}$.

Figure 6.3 gives the DAG representation of this model. The full conditionals for this model are special cases of the design matrix version of the constant diet model discussed in the next section and the full conditionals will be presented for that model in Appendix B.

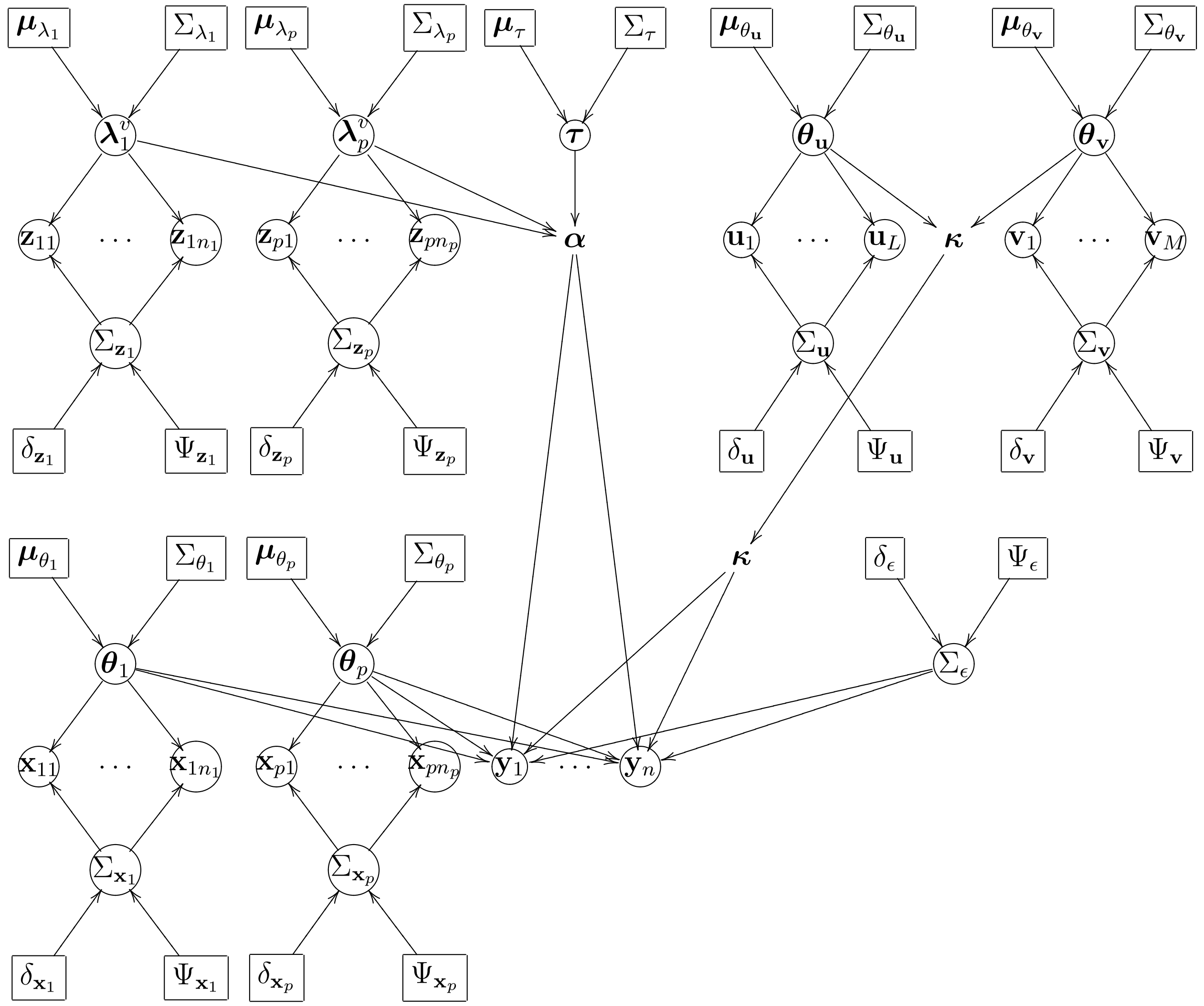

Figure 6.3: DAG: Constant Diet with calibration and fat content. Note $\boldsymbol{\kappa} = \boldsymbol{\theta}_\mathbf{v} \oplus \boldsymbol{\theta}_\mathbf{u}$, $\boldsymbol{\alpha} = \boldsymbol{\tau} \oplus \boldsymbol{\lambda}$ and $\mathbf{Y} = [(\Theta\Gamma) \oplus_c (\boldsymbol{\theta}_\mathbf{v} \ominus \boldsymbol{\theta}_\mathbf{u})] \oplus_c \mathbf{E}$. Nodes that are not contained in circles or squares are derived variables are used to simplify the graph.

6.3.3 Design Matrices

All the models considered thus far have dealt with a single population of predators, with a common diet composition $\boldsymbol{\tau}$. This section considers generalizing this to several populations and the possible presence of a continuous covariate. However, for our purposes, we restrict our attention to categorical variables, which divide the predators into distinct groups or populations. We construct synthetic data with three seasons (Spring, Summer, Fall/Winter) and sex to illustrate. The illustrations of the model will only have at most one factor.

The generalization is based on models developed Billheimer (2001) (briefly described in 2.6.1) which in turn is a generalization of some earlier work described in Aitchison (2003).

Consider the following model:

$$
\begin{aligned}
\underset{(a\times 1)}{\mathbf{y}_i} &= \underset{(a\times p)}{\boldsymbol{\Theta}}\underset{(p\times 1)}{\Gamma^i} \oplus (\underset{(a\times 1)}{\boldsymbol{\theta}_{\mathbf{v}}} \ominus \underset{(a\times 1)}{\boldsymbol{\theta}_{\mathbf{u}}}) \oplus \underset{(a\times 1)}{\boldsymbol{\epsilon}_i}, \quad i = 1, \ldots, n, \\
\underset{(p\times n)}{\Gamma} &= \boldsymbol{\phi}_c^{-1}\left(\boldsymbol{\phi}_c\left(\underset{(p\times w)}{\mathbf{T}}\right)\underset{(w\times n)}{\mathbf{W}}\right) \oplus_c \underset{(p\times 1)}{\boldsymbol{\lambda}} \\
\underset{(a\times 1)}{\mathbf{x}_{jk}} &= \underset{(a\times 1)}{\boldsymbol{\theta}_j} \oplus \underset{(a\times 1)}{\boldsymbol{\epsilon}^{\mathbf{x}}_{jk}}, \quad j = 1, \ldots, p, \ k = 1, \ldots, n_j, \\
\underset{(2\times 1)}{\mathbf{z}_{jk}} &= \underset{(2\times 1)}{\boldsymbol{\lambda}^v_j} \oplus \underset{(2\times 1)}{\boldsymbol{\epsilon}^{\mathbf{z}}_{jk}}, \quad j = 1, \ldots, p, \ k = 1, \ldots, n_j, \\
\underset{(a\times 1)}{\mathbf{u}_l} &= \underset{(a\times 1)}{\boldsymbol{\theta}_{\mathbf{u}}} \oplus \underset{(a\times 1)}{\boldsymbol{\epsilon}^{\mathbf{u}}_l}, \quad l = 1, \ldots, L, \\
\underset{(a\times 1)}{\mathbf{v}_m} &= \underset{(a\times 1)}{\boldsymbol{\theta}_{\mathbf{v}}} \oplus \underset{(a\times 1)}{\boldsymbol{\epsilon}^{\mathbf{v}}_m}, \quad m = 1, \ldots, M,
\end{aligned}
$$

where $\mathbf{W}$ is an $w \times n$ known design matrix, Γ^i indicates the ith column of the matrix Γ and $\mathbf{T} = [\boldsymbol{\tau}^p_1| \ldots |\boldsymbol{\tau}^p_w]$ is a $p \times w$ matrix of population diet compositions.

Expressing this model in matrix notation is relatively straightforward as follows:

$$\begin{aligned}
\underset{(a\times n)}{\mathbf{Y}} &= \left[\left(\underset{(a\times p)}{\Theta}\underset{(p\times n)}{\Gamma}\right)\oplus_c\left(\underset{(a\times 1)}{\boldsymbol{\theta}_{\mathbf{v}}}\ominus\underset{(a\times 1)}{\boldsymbol{\theta}_{\mathbf{u}}}\right)\right]\oplus_c\underset{(a\times n)}{\mathbf{E}}\\
\underset{(p\times n)}{\Gamma} &= \boldsymbol{\phi}_c^{-1}\left(\underset{((p-1)\times w)}{\mathbf{T}}\underset{(w\times n)}{\mathbf{W}}\right)\oplus_c\underset{(p\times 1)}{\boldsymbol{\lambda}}\\
\underset{(a\times n_j)}{\mathbf{X}_j} &= \underset{(a\times 1)}{\boldsymbol{\theta}_j}\underset{(1\times n_j)}{\mathbf{W}_{\mathbf{X}_j}}\oplus_c\underset{(a\times n_j)}{\mathbf{E}_{\mathbf{X}_j}},\ \ j=1,\dots,p,\\
\underset{(2\times n_j)}{\mathbf{Z}_j} &= \underset{(2\times 1)}{\boldsymbol{\lambda}_j^v}\underset{(1\times n_j)}{\mathbf{W}_{\mathbf{Z}_j}}\oplus_c\underset{(a\times n_j)}{\mathbf{E}_{\mathbf{Z}_j}},\ \ j=1,\dots,p,\\
\underset{(a\times L)}{\mathbf{U}} &= \underset{(a\times 1)}{\boldsymbol{\theta}_{\mathbf{u}}}\underset{(1\times L)}{\mathbf{W}_{\mathbf{U}}}\oplus_c\underset{(a\times L)}{\mathbf{E}_{\mathbf{U}}}\\
\underset{(a\times M)}{\mathbf{V}} &= \underset{(a\times 1)}{\boldsymbol{\theta}_{\mathbf{v}}}\underset{(1\times M)}{\mathbf{W}_{\mathbf{V}}}\oplus_c\underset{(a\times M)}{\mathbf{E}_{\mathbf{V}}}
\end{aligned}$$

We assign the following prior distributions for the location parameters

$$\begin{aligned}
\pi(\boldsymbol{\tau}_s|\boldsymbol{\mu}_{\boldsymbol{\tau}},\Sigma_{\boldsymbol{\tau}}) &\sim \mathcal{L}^p(\boldsymbol{\mu}_{\boldsymbol{\tau}_s},\Sigma_{\boldsymbol{\tau}_s}),\ \ s=1,\dots,w,\\
\pi(\boldsymbol{\theta}_j|\boldsymbol{\mu}_{\theta_j},\Sigma_{\theta_j}) &\sim \mathcal{L}^a(\boldsymbol{\mu}_{\theta_j},\Sigma_{\theta_j}),\ \ j=1,\dots,p,\\
\pi(\boldsymbol{\lambda}_j^v|\mu_{\lambda_j},\Sigma_{\lambda_j}) &\sim \mathcal{L}^2(\boldsymbol{\mu}_{\lambda_j},\Sigma_{\lambda_j}),\ \ j=1,\dots,p,\\
\pi(\boldsymbol{\theta}_{\mathbf{u}}|\boldsymbol{\mu}_{\theta_{\mathbf{u}}},\Sigma_{\theta_{\mathbf{u}}}) &\sim \mathcal{L}^a(\boldsymbol{\mu}_{\theta_{\mathbf{u}}},\Sigma_{\theta_{\mathbf{u}}}),\\
\pi(\boldsymbol{\theta}_{\mathbf{v}}|\boldsymbol{\mu}_{\theta_{\mathbf{v}}},\Sigma_{\theta_{\mathbf{v}}}) &\sim \mathcal{L}^a(\boldsymbol{\mu}_{\theta_{\mathbf{v}}},\Sigma_{\theta_{\mathbf{v}}}),
\end{aligned}$$

and for the covariance matrices

$$\begin{aligned}
\pi(\Sigma_{\boldsymbol{\epsilon}}|\delta_{\boldsymbol{\epsilon}},\Psi_{\boldsymbol{\epsilon}}) &\sim \mathcal{IW}^{a-1}(\delta_{\boldsymbol{\epsilon}},\Psi_{\boldsymbol{\epsilon}}),\\
\pi(\Sigma_{\mathbf{x}_j}|\delta_{\mathbf{x}_j},\Psi_{\mathbf{x}_j}) &\sim \mathcal{IW}^{a-1}(\delta_{\mathbf{x}_j},\Psi_{\mathbf{x}_j}),\ \ j=1,\dots,p,\\
\pi(\Sigma_{\mathbf{z}_j}|\delta_{\mathbf{z}_j},\Psi_{\mathbf{z}_j}) &\sim \mathcal{IW}^{1}(\delta_{\mathbf{z}_j},\Psi_{\mathbf{z}_j}),\ \ j=1,\dots,p,\\
\pi(\Sigma_{\mathbf{u}}|\delta_{\mathbf{u}},\Psi_{\mathbf{u}}) &\sim \mathcal{IW}^{a-1}(\delta_{\mathbf{u}},\Psi_{\mathbf{u}}),\\
\pi(\Sigma_{\mathbf{v}}|\delta_{\mathbf{v}},\Psi_{\mathbf{v}}) &\sim \mathcal{IW}^{a-1}(\delta_{\mathbf{v}},\Psi_{\mathbf{v}}).
\end{aligned}$$

The sampling distributions are given by

$$\begin{aligned}
\pi(\mathbf{y}_i|\boldsymbol{\Theta},\mathbf{T},\Sigma_{\boldsymbol{\epsilon}},\boldsymbol{\lambda},\boldsymbol{\theta}_{\mathbf{u}},\boldsymbol{\theta}_{\mathbf{v}}) &\sim \mathcal{L}^a(\boldsymbol{\phi}(\boldsymbol{\Theta}\Gamma^i \oplus (\boldsymbol{\theta}_{\mathbf{v}} \ominus \boldsymbol{\theta}_{\mathbf{u}})),\Sigma_{\boldsymbol{\epsilon}}), \;\; i=1,\ldots,n \\
\pi(\mathbf{x}_{jk}|\boldsymbol{\theta}_j,\Sigma_{\mathbf{x}_j}) &\sim \mathcal{L}^a(\boldsymbol{\phi}(\boldsymbol{\theta}_j),\Sigma_{\mathbf{x}_j}), \;\; j=1,\ldots,p; \;\; k=1,\ldots,n_j \\
\pi(\mathbf{u}_l|\boldsymbol{\theta}_{\mathbf{u}},\Sigma_{\mathbf{u}}) &\sim \mathcal{L}^a(\boldsymbol{\phi}(\boldsymbol{\theta}_{\mathbf{u}}),\Sigma_{\mathbf{u}}), \;\; l=1,\ldots,L; \\
\pi(\mathbf{v}_m|\boldsymbol{\theta}_{\mathbf{v}},\Sigma_{\mathbf{v}}) &\sim \mathcal{L}^a(\boldsymbol{\phi}(\boldsymbol{\theta}_{\mathbf{v}}),\Sigma_{\mathbf{v}}), \;\; m=1,\ldots,M; \\
\pi(\mathbf{z}_{jk}|\lambda_1,\Sigma_{\mathbf{z}_j}) &\sim \mathcal{L}^2(\boldsymbol{\phi}(\lambda_j),\Sigma_{\mathbf{z}_j}), \;\; j=1,\ldots,p; \;\; k=1,\ldots,n_j
\end{aligned}$$

For example, suppose we are interested in the diets of male and female grey seals. Let n_M represent the number of males and n_F represent the number of females ($n = n_M + n_F$). In this case $w = 2$ and we would have two obvious choices for the matrix $\mathbf{W}$. The more usual dummy variable coding would consist of a column of n ones, and a column consisting of n_M zeros followed by n_F ones. The second coding scheme would correspond to the matrix $\mathbf{W}$ having the first column consist of n_M ones followed by n_F zeros and the second column consisting of n_M zeros followed by n_F ones. The second parametrization seems to be the preferable one as suggested in Gelman (2004). See Hills and Smith (1993); Gelfand et al. (1995); Roberts and Sahu (1997); Gelman et al. (2008) for a general discussion about parametrizations in a Bayesian context.

The full conditionals and reversible systematic scan Metropolis–within–Gibbs algorithm is given in Appendix B. Appendix B also shows that the full posterior distribution is proper which is a key assumption for any MCMC sampler to be valid.

6.3.4 Synthetic Data

In this section we consider two synthetic diets constructed from the most recent Scotian shelf prey base. The Scotian shelf prey library consists of 28 species of which we chose 12 to illustrate the constant diet model with multiple populations. Table 6.1 gives the common name, the scientific name, sample size and the percentage fat for the 12 prey species used in the creation of synthetic diets given in table 6.2.

Recall that two vectors $\mathbf{x}$ and $\mathbf{y}$ are perpendicular (orthogonal) if an only if the angle between them, denoted by θ, is $90°$ or $270°$ in other words if $\cos(\theta) = 0$. Conversely, the two vectors are coincident if $\cos(\theta) = 1$ corresponding to an angle of $0°$. The $\cos(\theta)$ of two

Common name	Scientific name	n	% Fat
Forage Fish			
Capelin	*Mallotus villosus*	165	4.57
Herring	*Clupea harengus*	247	3.75
Northern sand lance	*Ammodytes dubius*	124	3.71
Gadids			
Cod	*Gadus morhua*	134	2.02
Pollock	*Pollachius virens*	57	2.18
Silver Hake	*Merluccius bilinearis*	70	1.44
White Hake	*Urophycis tenuis*	75	1.18
Flounders			
American plaice	*Hippoglossoides platessoides*	148	1.76
Yellowtail Flounder	*Limanda ferruginea*	118	2.03
Skates			
Thorny Skate	*Raja radiate*	74	2.02
Other fish			
Longhorn Sculpin	*Myoxocephalus Octodecemspinosus*	70	2.34
Redfish	*Sebastes sp.*	84	4.33

Table 6.1: Prey species used in the creation of the two synthetic diets.

vectors $\mathbf{x}$ and $\mathbf{y}$ is defined as follows

$$\cos(\theta) = \frac{\mathbf{x}'\mathbf{y}}{\sqrt{\mathbf{x}'\mathbf{x}}\sqrt{\mathbf{y}'\mathbf{y}}}.$$

The usual Pearson product moment correlation can be obtained if we center the vectors $\mathbf{x}$ and $\mathbf{y}$. Thus, we use $\cos(\theta)$ as a similarity measure between the two vectors $\mathbf{x}$ and $\mathbf{y}$ and define the standardized distance between $\mathbf{x}$ and $\mathbf{y}$ as $1 - \cos(\theta)$. The reason for using the non–centered data is that the source profiles are compositions and hence must remain in the simplex.

Figure 6.4 shows the 12 prey (source) fatty acid profiles for 32 fatty acids used in constructing the synthetic diet and a hierarchical clustering of $1 - \cos(\theta)$. The condition number of the source matrix of the source matrix is 220709, which indicates a large degree of linear dependence between the columns of the source matrix. Several species (sources) are very similar to each other for example Cod and Plaice, $\cos(\theta) = 0.9975$ and Pollock and Silver Hake, $\cos(\theta) = 0.9964$. As we have seen in the previous chapter high degrees of dependence among the sources can make correct apportionment much more difficult.

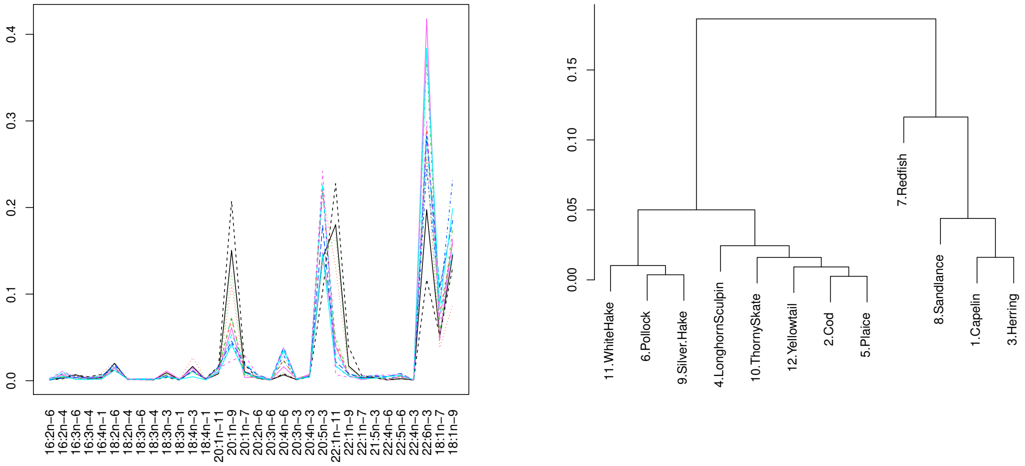

Figure 6.4: The left hand plot shows the source profile for the 12 species used in constructing the synthetic diets and the right hand plot shows a hierarchical clustering of the standardized distance using the average linkage method. The source matrix $\boldsymbol{\Theta}$ condition number is 202709.

The prey profiles shown in figure 6.4 all have similar patterns across the fatty acids which leads to the strong dependence among the species. With that in mind we considered choosing a distinct permutation of the fatty acid labels for each prey species. This led to a drastically reduced source matrix condition number, 26.909. Figure 6.5 gives the prey profiles for the permuted profiles and the corresponding hierarchical clustering. The largest degree of dependence occurs between Capelin and Silver Hake, $\cos(\theta) = 0.651$. For comparison purposes we generate synthetic data using the permuted prey to assess the effect of ill–conditioned source matrices.

Samples of fatty acid signatures and fat contents for the prey types denoted by $\mathbf{x}_{jk}$ and $\mathbf{z}_{jk}$ were generated as follows. We assume that $\mathbf{x}_{jk}$ and $\mathbf{z}_{jk}$ follow logistic normal distributions which are completely characterized by their mean and covariance matrices. We estimate

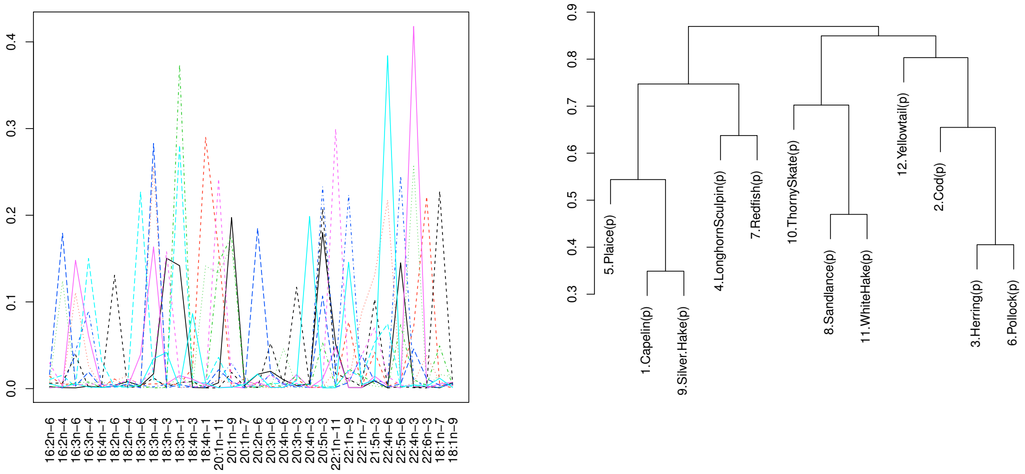

Figure 6.5: The left hand plot shows the source profile for the 12 species used in constructing the synthetic diets and the right hand plot shows a hierarchical clustering of the standardized distance using the average linkage method. The source matrix condition number is 26.909. Note for clarity we have denoted the permuted species with a “(p)” to indicate that it is a distince species from the original.

the mean and covariance matrices from the sample data and denote them as follows

$$\begin{aligned}
\widehat{\boldsymbol{\phi}(\boldsymbol{\theta}_j)} &= \frac{1}{n_j}\sum_{k=1}^{n_j}\boldsymbol{\phi}(\mathbf{x}_{jk}), \quad j=1,\ldots,p \\
\widehat{\boldsymbol{\phi}(\boldsymbol{\lambda}_j^v)} &= \frac{1}{n_j}\sum_{k=1}^{n_j}\boldsymbol{\phi}(\mathbf{z}_{jk}), \quad j=1,\ldots,p \\
\widehat{\Sigma}_{\mathbf{x}_j} &= \frac{1}{n_j}\sum_{k=1}^{n_j}\left(\boldsymbol{\phi}(\mathbf{x}_{jk})-\widehat{\boldsymbol{\phi}(\boldsymbol{\theta}_j)}\right)\left(\boldsymbol{\phi}(\mathbf{x}_{jk})-\widehat{\boldsymbol{\phi}(\boldsymbol{\theta}_j)}\right)', \quad j=1,\ldots,p \\
\widehat{\Sigma}_{\mathbf{z}_j} &= \frac{1}{n_j}\sum_{k=1}^{n_j}\left(\boldsymbol{\phi}(\mathbf{z}_{jk})-\widehat{\boldsymbol{\phi}(\boldsymbol{\lambda}_j)}\right)\left(\boldsymbol{\phi}(\mathbf{z}_{jk})-\widehat{\boldsymbol{\phi}(\boldsymbol{\lambda}_j)}\right)', \quad j=1,\ldots,p
\end{aligned}$$

The small sample sizes for the captive grey seal feeding experiment, only 8 grey seals and 30 herring, are not sufficient to estimate the covariance matrix. Therefore, we chose two species from the prey base, Haddock (M=140) to serve as the captive predator and Gaspereau (L=70) serve as the captive prey. We give a solution to this dilemma when we model the real predators. Our estimates are given by

$$\begin{aligned}
\widehat{\boldsymbol{\phi}(\boldsymbol{\theta}_{\mathbf{u}})} &= \frac{1}{L}\sum_{l=1}^{L}\boldsymbol{\phi}(\mathbf{u}_l) \\
\widehat{\boldsymbol{\phi}(\boldsymbol{\theta}_{\mathbf{v}})} &= \frac{1}{M}\sum_{m=1}^{M}\boldsymbol{\phi}(\mathbf{v}_m) \\
\widehat{\Sigma}_{\mathbf{u}} &= \frac{1}{L}\sum_{l=1}^{L}\left(\boldsymbol{\phi}(\mathbf{u}_l)-\widehat{\boldsymbol{\phi}(\boldsymbol{\theta}_{\mathbf{u}})}\right)\left(\boldsymbol{\phi}(\mathbf{u}_l)-\widehat{\boldsymbol{\phi}(\boldsymbol{\theta}_{\mathbf{u}})}\right)' \\
\widehat{\Sigma}_{\mathbf{v}} &= \frac{1}{M}\sum_{m=1}^{M}\left(\boldsymbol{\phi}(\mathbf{v}_m)-\widehat{\boldsymbol{\phi}(\boldsymbol{\theta}_{\mathbf{v}})}\right)\left(\boldsymbol{\phi}(\mathbf{v}_m)-\widehat{\boldsymbol{\phi}(\boldsymbol{\theta}_{\mathbf{v}})}\right)'
\end{aligned}$$

The compositional errors are generated from a logistic normal distribution with center zero and covariance matrix, $\Sigma_\epsilon = 0.1(\mathbf{I}+\mathbf{J})$, which corresponds to independent lognormal basis (see section 2.1). Table 6.2 gives the diet compositions used in the construction of the synthetic predator. We generated 300 predator profiles broken down according to the following table

Sex	Spring	Summer	Fall/Winter
M	50	50	50
F	50	50	50

The samples by are denoted $\mathbf{y}_i, i = 1,\ldots,300$ and were constructed as follows

$$\begin{aligned}
\underset{(a\times n)}{\mathbf{Y}} &= \left(\underset{(a\times p)}{\Theta}\underset{(p\times n)}{\Gamma}\right)\oplus_c\left(\underset{(a\times 1)}{\boldsymbol{\theta}_{\mathbf{v}}}\ominus\underset{(a\times 1)}{\boldsymbol{\theta}_{\mathbf{u}}}\right)\oplus_c\underset{(a\times n)}{\mathbf{E}} \\
\underset{(p\times n)}{\Gamma} &= \boldsymbol{\phi}_c^{-1}\left(\boldsymbol{\phi}_c\left(\underset{(p\times w)}{\mathbf{T}}\right)\underset{(w\times n)}{\mathbf{W}}\right)\oplus_c\underset{(p\times 1)}{\boldsymbol{\lambda}}
\end{aligned}$$

where $\mathbf{W}$ is a 6×300 matrix, each column has 5 zeros and 1 one corresponding to the population of interest; $\mathbf{T}$ is a 12×6 matrix whose columns correspond to the columns given in 6.2; $\boldsymbol{\lambda} = \left[\phi^{-1}(\widehat{\lambda}_1)|\ldots|\phi^{-1}(\widehat{\lambda}_{12})\right]_1$ that is, the first row of the matrix; $\boldsymbol{\Theta} = [\phi^{-1}(\widehat{\boldsymbol{\theta}}_1)|\ldots|\phi^{-1}(\widehat{\boldsymbol{\theta}}_{12})]$, $\boldsymbol{\theta}_{\mathbf{v}} = \phi^{-1}(\widehat{\boldsymbol{\theta}}_{\mathbf{v}})$ and $\boldsymbol{\theta}_{\mathbf{u}} = \phi^{-1}(\widehat{\boldsymbol{\theta}}_{\mathbf{u}})$.

Samples for $\mathbf{x}_{jk}, j = 1,\ldots,p, k = 1,\ldots,n_j$, $\mathbf{z}_{jk}, j = 1,\ldots,p, k = 1,\ldots,n_j$, $\mathbf{u}_l, l = 1,\ldots,L$ and $\mathbf{v}_m, m = 1,\ldots,M$ were generated from the appropriate logistic normal

	Spring		Summer		Winter	
Species	M	F	M	F	M	F
Capelin	0.3	3.0	1.5	0.005	3.0	1.8
Cod	3.0	1.0	5.5	4.0	5.0	1.5
Herring	5.0	6.0	4.0	2.0	4.0	8.0
Longhorn Sculpin	0.5	0.1	1.5	1.50	0.005	0.185
Plaice	5.0	1.5	3.0	0.50	1.0	0.005
Pollock	30.0	5.5	21.0	0.02	11.0	5.00
Redfish	33.0	35.0	33.0	23.0	35.0	38.0
Sandlance	7.0	41.0	22.0	67.0	32.0	43.0
Silver Hake	2.0	1.4	0.5	0.5	0.9	0.005
Thorny Skate	5.0	0.5	2.0	0.005	5.00	0.50
White Hake	6.0	1.0	4.0	0.50	0.095	0.005
Yellowtail	3.2	4.0	2.0	0.97	3.00	2.00

Table 6.2: Synthetic diet 1 compositions (in percent) for the 12 species and 6 populations (3 seasons times 2 sexes) used in to assess the constant diet model.

distribution using the estimated center and covariance matrix.

The original sample data serves as the population values from which we generate sample data using the sample estimates as the true value of the parameters. This is similar in spirit to the classical method of the parametric bootstrap to assess variability.

6.3.5 Choice of Prior Variances for Logistic Normal Distributions

Given the brief discussion in the previous about prior variances and practical considerations for the parameters of linear convex mixing models we consider the choice of prior variance more formally here. The discussion concluded that approximately plus or minus ten on the log–ratio scale is practically sufficient given how close these values are to one and zero respectively when viewed on the original scale via the inverse logistic transformation, denoted by ϕ^{-1}. Firstly, recall the functional form of the p–dimensional logistic normal distribution, denoted by $\mathcal{L}^p(\boldsymbol{\mu}, \Sigma)$

$$\pi(\mathbf{z}|\boldsymbol{\mu}, \Sigma) \propto |\Sigma|^{-1/2} \left(\prod_{i=1}^{p} z_i \right)^{-1} \exp\left\{ -\frac{1}{2}(\boldsymbol{\phi}(\mathbf{z}) - \boldsymbol{\mu})' \Sigma^{-1} (\boldsymbol{\phi}(\mathbf{z}) - \boldsymbol{\mu}) \right\}$$

where $\boldsymbol{\mu}$ is the location vector of length $p-1$ and Σ is the covariance matrix of dimension $(p-1) \times (p-1)$. This functional form is very closely related to the multivariate normal distribution and for our purposes we choose to exploit one of the well known properties of the multivariate normal distribution namely the distributional property of the quadratic form in the exponent. Specifically, a $(1 - a) \times 100$ % probability ellipsoid can be written as follows

$$(\boldsymbol{\phi}(\mathbf{z}) - \boldsymbol{\mu})^{'}\Sigma^{-1}(\boldsymbol{\phi}(\mathbf{z}) - \boldsymbol{\mu}) \leq \chi^2_{p-1}(a)$$

where $\chi^2_p(a)$ is the upper $(100a)$th percentile of the chi–square distribution with p degrees of freedom. The major–axis of the ellipse is given by

$$\pm c\sqrt{\lambda_1} e_1$$

where $c = \chi^2_{p-1}(a)$, λ_1 is the largest eigenvalue of Σ and e_1 is its corresponding eigenvector.

We restrict attention to the class of logistic normal priors that can be generated from p independent log–normal distributions with common variance σ^2, which have the following covariance structure

$$\underset{((p-1)\times(p-1))}{\Sigma} = \sigma^2 \left[\mathbf{I}_{p-1} + \mathbf{J}_{p-1}\right]$$

where $\mathbf{I}_p$ is the p–dimensional identity matrix and $\mathbf{J}_p$ is a $p \times p$ matrix of all ones. It can be shown that the largest eigenvalue is $\lambda_1 = p\sigma^2$ and associated eigenvector

$$e_1^{'} = \frac{\mathbf{j}_{p-1}}{\sqrt{p-1}}$$

where $\mathbf{j}_p$ is a vector of all ones of dimension p.

Therefore, to restrict the prior distribution to have most of its mass within the region of practical interest, that is, $\pm x_c$, we can solve the following equation for σ^2:

$$\sqrt{\chi^2_{p-1}(a)}\sqrt{p\sigma^2}\frac{\mathbf{j}_{p-1}}{\sqrt{p-1}} \leq x_c \mathbf{j}_{p-1}$$

giving

$$\sigma^2 < \frac{x_c^2(p-1)}{p\chi^2_{p-1}(a)}$$

for each component. For example, with $p = 12$ and restricting the range of practical interest to ± 10 with $a = 0.05$ choosing $\sigma^2 < 4.659$ gives the desired result. Note that this value gives a multi–modal prior distribution, see chapter 2.1. This analysis deals with the logistic normal distribution on the log–ratio scale.

This can be contrasted with "so called" box constraints on the components which would lead to $\sqrt{z_a 2\sigma^2} < 10$, yielding a solution of $\sigma^2 \leq 25$ for each component irrespective of the dimension of the problem under consideration.

We assign vague but proper priors to all of the location and scale parameters of the constant diet model. We center the prior distributions at the compositional zero and set the variances as described above setting $x_c = 10$ and use the 95% percentile of the Chi–squared distribution. In order to ensure the propriety of the prior inverse Wishart distributions we set the degrees of freedom parameter consistent with the dimension of the covariance matrix of interest.

6.3.6 Starting Values

Theoretically, starting values are not important provided the algorithm is Harris recurrent, ergodic, etc. However, in practice, starting values are an important issue as we only have limited computing resources. Gelman et al. (2003) suggest starting the MCMC algorithm at the posterior modes of the distribution. In fact, they suggest that one should sample from a mixture distribution of all posterior modes and run multiple chains each starting from a point sampled from this mixture distribution. The posterior distribution under consideration is difficult to handle analytically due to the presence of the term

$$\begin{aligned}
\underset{(a\times n)}{\mathbf{Y}} &= \left(\underset{(a\times p)}{\Theta}\underset{(p\times n)}{\Gamma}\right) \oplus_c \left(\underset{(a\times 1)}{\boldsymbol{\theta}_{\mathbf{v}}} \ominus \underset{(a\times 1)}{\boldsymbol{\theta}_{\mathbf{u}}}\right) \oplus_c \underset{(a\times n)}{\mathbf{E}} \\
\underset{(p\times n)}{\Gamma} &= \boldsymbol{\phi}_c^{-1}\left(\boldsymbol{\phi}_c\left(\underset{(p\times w)}{\mathbf{T}}\right)\underset{(w\times n)}{\mathbf{W}}\right) \oplus_c \underset{(p\times 1)}{\boldsymbol{\lambda}}
\end{aligned}$$

and the high dimensional nature of the space under consideration.

Gelman et al. (2003) suggest using conditional maximization or steepest ascent to address the issue of posterior modes. The method we employ here is a slight variant of this method but it also incorporates some ideas from Liu et al. (2009). Liu et al. (2009) considers the case where some parts (modules) of a Bayesian model may contain better quality data than others, they use partial likelihood as a motivating example. This is not the case here, however, the idea of a modulation approach is exploited to find starting values for the various modules.

Essentially, we modularize the model into five parts, $\boldsymbol{\theta}_j, j = 1, \ldots, p$, $\lambda_j, j = 1, \ldots, p$, $\boldsymbol{\theta}_{\mathbf{u}}$, $\boldsymbol{\theta}_{\mathbf{v}}$ and $\boldsymbol{\tau}_s$. We assume that the prior distributions are diffuse enough such that the posterior is dominated by the likelihood for these components. Effectively this means that the posterior distributions are logistic T distributions, (see Gelman et al. (2003), pg 88). For example, the starting values for $\boldsymbol{\theta}_j$ are sampled from the following distributions

$$\boldsymbol{\theta}_{j,0} \sim \mathcal{LT}^a(\overline{\boldsymbol{\phi}(\mathbf{x}_j)}, n_j(n_j - a + 1)SS_j^{-1}, n_j - a + 1)$$

where

$$\overline{\boldsymbol{\phi}(\mathbf{x}_j)} = \frac{1}{n_j} \sum_{k=1}^{n_j} \boldsymbol{\phi}(\mathbf{x}_{jk})$$

and

$$SS_j = \sum_{k=1}^{n_j} \left(\boldsymbol{\phi}(\mathbf{x}_{jk}) - \overline{\boldsymbol{\phi}(\mathbf{x}_j)}\right) \left(\boldsymbol{\phi}(\mathbf{x}_{jk}) - \overline{\boldsymbol{\phi}(\mathbf{x}_j)}\right)'$$

Similar distributions apply for λ_j, $\boldsymbol{\theta}_{\mathbf{u}}$ and $\boldsymbol{\theta}_{\mathbf{v}}$.

It is not possible to find starting values for the parameters $\boldsymbol{\tau}_s, s = 1, \ldots, w$ in this manner, so we start them at the compositional zero. That is, we assume that the diet compositions are all equally weighted among the prey types.

6.3.7 Adaptive Modifications

To avoid some of the pitfalls of adaptive algorithms we describe a variant, that by construction obeys the diminishing adaption condition required to be theoretically valid. One of the problems with adaptive algorithms is that they can get stuck when the initial solution is not near a posterior mode and in the application we address later this is concern. We propose to have two distinct stages of adaption followed by a stage of non–adaption. Note that, we only consider adaption on the Metropolis–Hastings steps of the Metropolis–within–Gibbs algorithm systematic scan algorithm. That is to say, we are not considering adaption in a random scan Metropolis–within–Gibbs algorithm where it is possible to adaptively update the probability of visiting a particular part of the posterior distribution. There are valid reasons for doing this, as some conditional distributions may not need to be sampled as often as other conditionals, however, as stated we implement a systematic scan Metropolis–within–Gibbs algorithm.

The first stage of our algorithm consists of adaptively updating the conditional distributions that require a Metropolis–Hastings algorithm with the following proposal distribution. Let $\boldsymbol{\nu}_i$ be the previous state of a given conditional distribution, and assume that the control parameter at the previous step was λ_i

$$\boldsymbol{\nu}^* \sim \mathcal{L}\left(\boldsymbol{\nu}_i, \lambda_i[\mathbf{I} + \mathbf{J}]\right)$$

where λ_{i+1} is then updated according to

$$\log \lambda_{i+1} = \log \lambda_i + \gamma_{i+1}(\rho(\boldsymbol{\nu}^*, \boldsymbol{\nu}_i) - \overline{\rho})$$

where γ_{i+1} is the adaption parameter which is updated in a systematic fashion, $\overline{\rho}$ is the target acceptance rate of the chain, and $\rho(\boldsymbol{\nu}^*, \boldsymbol{\nu}_{i-1})$ is the acceptance probability for the current point $\boldsymbol{\nu}^*$.

The second stage is more standard, along the lines of Andrieu and Thoms (2008); Roberts and Rosenthal (2009). The premise is that after running the stage one for a sufficient number of iterations, that we should now be close to posterior mode and can now implement the more standard methods. Specifically, we use the following algorithm from Andrieu and Thoms (2008) with a slight modification given in Roberts and Rosenthal (2009) to avoid a potentially singular covariance matrix in the initial start up.

Algorithm 2 Modified Adaptive Algorithm

```
Input: Initialize $X_0$, $\mu_0$, $\Sigma_0$ and $\lambda_0$
Output: An adaptive Markov chain $X_i, i = 1, \dots, n$
  At iteration $i + 1$, given $X_i$, $\mu_i$, $\Sigma_i$ and $\lambda_i$
  Sample $X^*_{i+1} \sim q^{SRWM}_{\mu_i,\Sigma_i}(X_i, .)$
  Update
      $\log(\lambda_{i+1}) = \log(\lambda_i) + \gamma_{i+1}\left(\rho(X_i, X^*_{i+1}) - \overline{\rho}\right)$
      $\mu_{i+1} = \mu_i + \gamma_{i+1}\left(X_i - \mu_i\right)$
      $\Sigma_{i+1} = \Sigma_i + \gamma_{i+1}\Big((X_{i+1} - \mu_i)(X_{i+1} - \mu_i)' - \Sigma_i\Big)$
```

where $\gamma_{i+1} = i^{-0.7}$.

The third stage of the algorithm follows the second stage, except we turn the adaption off, that is, $\gamma_{i+1} = 0$. This is done so that we can assess the convergence diagnostics on the final stage. To date, we are not aware of any convergence diagnostics that apply to adaptive algorithms.

We run stage 1 for 5×10^5 iterations, stage 2 for 1×10^6 iterations and the stage 3 for

1×10^6 iterations, with a thinning factor of 100 to save on storage space. The first two stages of the adaptive algorithm are not used to conduct posterior inference. However, they are used to monitor the chains.

6.3.8 Results

We now give the results of applying the constant diet model to the synthetic data generated using the diet composition given in table 6.2 for both the original prey and the permuted prey. Due to the large dimension of the compositions in question we only give detailed results for the diet composition and present various summary measures for the other parameters of the model.

Tables 6.3 and 6.4 give the results of the diet composition for the original prey and the permuted prey respectively. Perhaps the most striking feature of the tables is how much wider the component–wise credible intervals are for the original data compared to the permuted data. The total width of the 95% credible intervals for the original data on the percentage scale is 547.17 compared to 160.16 for the permuted data. Most of the credible intervals span the actual diet, 66 out of 72 for both cases, the exceptions are typically for the rarer diet items. These two facts indicate that we pay a substantial price in loss of certainty of the diet composition with high degree of collinearity among the diet items. The trace of the true error covariance matrix for this problem is 0.62, the posterior mean is 0.628 with 95% credible interval of $(0.579, 0.686)$ for the original data and 0.625 and $(0.577, 0.682)$ for the permuted data respectively. This result is consistent with the case of multicollinearity in regression contexts where the overall fit is not compromised, however, the individual regression coefficients are not well determined.

As a further comparison we compute the compositional distance introduced by Billheimer (2001) (see chapter 2.1) between the actual parameters and the 10,000 MCMC samples and the posterior mean. The results are given in figures 6.6 and 6.7 for the diet composition, $\boldsymbol{\tau}_w, w = 1, \ldots, 6$, the calibration predator $\boldsymbol{\theta}_v$, calibration prey $\boldsymbol{\theta}_u$, the prey (source) profiles $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1| \ldots |\boldsymbol{\theta}_{12}]$ and finally the fat composition, $\lambda_j, j = 1, \ldots, 12$.

Inspection of panel a) of both figures reiterates the difficulty in recovering the actual diet when there is a high degree of multicollinearity present. That is, the non–permuted example has much more difficulty reproducing the actual diet compared to the permuted version. However, both methods experience more difficulty reconstructing diets with rare components (summer females, and winter males and females, see table 6.2). Panels b)

		Spring		Summer		Winter				Spring		Summer		Winter	
		M	F	M	F	M	F			M	F	M	F	M	F
Capelin	pm	1.12	3.63	1.54	0.41	2.69	2.37	Redfish	pm	35.18	37.26	34.08	24.86	36.94	40.48
	lci	0.03	0.91	0.04	0.00	0.18	0.06		lci	30.23	31.64	29.12	19.96	31.56	34.41
	uci	3.69	6.71	4.48	1.91	5.99	5.74		uci	40.27	42.87	39.31	30.12	42.19	46.58
	act	0.3	3.0	1.5	0.005	3.00	1.80		act	33.0	35.0	33.0	23.0	35.0	38.0
Cod	pm	3.41	1.53	4.36	3.45	1.27	1.12	Sandlance	pm	3.65	36.09	18.39	62.7	27.58	38.76
	lci	0.05	0.01	0.05	0.02	0.02	0.01		lci	0.35	30.24	14.06	56.12	22.45	32.65
	uci	10.44	6.65	12.12	11.11	5.54	5.14		uci	7.35	42.35	23.29	69.01	33.32	45.26
	act	3.0	1.0	5.5	4.00	5.00	1.50		act	7.0	41.0	22.0	67.0	32.0	43.0
Herring	pm	3.19	0.91	3.01	0.88	3.44	4.06	Silver	pm	2.84	2.30	2.96	0.78	2.21	2.20
	lci	0.05	0.02	0.09	0.01	0.04	0.06	Hake	lci	0.04	0.03	0.05	0.01	0.03	0.02
	uci	9.24	4.32	8.66	4.27	10.53	11.45		uci	10.30	9.74	10.3	3.80	8.92	8.90
	act	5.0	6.0	4.0	2.00	4.00	8.00		act	2.0	1.4	0.5	0.50	0.90	0.005
Longhorn	pm	2.37	0.85	1.75	0.96	2.16	0.74	Thorny	pm	3.38	0.43	2.08	2.45	5.85	0.51
Sculpin	lci	0.06	0.01	0.03	0.01	0.05	0.01	Skate	lci	0.22	0.01	0.11	0.03	2.57	0.01
	uci	6.89	3.48	5.71	4.31	6.23	3.26		uci	7.01	1.82	5.26	6.02	9.09	2.23
	act	0.5	0.1	1.5	1.50	0.005	0.19		act	5.0	0.5	2.0	0.01	5.00	0.50
Plaice	pm	1.59	7.49	6.62	0.70	1.01	1.40	White	pm	6.86	1.76	3.17	1.77	1.49	0.76
	lci	0.02	1.32	0.71	0.01	0.02	0.02	Hake	lci	0.32	0.02	0.07	0.01	0.03	0.01
	uci	6.22	12.86	12.15	3.43	4.22	5.06		uci	15.28	6.99	9.77	7.81	6.22	3.47
	act	5.0	1.5	3.0	0.50	1.00	0.005		act	6.0	1.0	4.0	0.50	0.10	0.005
Pollock	pm	29.83	6.64	20.72	0.61	11.89	6.77	Yellowtail	pm	6.59	1.10	1.33	0.43	3.48	0.84
	lci	22.91	0.66	14.54	0.01	6.38	0.54		lci	3.19	0.02	0.03	0.00	0.11	0.01
	uci	36.86	11.91	27.00	3.20	17.45	12.25		uci	10.08	4.25	4.54	2.00	6.74	2.95
	act	30.0	5.5	21.0	0.02	11.0	5.00		act	3.2	4.0	2.0	0.97	3.00	2.00

Table 6.3: Posterior summaries of the diet composition vectors, $\boldsymbol{\tau}$'s, for the three seasons (Spring, Summer, Winter) and sex for the synthetic data generated with compositions given in table 6.2. Results are based on the 10,000 MCMC samples as described in the text in previous sections. The posterior mean is denoted by pm; the lower and upper 95% credible limits are denoted by lci and uci; and the actual diet is denoted by act.

		Spring		Summer		Winter				Spring		Summer		Winter	
		M	F	M	F	M	F			M	F	M	F	M	F
Capelin(p)	pm	0.39	3.34	1.59	0.09	3.23	1.98	Redfish(p)	pm	33.45	37.56	34.14	25.65	36.97	41.15
	lci	0.18	2.75	1.27	0.00	2.69	1.62		lci	30.13	33.54	30.65	21.97	33.20	36.77
	uci	0.64	3.96	1.94	0.27	3.83	2.39		uci	36.72	41.69	37.52	29.71	40.67	45.57
	act	0.3	3.0	1.5	0.005	3.00	1.80		act	33.0	35.0	33.0	23.0	35.0	38.0
Cod(p)p	pm	2.80	0.91	5.29	3.96	5.13	1.28	Sandlance(p)	pm	5.96	37.06	19.31	63.51	28.02	38.87
	lci	2.46	0.51	4.66	3.32	4.48	0.78		lci	4.90	32.38	16.39	58.54	24.14	33.96
	uci	3.18	1.33	5.97	4.68	5.84	1.79		uci	7.16	41.74	22.45	68.1	32.09	43.75
	act	3.0	1.0	5.5	4.00	5.00	1.50		act	7.0	41.0	22.0	67.0	32.0	43.0
Herring(p)	pm	4.57	5.82	3.75	2.13	3.85	7.92	Silver	pm	1.49	1.21	0.37	0.42	0.63	0.09
	lci	3.67	4.63	2.98	1.62	3.05	6.31	Hake(p)	lci	0.86	0.70	0.04	0.16	0.14	0.00
	uci	5.55	7.09	4.58	2.66	4.71	9.60		uci	2.20	1.78	0.89	0.69	1.14	0.34
	act	5.0	6.0	4.0	2.00	4.00	8.00		act	2.0	1.4	0.5	0.50	0.90	0.005
Longhorn(p)	pm	0.45	0.20	1.48	1.51	0.09	0.20	Thorny	pm	5.45	0.73	2.21	0.05	5.47	0.68
Sculpin	lci	0.23	0.03	1.19	1.21	0.01	0.03	Skate(p)	lci	4.82	0.56	1.91	0.00	4.79	0.53
	uci	0.68	0.43	1.81	1.87	0.26	0.41		uci	6.14	0.93	2.56	0.15	6.20	0.86
	act	0.5	0.1	1.5	1.50	0.005	0.19		act	5.0	0.5	2.0	0.01	5.00	0.50
Plaice(p)	pm	5.47	1.73	3.35	0.74	1.22	0.09	White	pm	5.90	1.00	3.97	0.72	0.05	0.13
	lci	4.55	1.38	2.70	0.53	0.87	0.01	Hake(p)	lci	5.18	0.75	3.44	0.43	0.00	0.01
	uci	6.47	2.11	4.03	0.98	1.60	0.29		uci	6.69	1.29	4.55	1.03	0.16	0.34
	act	5.0	1.5	3.0	0.50	1.00	0.005		act	6.0	1.0	4.0	0.50	0.10	0.005
Pollock(p)	pm	30.72	6.17	22.25	0.08	11.94	5.58	Yellowtail(p)	pm	3.34	4.26	2.29	1.14	3.40	2.03
	lci	27.91	5.22	19.90	0.00	10.43	4.62		lci	2.84	3.55	1.81	0.65	2.78	1.47
	uci	33.50	7.22	24.65	0.29	13.52	6.61		uci	3.91	5.02	2.83	1.69	4.04	2.65
	act	30.0	5.5	21.0	0.02	11.0	5.00		act	3.2	4.0	2.0	0.97	3.00	2.00

Table 6.4: Posterior summaries of the diet composition vectors, $\boldsymbol{\tau}$'s, for the three seasons (Spring, Summer, Winter) and sex for the synthetic data generated with compositions given in table 6.2 using the permuted prey. Results are based on the 10,000 MCMC samples as described in the text in previous sections.The posterior mean is denoted by pm; the lower and upper 95% credible limits are denoted by lci and uci; and the actual diet is denoted by act.

through d) of the figure do not indicate any appreciable difference in the ability of the two forms of the model to reconstruct the other model parameters.

The previous results allow communication of information from the predator (receptor) to influence the prey (sources). To see this consider the full conditional distribution for the jth prey type, $\boldsymbol{\theta}_j$:

$$\begin{aligned}\pi(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{-j}, \boldsymbol{\alpha}, \Sigma_{\boldsymbol{\epsilon}}, \mathbf{Y}, \mathbf{X}_j, \Sigma_{\mathbf{X}_j}) =&\pi(\boldsymbol{\theta}_j|\boldsymbol{\mu}_{\boldsymbol{\theta}_j}, \Sigma_{\boldsymbol{\theta}_j}) \prod_{i=1}^{n} \pi(\mathbf{y}_i|\boldsymbol{\alpha}, \boldsymbol{\Theta}, \Sigma_{\boldsymbol{\epsilon}}),\\ &\prod_{k=1}^{n_j} \pi(\mathbf{x}_{jk}|\boldsymbol{\theta}_j, \Sigma_{\mathbf{X}_j}).\end{aligned}$$

The full conditional depends on the prior distribution for $\boldsymbol{\theta}_j$, the sampling distribution for $\mathbf{x}_{jk}$ and more interestingly the sampling distribution of $\mathbf{y}_i$. That is, the predators actually contain information about the prey profiles. As was discussed in the previous chapter, it is interesting to see what the consequences of breaking this information flow is on the quality of the inference. The conditional distributions for λ_j, $\boldsymbol{\theta}_{\mathbf{u}}$ and $\boldsymbol{\theta}_{\mathbf{v}}$ also have similar functional forms, in that the sampling distribution is part of their full conditional distribution and hence has information about the plausible values for each of these parameters.

The Markov Chain Monte Carlo machinery gives samples from the full posterior distribution when the correct full conditional distributions are used in either Gibbs sampling or in our case, Metropolis–Hastings–within-Gibbs. If we change the full conditional distributions as suggested above, we are no longer generating samples from the full posterior distribution. At best, it might be an adequate approximation to the original posterior at worst it we could be sampling from a posterior distribution that isn't even a proper distribution and hence it would be impossible to draw even approximate posterior inference. For different reasons, Liu et al. (2009), considers this as well and concludes it can work in some situations. However, changing the functional form of the full conditional distributions when performing Gibbs sampling is not recommended in general.

With these caveats in mind we repeated the Gibbs sampling with the information flow between the predator and prey. Figures 6.8 and 6.9 give the distances between the model parameters and the true ones for the approximate method. Comparing panel a) of these figures with the previous ones, it is readily apparent that the diet composition is adversely

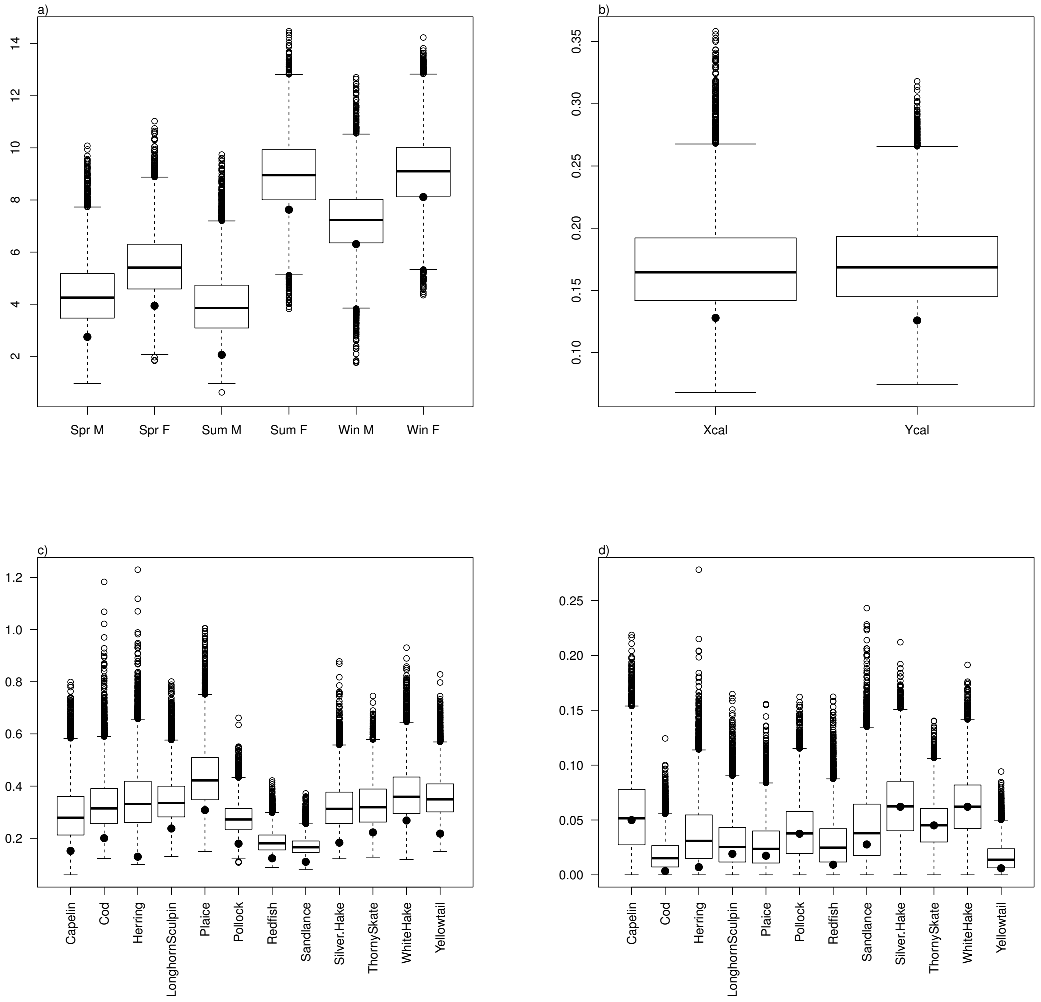

Figure 6.6: Boxplots of the compositional distance between the true parameters and the 10,000 MCMC samples for the non–permuted prey (the distance to the posterior mean is indicated by the solid circle). Panel a) $\boldsymbol{\tau}_w$ for each of the of the six season and sex combinations; panel b) $\boldsymbol{\theta}_{\mathbf{u}}$ the calibration predator and (Ycal in the figure), $\boldsymbol{\theta}_{\mathbf{v}}$ the calibration prey, (Xcal in the figure); panel c) $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1| \ldots |\boldsymbol{\theta}_{12}]$ the fatty acid profile of the 12 prey types; panel d) $\lambda_1, \ldots, \lambda_{12}$ the fat content of the 12 prey types.

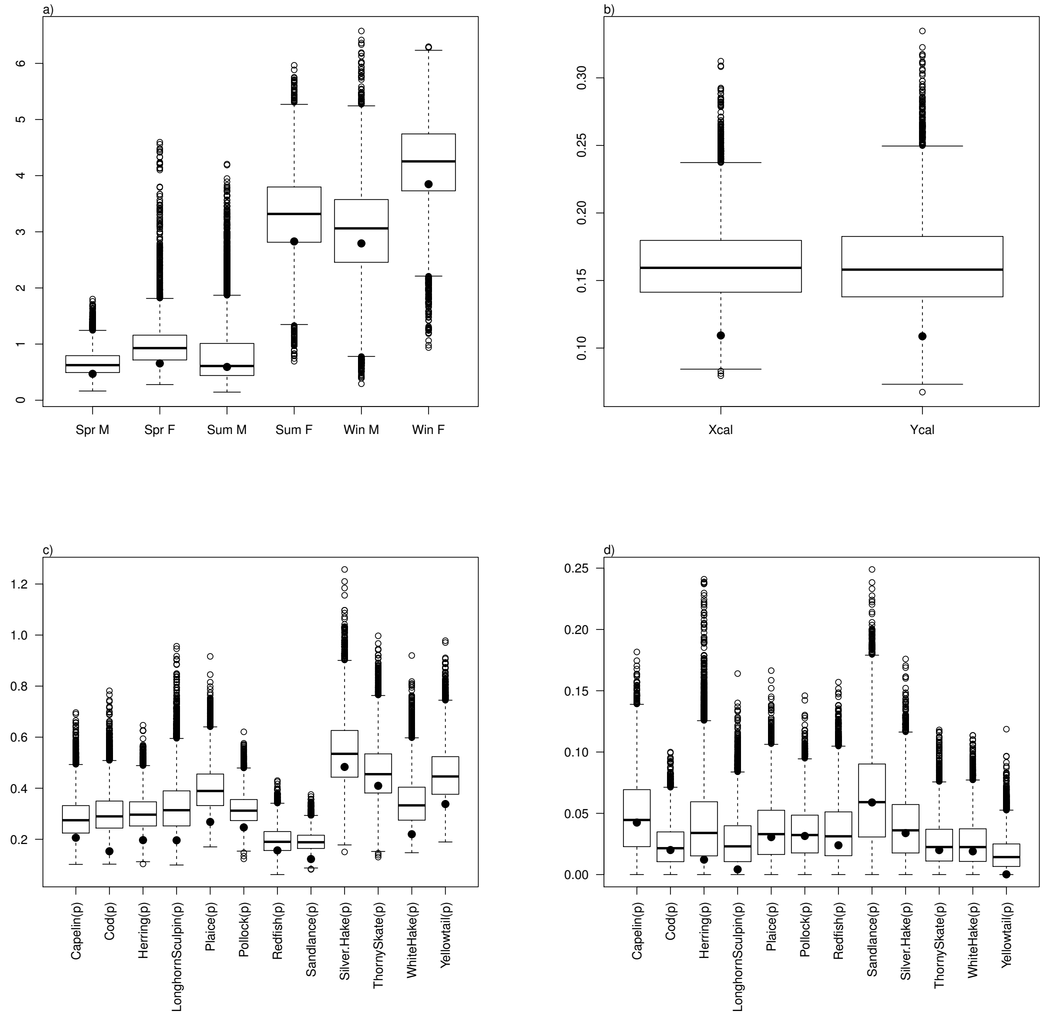

Figure 6.7: Boxplots of the compositional distance between the true parameters and the 10,000 MCMC samples for the permuted prey (the distance to the posterior mean is indicated by the solid circle). Panel a) $\boldsymbol{\tau}_w$ for each of the of the six season and sex combinations; panel b) $\boldsymbol{\theta}_\mathbf{u}$ the calibration predator and (Ycal in the figure), $\boldsymbol{\theta}_\mathbf{v}$ the calibration prey, (Xcal in the figure); panel c) $\Theta = [\boldsymbol{\theta}_1 | \ldots | \boldsymbol{\theta}_{12}]$ the fatty acid profile of the 12 prey types; panel d) $\lambda_1, \ldots, \lambda_{12}$ the fat content of the 12 prey types.

affected by removing the information flow from predator to prey. It is particularly problematic for the permuted version with the distances increasing by an order of magnitude. Panel b) indicates a slight problem with the calibration factors, in that, the prey calibration is further away from the its true value for both the permuted and non–permuted synthetic data. For the non–permuted example, several prey types have moved further away from their true values, indicating again some information loss by not allowing information to flow from the predator to the prey. Most notably this occurs for Pollock, Redfish and Sandlance the biggest contributors to the diet. A similar picture emerges for the permuted example, with the same species moving away from their actual profiles with the addition of Capelin. Not surprisingly, the fat content, denoted in panel d) is unaffected as there is no information in the predator as to the fat content of the prey types as the fat content does not get deposited in the predator.

Logically, it makes sense for information to flow from the predator to the prey, as only they contain the prey fatty acid compositions that were actually consumed. The samples on the prey themselves, at best, contain the actual prey items that could have potentially been consumed.

The quality of inference is affected by two aspects, the multicollinearity between the sources (prey types) and allowing information to flow from predator to prey as per the full conditional distribution. One other potential consequence of information flow between the predator and prey is that if an important prey type isn't present among the potential sources then the model can/will modify the existing prey types. In fact, this is the basis of the Billheimer (2001) model, he uses informative priors to help and allows the receptors to inform the sources. In our situation we can think of the samples on the prey (sources) as forming an informative prior on their fatty acid profiles which can then be modified by the predators.

One strategy to deal with the multicollinearity issue is to remove the potentially confounded prey types. Hoever, if the species that are removed are important constituents of the diet then other prey types may get modified to such an extent that they are no longer recognizable as the species of interest. With synthetic diet studies we can minimize this problem by keeping the larger diet constituents, however, in a practical application this is not feasible. Examining the prey species that are markedly different from their sample means after the removal of a candidate species gives evidence that the species removed

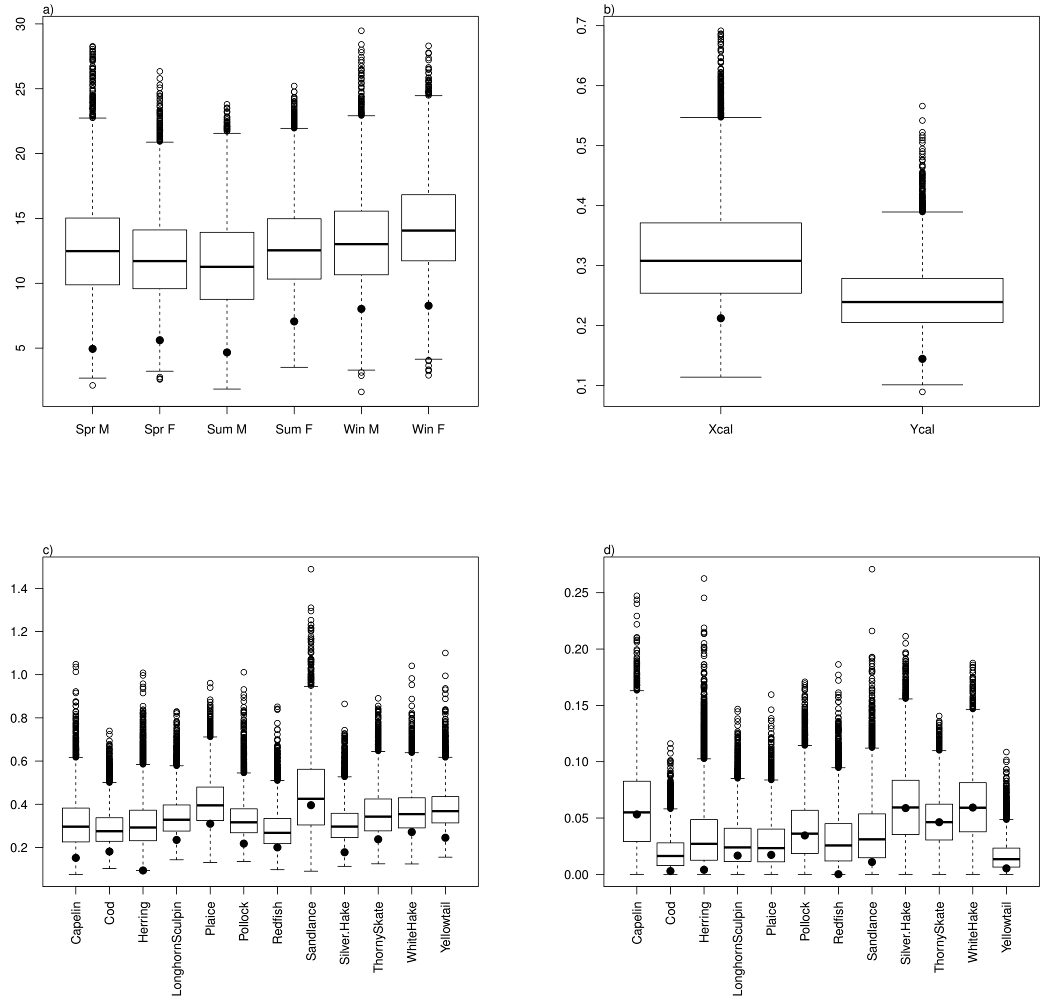

Figure 6.8: Boxplots of the compositional distance between the true parameters and the 10,000 MCMC samples for the non–permuted prey with no information flow between the predator and prey (the distance to the posterior mean is indicated by the solid circle). Panel a) $\boldsymbol{\tau}_w$ for each of the of the six season and sex combinations; panel b) $\boldsymbol{\theta}_{\mathbf{u}}$ the calibration predator and (Ycal in the figure), $\boldsymbol{\theta}_{\mathbf{v}}$ the calibration prey, (Xcal in the figure); panel c) $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1| \ldots |\boldsymbol{\theta}_{12}]$ the fatty acid profile of the 12 prey types; panel d) $\lambda_1, \ldots, \lambda_{12}$ the fat content of the 12 prey types.

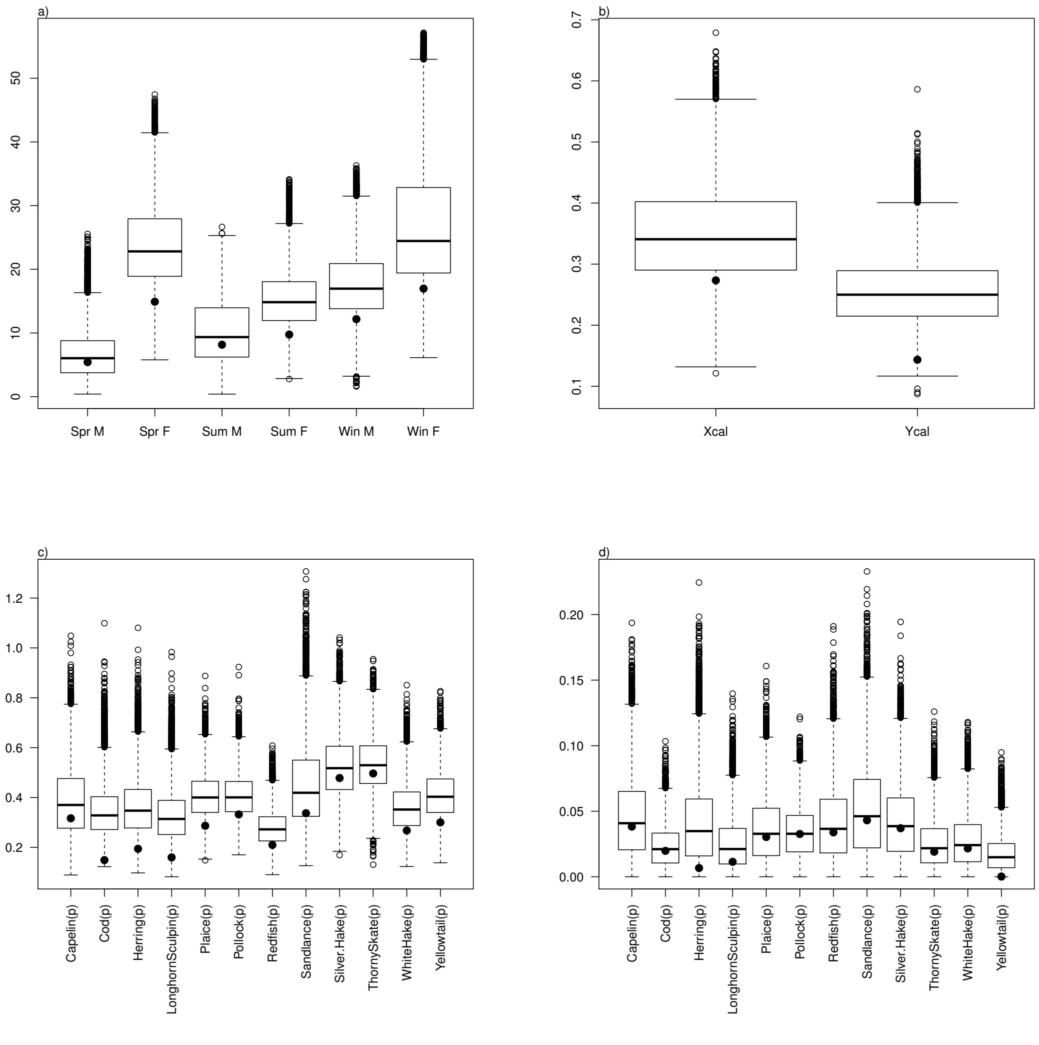

Figure 6.9: Boxplots of the compositional distance between the true parameters and the 10,000 MCMC samples for the permuted prey with no information flow between the predator and prey (the distance to the posterior mean is indicated by the solid circle). Panel a) $\boldsymbol{\tau}_w$ for each of the of the six season and sex combinations; panel b) $\boldsymbol{\theta}_{\mathbf{u}}$the calibration predator and (Ycal in the figure), $\boldsymbol{\theta}_{\mathbf{v}}$ the calibration prey, (Xcal in the figure); panel c) $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1|\ldots|\boldsymbol{\theta}_{12}]$ the fatty acid profile of the 12 prey types; panel d) $\lambda_1, \ldots, \lambda_{12}$ the fat content of the 12 prey types.

were an important part of the diet composition.

With these caveats in mind we try two simple methods of dealing with ill–conditioned source (prey) matrices, namely, fitting a reduced set of potential prey items and fitting the full suite of prey items and amalgamating after the fact on the original scale.

Figure 6.10 gives the results of fitting 5 species (Capelin, Cod, Pollock, Redfish, Sandlance), which were selected from the hierarchical clustering diagram (see figure 6.4). It looks as though the fully Bayesian method and the approximate method perform about equally well, however, the Bayesian approach modifies Cod quite substantially compared to the approximate method. That is one of the disadvantages of this approach, in that it will modify potential prey items so that they more resemble the predators. This can be used as a potential diagnostic tool to say when a species may be potentially missing from the diet library. However, we don't pursue this at the present time. We can also compute the compositional distance between the posterior means of $\boldsymbol{\tau}$ and the amalgamated true diet, they are 21.994 and 16.976 for the fully Bayes and approximate solution respectively. That is, we do a slightly better job of recovering the actual diet of the predator with the approximate solution. Additionally, the posterior mean of the trace of the covariance matrix is, 0.646, with 95% credible interval (0.591,0.713) for the fully Bayesian approach compared to a posterior mean of 0.908 and 95% credible interval of (0.722,1.269) for the approximate method. Thus, the fully Bayesian approach appears to fit better but it achieves this by modifying the species (sources) compared to the approximate method. Of course the difficulty with the Bayesian approach is that it produces a species that may not be recognizable.

Another simple method of dealing with multicollinearity issues are to fit the model ignoring the problem and then amalgamate after the fact. We illustrate this with the fully Bayesian approach as it was shown to be superior when all species are present in the fitting procedure. According to figure 6.4 the following five groups were selected (Capelin, Herring), (Cod, Longhorn Sculpin, Plaice, Thorny Skate, Yellowtail), (Pollock, Silver Hake, White Hake), Redfish and Sandlance each formed their own group. Figure 6.11 gives the amalgamated posterior estimates (note they were amalgamated on the original scale) and their 95% credible intervals, with the true value indicated by the closed circles. All the credible intervals contain the true values, and the total compositional distance is 5.078 which compares very favourably to selecting out a few prey items as previously discussed.

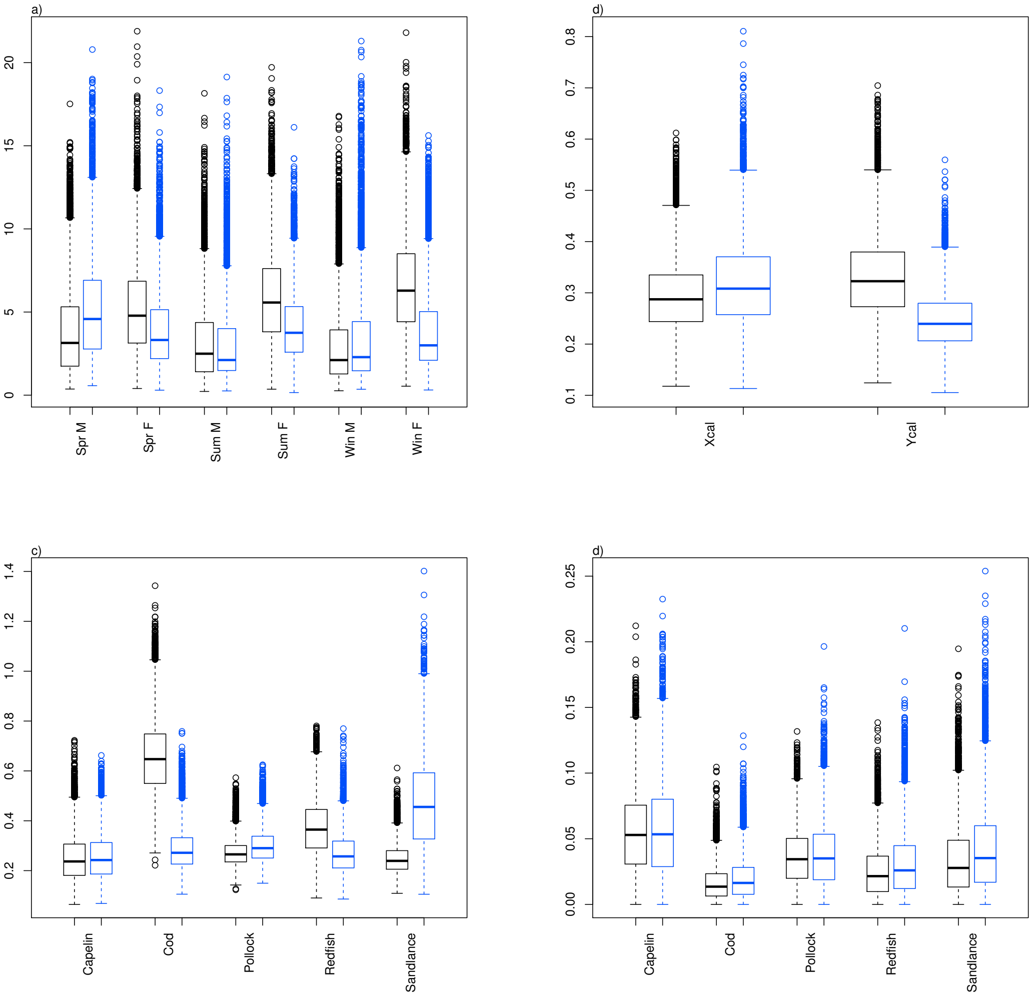

Figure 6.10: Boxplots of the compositional distance between the true parameters and the 10,000 MCMC samples for the original prey base using the following 5 species (Capelin, Cod, Pollock, Redfish, Sandlance) to ascertain the viability of fitting a reduced set of prey. Panel a) $\boldsymbol{\tau}_w$ for each of the of the six season and sex combinations; panel b) $\boldsymbol{\theta}_{\mathbf{u}}$ the calibration predator and (Ycal in the figure), $\boldsymbol{\theta}_{\mathbf{v}}$ the calibration prey, (Xcal in the figure); panel c) $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1| \ldots |\boldsymbol{\theta}_{12}]$ the fatty acid profile of the 12 prey types; panel d) $\lambda_1, \ldots, \lambda_{12}$ the fat content of the 12 prey types. Within each of the panels the first boxplot (black) corresponds to the fully Bayesian approach and the second boxplot (blue) to the approximate Bayesian solution.

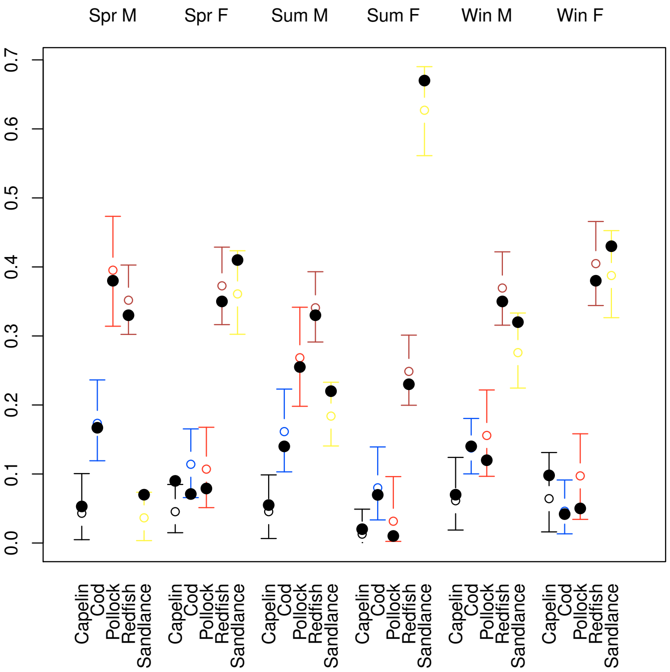

Figure 6.11: Posterior means and 95% credible intervals for the amalgamated diet composition based on figure 6.4 (see text for details). The open circle is the posterior mean and the closed circle is the amalgamated true diet.

6.4 Individual Diet Model

Most predators are opportunistic foragers, therefore, we would expect variations in the diet compositions of individuals within populations. The common diet model developed in the previous sections will not capture this nuance. Recall, in the original Billheimer (2001) model each day of air pollution had its own mixing coefficient. We pursue this generalization, known as the individual diet model, in this section.

To account for possibly multiple measurements per predator, we let $\mathbf{y}_{i1}, \ldots, \mathbf{y}_{ir}$ denote the r replicate fatty acid profiles on the ith predator. Consider the following generalization of the constant diet model with multiple populations

$$
\begin{aligned}
\underset{(a\times nr)}{\mathbf{Y}} &= \left[\left(\underset{(a\times p)}{\mathbf{\Theta}}\underset{(p\times n)}{\Gamma} \otimes \underset{(1\times r)}{\mathbf{U}}\right) \oplus_c \left(\underset{(a\times 1)}{\boldsymbol{\theta}_{\mathbf{v}}} \ominus \underset{(a\times 1)}{\boldsymbol{\theta}_{\mathbf{u}}}\right)\right] \oplus_c \underset{(a\times nr)}{\mathbf{E}}, \\
\underset{(p\times n)}{\Gamma} &= \boldsymbol{\phi}_c^{-1}\left(\boldsymbol{\phi}_c\left(\underset{(p\times w)}{\mathbf{T}}\right)\underset{(w\times n)}{\mathbf{W}} + \boldsymbol{\phi}_c\left(\underset{(p\times n)}{\Gamma^m}\right)\right) \oplus_c \underset{(p\times 1)}{\boldsymbol{\lambda}}, \\
\underset{(a\times n_j)}{\mathbf{X}_j} &= \underset{(a\times 1)}{\boldsymbol{\theta}_j}\underset{(1\times n_j)}{\mathbf{W}_{\mathbf{X}_j}} \oplus_c \underset{(a\times n_j)}{\mathbf{E}_{\mathbf{X}_j}}, \quad j = 1, \ldots, p, \\
\underset{(2\times n_j)}{\mathbf{Z}_j} &= \underset{(2\times 1)}{\boldsymbol{\lambda}_j^v}\underset{(1\times n_j)}{\mathbf{W}_{\mathbf{Z}_j}} \oplus_c \underset{(a\times n_j)}{\mathbf{E}_{\mathbf{Z}_j}}, \quad j = 1, \ldots, p, \\
\underset{(a\times L)}{\mathbf{U}} &= \underset{(a\times 1)}{\boldsymbol{\theta}_{\mathbf{u}}}\underset{(1\times L)}{\mathbf{W}_{\mathbf{U}}} \oplus_c \underset{(a\times L)}{\mathbf{E}_{\mathbf{U}}}, \\
\underset{(a\times M)}{\mathbf{V}} &= \underset{(a\times 1)}{\boldsymbol{\theta}_{\mathbf{v}}}\underset{(1\times M)}{\mathbf{W}_{\mathbf{V}}} \oplus_c \underset{(a\times M)}{\mathbf{E}_{\mathbf{V}}}
\end{aligned}
$$

where $\underset{(a\times nr)}{\mathbf{Y}} = [\mathbf{y}_{11}|\ldots|\mathbf{y}_{1r}|\ldots,\mathbf{y}_{n1}|\ldots|\mathbf{y}_{nr}]$ is an $a \times nr$ matrix of observations on the predators, $\underset{(a\times p)}{\mathbf{\Theta}} = [\boldsymbol{\theta}_1|\ldots|\boldsymbol{\theta}_p]$ is an $a \times p$ matrix of predator(source) profiles, $\underset{(p\times n)}{\Gamma}$ is an $p \times n$ matrix of individual diet(mixing) compositions adjusted for fat content, $\otimes$ represents the Kronecker product of two matrices defined in the previous chapter, $\underset{(1\times r)}{\mathbf{U}}$ is an $1 \times r$ matrix of ones which allows the predicted profile to be replicated r times to account for replicate measurements on the same predator, $\oplus_c$ is the perturbation operator applied column–wise for matrices of the same size (if the second argument is a vector, then the vector is replicated column–wise first then applied column wise), $\underset{(a\times 1)}{\boldsymbol{\theta}_{\mathbf{v}}}$ is an a–dimensional fatty acid profile of the calibration predator, $\underset{(a\times 1)}{\boldsymbol{\theta}_{\mathbf{u}}}$ is an a–dimensional fatty acid profile of the calibration prey, $\ominus$ is the inverse perturbation operator ($\mathbf{x} \ominus \mathbf{y} = \mathbf{x} \oplus \mathbf{y}^{-1}$), $\underset{(a\times nr)}{\mathbf{E}}$ is an $a \times nr$ dimensional matrix of compositional errors, $\boldsymbol{\phi}_c^{-1}$ is the logistic transformation applied column–wise, $\boldsymbol{\phi}_c$ is the log–ratio transformation, $\underset{(p\times w)}{\mathbf{T}} = [\boldsymbol{\tau}_1^p|\ldots|\boldsymbol{\tau}_w^p]$ is a $p \times w$ matrix of population diet compositions (table 6.2 gives the required matrices for the synthetic data), $\underset{(p\times n)}{\Gamma^m} = [\boldsymbol{\tau}_1^m|\ldots|\boldsymbol{\tau}_n^m]$ are samples from the multilevel distribution in our case the logistic normal with zero mean and covariance matrix $\Sigma_{\boldsymbol{\tau}}$, $\boldsymbol{\lambda}$ the p dimensional vector of fat contents for each prey type, $\underset{(a\times n_j)}{\mathbf{X}_j}$ is an $a \times n_j$ matrix of samples of the fatty acid profiles from the jth prey type, $\boldsymbol{\theta}_j$ is an a–dimensional vector consisting of the measure of location for the fatty acid profile of the jth prey type, $\underset{(1\times n_j)}{\mathbf{W}_{\mathbf{X}_j}}$ is a $1 \times n_j$ matrix of ones, $\underset{(a\times n_j)}{\mathbf{E}_{\mathbf{X}_j}}$ is an $a \times n_j$ dimensional matrix of compositional errors. We have similar definitions for the remaining parameters and observations for the model, however, note $\boldsymbol{\lambda}_j^v = (\lambda_j, 1-\lambda_j)'$ is the vector of fat and non–fat, similarly for the observations $\mathbf{Z}_j$. Let $\boldsymbol{\mu}_{\boldsymbol{\tau}_j} = \boldsymbol{\phi}(\boldsymbol{\tau}_j^p)$, or $\boldsymbol{\phi}_c(\mathbf{T}) = [\boldsymbol{\mu}_{\boldsymbol{\tau}_1}|\ldots|\boldsymbol{\mu}_{\boldsymbol{\tau}_w}]$, that is, $\boldsymbol{\mu}_{\boldsymbol{\tau}_j}$ are means of the logistic normal distributions of the individual populations. Note that, $\underset{(p\times n)}{\Gamma} = [\boldsymbol{\tau}_1|\ldots|\boldsymbol{\tau}_n$ consists of samples from the mixing distribution with zero mean added to the appropriate population mean.

To complete the model specification we assign the following prior distributions for the location parameters

$$\begin{aligned}
\pi(\boldsymbol{\mu}_{\boldsymbol{\tau}_r}|\boldsymbol{\eta}_r, \Sigma_{\boldsymbol{\mu}_{\boldsymbol{\tau}_r}}) &\sim \mathcal{MN}^{p-1}(\boldsymbol{\eta}_r, \Sigma_{\boldsymbol{\mu}_{\boldsymbol{\tau}_r}}), r = 1, \ldots, w \\
\pi(\boldsymbol{\theta}_j|\boldsymbol{\mu}_{\theta_j}, \Sigma_{\theta_j}) &\sim \mathcal{L}^a(\boldsymbol{\mu}_{\theta_j}, \Sigma_{\theta_j}), \;\; j = 1, \ldots, p, \\
\pi(\boldsymbol{\lambda}_j^v|\mu_{\boldsymbol{\lambda}_j}, \Sigma_{\boldsymbol{\lambda}_j}) &\sim \mathcal{L}^2(\boldsymbol{\mu}_{\boldsymbol{\lambda}_j}, \Sigma_{\boldsymbol{\lambda}_j}), \;\; j = 1, \ldots, p, \\
\pi(\boldsymbol{\theta}_{\mathbf{u}}|\boldsymbol{\mu}_{\theta_{\mathbf{u}}}, \Sigma_{\theta_{\mathbf{u}}}) &\sim \mathcal{L}^a(\boldsymbol{\mu}_{\theta_{\mathbf{u}}}, \Sigma_{\theta_{\mathbf{u}}}), \\
\pi(\boldsymbol{\theta}_{\mathbf{v}}|\boldsymbol{\mu}_{\theta_{\mathbf{v}}}, \Sigma_{\theta_{\mathbf{v}}}) &\sim \mathcal{L}^a(\boldsymbol{\mu}_{\theta_{\mathbf{v}}}, \Sigma_{\theta_{\mathbf{v}}}),
\end{aligned}$$

and for the covariance matrices

$$\begin{aligned}
\pi(\Sigma_{\boldsymbol{\tau}}|\delta_{\boldsymbol{\tau}}, \Psi_{\boldsymbol{\tau}})) &\sim \mathcal{IW}^{p-1}(\delta_{\boldsymbol{\tau}}, \Psi_{\boldsymbol{\tau}}), \\
\pi(\Sigma_{\mathbf{x}_j}|\delta_{\mathbf{x}_j}, \Psi_{\mathbf{x}_j}) &\sim \mathcal{IW}^{a-1}(\delta_{\mathbf{x}_j}, \Psi_{\mathbf{x}_j}), \;\; j = 1, \ldots, p, \\
\pi(\Sigma_{\mathbf{z}_j}|\delta_{\mathbf{z}_j}, \Psi_{\mathbf{z}_j}) &\sim \mathcal{IW}^{1}(\delta_{\mathbf{z}_j}, \Psi_{\mathbf{z}_j}), \;\; j = 1, \ldots, p, \\
\pi(\Sigma_{\boldsymbol{\epsilon}}|\delta_{\boldsymbol{\epsilon}}, \Psi_{\boldsymbol{\epsilon}}) &\sim \mathcal{IW}^{a-1}(\delta_{\boldsymbol{\epsilon}}, \Psi_{\boldsymbol{\epsilon}}), \\
\pi(\Sigma_{\mathbf{u}}|\delta_{\mathbf{u}}, \Psi_{\mathbf{u}}) &\sim \mathcal{IW}^{a-1}(\delta_{\mathbf{u}}, \Psi_{\mathbf{u}}), \\
\pi(\Sigma_{\mathbf{v}}|\delta_{\mathbf{v}}, \Psi_{\mathbf{v}}) &\sim \mathcal{IW}^{a-1}(\delta_{\mathbf{v}}, \Psi_{\mathbf{v}}).
\end{aligned}$$

The sampling distributions are given by

$$\begin{aligned}
\pi(\mathbf{y}_{is}|\boldsymbol{\Theta}, \boldsymbol{\tau}_i, \Sigma_{\boldsymbol{\epsilon}}, \boldsymbol{\lambda}, \boldsymbol{\theta}_{\mathbf{u}}, \boldsymbol{\theta}_{\mathbf{v}}) &\sim \mathcal{L}^a\left(\boldsymbol{\phi}\left(\left[\underset{(a\times p)(p\times 1)}{\boldsymbol{\Theta}\ \Gamma^i} \oplus_c \left(\underset{(a\times 1)}{\boldsymbol{\theta}_{\mathbf{v}}} \ominus \underset{(a\times 1)}{\boldsymbol{\theta}_{\mathbf{u}}}\right)\right]\right), \Sigma_{\boldsymbol{\epsilon}}\right), \quad \begin{array}{c} i = 1, \ldots, n \\ s = 1, \ldots, r \end{array} \\
\pi(\mathbf{x}_{jk}|\boldsymbol{\theta}_j, \Sigma_{\mathbf{x}_j}) &\sim \mathcal{L}^a(\boldsymbol{\phi}(\boldsymbol{\theta}_j), \Sigma_{\mathbf{x}_j}), \;\; j = 1, \ldots, p; \;\; k = 1, \ldots, n_j, \\
\pi(\mathbf{z}_{jk}|\boldsymbol{\lambda}^v_j, \Sigma_{\mathbf{z}_j}) &\sim \mathcal{L}^2(\boldsymbol{\phi}(\boldsymbol{\lambda}^v_j), \Sigma_{\mathbf{z}_j}), \;\; j = 1, \ldots, p; \;\; k = 1, \ldots, n_j, \\
\pi(\mathbf{u}_l|\boldsymbol{\theta}_{\mathbf{u}}, \Sigma_{\mathbf{u}}) &\sim \mathcal{L}^a(\boldsymbol{\phi}(\boldsymbol{\theta}_{\mathbf{u}}), \Sigma_{\mathbf{u}}), \;\; l = 1, \ldots, L, \\
\pi(\mathbf{v}_m|\boldsymbol{\theta}_{\mathbf{v}}, \Sigma_{\mathbf{v}}) &\sim \mathcal{L}^a(\boldsymbol{\phi}(\boldsymbol{\theta}_{\mathbf{v}}), \Sigma_{\mathbf{v}}), \;\; m = 1, \ldots, M,
\end{aligned}$$

and the mixing distribution is given by

$$\pi(\boldsymbol{\tau}_i|\mathbf{0}, \Sigma_{\boldsymbol{\tau}}) \sim \mathcal{L}^p(\mathbf{0}, \Sigma_{\boldsymbol{\tau}})$$

where $\mathbf{0}$ is a $p-1$ dimensional vector of zeros. We assume that the mixing distribution has the same covariance matrix in each population, although, this assumption could be easily adapted to allow a different mixing covariance per population. By comparison with more traditional analysis of variance models our individual diet model can be seen as a mixed model with the design matrix $\mathbf{W}$ giving the fixed effects and the covariance matrix $\Sigma_{\boldsymbol{\tau}}$ playing the role of controlling the amount of variability in the random effects.

Figure 6.12 gives the directed acyclic graph for this model. For the most part the DAG is similar to the DAG for the constant diet model, with some exceptions. As noted previously each predator now has its own diet composition and as a consequence we need an additional level of hierarchy for this part of the model. Namely, a random effects precision matrix and mean vector or in this case matrix to allow for the population level effects described in $\mathbf{W}$. The details of the reversible systematic scan Adaptive–Metropolis–within–Gibbs algorithm are described in Appendix B.

The full conditional distribution for $\boldsymbol{\mu}_{\boldsymbol{\tau}_s}, s = 1, \ldots, w$ is a multivariate normal distribution (see Rowe, 2003, for a derivation). Essentially, we have a multivariate regression where $\boldsymbol{\tau}_i$'s play the role of the observations and the regression parameters are $\mathbf{SW}$ where $\mathbf{S} = [\boldsymbol{\mu}_{\boldsymbol{\tau}_1}| \ldots |\boldsymbol{\mu}_{\boldsymbol{\tau}_w}]$ is a $(p-1) \times w$ matrix.

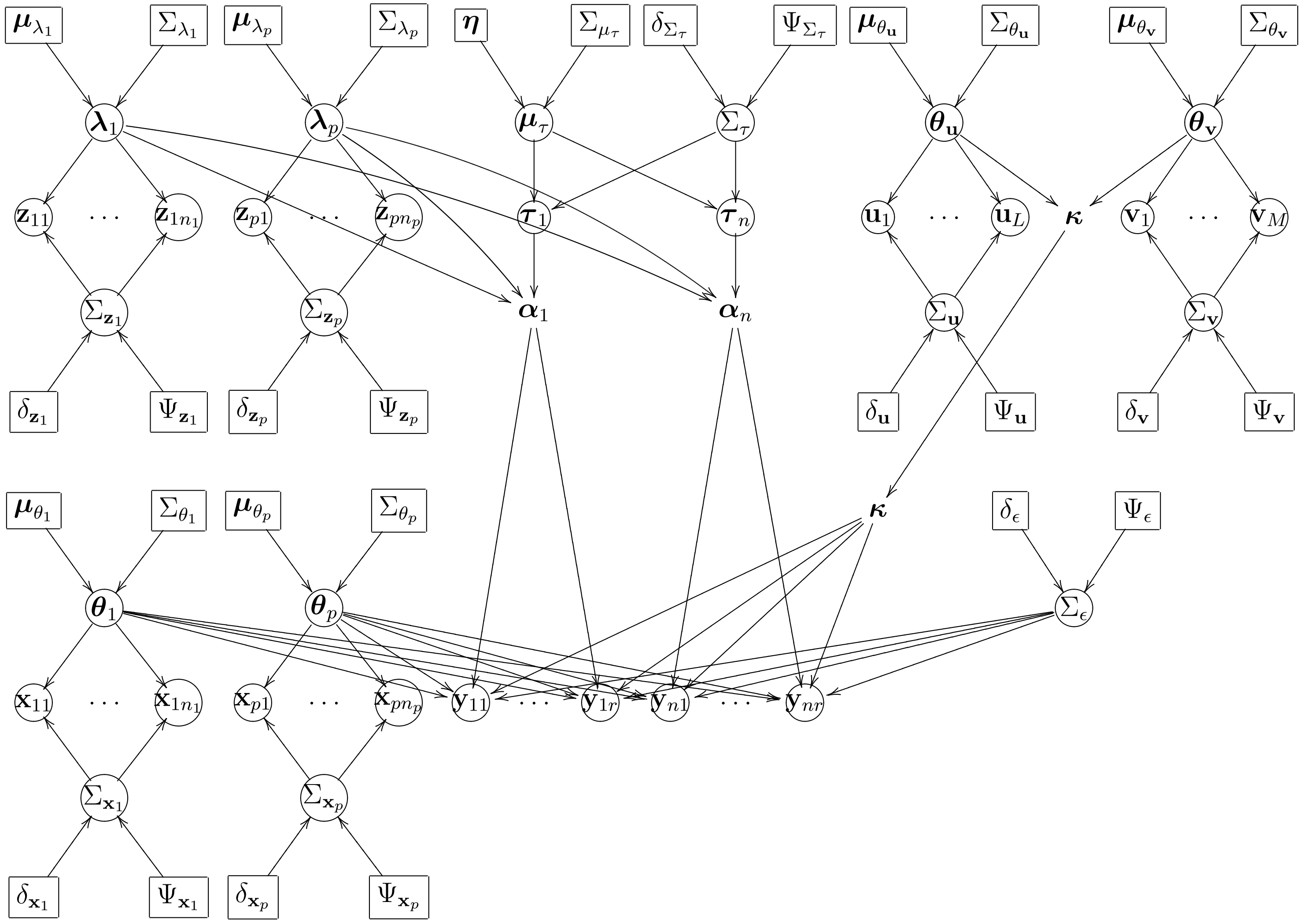

Figure 6.12: DAG: Individual Diet Model with calibration and fat content. Note $\boldsymbol{\kappa} = \boldsymbol{\theta}_\mathbf{v} \oplus \boldsymbol{\theta}_\mathbf{u}$, $\boldsymbol{\alpha}_i = \boldsymbol{\tau}_i \oplus \boldsymbol{\lambda}$ and $\mathbf{Y} = [(\boldsymbol{\Theta}\Gamma \otimes \mathbf{U}) \oplus_c (\boldsymbol{\theta}_\mathbf{v} \ominus \boldsymbol{\theta}_\mathbf{u})] \oplus_c \mathbf{E}$. Nodes that are not contained in circles or squares are derived variables are used to simplify the graph.

6.4.1 Synthetic Data

We illustrate individual diet level model by constructing synthetic data in a similar fashion to the constant data model and then carry out Bayesian inference to assess how well we can reconstruct the true parameters. We are also interested in determining whether multiple replicates on the same individual predator improve the inference. To explore this we generate two replicate samples for each predator and use the first observation for the case of one replicated and both for the two replicate case.

The mixing distribution was logistic normal with two settings for $\sigma^2_{\boldsymbol{\tau}}$ corresponding to the following structured covariance matrix:

$$\Sigma_{\boldsymbol{\tau}} = \sigma^2_{\boldsymbol{\tau}} \left(\mathbf{I}_{p-1} + \mathbf{J}_{p-1} \right)$$

with $\sigma^2_{\boldsymbol{\tau}} = 0.5, 1.0$. All other settings were exactly the same as in the constant diet model. The same permutation of the species labels was used to assess the affect of an ill–conditioned source matrix, Θ.

Prior parameters were assigned using the same algorithm as the constant model to give relative vague but proper priors. Ideally maximum entropy priors would be assigned, however, this was not carried out for the thesis.

Starting values were assigned in a similar way as for the constant model with the exception of the mixing distribution. To assess the affect of burn in for the mixing vectors $\boldsymbol{\tau}_i$'s we started the chains at two settings: starting at the “true” values and also initializing them at the compositional center.

The MCMC algorithm was run in exactly the same way as the constant diet model with appropriate modifications including the same three part adaptive scheme. See Appendix B for the full individual diet reversible systematic scan Metropolis–within–Gibbs algorithm.

6.4.2 Results

In this section we give the results of applying the individual diet model to the synthetic data sets discussed in the previous section. We have two levels of mixing variability, $\sigma^2_{\boldsymbol{\tau}} = 0.5, 1.0$, whether or not the fatty acids have been permuted or not and one or two replicates. We also give results for the fully Bayesian approach and the approximate method where we turn off the information flow between the predator and prey. Though as discussed in the constant diet model section it isn't clear what posterior distribution we are actually sampling from. More work needs to be done in this area. We also give a comparison with the original QFASA method developed by Iverson et al. (2004) as it gives individual diet estimates.

Figure 6.13 gives boxplots of the true diet compositions, $\boldsymbol{\tau}_i$, the posterior means computed on the log–ratio scale and transformed back via the logistic transformation for the fully Bayesian approach and the approximate approach as well as the QFASA point estimates. We only give the results for one of the diets (Spring Males), the other diets have similar patterns. Panels a) and c) give the results for the permuted fatty acid case. The fully Bayesian method agrees quite closely with the actual diets, suggesting that the method performs quite well with well conditioned source matrices. The approximate Bayesian method, also does quite well but there are slight departures from the actual diet, though they are hard to detect on the original proportional scale. The original QFASA method also does quite well, however, it tends to under estimate the amount of Pollock in the diet and over estimate the amount of Redfish. However, the picture is quite different in panels b) and d), for the case of the original fatty acid profiles, that is, the case where the condition number of the prey fatty acid profiles is quite high, indicating a large amount of dependence among the profiles. There are several discrepancies between the fully Bayesian approach and the actual diets, particularly for some of the rarer prey items in diet 1. The picture is worse for the approximate case with larger discrepancies even for the larger diet items. The picture for the original QFASA is not nearly as good, with large discrepancies between the QFASA estimates and the actual diet.

To make the comparison between the fully Bayesian method and the original QFASA model more fair we also initialized the MCMC algorithm at the compositional center. The results are given in figure 6.14. It is clear that starting values are not a major factor in having the Bayesian method outperform the original QFASA model as the Bayesian posterior means are still much closer to the true $\boldsymbol{\tau}_i$'s than the QFASA method.

Figure 6.15 gives another comparison between the actual individual diets and their posterior mean values and the QFASA diet estimates. It gives boxplots of the compositional distance between the actual diet and each of the three estimates discussed previously. Note as in the previous case, we are only considering the case with one replicate. The figure shows boxplots for all 300 diet estimates for each of the three methods and the actual $\boldsymbol{\tau}_i$. Recall the compositional distance compares the diets on the log–ratio scale, therefore,

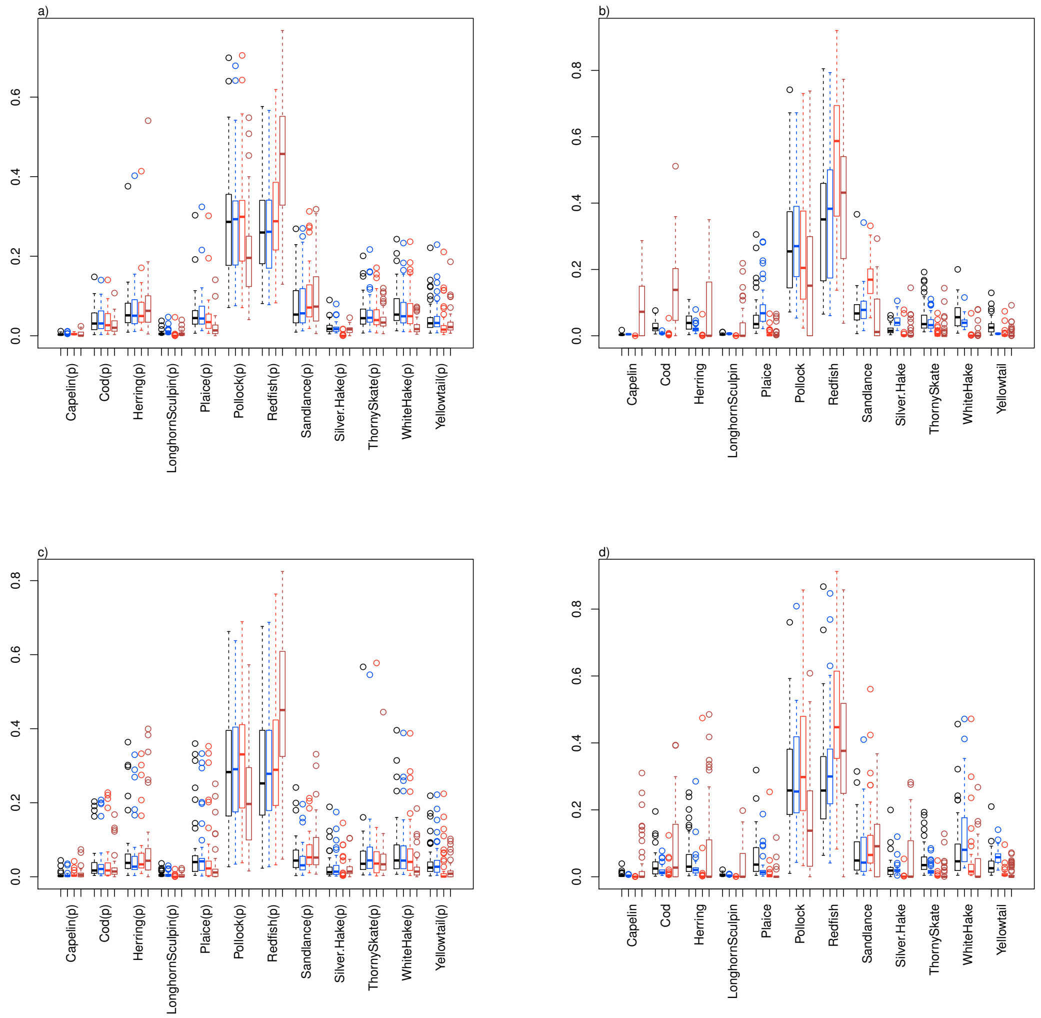

Figure 6.13: Boxplots of $\boldsymbol{\tau}_i$ for Males sampled in the Spring (other 5 diet combinations are similar but not shown) for the synthetic data from diet 1 (see table 6.2). Panels a) through d) represent the following combinations respectively: $\sigma^2_{\boldsymbol{\tau}} = 0.5$ and permuted fatty acids; $\sigma^2_{\boldsymbol{\tau}} = 0.5$ and original fatty acids; $\sigma^2_{\boldsymbol{\tau}} = 1.0$ and permuted fatty acids; $\sigma^2_{\boldsymbol{\tau}} = 1.0$ and original fatty acids. For each panel and species the individual boxplots represent: the true $\boldsymbol{\tau}_i$'s (in black), posterior means for the fully Bayesian approach (in blue), posterior means for the approximate Bayesian approach (in red) and the QFASA estimates from Iverson et al. (2004) (in brown). The Bayesian approaches are based on 10,000 posterior samples, see the text for details.

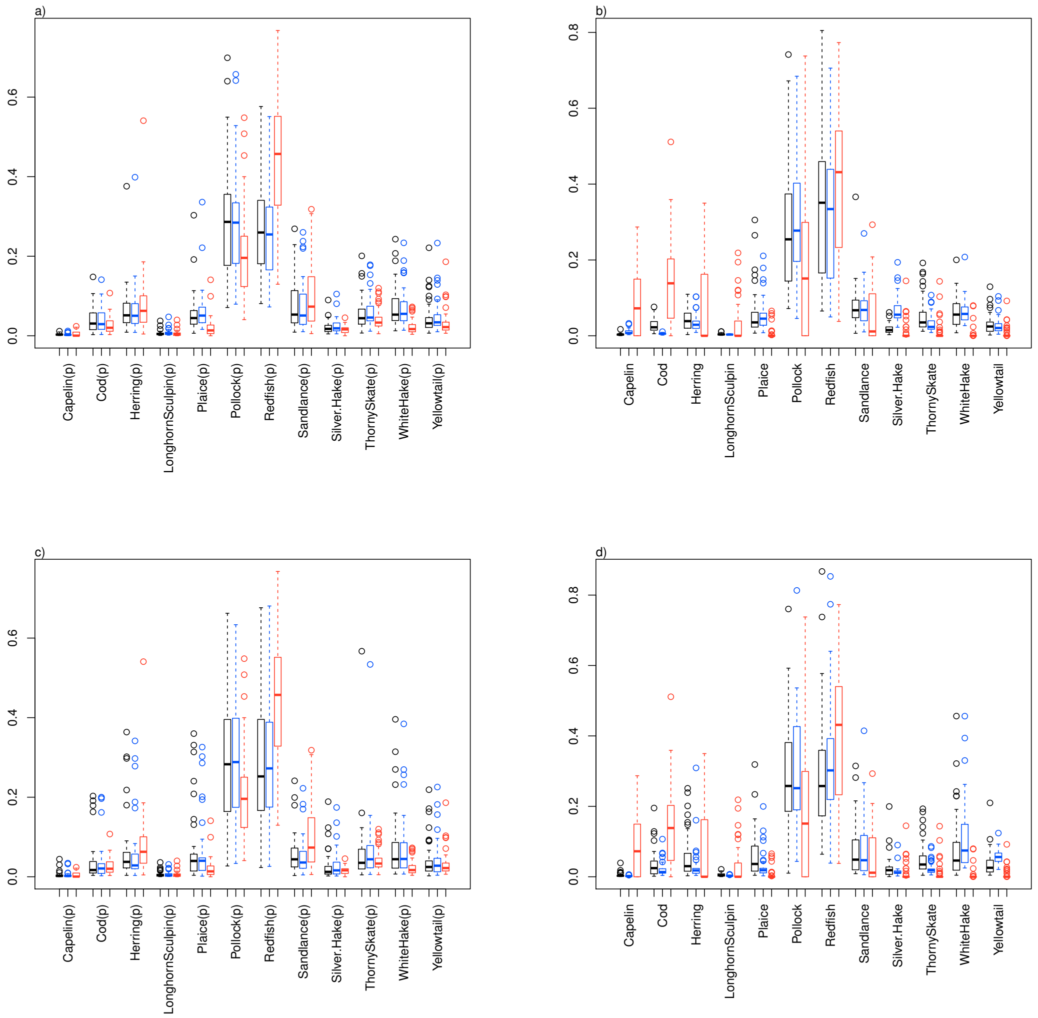

Figure 6.14: Boxplots of $\boldsymbol{\tau}_i$ for Males sampled in the Spring (other 5 diet combinations are similar but not shown) for the synthetic data from diet 1 (see table 6.2) starting the Bayesian method at the compositional zero. Panels a) through d) represent the following combinations respectively: $\sigma^2_{\boldsymbol{\tau}} = 0.5$ and permuted fatty acids; $\sigma^2_{\boldsymbol{\tau}} = 0.5$ and original fatty acids; $\sigma^2_{\boldsymbol{\tau}} = 1.0$ and permuted fatty acids; $\sigma^2_{\boldsymbol{\tau}} = 1.0$ and original fatty acids. For each panel and species the individual boxplots represent: the true $\boldsymbol{\tau}_i$'s (in black), posterior means for the fully Bayesian approach (in blue) and the QFASA estimates from Iverson et al. (2004) (in red). The Bayesian approaches are based on 10,000 posterior samples, see the text for details.

differences in small diet components get magnified compared to the original scale. For the case with little dependence (panels a and c) the fully Bayesian approach has the best overall estimates followed by QFASA and the approximate Bayesian solution having the worst performance, which is a little surprising. The ordering is slightly different for the non–permuted data, with the QFASA and approximate Bayesian method reversing positions. A similar pattern emerges for the Bayesian solution versus QFASA when the Bayesian solution is started at the compositional center as shown in figure 6.16

Figure 6.17 gives boxplots of the compositional distance between the true mean diet composition $\boldsymbol{\mu}_{\boldsymbol{\tau}_j}$ for each season and sex combination and the 10,000 MCMC samples for the fully Bayesian approach versus the approximate solution for the one and two replicate cases. It is clear, that the fully Bayesian approach does better in reconstructing the diet composition in all situations. When the source matrix is well conditioned panels a) and c) adding a second observation improves the Bayesian solution, however, the situation is not as clear for the ill–conditioned case, though it most cases it does improve the inference. The approximate solution, doesn't seem to make use of the second observation well at all, in almost all situations it moves the posterior means further away from the true mean compositions. This is even more pronounced in the original non–permuted prey profiles.

Figures 6.18 and 6.19 give boxplots of the trace of the covariance matrix of the mixing distribution,$\Sigma_{\boldsymbol{\tau}}$, and the covariance of the compositional error distribution , $\Sigma_{\boldsymbol{\epsilon}}$, respectively. Note the logarithm scale on the trace of the gamma plots. It is quite clear that the approximate method does not recover the covariance matrix breakdown between the mixing distribution and the error distribution, while the fully Bayesian approach does quite well in recovering the correct covariance matrix breakdown as measured by the trace. Note the large discrepancies in the scale of the two variances, this due to the fact that the mixing distribution is parametrized on the log–ratio scale, while the mixing actually occurs on the original scale.

As with the constant diet model, we are still in a quandary as to how to optimally proceed in the presence of ill–conditioned prey matrices $\boldsymbol{\Theta}$. One possibility is to remove species (columns) of the matrix to make it better conditioned, however, we might remove an ecologically important prey item. If we happen to remove an important prey item the the fully Bayesian approach may in fact modify one of the remaining prey items to compensate. Kashiwagi (2004) presents a model which allows the incorporation of a

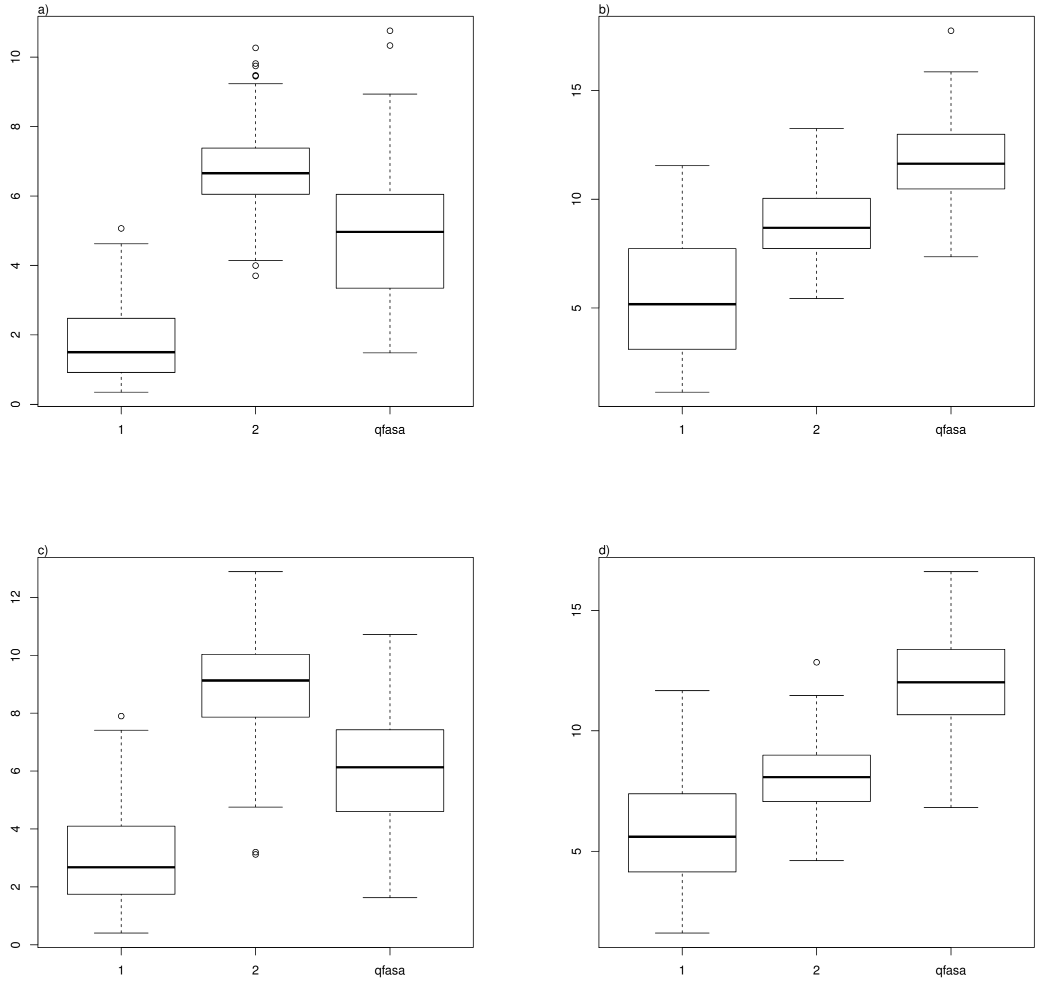

Figure 6.15: Boxplots of the compositional distance between the true $\boldsymbol{\tau}_i$ parameters and the posterior means of 10,000 MCMC samples. Panels a) through d) represent the following combinations respectively: $\sigma^2_{\boldsymbol{\tau}} = 0.5$ and permuted fatty acids; $\sigma^2_{\boldsymbol{\tau}} = 0.5$ and original fatty acids; $\sigma^2_{\boldsymbol{\tau}} = 1.0$ and permuted fatty acids; $\sigma^2_{\boldsymbol{\tau}} = 1.0$ and original fatty acids. The boxplots in each panel represent the following: 1) the fully Bayesian solution with one replicate, 2) the approximate Bayesian solution with one replicate and qfasa) the QFASA solution. Note each boxplot represents the 300 $\boldsymbol{\tau}_i$'s (50 for each of 6 diet combinations).

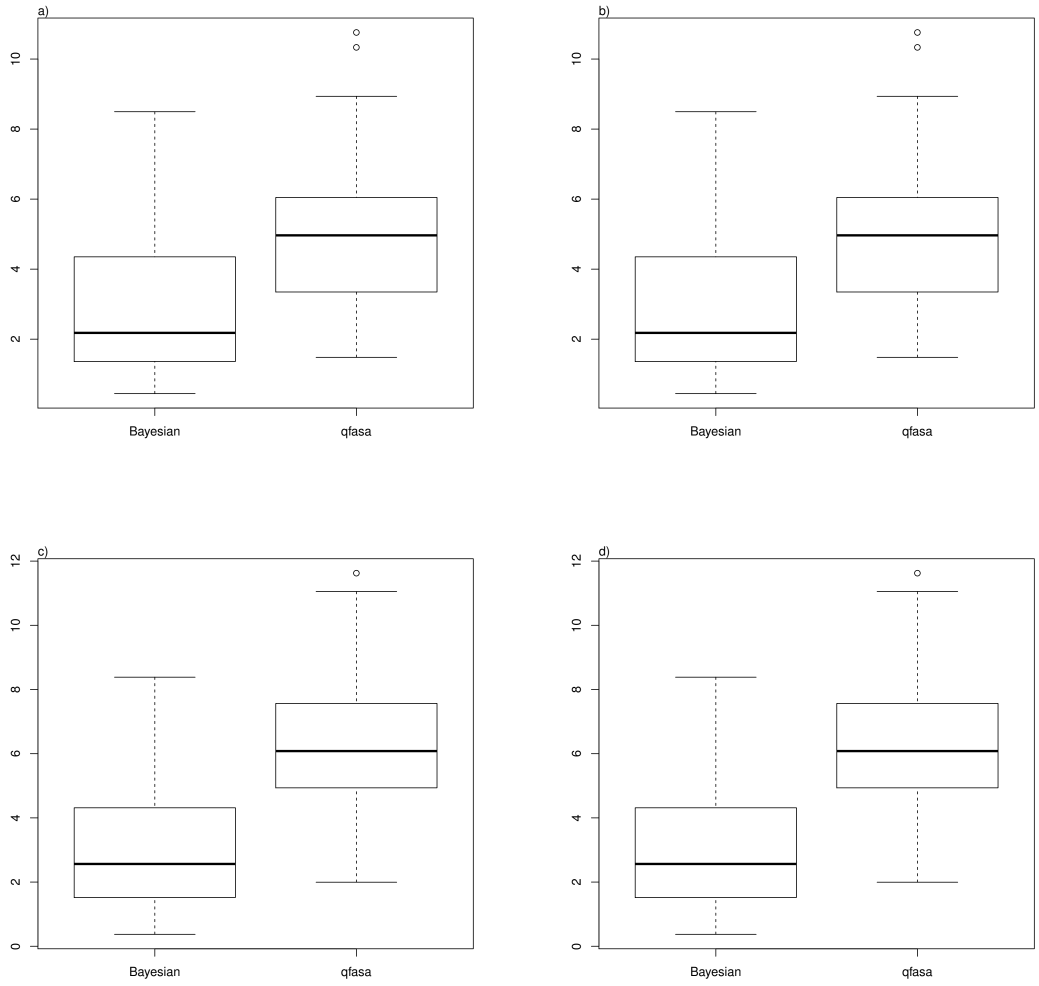

Figure 6.16: Boxplots of the compositional distance between the true $\boldsymbol{\tau}_i$ parameters and the posterior means of 10,000 MCMC samples with the fully Bayesian approach initialized at the compositional center. Panels a) through d) represent the following combinations respectively: $\sigma^2_{\boldsymbol{\tau}} = 0.5$ and permuted fatty acids; $\sigma^2_{\boldsymbol{\tau}} = 0.5$ and original fatty acids; $\sigma^2_{\boldsymbol{\tau}} = 1.0$ and permuted fatty acids; $\sigma^2_{\boldsymbol{\tau}} = 1.0$ and original fatty acids. The boxplots in each panel represent the following: 1) the fully Bayesian solution with one replicate and qfasa) the QFASA solution. Note each boxplot represents the 300 $\boldsymbol{\tau}_i$'s (50 for each of 6 diet combinations).

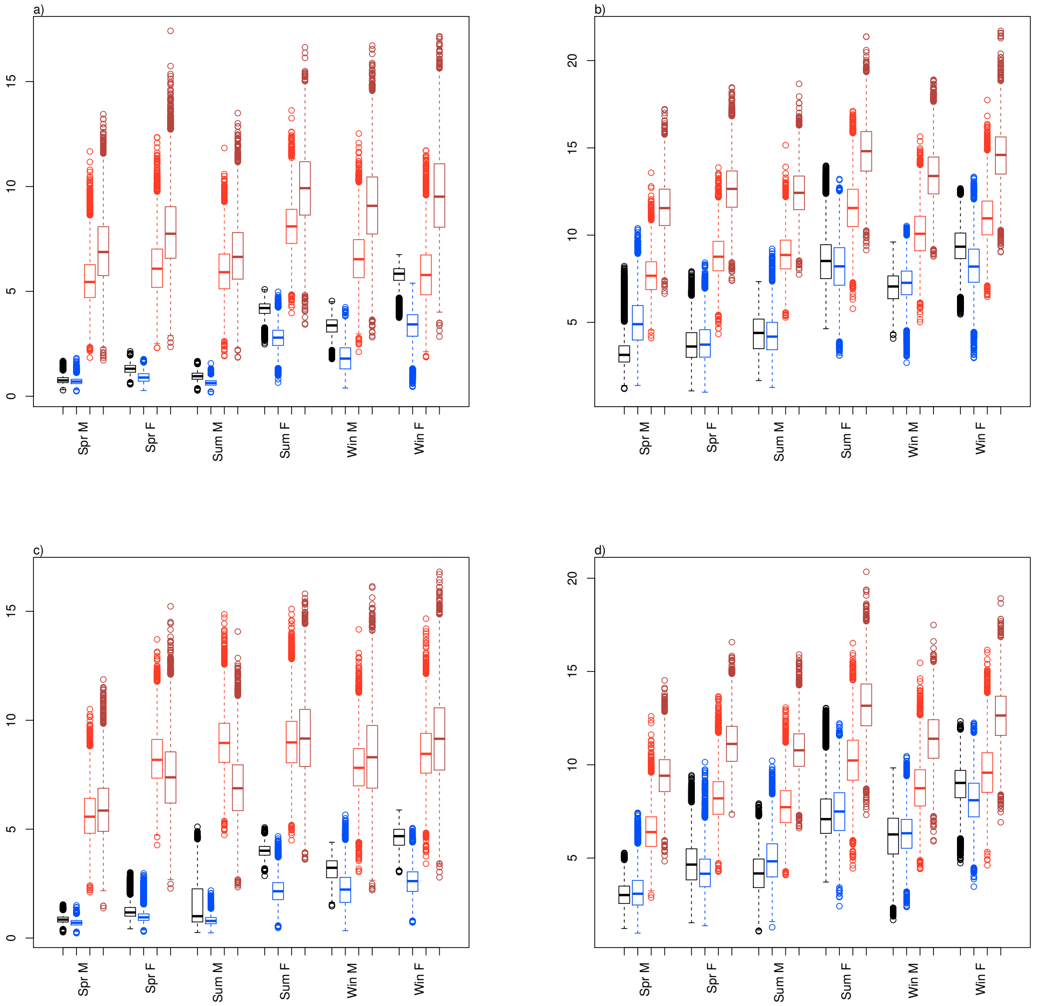

Figure 6.17: Boxplots of the compositional distance between the true mean diet composition $\boldsymbol{\mu}_{\boldsymbol{\tau}_j}$ and the 10,000 MCMC samples. Panels a) through d) represent the following combinations respectively: $\sigma^2_{\boldsymbol{\tau}} = 0.5$ and permuted fatty acids; $\sigma^2_{\boldsymbol{\tau}} = 0.5$ and original fatty acids; $\sigma^2_{\boldsymbol{\tau}} = 1.0$ and permuted fatty acids; $\sigma^2_{\boldsymbol{\tau}} = 1.0$ and original fatty acids. For each panel and diet composition the individual boxplots represent: the fully Bayesian approach with one replicate (in black), the fully Bayesian approach with two replicates (in blue), the approximate Bayesian solution with one replicate (in red) and the approximate Bayesian solution with two replicates (in brown).

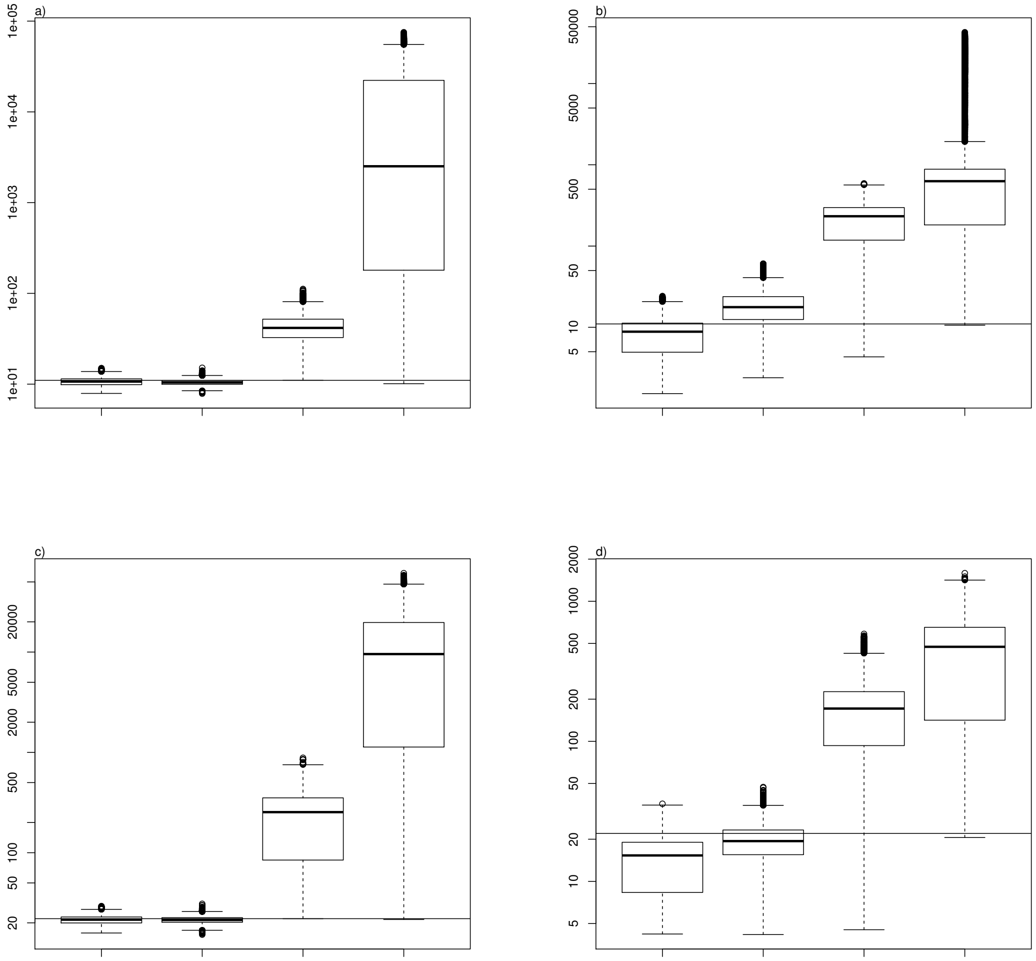

Figure 6.18: Boxplots of the trace of the $\Sigma_{\boldsymbol{\tau}}$ mixing distribution for the 10,000 MCMC samples. Panels a) through d) represent the following combinations respectively: $\sigma^2_{\boldsymbol{\tau}} = 0.5$ and permuted fatty acids; $\sigma^2_{\boldsymbol{\tau}} = 0.5$ and original fatty acids; $\sigma^2_{\boldsymbol{\tau}} = 1.0$ and permuted fatty acids; $\sigma^2_{\boldsymbol{\tau}} = 1.0$ and original fatty acids. For each panel the individual boxplots represent: the fully Bayesian approach with one replicate, the fully Bayesian approach with two replicates, the approximate Bayesian solution with one replicate and the approximate Bayesian solution with two replicates. Note the logarithm scale on plots.

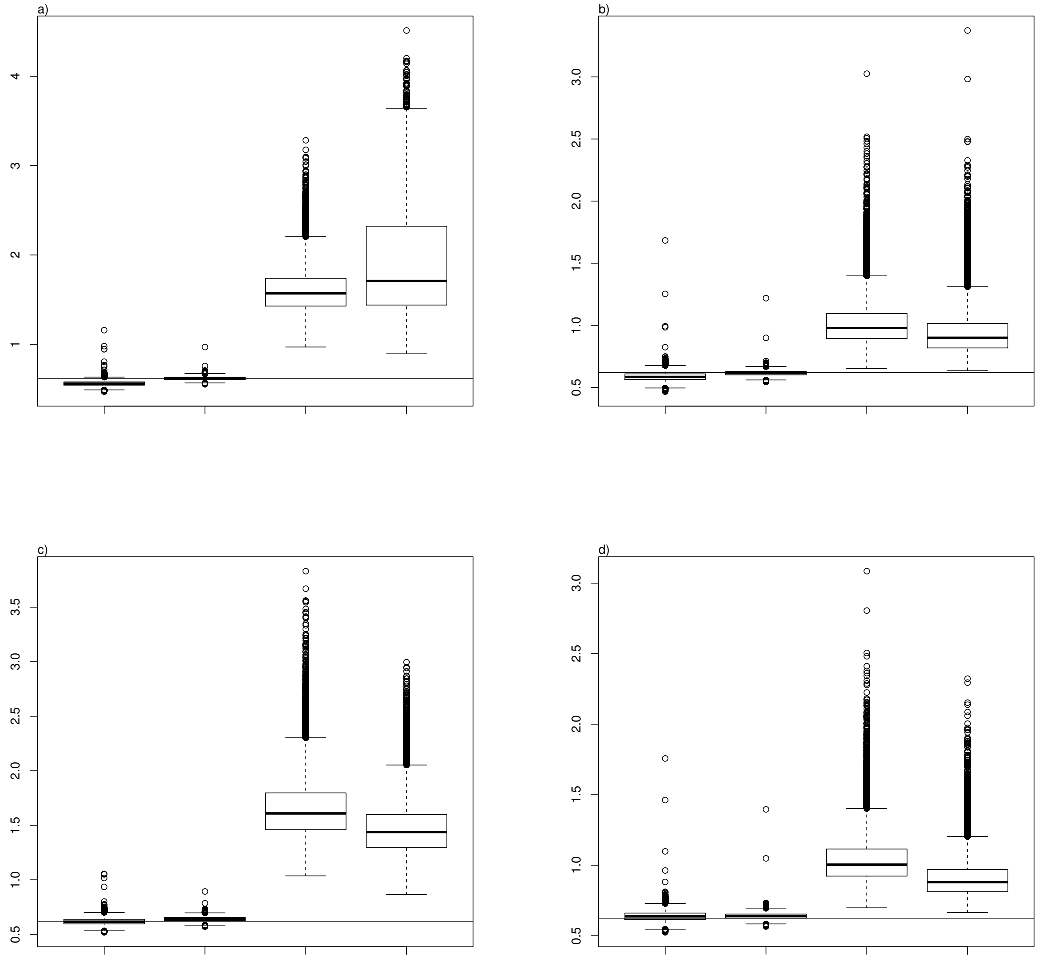

Figure 6.19: Boxplots of the trace of the Σ_ϵ for the 10,000 MCMC samples. Panels a) through d) represent the following combinations respectively: $\sigma^2_{\boldsymbol{\tau}} = 0.5$ and permuted fatty acids; $\sigma^2_{\boldsymbol{\tau}} = 0.5$ and original fatty acids; $\sigma^2_{\boldsymbol{\tau}} = 1.0$ and permuted fatty acids; $\sigma^2_{\boldsymbol{\tau}} = 1.0$ and original fatty acids. For each panel the individual boxplots represent: the fully Bayesian approach with one replicate, the fully Bayesian approach with two replicates, the approximate Bayesian solution with one replicate and the approximate Bayesian solution with two replicates.

potentially unknown but unobserved source. However, as noted by Kashiwagi (2004) there are some identifiability issues with his model and due to time constraints we were not able to implement his approach.

Another possibility is to amalgamate prey items after the fact, however, due to the nature of the log–ratio transformation it isn't entirely clear how one should do this (see Aitchison, 2003). We amalgamate on the original scale, rather than the log–ratio scale, though there is no theoretical reason for choosing one scale over another. Rather than repeat the amalgamation exercise again we refer the reader to the constant data section for an illustration.

Still another possibility, which in some sense would be preferable, is to have a variable selection procedure to eliminate prey items that were not contributing to the predator and hence reduce the dimension. For instance using the Reversible Jump/trans–dimensional MCMC approach suggested by Green (1995). However, this was not approached in the thesis, due to time constraints.

6.5 Applications

The previous discussions on synthetic data illustrated the flexibility of the constant diet model and the individual diet model, particularly the fully Bayesian approach. That is, the predator has valuable information about the prey profiles in its own fatty acid signature. The prey samples can be thought of as a prior for the source profiles consumed by the predator, that is, similar to the approach taken by Billheimer (2001). However, the interpretation can be problematic, as the predator may substantially modify the source profiles to an extent that they are no longer recognizable as a known species. For this reason, we introduced the approximate Bayesian method of not allowing information to flow from the predator to the prey source profiles, the calibration coefficients and the fat contents. However, the validity of this approach is still in question, as is isn't clear what posterior distribution we are sampling from. Our approximate technique is similar in spirit to the approach modularization technique taken by Liu et al. (2009).

We further illustrate the applicability of the constant and individual diet models with one captive study and one field study. The captive study was conducted on Murres and Kittiwakes collected from the Pribilofs Islands, Alaska (see Iverson et al., 2007) . The field study was conducted on Harbour Seals from Sable Island, Nova Scotia using critter cams (see Iverson et al., 2004). We use the common names for species rather than their scientific counterparts where available. Also note that all samples of both predators and potential prey items were analyzed for fat content and their constituent fatty acids using the methods described in Iverson et al. (2004).

Our strategy is to fit both the fully Bayesian method and the approximate Bayesian method and compare the results. But before proceeding to the real data sets we need to address some other model complications. The sample sizes of both the predators and prey are typically smaller than we considered in the synthetic data discussions making the inference of the various covariance matrices problematic. With this in mind we considered two parametrized/patterned covariance matrices, one of which we have discussed previously.

The first covariance matrix is given by

$$\underset{(a-1)\times(a-1)}{\Sigma_1} = \sigma^2\left[\mathbf{I}_{a-1} + \mathbf{J}_{a-1}\right] = \begin{bmatrix} 2\sigma^2 & \sigma^2 & \cdots & \sigma^2 & \sigma^2 \\ \sigma^2 & 2\sigma^2 & \cdots & \sigma^2 & \sigma^2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \sigma^2 & \sigma^2 & \cdots & 2\sigma^2 & \sigma^2 \\ \sigma^2 & \sigma^2 & \cdots & \sigma^2 & 2\sigma^2 \end{bmatrix} \tag{6.1}$$

where $\mathbf{I}_{a-1}$ is the $(a-1)\times(a-1)$ identity matrix and $\mathbf{J}_{a-1}$ is the $(a-1)\times(a-1)$ matrix of ones. This patterned covariance matrix can be thought of as the closure of a independent log–normal variables with common variance, σ^2. The second patterned covariance matrix is given by

$$\begin{aligned} \underset{((a-1)\times(a-1)}{\Sigma_2} &= \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 & 0 \\ 0 & \sigma_2^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma_{d-2}^2 & 0 \\ 0 & 0 & \cdots & 0 & \sigma_{d-1}^2 \end{bmatrix} + \sigma_a^2\mathbf{J} \\ &= \begin{bmatrix} \sigma_1^2+\sigma_a^2 & \sigma_a^2 & \cdots & \sigma_a^2 & \sigma_a^2 \\ \sigma_a^2 & \sigma_2^2+\sigma_a^2 & \cdots & \sigma_a^2 & \sigma_a^2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \sigma_a^2 & \sigma_a^2 & \cdots & \sigma_{a-2}^2+\sigma_a^2 & \sigma_a^2 \\ \sigma_a^2 & \sigma_a^2 & \cdots & \sigma_a^2 & \sigma_{a-1}^2+\sigma_a^2 \end{bmatrix} \end{aligned} \tag{6.2}$$

and can be thought of as the closure of a independent log–normals with variances $\sigma_1^2, \ldots, \sigma_a^2$. The first matrix was used to represent our prior knowledge of scale matrices and covariance matrices. Both covariance matrices imply a very strong form of independence among the elements of the composition, however, with limited data they are necessary, unless one is willing to impose a lot of prior information.

6.5.1 Captive Murre and Kittiwakes

In this section we apply the constant diet model to a captive seabird study described in Iverson et al. (2007). We apply the fully Bayesian approach and the approximate version of the model.

The experiment consisted of collecting twenty six partially incubated eggs of Common Murres (COMU) and thirteen Red Legged Kittwakes (RLKI) from Cook Inlet and the Pribilof Islands, Alaska respectively. Once hatched, the chicks were raised in individual nests and hand fed controlled diets. The Common Murres were fed for 45 days and a synsacral adipose tissue was obtained on the last day. Half of the COMU were fed silverside for the full 45 days. The other half was fed silverside for the first 11 days and then switched to a diet of rainbow smelt. Red legged Kittiwakes were fed for 42 days and biopsied on the final day. All the RLKI were fed the same diet for the first 15 days which consisted of 8 parts herring to 2 parts silverside. Seven of the chicks where then fed silverside for remainder of the experiment while the other 6 chicks were fed rainbow smelt. Additionally samples of the prey were collected from the lots that were experimentally fed: fifteen Silverside, fifteen rainbow smelt and ten herring.

Since $n = 13$ Common Murres were fed a constant diet of silverside for 45 days we used these samples to construct calibration coefficients which mimic predator biosynthesis as previously described. Iverson et al. (2007) also modeled the thirteen Common Murres that were used in the calibration study and we follow that approach here. Note that the same silverside samples were used for both calibration and as part of the prey base. Ideally this would not be the case, but we follow their approach.

There are four distinct diet scenarios: two levels of species by two levels of diets fed during the second half of the feeding regime. Accordingly, we model the (n=39) specimens with a constant diet model with four populations. We used a slightly modified version of extended dietary set considered in Iverson et al. (2007) which consists of 40 fatty acids for see figure 6.20, we omitted two fatty acids 16:3n-1 and 22:2n-6 as they were constant across prey species.

Given the large number of fatty acids relative to the sample size of the prey and predator, we chose to model the prey covariance matrices $\Sigma_{\mathbf{x}_i}, i = 1, 2, 3$, the predator covariance matrix $\Sigma_{\boldsymbol{\epsilon}}$ and the calibration covariance matrices $\Sigma_{\mathbf{u}}$ and $\Sigma_{\mathbf{v}}$ by the patterned covariance matrix given in equation 6.1 with a distinct σ^2 for each of the prey types, predators and the calibration data. It is relatively well known in the biological literature that the fat content of prey item can vary from approximately 0.22% to approximately 28.0% fat, we therefore further restrict the fat content to this range by using the methods of assigning priors in the synthetic data sections and setting the mean of the prior distribution to be the center of the

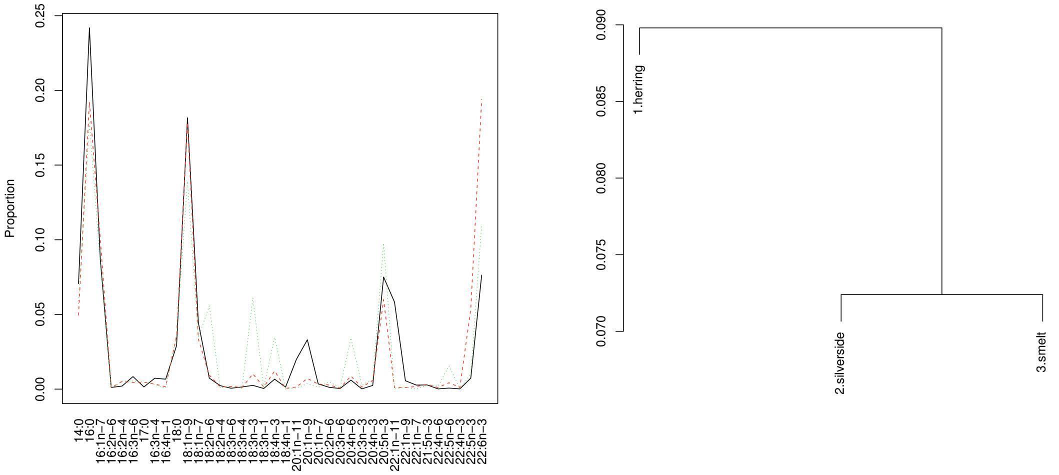

Figure 6.20: The left hand panel shows the source profiles for the three species fed in the captive seabird study and the right hand panel shows a hierarchical clustering of the standardized distance using the average linkage method. Source matrix condition number is 43.96.

range.

Figure 6.20 depicts the three source profiles for Herring, silverside and smelt and the hierarchical clustering of $1 - \cos(\theta)$. The condition number of the source matrix 43.96939, which is much more manageable than the non–permuted synthetic data set. This is also confirmed by examining the individual source profiles.

To compare the fully Bayesian approach and the approximate method we monitor the compositional distance between the posterior distribution (in steady state) and the sample means of the prey fatty acid profiles, fat content and the calibration profiles (both predator and prey). If one of the prey profiles is drastically different from its sample value, this could indicate one of several possibilities: there isn't enough sample information to dominate the "vague" prior used; one of the prey items that was sampled does not represent what the predator actually consumed and the fully Bayesian method is modifying an existing one to compensate for the discrepancy. Calibration experiments are typically quite small and potentially variable due to inherent variability in the prey samples, deviations in the calibration parameters and their respective sample means would indicate substantial variability in the calibration coefficients or alternatively could indicate that the present calibration coefficients are not suitable to the current problem. The most likely cause of divergence between the various parameters and their sample means is a small sample size, as we didn't experience any large departures in the synthetic data.

More investigation is needed into the potential causes, however, we adopt the approach that if the fully Bayesian approach deviates substantively from the sample means, we take this as evidence that the fully Bayesian model is not identifiable without stronger prior information or equivalently larger number of observations on its component parts. This is similar to the way of addressing the identifiability issue proposed by Billheimer (2001).

6.5.1.1 Starting Values and Prior Assignment

The covariance matrices are now parametrized in terms of equation 6.1, that is, one parameter. We assign inverse scaled Chi–squared distributions (or Inverse–Wishart distributions) with one degree of freedom to ensure a proper prior distribution with scale parameter of 100. The prior distributions for the location parameters were set in a similar way to the synthetic data. That is, the $\boldsymbol{\theta}_j$'s, $\boldsymbol{\theta}_{\mathbf{u}}$ and $\boldsymbol{\theta}_{\mathbf{v}}$ were assigned a logistic normal distribution centered at the compositional zero with covariance matrices chosen to limit the range of the prior as described in section 6.3.5. The $\boldsymbol{\lambda}_j^v$ were further restricted to lie in the range of plausible values of fat content. Note we could have pursued more stringent priors for the other parameters as well, but that avenue was not pursued here.

We started the adaptive Metropolis–Hastings–within–Gibbs algorithm in a similar fashion for the synthetic data, starting the $\boldsymbol{\theta}_j$, $\boldsymbol{\lambda}_j^v$, $\boldsymbol{\theta}_{\mathbf{u}}$, $\boldsymbol{\theta}_{\mathbf{v}}$ at the sample means. We started the, mixing vector $\boldsymbol{\tau}$, at the compositional zero. Note this is a slight departure from what Gelman et al. (2003) recommends as we didn't sample from an overdispersed distribution to generate starting values.

6.5.1.2 Convergence Diagnostics

The convergence of the chain was assessed using the standard techniques in the R–package, CODA. As the standard techniques are meant to apply to a chain that isn't adaptive we only use and assess the convergence of the third part of our hybrid strategy.

6.5.1.3 Results

After running two chains through our hybrid algorithm for stage lengths of 5×10^5, 1×10^6 and 1×10^6, there was no evidence of non–convergence to the stationary distribution for the third stage. Note that all stages were thinned by a factor of 100 to reduce storage requirements.

Figure 6.21 gives the posterior mean and 95% credible intervals for the diet composition vector $\boldsymbol{\tau}_w$ for each of the four treatments (COMU fed silverside/silverside, COMU fed silverside/rainbow smelt, RLKI fed silverside&herring/silverside, RLKI fed silverside&herring/rainbow smelt). Both the fully Bayesian approach and the approximate approach are presented.

The posterior means and credible intervals agree quite well between the fully Bayesian approach and the approximate method, with the exception that fully Bayesian approach indicates more smelt in the RLKI fed herring/smelt initially then fed silverside. However, both methods indicate some evidence of smelt in the COMU group fed Silverside for the duration of the experiment, but the amount in both cases is slight. Note that neither variant shows any evidence of a herring signature in either RLKI group. Both methods agree quite well with the diet consumed in each group, with the exception of not having much residual traces of the initial diet as we would expect.

Recall, $\Sigma_{\boldsymbol{\epsilon}} = \sigma^2_{\boldsymbol{\epsilon}}(\mathbf{I}_{a-1} + \mathbf{J}_{a-1})$ is the covariance matrix of the compositional errors, which allows us to compare the overall fits of the fully Bayes approach and the approximate approach. The 95% credible intervals for the fully Bayesian and approximate approach are $\sigma^2_{\boldsymbol{\epsilon}}$ are (0.138,0.161) and (0.252,0.318) respectively.

To explore the discrepancy in fits between the two approaches, we compared how the fully Bayesian and approximate approaches deviated from the sample means. Examination of the compositional distances between the sample means and the posterior showed that the fully Bayesian approach modified the fatty acid profile of rainbow smelt as shown in figure 6.22. The biggest discrepancies appear on fatty acids 16:0, 18:1n-7, 20:4n-6, 20:5n-3, 22:5n-6, and 22:6n-3.

The slight modification of rainbow smelt results in an improved fit of the Bayesian versus approximate method as measured by the lower $\sigma^2_{\boldsymbol{\epsilon}}$ and the $\boldsymbol{\mu}_{\boldsymbol{\tau}}$'s being closer to the diet compositions consumed by the captive sea birds.

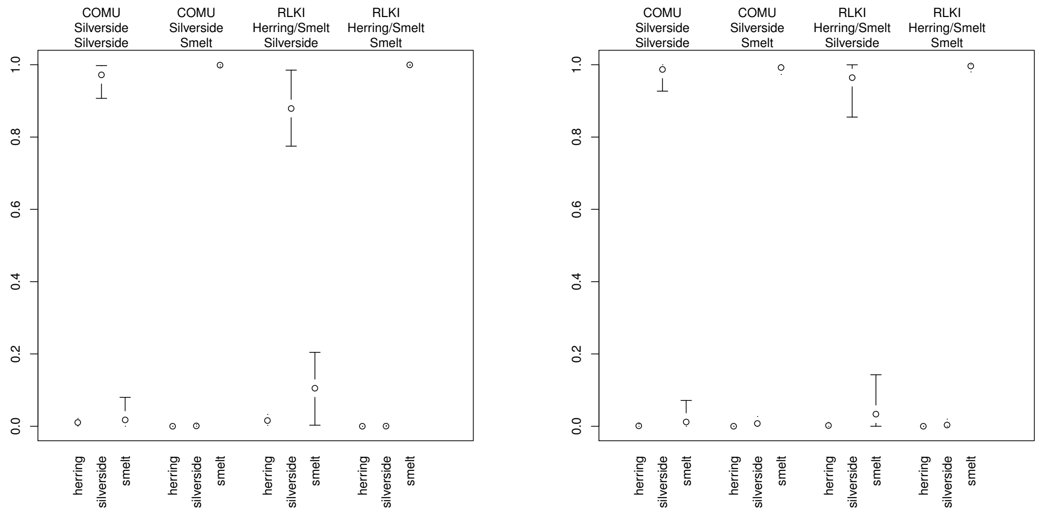

Figure 6.21: Posterior Means and component–wise 95% credible intervals for the diet composition, $\boldsymbol{\tau}_w$, for the four treatment combinations of captive birds. The fully Bayesian solution is given in the left panel, while the approximate method is given in the right hand panel. Note the labels at the top of the panels indicate the species, followed by the initial diet (see text for duration) and main diet. Note if the credible interval overlaps with the posterior mean then the credible interval is not shown.

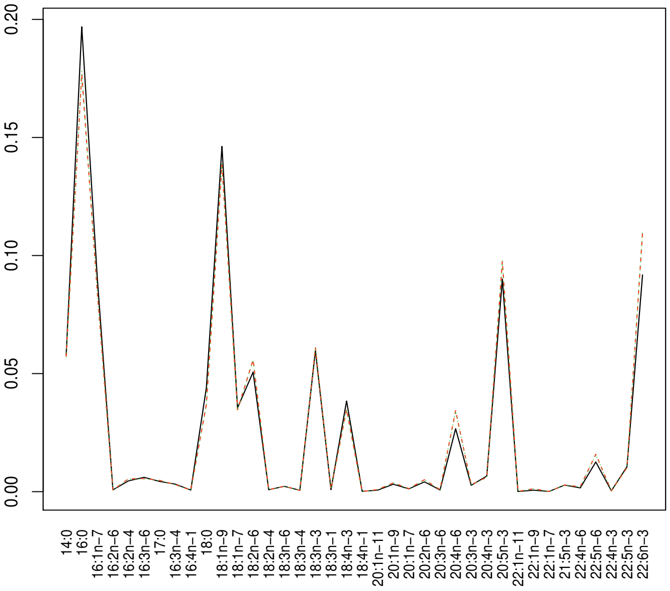

Figure 6.22: Posterior mean profile of rainbow smelt. The fully Bayesian is given by the solid line, the profiles for the approximate approach and the sample mean of rainbow smelt are superimposed.

6.5.2 Harbour Seals

Iverson et al. (2004) illustrated the original QFASA model using a sample of 23 free–ranging adult male harbour seals collected on Sable Island, Nova Scotia in 1997 during the breeding season (May–June). The harbour seals are known to make foraging trips returning to Sable every few days. Each animal was equipped with an animal–borne video system, or critter cam [National Geographic Television, Washington, D.C., USA] for at least 3 days. The camera was fitted to the animals head and was programmed to film for 10 minute segments every 45min during daylight, allowing prey encounters to be recorded (Bowen et al., 2002). Blubber biopsies were taken during each return to Sable island, allowing the comparison of the filmed encounters and the estimated diet via QFASA. To illustrate the individual diet model we analyzed the 23 harbour seals using the same 28 species (see table 6.5.2) used in Iverson et al. (2004).

Common name	Category	n	%Fat	Common name	Category	n	%Fat
Argentine	Other	10	6.64	Sandlance	Forage	71	5.61
Butterfish	Other	10	7.22	Sculpin	Other	20	1.37
Capelin	Other	56	8.27	Sea raven	Other	6	0.76
Cod	Gadiod	84	2.14	Silver hake	Gadiod	38	2.15
Gaspereau	Forage	41	12.56	Smooth skate	Skate	5	1.40
Haddock	Gadiod	54	1.39	Thorny skate	Skate	12	1.14
Halibut	Flat	8	1.10	White hake	Gadiod	46	1.29
Herring	Forage	74	7.73	Winter flounder	Flat	25	1.88
Mackerel	Forage	10	3.38	Winter skate	Skate	15	1.55
Ocean Pout	Other	18	1.99	Yellowtail flounder	Flat	92	2.69
Plaice	Flat	99	2.20	Lobster	Invertebrate	9	1.98
Pollock	Gadiod	25	3.02	Red Crab	Invertebrate	14	1.84
Red Hake	Gadiod	7	1.71	Rock Crab	Invertebrate	10	0.75
Redfish	Other	49	6.33	Shrimp	Invertebrate	46	2.58

Table 6.5: Potential diet items of harbour seals. Category refers to the level of identification that can be differentiated with the video from the critter cams, n is the number of samples, % fat is the average fat content.

Figure 6.23 shows the mean fatty acid profile of the 28 species along with a hierarchical clustering of the standardized distance measure of the profiles. Note that three fatty acids (16:3n-1, 16:2n-6, 22:2n-6) were removed from the extended dietary set considered in Iverson et al. (2004) due to lack of variability on several prey species. The condition number of the source matrix is extremely large, which indicates strong similarities among the prey

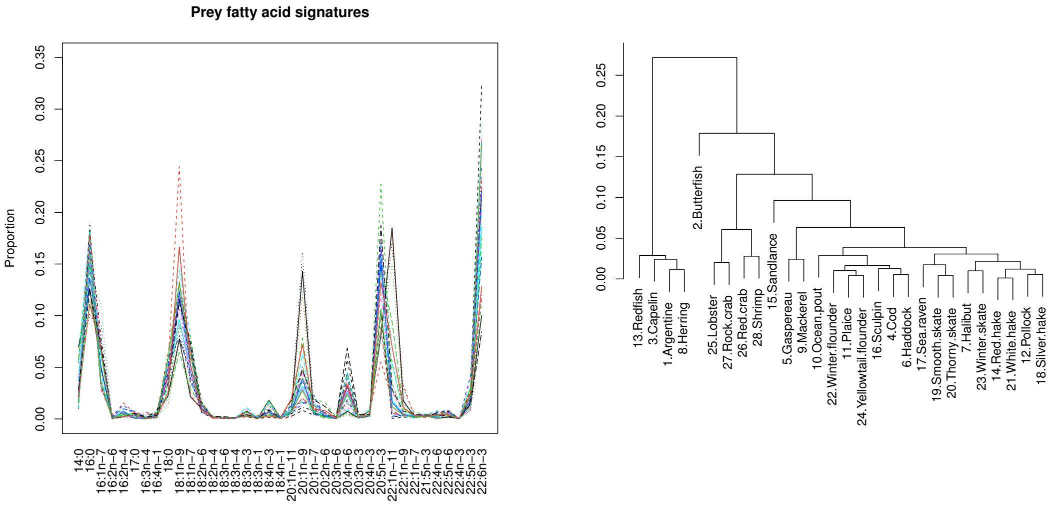

Figure 6.23: The left hand panels shows the mean fatty acid signatures for the 28 species of potential prey of Harbour seals used in Iverson et al. (2004) and the right hand panel shows a hierarchical clustering of the standardized distance using the average linkage method. The source matrix, $\mathbf{\Theta}$, condition number is 8592067. Note that three fatty acids (16:3n-1, 16:2n-6, 22:2n-6) were removed from the extended dietary set considered in Iverson et al. (2004).

species. For instance the cosine of the angle between Red hake and White hake is 0.9988, indicating that are almost coincident. Several other pairs of species have $\cos(\theta)$'s in excess of 0.99 (Plaice & Yellowtail flounder, Smooth skate & Thorny skate, Cod & Haddock, Pollock & Silver hake and Halibut & Winter skate). The smallest degree of association was observed between Halibut and Red hake, $\cos(\theta) = 0.547$. This example can be seen to be an extreme case of an ill–conditioned source matrix.

A long term calibration study was carried out (see Iverson et al., 2004, for further details) on eight juvenile (2-3 year) grey seals housed in large indoor seawater tanks at Dalhousie University's Aquatron facilities. The grey seals were kept on a diet of Atlantic herring for at least 5 months. Biopsies were taken on each of the eight animals at the end of the study period. Thirty herring were also collected during the study and analyzed. See Iverson et al. (2004) for more details on the laboratory methods that were used to analyze the blubber and herring samples.

As with the captive sea bird study in the previous section, the covariance matrices for the predator and calibration components of the model were not modeled by unconstrained covariance matrices, due to the small number of samples collected. We model the predator covariance matrix and the calibration covariance matrices by the pattern covariance matrix given in equation (6.1), with different values for the multiplier σ^2. Our strategy for the prey species covariance matrices is the following: sample sizes less than 30, model the covariance using equation (6.1), for sample sizes between 30 and 60 we use equation (6.2) and for sample sizes greater than 60 we use a non–pattered covariance matrix. Additionally, we model the mixing distribution covariance matrix with the patterned covariance matrix given in equation (6.1) as we only have 23 predators.

Despite the difficulties that the original QFASA method has when dealing with ill-conditioned source matrices we use the estimates as starting values for the $\boldsymbol{\tau}_i$'s. The hope is that they are in approximately the right region of posterior space. However, we modify the original QFASA method slightly. Firstly the distance was modified from the Kulback–Liebler distance used originally to the compositional distance as suggested in Stewart (2005). Secondly, the log–ratio transformation was applied to the composition to assure non–zero components which are crucial for our purposes. Thirdly, the optimization algorithm was adapted to have a simulated annealing step to avoid getting stuck in local optima.

6.5.3 Results

Video footage was collected on 30 adult harbour seals, including the 23 that had blubber biopsies taken, for an average of 3 days. All males, save one, was observed foraging on sandlance during the 3 day observation period. In total, there were 223, 10–minute video segments where identifiable prey captures occurred, 91% were on sandlance, 7% on flatfish and 2% on gadoids and other prey (Bowen et al., 2002).

Figure 6.24 gives 95% component–wise credible intervals for $\boldsymbol{\mu}_{\tau}$, the mean of the mixing distribution, for the fully Bayesian approach and the approximate method. Both methods identify sandlance as the dominant component of the diet, however, the fully Bayes approach has a lower portion compared to the approximate method. Interestingly, the approximate method identifies only three species with an upper credible limit greater than 5% (capelin, redfish and sadlance, while the Bayesian method identifies six species with an upper credible limit greater than 10% (capelin, cod, herring, pollock, redfish, sandlance).

As previously discussed, one of the features of the fully Bayesian approach is that the predator has information about the fatty acid profiles of the prey actually consumed and for this reason it is not surprising if the fully Bayesian approach modifies a given prey's fatty

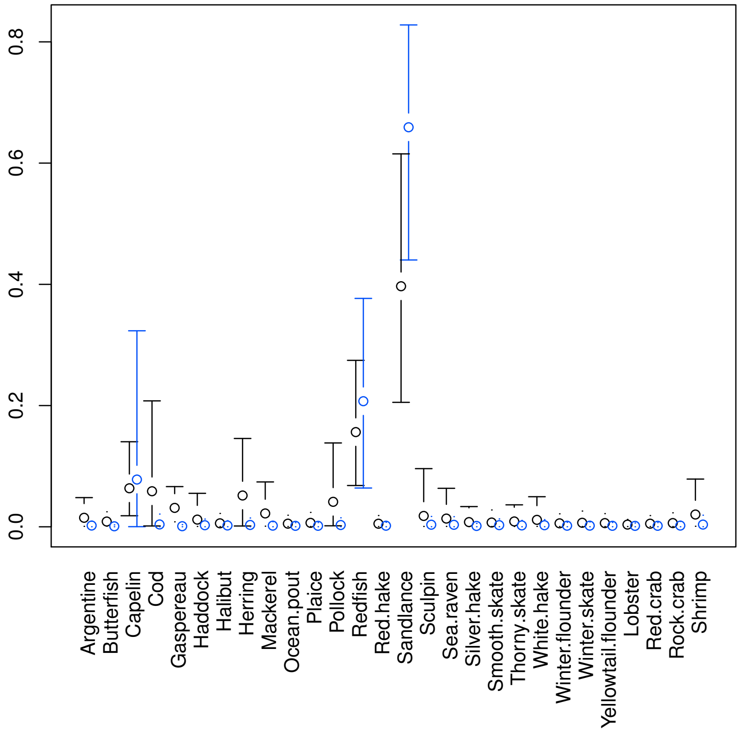

Figure 6.24: 95% credible intervals for the mean diet composition for the 23 adult male harbour seals. The black credible intervals correspond to the fully Bayesian model while the blue credible interval corresponds to the approximate method.

acid signature away from its sample mean. The species that the fully Bayesian approach modifies most is Silver hake, which turns out to be one of the most variable species, however, it doesn't appear to be a major diet component in either method.

Figure 6.25 gives boxplots of the posterior mean for the 23 individual estimates for the fully Bayesian approach, the approximate approach and slight modification of the original QFASA method. The original QFASA was modified in two ways: firstly the distance was modified from the Kulback–Liebler distance used originally to the compositional distance as suggested in Stewart (2005). Additionally, the log–ratio transformation was applied to the composition to assure non–zero components. Both of these modifications explain the departures from the estimates presented in Iverson et al. (2004).

As both the error covariance matrix Σ_ϵ and the mixing distribution covariance matrix $\Sigma_{\boldsymbol{\tau}}$ were modeled with the simple patterned structure given in equation (6.1), we present 95% credible intervals for both parameters. The credible intervals for the fully Bayesian

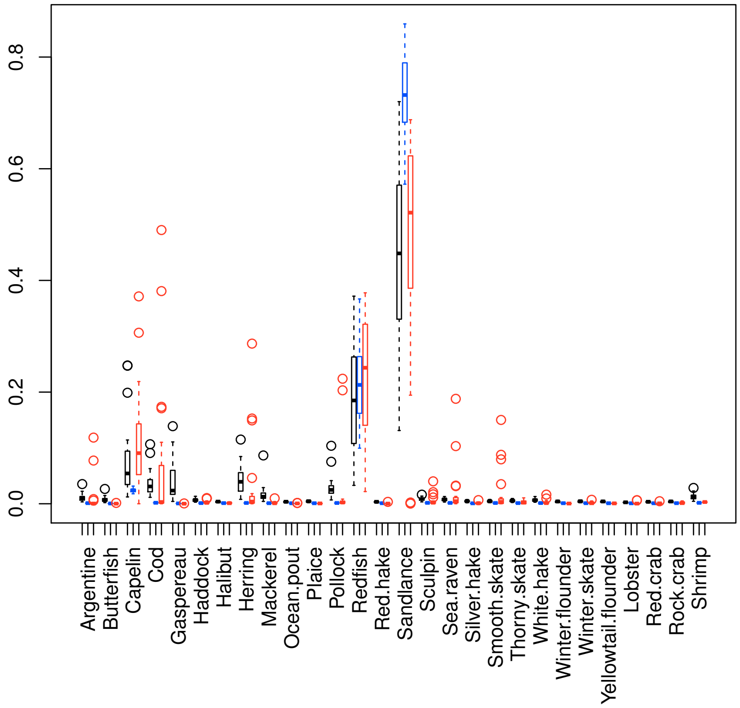

Figure 6.25: Boxplots of the 23 individual estimates of the diet composition: The fully Bayesian approach is given in black, the approximate method in blue and the modified QFASA approach (see text for details) is given in red.

approach and the approximate method are given in the following table:

Parameter	Fully Bayesian	Approximate
$\sigma^2_{\boldsymbol{\epsilon}}$	(0.0097,0.0131)	(0.1512,0.2189)
$\sigma^2_{\boldsymbol{\tau}}$	(0.5172,1.5102)	(0.2242,1.6645)

The fully Bayesian approach fits the 23 Harbour seals much better as indicated by the credible intervals for the $\Sigma_{\boldsymbol{\epsilon}}$ parameter. The difference between the mixing distribution parameter $\sigma^2_{\boldsymbol{\tau}}$ for the two methods is less clear from the credible intervals, however, the distribution for the approximate method is more right skewed and the credible interval is much longer.

The results indicate substantial variability in the individual diets of adult male harbour seals but they are dominated by sandlance, though the methods disagree as to the extent of sandlance in the various individual diets. The diet compositions given in Iverson et al. (2004) did not have redfish present in the diet, however, all three methods, including the modification of QFASA, identify redfish as a major component of the diet. This suggests that the choice of distance is a critical factor in the estimation of predator diets, however, due to the nature of the distributions considered we feel the compositional distance is most appropriate.

It is known that redfish are a relatively deep–water species and as a result may not be an item on the harbour seal menu. This brings up an interesting point about the prey library, one should remove species that are not potential prey of the predator due to physical size constraints, predator range considerations, etc.

The removal of redfish from the prey library of harbour seals results in different behaviour for the Bayesian approach and the approximate method. The approximate method behaves as one would expect and switches the redfish portion of the diet to capelin in this case while the amount of sandlance remains largely unchanged. The inference for the residual covariance parameter $\sigma_{\boldsymbol{\epsilon}}$ indicates a slight worsening of the fit as indicated by the 95% credible interval shifting to the right $(0.1631, 0.2301)$.

The behaviour of the fully Bayesian method is more interesting. It shifts the diet away from sandlance and capelin completely in favour of Silver hake and Winter skate two items that were not part of the diet before, with Silver hake dominating at nearly 85% without suffering any increase in variability of the residuals (95% credible interval (0.0096,0.0131)). Upon further investigation the Bayesian approach modifies the fatty acid signature of silver hake quite substantially away from the sample mean of silver hake and as a result is able to fit the data equally well without using sandlance or capelin. The resulting signature is a combination of sandlance and the original redfish signature that was removed. This suggests a lack of identifiability in the prey source matrix, which is not surprising given the large degree of multicollinearity in the original matrix.

When there is a large discrepancy between the approximate method and the fully Bayesian method it suggests that there is either an important prey item missing from the diet or there is an ill–conditioned source matrix.

As a further followup to the consequences of removing a potentially important prey item from the prey library we return to the synthetic data runs of the previous sections. We chose to run the individual diet model to assess the effect of dropping out a major element of the diet to determine how the fully Bayesian method copes with a non-existent prey item. We removed sandlance from the prey library and ran the individual diet model with $\Sigma_\tau = 0.5(\mathbf{I} + \mathbf{J})$ for both the permuted and non–permuted source matrix. For simplicity of presentation we only present the posterior means for each of the diet compositions.

Tables 6.6 and 6.7 give the posterior means of $\phi^{-1}(\boldsymbol{\mu}_{\boldsymbol{\tau}_j})$ for each of the six populations given in table 6.2 for the permuted and non–permuted synthetic data respectively. Recall the permuted data is very well conditioned, meaning that no two prey types are that similar and in this case the model apportions the amount of sandlance to redfish though some of its contribution gets apportioned to other species as well. By contrast, for the non–permuted synthetic data, the missing sandlance portion gets apportioned to capelin and herring which are the two species that are most similar to sandlance (see figure 6.5). Interestingly, the permuted version doesn't choose the most similar species to apportion the missing sandlance to, rather it chooses redfish. As a result of not having species close to sandlance in the prey library, the permuted residual variance suffers drastically as evidence by the 95% credible interval for the trace of $\Sigma_{\boldsymbol{\epsilon}}$ increasing to (14.89,20.94) compared to the non–permuted data (0.51,1.47).

The interesting feature of the synthetic data is that the sample sizes for all the components are much larger than in either of the two applications considered. The result of the larger sample sizes is that the various components are much better identified and as a result do not move away from their prior values as the application examples shows. This suggests that there is a potential identifiability issue for the fully Bayesian approach when the samples

	Spring		Summer		Winter	
Species	M	F	M	F	M	F
Capelin	0.04	3.02	1.29	0.02	2.81	1.94
Cod	4.05	4.30	5.80	6.62	6.30	4.54
Herring	5.68	4.65	3.04	1.47	3.01	7.96
Longhorn Sculpin	1.15	1.03	2.39	4.10	0.84	1.24
Plaice	5.63	2.96	3.83	1.79	3.16	1.32
Pollock	26.73	7.98	16.59	2.77	9.35	7.79
Redfish	35.50	61.6	51.16	73.25	57.09	64.77
Silver.Hake	2.46	3.14	1.67	1.45	2.13	1.28
Thorny Skate	5.71	0.74	2.77	0.50	5.68	0.88
White Hake	7.30	4.14	6.03	3.59	4.23	3.90
Yellowtail	5.74	6.44	5.43	4.42	5.39	4.39

Table 6.6: Posterior means of $\phi^{-1}(\boldsymbol{\mu}_{\boldsymbol{\tau}_j})$'s for the individual diet model for the permuted synthetic data without sandlance as a prey item. The true diet compositions are given in table 6.2.

size are small. Billheimer (2001) dealt with this issue with much stronger prior information and that would be one way to deal with potential identifiability issues if they arise. That is, additional prior information available combined with the information contained in the observations on the various components would make the model more identifiable.

The approximate method does not appear to suffer from the identifiability issue as much as the fully Bayesian method. Our conjecture is that by turning off the information flow between the predator and prey doesn't allow the predator to influence the prey and cause problems in the identification of $\Theta\boldsymbol{\alpha}$. However, it isn't entirely clear what properties the approximate method has. Particularly with regards to what joint distribution the MCMC sampler is actually sampling from.

	Spring		Summer		Winter	
Species	M	F	M	F	M	F
Capelin	5.59	12.1	4.82	14.66	9.46	13.18
Cod	1.39	1.72	4.87	1.90	5.59	2.45
Herring	12.25	36.49	25.41	41.32	30.08	31.98
LonghornSculpin	0.68	0.05	1.17	1.94	2.63	0.88
Plaice	6.02	1.98	1.00	2.45	1.01	0.90
Pollock	25.00	3.60	22.42	0.43	7.59	7.26
Redfish	26.10	31.65	25.88	33.56	30.77	39.80
Silver.Hake	7.88	5.21	1.09	0.55	4.20	0.30
ThornySkate	4.79	0.30	0.54	0.64	2.26	0.59
WhiteHake	8.21	1.22	7.74	0.32	2.37	0.27
Yellowtail	2.09	5.22	5.07	2.23	4.03	2.38

Table 6.7: Posterior means of $\phi^{-1}(\boldsymbol{\mu}_{\boldsymbol{\tau}_j})$'s for the individual diet model for the non–permuted synthetic data without sandlance as a prey item. The true diet compositions are given in table 6.2.

6.6 Conclusion

The applicability of the modification of the linear mixing model to the diet composition problem was demonstrated in this chapter. We saw excellent results with the synthetic data for both the constant and individual diet models when the source matrix was well conditioned. However, the performance of the model deteriorated in the presence of strongly linearly related prey fatty acid profiles.

Clearly, a method for dealing with this collinearity is needed. Several methods suggest themselves: a priori selection of the prey species before modeling the predator diets; a variable selection procedure akin to that suggested by Green (1995) and others. However, the variable selection procedure will problematic due to the fact that the fully Bayesian method will potentially modify prey fatty acid profiles.

Additionally, further work is needed along the lines of Kashiwagi (2004) to allow for the potential for unknown prey species to be missing from the prey library. The consequences of this are well understood in regression contexts, but less so in the present modeling situation.

Finally, more work is also needed to understand the strengths and limitations of our approximate method. It has the attractive feature of interpretability, in that, the prey species' fatty acid signatures are not typically altered from their mean values, though this is not always the case. However, it doesn't typically fit as well as the fully Bayesian approach, which is not surprising as it ignores the potentially valuable information contained in the predator about the fatty acid profiles it actually consumed. Also, more theoretical work needs to be done to understand actually what posterior distribution the approximate method is sampling from. At best it is a close approximation to the true joint posterior; at worst it may not correspond to any joint distribution.

CHAPTER 7

CONCLUSION

In this thesis we have argued for the Bayesian approach to inference for the diet composition problem as it allows for incorporation of all sources of uncertainty. Additionally, it allows one to carry out inference on population level compositions as well as individual predator compositions as required.

The method performs better than the original QFASA method of Iverson et al. (2004) as it gives diet compositions that are closer to generated ones in the synthetic data sets considered. It also doesn't appear to be as affected by multicollinearity (see section 6.4.1 for details). However, the fully Bayesian approach does appear to have some issues with non–identifiability when both the sources and diet compositions are unknown and not a lot of sample or prior information is available on the sources. The original QFASA approach avoids this issue by conditioning on the sources. The approximate Bayesian method doesn't appear to suffer the same non-identifiability problems as the fully Bayesian approach, however, it is not entirely clear what distribution we are actually sampling from. Though it appears to work reasonably well in the situations considered, this is far from an exhaustive study.

The multi–modality of the logistic normal distribution was not reported in the literature, to our knowledge, and has some implications for modeling compositional data. We chose a very restrictive covariance structure to study and more work is needed to determine if similar patterns hold with different covariance matrices and mean vector combinations.

The work of Priestley and Subba Rao (1969) and our extensions on a frequency domain test of stationarity of time series could be a promising addition to the MCMCist's tool box for assessing when a chain has reached its stationary distribution. The modification of the original test is based on the properties of the spectra matrix given in Brillinger (1981), specifically the eigenvalues of the spectral matrix and their asymptotic variances. It gives the analyst a relatively objective way of determining when the chain(s) is its stationary distribution, that is, it has forgotten its initial conditions.

Robert and Casella (2004) give the following quote against the use of spectra methods and non–parametric methods:

> A global criticism of the spectral approach also applies to all the methods using a nonparametric intermediary step to estimate a parameter of the model, namely that they necessarily induce losses in efficiency in the processing of the problem (since they are based on a less constrained representation of the

> model). Moreover, the calibration of nonparametric estimation methods (as the choice of the window in the kernel method) is always delicate since it is non–standardized.

Though we agree with the first part of this statement, we disagree with the second as MCMC methods by their very nature are delicate and non-standardized. We feel it is a promising avenue to approach the issue of when MCMC chains have reached their stationary distribution.

In the thesis we considered the case were we had partial knowledge of the source matrix through observations collected on the potential prey of a given predator. We can consider those observations as forming a prior on the various prey profiles and in that sense our approach can be seen as a direct application of the Billheimer (2001) model, with modifications for fat content and predator biosynthesis which were not present in the air pollution data considered by Billheimer (2001).

Analysis of synthetic data and two experimental studies, one on captive sea birds and the other on Harbour seals equipped with National Geographic critter cams, illustrated the validity and some of the pitfalls of the constant diet and individual diet models for inference on the diet composition of predators. The models performed best when the source matrix was well conditioned but this is hardly a new result as it is well known in the chemical mass balance/receptor literature.

The compositional constant mixing and the compositional multilevel models perform well in situations where there is little dependence among the sources. This is problematic for the diet composition problem as the species of interest have similar fatty acid profiles and proper apportionment in this situation is difficult.

This was not fully known or appreciated in the previous work on the diet composition using fatty acids as there was no diagnostic procedures available to alert the user when the optimizer may have encountered a ridge in the distance surface that was the basis of the previous methods. By contrast careful examination of MCMC samples indicate that in the presence of highly linearly related prey fatty acid signatures, ridges in the posterior surface occur. This was evident in the preliminary investigations in chapter 5. However, these ridges can be difficult to detect in higher dimensional posterior surfaces.

An interesting feature of the models is that they are highly non–linear in the case where the source profiles are not known, and it means that the only way to approach the models in a sensible fashion is through the Bayesian conditioning machinery. Thus the models are difficult to analyze analytically other than with approximations like the delta method as Aitchison and Bacon-Shone (1999) considered. By contrast the Bayesian MCMC machinery and the local nature of the model as given in its DAG representation allow for relatively straightforward Metropolis–within–Gibbs algorithms.

Consider the basic version of the model with no predator biosynthesis or fat content, that is,

$$\mathbf{y}_i = \Theta\alpha_i \oplus \boldsymbol{\epsilon}_i$$

where the columns of the source matrix $\boldsymbol{\Theta}$ are compositions as is the mixing vector α_i, *i.e.*, the compositional mixing model. As the mixing occurs on the scale of compositions we have argued that it is the linear properties of the source matrix, Θ , that are important in determining how well separated each species is. We measure the total contribution of all the linear associations among the species fatty acid profiles by the condition number of the source matrix, Θ. This appears to violate the recommendations of Aitchison (2003) where he argues that ignoring the unit sum constraint can lead to spurious correlations between components of the individual elements of the composition. However, we are arguing here that the correlation/linear dependence is obtained by using the sources as the variables and the fatty acids as the observations. Of course, the condition number is directly related in both cases as all we are considering here is a matrix versus its transpose. This is something that needs to be explored further as it is crucial for reducing ill–conditioned source matrices.

A final note on the diet models, is that implicitly we are assuming that the predators consume the "average" prey of a given type. That is, if they were feeding on one prey type, the fatty acid profile of the predator would resemble the average of that prey type ignoring predator biosynthesis. Thus, if the predator happens to prefer a size class of a certain prey type that may be different in its fatty acid profile from the overall average, then this would be problematic for the diet models considered here, as well as, for the original QFASA model.

7.1 Future Work

We give a brief account in point form (in no particular order) of the outstanding issues not yet addressed in the current thesis .

- Improved posterior summaries of the high dimensional mixing vectors and other parameters. That is, how do we sensibly summarize a high dimensional posterior surface? To this end we need to address the issue of multivariate percentiles or quantiles, see for example Chakraborty (2001).

- Multicollinearity of prey sources is a key issue. It can have potentially drastic effects on the quality of the inference as shown in the synthetic data studies with well conditioned and ill–conditioned source matrices. Improving the condition number of the source matrix when the data is compositional is not straightforward, as the usual methods will not necessarily give rise to sources that remain in the simplex. Also, a further fleshing out of the dependence properties of the source matrix Θ is required. Compositional principal components Aitchison (2003) hold some promise, though as with traditional principal components or factor analytic solutions, the resulting factors may not be directly interpretable. A potential possibility is discussed in the next point. However, perhaps the best tactic is more a judicious choice of potential prey is needed before proceeding with the linear mixing model.

- In other areas of statistics, variable selection techniques can be used to reduce the effects of multicollinearity. That is, if two species are similar and one is currently in the model then adding the second species will give little improvement in the overall fit of the model. However, Hamilton (1987) shows that this isn't always the case as he demonstrated in the multiple regression context, thus any forward selection technique must be interpreted with caution. The methods of Green (1995) on transdimensional Markov chains would be essential here. However, there are several complications compared to traditional variable selection in linear regression for example. Most importantly is that we don't have complete knowledge of the source profiles. Whereas in linear regression variable selection we assume the predictor variables are measured without error. This lack of certainty in the source profiles, fat contents or calibration components of the model are due to insufficient prior information and/or lack of adequate samples of the various model components to sufficiently determine the various portions of the model. In other words, when the model is only partially/conditionally identifiable, then the model selection problem is even more difficult as we saw some evidence of in the Harbour seal example when redfish was removed from the prey library and the fully Bayesian method changed the profile of Silver hake to resemble a combination of redfish and sandlance. This lack of certainty on the prey profiles will make the model selection very difficult indeed, in some cases that is.

- As previously mentioned, an added bonus of the multilevel or individual diet model is that we get an individual level inference as well as a population level model. However, as is well known in the generalized linear mixed models literature (see Diggle et al., 2002), some interpretation problems exist between the various types of models. Our model would most closely resemble generalized linear mixed models rather than population averaged models, which means that careful consideration must be given to the interpretation of the individual level compositions.

- Further refinements of the convergence test based on the evolutionary spectra are needed to assess how well it does in other situations. Also, need to address the concern of Robert and Casella (2004) on the use of convergence methods that require the choice of window width. There are surprisingly few methods based on the spectra of the MCMC output to assess convergence and this may be the reason.

- Need to do some further research into the applicability of the so called approximate Bayesian technique where the flow of information between the predator and prey is turned off by removing the contribution of the predator from the full conditional of the other various parts of the model. Recall that the full conditional distribution for the fatty acid profile for the jth source which we denote by $\boldsymbol{\theta}_j$ is

$$\begin{aligned}\pi(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{-j},\boldsymbol{\alpha},\Sigma_{\boldsymbol{\epsilon}},\mathbf{Y},\mathbf{X}_j,\Sigma_{\mathbf{X}_j}) =& \pi(\boldsymbol{\theta}_j|\boldsymbol{\mu}_{\boldsymbol{\theta}_j},\Sigma_{\boldsymbol{\theta}_j})\prod_{i=1}^{n}\pi(\mathbf{y}_i|\boldsymbol{\alpha},\boldsymbol{\Theta},\Sigma_{\boldsymbol{\epsilon}}),\\ &\prod_{k=1}^{n_j}\pi(\mathbf{x}_{jk}|\boldsymbol{\theta}_j,\Sigma_{\mathbf{X}_j}).\end{aligned}$$

  The approximate method removes the term $\prod_{i=1}^{n}\pi(\mathbf{y}_i|\boldsymbol{\alpha},\boldsymbol{\Theta},\Sigma_{\boldsymbol{\epsilon}})$, however, by modifying the full conditionals in this way, we are no longer sampling from the joint posterior with our MCMC algorithms. We need to determine what distribution the approximate method is actually sampling from.

- Exploring different distributional assumptions for the various parts of the model. The most natural one being the skew–logistic distribution mentioned in chapter 2 given by

Mateu-Figueras et al. (2005) which generalized the work of Azzalini and DallaValle (1996); Azzalini and Capitanio (1999) on the skew–normal distribution.

- Further modifications of the MCMC sampler can be explored as well: different proposal distributions, different adaption algorithms, random scans versus the reversible systematic scan method that was implemented. The reversible system scan is quite wasteful of resources as it only uses one out of every two samples generated.

- The constant and individual diet models were written in Fortran 90 for computational reasons, and further refinements of the code are needed for both readability and efficiency reasons.

- Modeling the uncertainty associated with the rounded zeros which are generated by the absence of peaks in the gas chromatograph. Currently, we arbitrarily assign a value half way between the lower detection limit of the gas chromatograph and zero. There are several approaches in the compositional literature that can be tried to ascertain their affect on the diet estimates and their associated uncertainty. See for example Martín-Fernández et al. (2003); Palarea-Albaladejo et al. (2007)

Appendix A

The Bayesian Paradigm and Directed Acyclic Graphs

In this chapter we briefly present some of the arguments for the Bayesian approach to statistical inference. Inference as an extension of logic and probability theory is the approach advocated by Cox (1946); Jaynes (2003), we present those arguments briefly in the first section. We then follow those with the likelihood principle argument first presented by Birnbaum (1962). Another argument in favour of Bayesian principles is the Decision theoretic approach. We end this chapter with the a brief introduction to Directed Acyclic Graph, which are a powerful way to represent hierarchical models and are particularly important in implementing the Markov Chain Monte Carlo approach to approximating the posterior distribution.

A.1 Probabilistic Approach to Inference

The probabilistic approach to statistical inference is very appealing for many reasons not the least of which is the numerous interpretation problems that arise with the classic/orthodox paradigm. Specifically, the failure of the classical approach to follow the likelihood principle which follows directly from two seemingly innocuous principles, the sufficiency principle and the conditionality principle (see Birnbaum, 1962). Physicist Richard T. Cox (see Cox, 1946, 1958, 1962) laid out some basic rules or desiderata that any system of inference should follow. He went on to show, by way of variational calculus techniques, that any such system must be based on the product and sum rule. In other words, Bayes theorem. This work went largely ignored in the literature, except by fellow physicist E.T. Jaynes who published numerous articles on statistical inference and maximum entropy methods, culminating in his life's work (see Jaynes, 2003), which was edited and published by former student Larry Bretthorst. Jaynes advocated the so called objective Bayesian approach to statistical inference which is consistent with the approach taken by Jeffreys (1961) but is contrasted with the subjective approach taken by other others most notably, Savage (1954); Lindley (1965a,b); de Finetti (1990a,b); Berger (1985). That is not to say that they all had the same view on subjective probability, in fact, several of them built an axiomatic approach to utility as a way of overcoming the wholly subjective approach. It is not our purpose to expose the virtues of a given approach only to say that the objective approach, when possible, should be taken.

We briefly mention the Cox and Jaynes desiderata:

1. Degrees of Plausibility are represented by real numbers.
2. (a) If a conclusion can be reasoned out in more than one way, then every possible way must lead to the same result.

   (b) All relevant evidence/information is taken into account. That is, information is not ignored or taken into account arbitrarily.

   (c) Equivalent states of knowledge are represented by equivalent plausibility assignments. That is, if in two problems our state knowledge is the same (except perhaps for the labeling of the propositions), then the same plausibilities must be assigned in both.
3. Qualitative Correspondence with common sense.

Theorem A.1 (Cox's Theorem)**.** *Cox (1946, 1962) A measure of plausibility that satisfy the above desiderata must be a monotonic function of* $p(.)$ *that satisfies the following rules (the product and sum rule)*

(i) $p(AB|C) = p(A|BC)p(B|C) = p(B|AC)p(A|C)$

(ii) $p(A|B) + p(\bar{A}|B) = 1$

Proof. Cox (1946, 1962) provided detailed proofs, and Jaynes (2003) gives a more general proof. They both rely on functional analysis and the calculus of variations. □

A.2 Classical Versus Bayesian Paradigms

Arguably the fundamental distinction between the classical (Neyman–Pearson) approach and the Bayesian approach is their view on parameters. In the classical approach unknown parameters are considered as fixed constants and probability statements about the unknown constants are generated by considering all possible realizations of the experiment, the so called sampling based approach. The Bayesian approach, by contrast, considers the unknown parameters on an equal footing with the data, that is, the uncertainty or lack of knowledge about the unknown parameter(s) is described by a probability distribution. Thus we consider the joint distribution of the parameters and the data, using basic rules of probability theory and then once the data are observed we base our inference on the conditional distribution of the parameters given the data.

To highlight the difference between the two approaches consider the following simple example. Assume we have a parametric model with an single unknown parameter θ, that is, we assume our data $y_1, \ldots, y_n$ are generated from $\pi(y_i|\theta)$, where π is our parametric model. The classical approach to inference, would then use the likelihood function to construct a maximum likelihood estimator of θ, say $\hat{\theta}$. Now comes the interesting part, since the classical approach assumes that the parameter is a fixed constant, how do they go about assigning probabilities. To do this, they construct confidence intervals, where they consider all possible realizations of the data, $y_1, \ldots, y_n$ and construct an form intervals of the form

$$\hat{\theta} \pm c(\alpha)\text{s.e}(\theta)$$

where $c(\alpha)$ is an appropriately chosen quantile of the sampling distribution and s.e(θ) is a measure of the spread of the sampling distribution. The constructed confidence interval then has the following interpretation: Overall all possible realizations of the experiment intervals constructed in the above fashion will have $(1 - \alpha) * 100\%$ coverage. That is, confidence intervals have a perfectly valid interpretation right up until the data is collected. Confidence and coverage probabilities tell us about how we can expect our procedures to behave before we actually collect the data. However, they don't tell us anything about a particular data set as their probability is derived from the collection of all possible outcomes.

The Bayesian approach to the same problem, requires one more piece of information before the analysis can proceed, namely a distribution which gives our prior knowledge about the unknown parameter, θ. We shall return to the discussion of the choice of prior later.

We then use basic rules of conditional probability theory to arrive at the joint distribution of $\theta, y_1, \ldots, y_n$ as follows:

$$\pi(\theta)p(y_1, \ldots, y_n|\theta) = \pi(\theta, y_1, \ldots, y_n)$$

Once the data $y_1, \ldots, y_n$ are observed we then use Bayes theorem to write the conditional distribution of $\pi(\theta|y_1, \ldots, y_n)$ as follows

$$\pi(\theta|y_1, \ldots, y_n) = \frac{\pi(\theta, y_1, \ldots, y_n)}{\int \pi(\theta, y_1, \ldots, y_n)d\theta} = \frac{\pi(\theta)\pi(y_1, \ldots, y_n|\theta)}{\int \pi(\theta)\pi(y_1, \ldots, y_n|\theta)d\theta}.$$

Inference is then based on $pi(\theta|y_1, \ldots, y_n)$, the conditional distribution of the unknown parameter given the observed data. We can then construct probability intervals or any other summary measures of interest of the conditional distribution.

Classical statisticians claim that the Bayesian approach isn't objective given that one has to assign a prior distribution. However, the method of maximum entropy gives one an objective way to assign prior distributions (see Jaynes, 2003, and the references therein).

Once we assign a prior distribution everything flows from the basic rules of probability theory, there is no need for any other principles that are common in the classical approach. The prior and likelihood lead to a posterior and that is the end of the story, at least from a theoretical perspective.

Bayesian calculations are more complex than their classical counterpart and in all but the simplest situations exact computations are not possible. The computational difficulty caused the Bayesian approach to lag behind its classical counterpart in applied settings. However, that changed drastically with the explosion of Markov Chain Monte Carlo (MCMC) methods in the early 1990s. This meant that many more problems could be tackled from a Bayesian perspective.

A.3 Likelihood Principle

The likelihood principle states that inferences involving proportional likelihoods should reach the same conclusions. We begin this section with an example to illustrate the essential ideas and then provide more details as to the actual principle and how it was derived. The original example was given in Lindley and Phillips (1976), however, we use the form presented in Berger (1985).

Example A.1. *We are given a coin and are interested in the parameter θ, the probability of heads when flipped. It is desired to test $H_0 : \theta = 1/2$ versus $H_1 : \theta > 1/2$. An experiment is conducted by flipping the coin in a series of trials, the result of which is the observation of 9 heads and 3 tails. We assume that the trials are exchangeable. That is, the order in which the heads and tails occur gives no information as to the value of θ.*

This is not yet enough information to specify $f(x|\theta)$, since the series of trials was not explained. Two possibilities are: (1) the experiment consisted of a predetermined 12 flips, so that $\mathbf{X}$, the number of heads, would have a binomial distribution denoted by $\mathcal{B}(12, \theta)$; or (2) the experiment consisted of flipping the coin until 3 tails were observed, thus $\mathbf{X}$ would have a negative binomial distribution, denoted by $\mathcal{NB}(3, \theta)$. The sampling distributions are given by

$$\pi_1(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} = 220\theta^9(1-\theta)^3$$

and

$$\pi_2(x|\theta) = \binom{y+x-1}{x} \theta^x (1-\theta)^y = 55\theta^9(1-\theta)^3.$$

Recall that the likelihood functions $l_1(\theta + x)$ and $l_2(\theta|x)$, are proportional to their respective sampling distributions.

Traditional hypothesis testing relies on the computing the observed significance levels or p–values for each of the experimental situations as they obviously have different sample spaces. Therefore, the significance level for the binomial version of the experiment is:

$$\begin{aligned} \textit{p–value}_1 &= p(X \geq 9|\theta = 0.5) = \sum_{x=9}^{12} \pi_1(x|\theta = 0.5) \\ &= 0.075. \end{aligned}$$

For the negative binomial model, the significance level is,

$$\begin{aligned} \textit{p–value}_2 &= p(X \geq 9|\theta = 0.5) = \sum_{x=9}^{\infty} \pi_2(x|\theta = 0.5) \\ &= 0.0325. \end{aligned}$$

Thus, depending on which sampling design we employ we get a very different picture of the significance level of the hypothesis test. This result seems counter–intuitive, in the sense *that why should it matter what our intentions were before the experiment. If our data is identical in two experiments, shouldn't the conclusions be the same? According to the likelihood principle, this should in fact be the case.*

There are two readily accepted principles that we need to discuss before proceeding to the discussion of the likelihood principle.

Conditionality Principle. Suppose one can perform either of two experiments E_1 or E_2, both pertaining to θ, and that the actual experiment conducted is the mixed experiment of first choosing $J = 1$ or 2 with probability p (logically independent of θ), then performing experiment E_J. Then the actual information about θ obtained from the overall experiment should depend only on the experiment E_j actually performed.

The following examples from Berger (1985); Cox (1958) respectively illustrate the weak conditionality principle.

Example A.2. *(Berger, 1985) Suppose a substance to be analyzed can be sent either to a laboratory in New York or a laboratory in California. The two labs seem equally good, so a fair coin is flipped to choose between them, with heads denoting that the lab in New York will be chosen. The coin is flipped and comes up tails, so the California lab is used. After a while, the experimental results come back and a conclusion and report must be developed. Should this conclusion take into account the fact the coin could have been heads, and hence that the experiment in New York might have been performed instead? Common sense (and the conditional viewpoint) cries no, that only the experiment actually performed is relevant, but frequentist reasoning would call for averaging over all possible data, even the possible New York data.*

Example A.3. *(Cox, 1958) In a research laboratory, a physical quantity θ can be measured by a precise but often busy machine, which provides a measurement, $x_1 \sim N(\theta, \sigma^2 = 1)$, with probability $p = 0.5$, or through a less precise but always available machine, which gives $x_2 \sim N(\theta, \sigma^2 = 9)$. The machine being selected at random, depending on the availability of the more precise machine, the inference on θ when it has been selected should not depend on the fact that the alternative machine could have been selected.*

Firstly, its clear that if we condition on the toss of the coin, the coverage probabilities are 0.95. However, if we were to consider the coverage of the more precise interval (that is, based on $\sigma^2 = 1$) under repeated sampling without conditioning on the coin toss then this *interval has coverage given by*

$$\begin{aligned}
P(x-1.96<\theta<x+1.96) &= P(x-1.96<\theta<x+1.96|Y=1)P(Y=1) \\
&+ P(x-1.96<\theta<x+1.96|Y=0)P(Y=0) \\
&= 0.95\frac{1}{2}+0.243\frac{1}{2} \\
&= 0.718
\end{aligned}$$

By a similar argument the coverage of the less precise instrument would have coverage of 0.975. Finally, if we wanted to have a classical interval with coverage 0.95, the interval would be (-4.935,4.935).

Thus, by not conditioning on the coin toss, we have to suffer a substantial decrease in precision in order to get the required coverage probability

Definition A.1. *When $x \sim f(x|\theta)$, a function T of x (also called a statistic) is said to be sufficient if the distribution of x conditional upon $T(x)$ does not depend on (θ)*

Thus a sufficient statistic contains all the information about θ contained in the sample. The factorization theorem allows for easy identification of sufficient statistics, that is,

$$f(x|\theta) = g(T(x)|\theta)h(x|T(x))$$

Example A.4. *Let $x_1,\ldots,x_n$ be a sample from a Poisson distribution with parameter θ, thus the joint density can be written as follows*

$$\begin{aligned}
p(x_1,\ldots,x_n|\theta) &= \prod_{i=1}^{n}\frac{e^{-\theta}\theta^{x_i}}{x_i!} \\
&= \frac{e^{-n\theta}\theta^{\sum x_i}}{\prod x_i!}
\end{aligned}$$

Thus if we let $T(x)=\sum_{i=1}^{n}$ then we have

$$p(x_1,\ldots,x_n|\theta) = \frac{e^{-n\theta}\theta^{T(x)}}{\prod x_i!}.$$

Now if we let $g(T(x)|\theta)=e^{-n\theta}\theta^{T(x)}$ and $h(x|T(x))=1/\prod x_i!$ we can apply the factorization theorem and see that $T(x)$ is sufficient for θ.

Sufficiency Principle Two observations x and y factorizing through the same value of a sufficient statistic T, that is, such that $T(x) = T(y)$, must lead to the same inference on θ.

Almost all classically trained statisticians ascribe to the sufficiency principle and if one restricts attention to the exponential family sufficient statistics exist. Rao-Blackwell theorem states that if you have an estimator and then condition on a sufficient statistic then that's the best you can do Bickel and Doksum (2000).

Likelihood Principle The information in the observation x about θ is entirely contained in the likelihood function $l(\theta|x)$. Moreover, if x_1 and x_2 are two observations depending on the same parameter θ, such that there exists a constant c satisfying

$$l_1(\theta|x_1) = cl_1(\theta|x_2)$$

for every θ, thus they have the same information concerning θ and must lead to identical inferences.

Note that the likelihood principle only applies when the parameter θ is same in both models.

Birnbaum (1962) shows that the likelihood principle follows from the conditionality principle and the sufficiency principle. Robert (2004) also gives a proof of this result.

Example A.5. *The coin tossing experiment discussed previously, is the standard example of how the likelihood principle can be violated. To see this note, the classical analysis has a different p–value depending on which sampling design was used. However, the parameter θ is exactly the same in both experiments and the likelihood functions are proportional, therefore the inferences reached have to be the same according to the likelihood principle.*

Thus, the seemingly innocuous conditionality principle and the generally accepted sufficiency principle leads to a powerful principle that as the above example shows, is violated by one of the cornerstones of classical statistical inference, the p–value. As Jeffreys (1961) so eloquently stated

> "... a hypothesis which may be true may be rejected because it has not predicted results which have not occurred"

The likelihood principle does not indicate how inference should be carried out, only that datasets with identical likelihood functions for the same parameter should yield the same inference. Bayesian inference obeys the likelihood principle as a matter of course, as it conditions on the observed data.

A.4 Decision Theory

Statistical decision theory is concerned with how to make decisions on the basis of statistical knowledge (ie a sample of data) which presumably sheds light on the uncertainties associated with the unknown quantities in the problem of interest. For example, consider making a decision as to whether or not to put a street light at a particular intersection in a city, given a sample of the number of accidents at that particular intersection. Parameter estimation, hypothesis testing and confidence intervals as well as their Bayesian counterparts can all be thought of as decision problems with particular loss functions. However, the class of decision problems is also much richer than this.

The following development is taken from Berger (1985) and supplemented from Robert (2004); Jaynes (2003). We assume that the state of nature or unknown quantity is given by θ and the set of all possible values is given by Θ. In most settings of interest, θ represents the unknown parameters of some statistical model and Θ represents the set of all possible values of said parameter(s) also known as the parameter space. In our example above, θ would be the true accident rate at the intersection of interest and Θ would typically be the positive real line, but of course practical considerations would give a upper limit to the parameter space. Let a represent a particular action (decision) and $\mathcal{A}$ be the collection of all possible actions.

The next ingredient in the decision theory pie, is that of a loss or utility function. The loss function, denoted by $L(\theta, a)$ gives the penalty/loss for decision a when the state of nature is θ. The loss function $L(\theta, a)$ is defined for all $(\theta, a) \in \Theta \times \mathcal{A}$ and gives a mapping onto positive real line. The utility function is denoted by $U(\theta, a)$ and gives the amount of gain for decision a when the state of nature is θ. Loss and utility functions are related by $L(\theta, a) = -U(\theta, a)$. Most of the statistical literature deals with loss functions while economists typically deal with utilities. Berger (1985); Robert (2004) give some technical details on the existence of utility/loss functions. For our purposes, we consider standard loss functions and investigate the properties of the decisions they imply.

The data are denoted by $\mathbf{X} = (X_1, \ldots, X_n)$ and joint their distribution is given by $\pi(\mathbf{X}|\theta)$. The observations are typically assumed to be conditionally independent given θ, that is,

$$\pi(\mathbf{X}|\theta) = \prod_{i=1}^{n} \pi(X_i|\theta).$$

Finally, our prior knowledge of θ is summarized in a prior $p(\theta)$. Using Bayes theorem, we can write our updated knowledge of θ using the posterior distribution

$$\pi(\theta|\mathbf{X}) \propto \pi(\theta)p(\mathbf{X}|\theta)$$

Consider the loss function $L(\theta, a)$ which gives, the loss for action(decision) a when the actual state of nature is θ. As θ is the only unknown, after the data is collected, it is natural to consider our expected loss over different values of θ and choose a decision on this basis.

Definition A.2 (Bayes Posterior Expected Loss)**.** *If $\pi(\theta|\mathbf{x})$ is the posterior distribution of θ, then the Bayesian expected loss for action a is*

$$\rho(\pi(\theta|\mathbf{x}), a) = E_{\pi(\theta|\mathbf{x})}L(\theta, a) = \int_{\Theta} L(\theta, a)\pi(\theta|\mathbf{x})d\theta.$$

The conditional Bayes Principle states the following: choose an action $a \in \mathcal{A}$ which minimizes the Bayes expected loss $\rho(\theta, a)$, assuming a minimum is obtained. We denote such actions by $\delta^{\pi}(x)$ and call them posterior Bayes actions. Consider the following example.

Example A.6. *Let $\pi(X|\theta) \sim N(\theta, 1)$ and consider the squared error loss $L(\theta, a) = (\theta - a)^2$. Now suppose that our prior for θ is given by $\pi(\theta) \sim N(0, \tau^2)$, for some known τ . Assume we have observed a sample $x = (x_1, \ldots, x_n)$. Firstly, we need to derive the posterior distribution of θ. Bayes theorem gives the following*

$$\begin{aligned} \pi(\theta|\mathbf{x}) &\propto \pi(\theta)\pi(\mathbf{x}|\theta) \\ &\propto N\left(\frac{n\tau^2\bar{x}}{n\tau^2+1}, \frac{n\tau^2+1}{\tau^2}\right) \end{aligned}$$

this follows as this is a conjugate family. the Bayes expected loss is

$$\begin{aligned}
\rho(\theta, a) &= \int_{-\infty}^{\infty} L(\theta, a)\pi(\theta|\mathbf{x}) \\
&= \int_{-\infty}^{\infty} (\theta - a)^2 \pi(\theta|\mathbf{x}) \\
&= \int_{-\infty}^{\infty} ((\theta - E(\theta|\mathbf{x})) + (E(\theta|\mathbf{x}) - a))^2 \pi(\theta|\mathbf{x}) \\
&= \int_{-\infty}^{\infty} \left\{(\theta - E(\theta|\mathbf{x}))^2 + (E(\theta|\mathbf{x}) - a))^2\right\} \pi(\theta|\mathbf{x})
\end{aligned}$$

The cross product term vanishes, the first term is the posterior variance of θ and the second term is minimized by choosing $a = E(\theta|\mathbf{x})$). That is, the conditional Bayes principle says estimate the unknown location parameter with the posterior mean. Note that, since we did not use anything about the form of the posterior other than having finite first two moments, this result generalizes to any posterior distribution that has finite first moments. That is, if the loss function is squared error loss, then the conditional Bayes principle gives the posterior mean as estimator that minimizes the Bayes expected loss.

In this particular case, the posterior mean is

$$\begin{aligned}
E(\theta|\mathbf{x}) &= \frac{n\tau^2 \bar{x}}{n\tau^2 + 1} \\
&= \frac{\bar{x}}{1 + 1/(n\tau^2)}.
\end{aligned}$$

Thus, as the prior knowledge gets more vague, in other words, larger prior variance τ^2 or the sample size gets larger, the estimator becomes closer to the usual sample mean. That is, in the limit the posterior mean is dominated by the likelihood function.

A.4.1 Classical Decision Theory

As the classical statistician does not assign probabilities to unknown constants (parameters), the concept of expected loss is very different. Following Berger (1985) we define a decision rule $\delta(x)$ as follows.

Definition A.3. *A decision rule $\delta(x)$ is a function from the sample space, $\mathcal{X}$ into the action space, $\mathcal{A}$. If $X = x$ is observed then $\delta(x)$ is the action taken. Two decision rules δ_1 and δ_2, are equivalent if $P_\theta(\delta_1(X) = \delta_2(X)) = 1$ for all θ.*

Rather than take expectations over θ, the frequentist prefers to measure his expected loss over all possible samples X.

Definition A.4. *The risk function of a decision rule $\delta(x)$ is given by*

$$R(\theta, \delta) = E_\theta^X[L(\theta, \delta(X))] = \int_{\mathcal{X}} L(\theta, \delta(x))\pi(x|\theta)d\theta$$

However, note that unlike the Bayes expected loss, which is a single number as the expectation was taken over θ the risk function is defined for all θ. Which naturally leads to some potential problems as it is clear that any one decision rule will be not be "best" for all θ. To see this, just consider a trivial decision rule of $\delta(x) = \theta_0$, this rule is best when $\theta = \theta_0$ and no other rule could beat it. This leads us to consider some other types of rules.

Definition A.5. *A decision rule δ_1 is R–better than a decision rule δ_2 if $R(\theta, \delta_1) \leq R(\theta, \delta_2)$ for all $\theta \in \Theta$, with strict inequality for some θ. A rule δ_1 is said to be R–equivalent to $delta_2$ if $R(\theta, \delta_1) = R(\theta, \delta_2)$ for all $\theta \in \Theta$.*

Definition A.6. *A decision rule is admissible if there exists no R–better decision rule. A decision rule is inadmissible if there does exist an R–better decision rule.*

We now define Bayes risk where the expectation is taken both of the sample space and the parameter space as seen in the following definition.

Definition A.7. *The Bayes risk of a decision rule δ, with respect to a prior distribution $\pi(\theta)$ on Θ is defined as*

$$\begin{aligned} r\left(\pi(\theta), \delta\right) &= E^\theta[R(\theta, \delta)] \\ &= \int_\Theta \int_{\mathcal{X}} L(\theta, \delta(x))\pi(x|\theta)\pi(\theta)dxd\theta \end{aligned}$$

Since we have taken expectation over all possible θ values with respect to the prior $\pi(\theta)$, the Bayes risk is now a single number and we can seek rules which minimize this risk.

The Bayes Risk Principle. A decision rule δ_1, is preferred to a rule δ_2 if

$$r\left(\pi(\theta), \delta_1\right) < r\left(\pi(\theta), \delta_2\right)$$

A decision rule that minimizes $r\left(\pi(\theta), \delta\right)$ is optimal; it is called a Bayes rule, and will be denoted by δ^π. The quantity $r(\pi(\theta)) = r(\pi(\theta), \delta^p)$ is called the Bayes risk for $\pi(\theta)$.

Example A.7. *Consider the example from the previous section. First, we find the risk and Bayes risk for the squared error loss function. Consider decision rules of the form* $\delta_c(x) = cx$*. The risk is given by*

$$\begin{aligned} R(\theta, \delta_c) &= E_\theta^X L(\theta, \delta_c(X)) = E_\theta^X (\theta - cX)^2 \\ &= E_\theta^X (\theta - c\theta + c\theta - cX)^2 \\ &= E_\theta^X (\theta(1-c) + c(\theta - X))^2 \\ &= c^2 E_\theta^X [\theta - X]^2 + 2c(1-c)\theta E_\theta^X [\theta - X] + (1-c)^2\theta^2 \\ &= c^2 + (1-c)^2\theta^2 \end{aligned}$$

$$R(\theta, \delta_1) = 1 < c^2 + (1-c)^2\theta^2 = R(\theta, \delta_c)$$

for $c > 1$*, which implies that* δ_1 *is R–better than any rule* δ_c*. It is less clear when* $0 \leq c \leq 1$*, for instance, the rules* δ_1 *and* $\delta_{1/2}$ *cross due to the quadratic nature of the loss function for* $0 \leq c \leq 1$*, so the rules are not really comparable. Thus, also admissibility is a desirable property it cannot always be achieved even for simple examples.*

The Bayes risk is given by

$$\begin{aligned} r(\pi(\theta|X), \delta) &= E^\pi[R(\theta, \delta)] = E^\pi[c^2 + (1-c)^2\theta^2] \\ &= c^2 + (1-c)^2 E^\pi[\theta^2] = c^2 + (1-c)^2\tau^2. \end{aligned}$$

Applying the Bayes risk Principle to this example, by differentiating with respect to c *yields*

$$c_0 = \frac{\tau^2}{1+\tau^2}$$

Note that by application of the following proposition, δ_{c_0} *is optimal in the sense that it also minimizes the Bayes risk.*

Berger (1985) states the following two results:

Proposition A.1. *A Bayes rule* δ^p *(a rule minimizing* $r(\pi(\theta|X), \delta(x))$ *) , the bayes risk) can be found by choosing, for each* x *such that the* $\pi(x) > 0$*, an action which minimizes the posterior expected loss. The rule can be defined arbitrarily when* $\pi(x) = 0$*.*

Proposition A.2. *If δ is a non–randomized estimator, then*

$$r(\theta, \delta) = \int_{x:\pi(x)>0} \rho(\pi(\theta|x), \delta(x))$$

Proposition A.1 follows directly from proposition A.2 and proposition A.2 is a direct consequence of Fubini's theorem which allows for a change of the order of integration.

A.4.2 Other Common Loss Function Results

In this section, we give results for several common loss functions and discuss their implementation in terms of minimizes Bayes expected loss and hence Bayes risk. We consider, the squared–error loss, the absolute loss function, and the 0–1 loss function. As we have seen before minimizing $\rho(\pi(\theta|x), \delta(x))$ is all that is required, that is,

$$\rho(\pi(\theta|x), \delta(x)) = \int_{\Theta} L(\theta, \delta(x))\pi(\theta|x)d\theta$$

If $L(\theta, \delta(x)) = (\theta - \delta(x))^2$ the squared–error loss then the Bayes estimator is

$$\delta^{\pi}(x) = E^{\pi(\theta|x)}[\theta]$$

the posterior mean.

If $L(\theta, \delta(x)) = |\theta - \delta(x)|$ the absolute-error loss then the Bayes estimator is any median of $\pi(\theta|x)$

If

$$L(\theta, \delta(x)) = \begin{cases} 0 & \delta(x) = \theta \\ 1 & \text{otherwise} \end{cases}$$

the zero-one loss, this corresponds to the highest posterior mode of $\pi(\theta|x)$.

Consider the following non-technical proof. The Bayes risk is given by

$$\begin{aligned} \rho(\pi(\theta|x), \delta(x)) &= E(L(\theta, \delta(x))) \\ &= \int_{\Theta} L(\theta, \delta(x))\pi(\theta|x) \\ &= \int_{\theta \neq \delta(x)} \pi(\theta|x) \\ &= 1 - \pi(\theta = \delta(x)|x) \end{aligned}$$

The issue here is how to define a posterior mode for a continuous parameter spaces.. Not sure.. have to do some reading. The proof follows very easily in the discrete parameter case,though not very interesting in practice.

Thus, the Bayes risk is minimized if we allow $\delta(x)$ the posterior mode.

Jaynes (2003) gives an interesting interpretation of this loss function in the case that the prior distribution is relatively flat in the high likelihood region and not much greater else where, then we essentially have the maximum likelihood estimator in this case. To quote Jaynes (2003)

> In this result we see finally just what maximum likelihood accomplishes, and under what circumstances it is the appropriate method to use. The maximum–likelihood criterion is the one in which we care only about the chance of being exactly right; and, if we are wrong, we don't care how wrong we are. This is just the situation we have in shooting at a small target, where 'a miss is as good as a mile'. But it is clear that there are few other situations where this would be a rational way to behave; almost always, the amount of error is of some concern to us, and so maximum likelihood is not the best estimation criterion

Several of the theorems mentioned can be found in the classic frequentist text on point estimation by Lehmann (1983). Thus, even the most staunch classicalist acknowledges the existence of Bayes estimators and their optimality properties. They use the Bayes methods to derive classes of estimators, however, they don't take the next logical step and perform inference in a conditional fashion once the data has been collected.

A.5 Graphical Models

Directed acyclic graphs or DAGs are very useful for depicting hierarchical models and their associated dependencies. We briefly review some of their properties and note their usefulness to Markov Chain Monte Carlo methods for finding noting full conditional distributions and their potential simplification.

A graph is a collection of vertices or nodes, denoted by $\mathcal{V}$ and edges, denoted by $\mathcal{E}$, between vertices. We say the graph is directed if all the edges are represented by arrows, implying a particular path through the nodes. We say the graph is acyclic if no cycles are allowed and call these graphs directed acyclic graphs or DAGs. Figure A.1 gives examples of three commonly occurring graphs, however, note that we are only interested in the DAGs for the subsequent discussion.

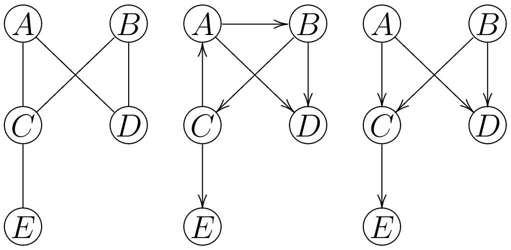

Figure A.1: Examples of three graphs. The first graph is an undirected graph, note it has no directed edges. The center graph is a directed graph, note it has a cycle, (A,B,C). The final graph is a directed acyclic graph (DAG) (taken from Spiegelhalter, 1998).

We say that vertex B is a child of the vertex A if a directed edge (arrow) connects A to B or A is a parent of B. It is convenient to label all the parents of a node B by pa(B) or when needed the children of node A by ch(A).

Consider the collection of all nodes, $\mathcal{V}$, as represent unknown quantities which we want to model the joint distribution of. As pointed by Spiegelhalter (1998), it seems natural to assume that any node or vertex is conditionally independent of any other non–descendants, given that nodes parents. He goes to assert that this formulation obeys the conditionally independence axioms given in Dawid (1979b). Also, one can read off the conditional independence properties directly off the DAG which he states can be used as a "theorem prover". Most importantly it enables one to write down the joint distribution for the set of nodes $\mathcal{V}$ as follows

$$p(\mathcal{V}) = \prod_{\nu \in \mathcal{V}} \pi(\nu | \text{pa}[\nu]).$$

Thus, we can write down the full joint distribution in terms of the local conditional properties of the graph. This is similar in nature to the Hammersley–Clifford theorem that we shall see in the chapter on Markov Chain Monte Carlo methods. Applying this rule to the example DAG of (Spiegelhalter, 1998) gives

$$p(A, B, C, D, E) = p(A)p(B)p(C|A, B)p(D|A, B)p(E|C).$$

Spiegelhalter (1998) also points out that the directed local Markov property is equivalent to other Markov properties given in Lauritzen et al. (1990) in more general context.

There is a connection between DAGs and Gibbs and hence its variants of Metropolis–Hastings–within–Gibbs which was pointed out by Pearl (1987) in the artificial intelligence literature but it bears pointing out here. In essence it allows one to read the full conditional distributions, essential in Gibbs sampling, off the DAG. The connection between the full conditional distributions and the local dependence properties of the DAG can be see by the following argument.

Consider a node ν a given node of the set $\mathcal{V}$ and let $\mathcal{V} \setminus \nu$ represent the remaining nodes. The full conditional for node ν is then given by

$$\begin{aligned} p(\nu|\mathcal{V} \setminus \nu) &\propto p(\nu, \mathcal{V} \setminus \nu) \\ &\propto \text{terms in } \; p(\mathcal{V}) \text{ containing } \nu \\ &= p(\nu|\text{pa}(\nu)) \prod_{\omega \in \text{ch}(\nu)} p(\omega|\text{pa}(\omega)) \end{aligned}$$

That is, to read off the full conditional distribution of a particular unobservable we only have to consider the co-parents and any children of the node ν.

For our example DAG the full conditional distributions are :

$$\begin{aligned} p(C|A, B, D, E) &\propto \text{terms in } p(V) \text{ containing } C \\ &\propto p(C|A, B)p(E|C) \\ p(A|B, C, D, E) &\propto p(A)p(C|A, B)p(D|A, B) \\ p(B|A, C, D, E) &\propto p(B)p(C|A, B)p(D|A, B) \\ p(D|A, B, C, E) &\propto p(D|A, B) \\ p(E|A, B, C, D) &\propto p(E|C). \end{aligned}$$

The local properties of the graph enable one to write down the full conditional distributions in a relatively simple and systematic way. Therefore, if you can write the model of interest as DAG then doing Gibbs sampling algorithm or Metropolis–Hastings–within–Gibbs algorithm is, in theory at least, straightforward.

For further details on conditional independence and graphical models the reader is referred to Dawid (1979a); Pearl (1987); Lauritzen et al. (1990); Lauritzen (1996); Spiegelhalter and Lauritzen (1990); Spiegelhalter (1998); Cowell et al. (1999); Consonni and Leucari (2001) plus the references therein.

APPENDIX B

REVERSIBLE SYSTEMATIC SCAN METROPOLIS–WITHIN–GIBBS ALGORITHMS

B.1 Constant Diet Model

This section gives details for the reversible systematic scan Metropolis–within-Gibbs algorithm for the constant diet model.

The model is:

$$
\begin{aligned}
\underset{(a\times1)}{\mathbf{y}_i} &= \underset{(a\times p)}{\boldsymbol{\Theta}}\underset{(p\times1)}{\Gamma^i} \oplus (\underset{(a\times1)}{\boldsymbol{\theta}_{\mathbf{v}}} \ominus \underset{(a\times1)}{\boldsymbol{\theta}_{\mathbf{u}}}) \oplus \underset{(a\times1)}{\boldsymbol{\epsilon}_i}\,, \quad i = 1, \ldots, n, \\
\underset{(p\times n)}{\Gamma} &= \boldsymbol{\phi}_c^{-1}\left(\boldsymbol{\phi}_c\left(\underset{(p\times w)}{\mathbf{T}}\right)\underset{(w\times n)}{\mathbf{W}}\right) \oplus_c \underset{(p\times1)}{\boldsymbol{\lambda}} \\
\underset{(a\times1)}{\mathbf{x}_{jk}} &= \underset{(a\times1)}{\boldsymbol{\theta}_j} \oplus \underset{(a\times1)}{\boldsymbol{\epsilon}^{\mathbf{x}}_{jk}}\,, \quad j = 1, \ldots, p,\ k = 1, \ldots, n_j, \\
\underset{(2\times1)}{\mathbf{z}_{jk}} &= \underset{(2\times1)}{\boldsymbol{\lambda}^v_j} \oplus \underset{(2\times1)}{\boldsymbol{\epsilon}^{\mathbf{z}}_{jk}}\,, \quad j = 1, \ldots, p,\ k = 1, \ldots, n_j, \\
\underset{(a\times1)}{\mathbf{u}_l} &= \underset{(a\times1)}{\boldsymbol{\theta}_{\mathbf{u}}} \oplus \underset{(a\times1)}{\boldsymbol{\epsilon}^{\mathbf{u}}_l}\,, \quad l = 1, \ldots, L, \\
\underset{(a\times1)}{\mathbf{v}_m} &= \underset{(a\times1)}{\boldsymbol{\theta}_{\mathbf{v}}} \oplus \underset{(a\times1)}{\boldsymbol{\epsilon}^{\mathbf{v}}_m}\,, \quad m = 1, \ldots, M,
\end{aligned}
$$

where $\mathbf{W}$ is an $w \times n$ known design matrix, $\mathbf{T} = [\boldsymbol{\tau}_1|\boldsymbol{\tau}_2; \ldots|\boldsymbol{\tau}_w]$ is an $p \times w$ matrix.

We assign the following prior distributions for the location parameters

$$\begin{aligned}
\pi(\boldsymbol{\tau}_s|\boldsymbol{\mu}_{\boldsymbol{\tau}}, \Sigma_{\boldsymbol{\tau}}) &\sim \mathcal{L}^p(\boldsymbol{\mu}_{\boldsymbol{\tau}_s}, \Sigma_{\boldsymbol{\tau}_s}), \ \ s = 1, \ldots, w, \\
\pi(\boldsymbol{\theta}_j|\boldsymbol{\mu}_{\theta_j}, \Sigma_{\theta_j}) &\sim \mathcal{L}^a(\boldsymbol{\mu}_{\theta_j}, \Sigma_{\theta_j}), \ \ j = 1, \ldots, p, \\
\pi(\boldsymbol{\lambda}_j^v|\mu_{\lambda_j}, \Sigma_{\lambda_j}) &\sim \mathcal{L}^2(\boldsymbol{\mu}_{\lambda_j}, \Sigma_{\lambda_j}), \ \ j = 1, \ldots, p, \\
\pi(\boldsymbol{\theta}_{\mathbf{u}}|\boldsymbol{\mu}_{\theta_{\mathbf{u}}}, \Sigma_{\theta_{\mathbf{u}}}) &\sim \mathcal{L}^a(\boldsymbol{\mu}_{\theta_{\mathbf{u}}}, \Sigma_{\theta_{\mathbf{u}}}), \\
\pi(\boldsymbol{\theta}_{\mathbf{v}}|\boldsymbol{\mu}_{\theta_{\mathbf{v}}}, \Sigma_{\theta_{\mathbf{v}}}) &\sim \mathcal{L}^a(\boldsymbol{\mu}_{\theta_{\mathbf{v}}}, \Sigma_{\theta_{\mathbf{v}}}),
\end{aligned}$$

and for the covariance matrices

$$\begin{aligned}
\pi(\Sigma_{\boldsymbol{\epsilon}}|\delta_{\boldsymbol{\epsilon}}, \Psi_{\boldsymbol{\epsilon}}) &\sim \mathcal{IW}^{a-1}(\delta_{\boldsymbol{\epsilon}}, \Psi_{\boldsymbol{\epsilon}}), \\
\pi(\Sigma_{\mathbf{x}_j}|\delta_{\mathbf{x}_j}, \Psi_{\mathbf{x}_j}) &\sim \mathcal{IW}^{a-1}(\delta_{\mathbf{x}_j}, \Psi_{\mathbf{x}_j}), \ \ j = 1, \ldots, p, \\
\pi(\Sigma_{\mathbf{z}_j}|\delta_{\mathbf{z}_j}, \Psi_{\mathbf{z}_j}) &\sim \mathcal{IW}^{1}(\delta_{\mathbf{z}_j}, \Psi_{\mathbf{z}_j}), \ \ j = 1, \ldots, p, \\
\pi(\Sigma_{\mathbf{u}}|\delta_{\mathbf{u}}, \Psi_{\mathbf{u}}) &\sim \mathcal{IW}^{a-1}(\delta_{\mathbf{u}}, \Psi_{\mathbf{u}}), \\
\pi(\Sigma_{\mathbf{v}}|\delta_{\mathbf{v}}, \Psi_{\mathbf{v}}) &\sim \mathcal{IW}^{a-1}(\delta_{\mathbf{v}}, \Psi_{\mathbf{v}}).
\end{aligned}$$

The sampling distributions are given by

$$\begin{aligned}
\pi(\mathbf{y}_i|\boldsymbol{\Theta}, \mathbf{T}, \Sigma_{\boldsymbol{\epsilon}}, \boldsymbol{\lambda}, \boldsymbol{\theta}_{\mathbf{u}}, \boldsymbol{\theta}_{\mathbf{v}}) &\sim \mathcal{L}^a(\boldsymbol{\phi}(\boldsymbol{\Theta}\Gamma^i \oplus (\boldsymbol{\theta}_{\mathbf{v}} \ominus \boldsymbol{\theta}_{\mathbf{u}})), \Sigma_{\boldsymbol{\epsilon}}), \ \ i = 1, \ldots, n \\
\pi(\mathbf{x}_{jk}|\boldsymbol{\theta}_j, \Sigma_{\mathbf{x}_j}) &\sim \mathcal{L}^a(\boldsymbol{\phi}(\boldsymbol{\theta}_j), \Sigma_{\mathbf{x}_j}), \ \ j = 1, \ldots, p; \ \ k = 1, \ldots, n_j \\
\pi(\mathbf{u}_l|\boldsymbol{\theta}_{\mathbf{u}}, \Sigma_{\mathbf{u}}) &\sim \mathcal{L}^a(\boldsymbol{\phi}(\boldsymbol{\theta}_{\mathbf{u}}), \Sigma_{\mathbf{u}}), \ \ l = 1, \ldots, L; \\
\pi(\mathbf{v}_m|\boldsymbol{\theta}_{\mathbf{v}}, \Sigma_{\mathbf{v}}) &\sim \mathcal{L}^a(\boldsymbol{\phi}(\boldsymbol{\theta}_{\mathbf{v}}), \Sigma_{\mathbf{v}}), \ \ m = 1, \ldots, M; \\
\pi(\mathbf{z}_{jk}|\lambda_1, \Sigma_{\mathbf{z}_j}) &\sim \mathcal{L}^2(\boldsymbol{\phi}(\lambda_j), \Sigma_{\mathbf{z}_j}), \ \ j = 1, \ldots, p; \ \ k = 1, \ldots, n_j
\end{aligned}$$

Before we can describe the MCMC algorithm to sample the posterior, we need to establish that the posterior is proper.

Proposition B.1. *The posterior distribution for the constant diet model is proper. That is:*

$$\int \pi(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p, \Sigma_{\mathbf{x}_1}, \ldots, \Sigma_{\mathbf{x}_p}, \boldsymbol{\lambda}_1^v, \ldots, \boldsymbol{\lambda}_p^v, \Sigma_{\mathbf{z}_1}, \ldots, \Sigma_{\mathbf{z}_p}, \boldsymbol{\theta}_{\mathbf{u}}, \Sigma_{\mathbf{u}}, \boldsymbol{\theta}_{\mathbf{v}}, \Sigma_{\mathbf{v}}, \boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_w, \Sigma_{\boldsymbol{\epsilon}}|\mathcal{D})$$

$$d\boldsymbol{\theta}_1 \ldots, d\boldsymbol{\theta}_p, d\Sigma_{\mathbf{x}_1}, \ldots, d\Sigma_{\mathbf{x}_J}, d\boldsymbol{\lambda}_1^v, \ldots, d\boldsymbol{\lambda}_p^v, d\Sigma_{\mathbf{z}_1}, \ldots, d\Sigma_{\mathbf{z}_p}, d\boldsymbol{\theta}_{\mathbf{u}}, d\Sigma_{\mathbf{u}}, d\boldsymbol{\theta}_{\mathbf{v}}, d\Sigma_{\mathbf{v}}, d\boldsymbol{\tau}, d\Sigma_{\boldsymbol{\epsilon}} < \infty$$

where the range of integration is over the simplex for $\boldsymbol{\theta}_j$, $\boldsymbol{\lambda}_j^v$, $\boldsymbol{\theta}_{\mathbf{u}}$, $\boldsymbol{\theta}_{\mathbf{v}}$, $\boldsymbol{\tau}_s$,, and the range of positive definite matrices for $\Sigma_{\mathbf{x}_j}$, $\Sigma_{\mathbf{z}_j}$, $\Sigma_{\mathbf{u}}$, $\Sigma_{\mathbf{v}}$ and $\Sigma_{\boldsymbol{\epsilon}}$ and $\mathcal{D}$ is the collection of all observed data.

Proof. The posterior is given below:

$$
\begin{aligned}
&\pi(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p, \Sigma_{\mathbf{x}_1}, \ldots, \Sigma_{\mathbf{x}_p}, \boldsymbol{\lambda}_1^v, \ldots, \boldsymbol{\lambda}_p^v, \Sigma_{\mathbf{z}_1}, \ldots, \Sigma_{\mathbf{z}_p}, \boldsymbol{\theta}_{\mathbf{u}}, \Sigma_{\mathbf{u}}, \boldsymbol{\theta}_{\mathbf{v}}, \Sigma_{\mathbf{v}}, \boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_w, \Sigma_{\boldsymbol{\epsilon}}|\mathcal{D}) \\
&= \prod_{j=1}^{p} \left\{ \mathcal{L}^{a-1}(\boldsymbol{\theta}_j|\boldsymbol{\mu}_{\boldsymbol{\theta}_j}, \Sigma_{\boldsymbol{\theta}_j}) \times \mathcal{IW}^{a-1}(\Sigma_{\mathbf{x}_j}|\delta_{\mathbf{x}_j}, \Psi_{\mathbf{x}_j}) \times \prod_{k=1}^{n_j} \mathcal{L}^{a-1}(\mathbf{x}_{jk}|\boldsymbol{\theta}_j, \Sigma_{\mathbf{x}_j}) \right\} \\
&\times \prod_{j=1}^{p} \left\{ \mathcal{L}^{1}(\boldsymbol{\lambda}_j^v|\boldsymbol{\mu}_{\boldsymbol{\lambda}_j^v}, \Sigma_{\text{-}j}) \times \mathcal{IW}^{1}(\Sigma_{\mathbf{z}_j}|\delta_{\mathbf{z}_j}, \Psi_{\mathbf{z}_j}) \times \prod_{k=1}^{n_j} \mathcal{L}^{1}(\mathbf{z}_{jk}|\boldsymbol{\lambda}_j^v, \Sigma_{\mathbf{z}_j}) \right\} \\
&\times \mathcal{L}^{a-1}(\boldsymbol{\theta}_{\mathbf{u}}|\boldsymbol{\mu}_{\boldsymbol{\theta}_{\mathbf{u}}}, \Sigma_{\boldsymbol{\theta}_{\mathbf{u}}}) \times \mathcal{IW}^{a-1}(\Sigma_{\mathbf{u}}|\delta_{\mathbf{u}}, \Psi_{\mathbf{u}}) \times \prod_{l=1}^{L} \mathcal{L}^{a-1}(\mathbf{u}_l|\boldsymbol{\theta}_{\mathbf{u}}, \Sigma_{\mathbf{u}}) \\
&\times \mathcal{L}^{a-1}(\boldsymbol{\theta}_{\mathbf{v}}|\boldsymbol{\mu}_{\boldsymbol{\theta}_{\mathbf{v}}}, \Sigma_{\boldsymbol{\theta}_{\mathbf{v}}}) \times \mathcal{IW}^{a-1}(\Sigma_{\mathbf{v}}|\delta_{\mathbf{v}}, \Psi_{\mathbf{v}}) \times \prod_{m=1}^{M} \mathcal{L}^{a-1}(\mathbf{v}_m|\boldsymbol{\theta}_{\mathbf{v}}, \Sigma_{\mathbf{v}}) \\
&\times \prod_{s=1}^{w} \mathcal{L}^{p-1}(\boldsymbol{\tau}_s|\boldsymbol{\mu}_{\boldsymbol{\tau}}, \Sigma_{\boldsymbol{\tau}}) \times \mathcal{IW}^{a-1}(\Sigma_{\boldsymbol{\epsilon}}|\delta_{\boldsymbol{\epsilon}}, \Psi_{\boldsymbol{\epsilon}}) \\
&\times \prod_{i=1}^{n} \mathcal{L}^{a-1}(\mathbf{y}_i|\boldsymbol{\Theta}, \boldsymbol{\lambda}, \boldsymbol{\theta}_{\mathbf{u}}, \boldsymbol{\theta}_{\mathbf{v}}, \boldsymbol{\tau}, \Sigma_{\boldsymbol{\epsilon}})
\end{aligned}
$$

With the exception of the term $\prod_{i=1}^{n} \mathcal{L}(\mathbf{y}_i|\Theta, \boldsymbol{\lambda}, \boldsymbol{\theta}_{\mathbf{u}}, \boldsymbol{\theta}_{\mathbf{v}}, \boldsymbol{\tau}, \Sigma_{\boldsymbol{\epsilon}})$ there is complete separation between the terms of the posterior distribution. Consider the functional form for this term

$$
|\Sigma_{\boldsymbol{\epsilon}}|^{-n/2} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} (\mathbf{y}_i - \boldsymbol{\mu}_{\mathbf{y}_i})' \Sigma_{\epsilon}^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_{\mathbf{y}_i}) \right\}
$$

where

$$
\boldsymbol{\mu}_{\mathbf{y}_i} = \underset{(a\times p)}{\boldsymbol{\Theta}} \underset{(p\times 1)}{\Gamma^i} \oplus (\underset{(a\times n)}{\boldsymbol{\theta}_{\mathbf{v}}} \ominus \underset{(a\times 1)}{\boldsymbol{\theta}_{\mathbf{u}}})
$$

and Γ^i means the ith column of the matrix Γ.

Since $\Sigma_{\boldsymbol{\epsilon}}$ is positive definite, $\sum_{i=1}^{n} (\mathbf{Y}_i - \Theta\boldsymbol{\alpha})' \Sigma_{\epsilon}^{-1} (\mathbf{Y}_i - \Theta\boldsymbol{\alpha}) \geq 0$, therefore, the term is bounded by $|\Sigma_{\boldsymbol{\epsilon}}|^{-n/2}$.

Using a similar bounding technique we can bound each of the sampling distributions $\mathcal{L}^{a-1}(\mathbf{x}_{jk}|\boldsymbol{\theta}_j, \Sigma_{\mathbf{x}_j})$, $\mathcal{L}^{1}(\mathbf{z}_{jk}|\boldsymbol{\lambda}_j^v, \Sigma_{\mathbf{z}_j})$, $\mathcal{L}^{a-1}(\mathbf{u}_l|\boldsymbol{\theta}_{\mathbf{u}}, \Sigma_{\mathbf{u}})$ $\mathcal{L}^{a-1}(\mathbf{v}_m|\boldsymbol{\theta}_{\mathbf{v}}, \Sigma_{\mathbf{v}})$ by $\prod_{j=1}^{p} |\Sigma_{\mathbf{x}_j}|^{n_j/2}$,

$\prod_{j=1}^{p} |\Sigma_{\mathbf{z}_j}|^{n_j/2}$, $|\Sigma_{\mathbf{u}}|^{L/2}$ and $|\Sigma_{\mathbf{v}}|^{M/2}$ respectively. Therefore we can bound our original integral by

$$\begin{aligned}
&\int \prod_{j=1}^{p} \Big\{ \mathcal{L}^{a-1}(\boldsymbol{\theta}_j|\boldsymbol{\mu}_{\boldsymbol{\theta}_j}, \Sigma_{\boldsymbol{\theta}_j}) \times \mathcal{IW}^{a-1}(\Sigma_{\mathbf{x}_j}|\delta_{\mathbf{x}_j}, \Psi_{\mathbf{x}_j}) \times |\Sigma_{\mathbf{x}_j}|^{-n_j/2} \Big\} \\
&\times \prod_{j=1}^{p} \Big\{ \mathcal{L}^{1}(\boldsymbol{\lambda}_j^v|\boldsymbol{\mu}_{\boldsymbol{\lambda}_j^v}, \Sigma_{\text{-}j}) \times \mathcal{IW}^{1}(\Sigma_{\mathbf{z}_j}|\delta_{\mathbf{z}_j}, \Psi_{\mathbf{z}_j}) \times |\Sigma_{\mathbf{x}_j}|^{-n_j/2} \Big\} \\
&\times \mathcal{L}^{a-1}(\boldsymbol{\theta}_{\mathbf{u}}|\boldsymbol{\mu}_{\boldsymbol{\theta}_{\mathbf{u}}}, \Sigma_{\boldsymbol{\theta}_{\mathbf{u}}}) \times \mathcal{IW}^{a-1}(\Sigma_{\mathbf{u}}|\delta_{\mathbf{u}}, \Psi_{\mathbf{u}}) \times |\Sigma_{\mathbf{u}}|^{-L/2} \\
&\times \mathcal{L}^{a-1}(\boldsymbol{\theta}_{\mathbf{v}}|\boldsymbol{\mu}_{\boldsymbol{\theta}_{\mathbf{v}}}, \Sigma_{\boldsymbol{\theta}_{\mathbf{v}}}) \times \mathcal{IW}^{a-1}(\Sigma_{\mathbf{v}}|\delta_{\mathbf{v}}, \Psi_{\mathbf{v}}) \times |\Sigma_{\mathbf{v}}|^{-M/2} \\
&\times \prod_{s=1}^{w} \mathcal{L}^{p-1}(\boldsymbol{\tau}_s|\boldsymbol{\mu}_{\boldsymbol{\tau}}, \Sigma_{\boldsymbol{\tau}}) \times \mathcal{IW}^{a-1}(\Sigma_{\boldsymbol{\epsilon}}|\delta_{\boldsymbol{\epsilon}}, \Psi_{\boldsymbol{\epsilon}}) \times |\Sigma_{\boldsymbol{\epsilon}}|^{-n/2} \\
&d\boldsymbol{\theta}_1 \ldots, d\boldsymbol{\theta}_p, d\Sigma_{\mathbf{x}_1}, \ldots, d\Sigma_{\mathbf{x}_J}, d\boldsymbol{\lambda}_1^v, \ldots, d\boldsymbol{\lambda}_p^v, d\Sigma_{\mathbf{z}_1}, \ldots, d\Sigma_{\mathbf{z}_p}, d\boldsymbol{\theta}_{\mathbf{u}}, d\Sigma_{\mathbf{u}}, d\boldsymbol{\theta}_{\mathbf{v}}, d\Sigma_{\mathbf{v}}, d\boldsymbol{\tau}, d\Sigma_{\boldsymbol{\epsilon}}
\end{aligned}$$

Since all the integrands are positive we can rearrange the order of integral by Fubini's theorem.

$$\begin{aligned}
&\prod_{j=1}^{p} \Big\{ \int \mathcal{L}^{a-1}(\boldsymbol{\theta}_j|\boldsymbol{\mu}_{\boldsymbol{\theta}_j}, \Sigma_{\boldsymbol{\theta}_j}) d\boldsymbol{\theta}_j \times \int \mathcal{IW}^{a-1}(\Sigma_{\mathbf{x}_j}|\delta_{\mathbf{x}_j}, \Psi_{\mathbf{x}_j}) \times |\Sigma_{\mathbf{x}_j}|^{-n_j/2} d\Sigma_{\mathbf{x}_j} \Big\} \\
&\times \prod_{j=1}^{p} \Big\{ \int \mathcal{L}^{1}(\boldsymbol{\lambda}_j^v|\boldsymbol{\mu}_{\boldsymbol{\lambda}_j^v}, \Sigma_{\text{-}j}) d\boldsymbol{\lambda}_j^v \times \int \mathcal{IW}^{1}(\Sigma_{\mathbf{z}_j}|\delta_{\mathbf{z}_j}, \Psi_{\mathbf{z}_j}) \times |\Sigma_{\mathbf{x}_j}|^{-n_j/2} d\Sigma_{\mathbf{z}_j} \Big\} \\
&\times \int \mathcal{L}^{a-1}(\boldsymbol{\theta}_{\mathbf{u}}|\boldsymbol{\mu}_{\boldsymbol{\theta}_{\mathbf{u}}}, \Sigma_{\boldsymbol{\theta}_{\mathbf{u}}}) d\boldsymbol{\theta}_{\mathbf{u}} \times \int \mathcal{IW}^{a-1}(\Sigma_{\mathbf{u}}|\delta_{\mathbf{u}}, \Psi_{\mathbf{u}}) \times |\Sigma_{\mathbf{u}}|^{-L/2} d\Sigma_{\mathbf{u}} \\
&\times \int \mathcal{L}^{a-1}(\boldsymbol{\theta}_{\mathbf{v}}|\boldsymbol{\mu}_{\boldsymbol{\theta}_{\mathbf{v}}}, \Sigma_{\boldsymbol{\theta}_{\mathbf{v}}}) d\boldsymbol{\theta}_{\mathbf{v}} \times \int \mathcal{IW}^{a-1}(\Sigma_{\mathbf{v}}|\delta_{\mathbf{v}}, \Psi_{\mathbf{v}}) \times |\Sigma_{\mathbf{v}}|^{-M/2} d\Sigma_{\mathbf{v}} \\
&\times \prod_{s=1}^{w} \int \mathcal{L}^{p-1}(\boldsymbol{\tau}_s|\boldsymbol{\mu}_{\boldsymbol{\tau}}, \Sigma_{\boldsymbol{\tau}}) d\boldsymbol{\tau} \times \int \mathcal{IW}^{a-1}(\Sigma_{\boldsymbol{\epsilon}}|\delta_{\boldsymbol{\epsilon}}, \Psi_{\boldsymbol{\epsilon}}) \times |\Sigma_{\boldsymbol{\epsilon}}|^{-n/2} d\Sigma_{\boldsymbol{\epsilon}}
\end{aligned}$$

We assume that all prior distributions are proper, though diffuse. The integrals involving the logistic normal distribution are all finite since the priors are proper. Thus all that remains to be shown is that the integrals involving the Inverse Wishart distributions are finite. The kernel of a Wishart distribution with scale matrix M and degrees of freedom n is given by

$$|\Sigma|^{-(n+d+1)/2} e^{-\frac{1}{2} tr(M\Sigma^{-1})}.$$

We focus on the term for $\Sigma_{\boldsymbol{\epsilon}}^{-1}$ and the other terms follow analogously.

$$\int |\Sigma_{\boldsymbol{\epsilon}}|^{-(\delta_{\boldsymbol{\epsilon}}+(a-1)+1)/2} e^{-\frac{1}{2}tr(M\Sigma^{-1})} |\Sigma_{\boldsymbol{\epsilon}}|^{-n/2} d\Sigma_{\epsilon}$$
$$= \int |\Sigma_{\boldsymbol{\epsilon}}|^{-(\delta_{\boldsymbol{\epsilon}}+n+(a-1)+1)/2} e^{-\frac{1}{2}tr(M^{-1}\Sigma)} d\Sigma_{\epsilon}$$

where $M = \Psi_{\boldsymbol{\epsilon}}$. Which is the kernel of an Inverse Wishart distribution with degrees of freedom $\delta_{\boldsymbol{\epsilon}} + n$. Thus all of the Wishart integrals are finite since the priors were assumed to be proper and the result follows. □

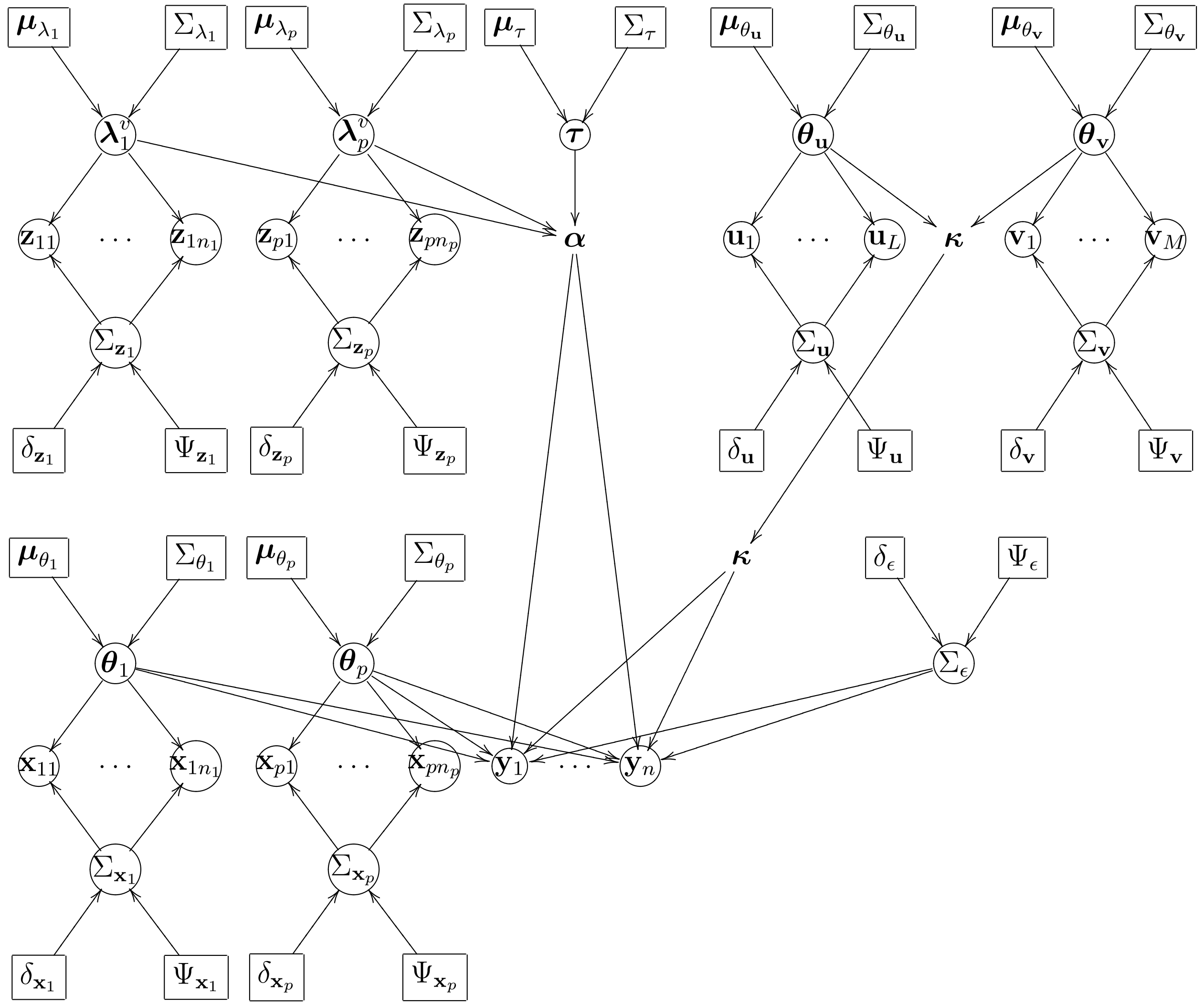

Figure B.1: DAG: Constant Diet with calibration and fat content. Note $\boldsymbol{\kappa} = \boldsymbol{\theta}_{\mathbf{v}} \oplus \boldsymbol{\theta}_{\mathbf{u}}$, $\boldsymbol{\alpha} = \boldsymbol{\tau} \oplus \boldsymbol{\lambda}$ and $\mathbf{Y} = [(\Theta\Gamma) \oplus_c (\boldsymbol{\theta}_{\mathbf{v}} \ominus \boldsymbol{\theta}_{\mathbf{u}})] \oplus_c \mathbf{E}$. Nodes that are not contained in circles or squares are derived variables are used to simplify the graph.

The reversible systematic scan Metropolis–within–Gibbs sampler for the constant diet model is given below

0. Choose starting values for $\boldsymbol{\tau}_{s,0}, s = 1, \ldots, w$, $\boldsymbol{\theta}_{j,0}, j = 1, \ldots, p$, $\lambda_{j,0}, j = 1, \ldots, p$, $\boldsymbol{\theta}_{\mathbf{u},0}$ and $\boldsymbol{\theta}_{\mathbf{v},0}$

1. Sample $\Sigma^*_{\boldsymbol{\epsilon}}$ from
$$\mathcal{IW}^{a-1}(\Psi_{\boldsymbol{\epsilon}} + S(\mathbf{y}_i, \boldsymbol{\mu}_{\mathbf{y}_i}), \delta_{\boldsymbol{\epsilon}} + n)$$
where
$$S(\mathbf{y}_i, \boldsymbol{\mu}_{\mathbf{y}_i}) = \sum_{i=1}^{n} (\phi(\mathbf{y}_i) - \phi(\boldsymbol{\mu}_{\mathbf{y}_i}))(\phi(\mathbf{y}_i) - \phi(\boldsymbol{\mu}_{\mathbf{y}_i}))',$$
and
$$\begin{aligned} \boldsymbol{\mu}_{\mathbf{y}_i} &= \boldsymbol{\Theta}_{t-1}\Gamma^i_{t-1} \oplus (\boldsymbol{\Theta}_{\mathbf{v},t-1} \ominus \boldsymbol{\theta}_{\mathbf{u},t-1}) \\ \Gamma_{t-1} &= \boldsymbol{\phi}_c^{-1} \left(\boldsymbol{\phi}_c \left(T_{t-1} \right) \mathbf{W} \right) \oplus_c \boldsymbol{\lambda}_{t-1}, \end{aligned}$$

   $\boldsymbol{\Theta}_{t-1} = [\boldsymbol{\theta}_{1,t-1}| \ldots |\boldsymbol{\theta}_{p,t-1}]$, Γ^i means the ith column of the matrix Γ, $\mathbf{T}_{t-1} = [\boldsymbol{\tau}_{1,t-1}| \ldots |\boldsymbol{\tau}_{w,t-1}]$ is an $p \times w$ matrix and $\boldsymbol{\lambda}_{t-1} = (\lambda_{1,t-1}, \ldots, \boldsymbol{\lambda}_{p,t-1})$.

2. For $j = 1, \ldots, p$ sample $\Sigma^*_{\mathbf{x}_j}$ from
$$\mathcal{IW}^{a-1}(\Psi_{\mathbf{x}_j} + S(\mathbf{x}_{jk}, \boldsymbol{\theta}_{j,t-1}), \delta_{\mathbf{x}_j} + n_j)$$
where
$$S(\mathbf{x}_{jk}, \boldsymbol{\theta}_{j,t-1}) = \sum_{k=1}^{n_j} (\phi(\mathbf{x}_{jk}) - \phi(\boldsymbol{\theta}_{j,t-1}))(\phi(\mathbf{x}_{jk}) - \phi(\boldsymbol{\theta}_{j,t-1}))'$$

3. Sample $\Sigma^*_{\mathbf{u}}$ from
$$\mathcal{IW}^{a-1}(\Psi_{\mathbf{u}} + S(\mathbf{u}_l, \boldsymbol{\theta}_{\mathbf{u},t-1}), \delta_{\mathbf{u}} + L)$$
where
$$S(\mathbf{u}_l, \boldsymbol{\theta}_{\mathbf{u},t-1}) = \sum_{l=1}^{L} (\phi(\mathbf{u}_l) - \phi(\boldsymbol{\theta}_{\mathbf{u},t-1}))(\phi(\mathbf{u}_l) - \phi(\boldsymbol{\theta}_{\mathbf{u},t-1}))'$$

4. Sample $\Sigma^*_{\mathbf{v}}$ from

$$\mathcal{IW}^{a-1}(\Psi_{\mathbf{v}} + S(\mathbf{v}_m, \boldsymbol{\theta}_{\mathbf{v},t-1}), \delta_{\mathbf{v}} + M)$$

   where

$$S(\mathbf{v}_m, \boldsymbol{\theta}_{\mathbf{v},t-1}) = \sum_{m=1}^{M} (\phi(\mathbf{v}_m) - \phi(\boldsymbol{\theta}_{\mathbf{v},t-1}))(\phi(\mathbf{v}_m) - \phi(\boldsymbol{\theta}_{\mathbf{v},t-1}))'$$

5. Generate $\boldsymbol{\theta}^*_{\mathbf{u}}$ from the following Metropolis–Hastings algorithm with proposal distribution $q_{\boldsymbol{\theta}_{\mathbf{u}}}(.|\boldsymbol{\theta}_{\mathbf{u},t-1})$

   (a) Generate $\boldsymbol{\nu} \sim q_{\boldsymbol{\theta}_{\mathbf{u}}}(\boldsymbol{\nu}|\boldsymbol{\theta}_{\mathbf{u},t-1})$

   (b)

$$\boldsymbol{\theta}^*_{\mathbf{u}} = \begin{cases} \boldsymbol{\nu} & \text{with probability} \rho_{\boldsymbol{\theta}_{\mathbf{u}}}(\boldsymbol{\theta}_{\mathbf{u},t-1}, \boldsymbol{\nu}) \\ \boldsymbol{\theta}_{\mathbf{u},t-1} & \text{with probability} 1 - \rho_{\boldsymbol{\theta}_{\mathbf{u}}}(\boldsymbol{\theta}_{\mathbf{u},t-1}, \boldsymbol{\nu}) \end{cases}$$

   where

$$\rho_{\boldsymbol{\theta}_{\mathbf{u}}}(\boldsymbol{\nu}^o, \boldsymbol{\nu}^n) = \min\left\{ \frac{f_{\boldsymbol{\theta}_{\mathbf{u}}}(\boldsymbol{\nu}^n) q_{\boldsymbol{\theta}_{\mathbf{u}}}(\boldsymbol{\nu}^o|\boldsymbol{\nu}^n)}{f_{\boldsymbol{\theta}_{\mathbf{u}}}(\boldsymbol{\nu}^o) q_{\boldsymbol{\theta}_{\mathbf{u}}}(\boldsymbol{\nu}^n|\boldsymbol{\nu}^o)}, 1 \right\}$$

6. Generate $\boldsymbol{\theta}^*_{\mathbf{v}}$ from the following Metropolis–Hastings algorithm with proposal distribution $q_{\boldsymbol{\theta}_{\mathbf{v}}}(.|\boldsymbol{\theta}_{\mathbf{v},t-1})$

   (a) Generate $\boldsymbol{\nu} \sim q_{\boldsymbol{\theta}_{\mathbf{v}}}(\boldsymbol{\nu}|\boldsymbol{\theta}_{\mathbf{v},t-1})$

   (b)

$$\boldsymbol{\theta}^*_{\mathbf{v}} = \begin{cases} \boldsymbol{\nu} & \text{with probability} \rho_{\boldsymbol{\theta}_{\mathbf{v}}}(\boldsymbol{\theta}_{\mathbf{v},t-1}, \boldsymbol{\nu}) \\ \boldsymbol{\theta}_{\mathbf{v},t-1} & \text{with probability} 1 - \rho_{\boldsymbol{\theta}_{\mathbf{v}}}(\boldsymbol{\theta}_{\mathbf{v},t-1}, \boldsymbol{\nu}) \end{cases}$$

   where

$$\rho_{\boldsymbol{\theta}_{\mathbf{v}}}(\boldsymbol{\nu}^o, \boldsymbol{\nu}^n) = \min\left\{ \frac{f_{\boldsymbol{\theta}_{\mathbf{v}}}(\boldsymbol{\nu}^n) q_{\boldsymbol{\theta}_{\mathbf{v}}}(\boldsymbol{\nu}^o|\boldsymbol{\nu}^n)}{f_{\boldsymbol{\theta}_{\mathbf{v}}}(\boldsymbol{\nu}^o) q_{\boldsymbol{\theta}_{\mathbf{v}}}(\boldsymbol{\nu}^n|\boldsymbol{\nu}^o)}, 1 \right\}$$

7. For $j = 1, \ldots, p$ generate $\boldsymbol{\theta}^*_j$ from the following Metropolis–Hastings algorithm with proposal distribution $q_{\boldsymbol{\theta}_j}(.|\boldsymbol{\theta}_{j,t-1})$

   (a) Generate $\boldsymbol{\nu} \sim q_{\boldsymbol{\theta}_j}(\boldsymbol{\nu}|\boldsymbol{\theta}_{j,t-1})$

   (b)

$$\boldsymbol{\theta}^*_j = \begin{cases} \boldsymbol{\nu} & \text{with probability} \rho_{\boldsymbol{\theta}_j}(\boldsymbol{\theta}_{j,t-1}, \boldsymbol{\nu}) \\ \boldsymbol{\theta}_{j,t-1} & \text{with probability} 1 - \rho_{\boldsymbol{\theta}_j}(\boldsymbol{\theta}_{j,t-1}, \boldsymbol{\nu}) \end{cases}$$

where

$$\rho_{\boldsymbol{\theta}_j}(\boldsymbol{\nu}^o, \boldsymbol{\nu}^n) = \min\left\{\frac{f_{\boldsymbol{\theta}_j}(\boldsymbol{\nu}^n) q_{\boldsymbol{\theta}_j}(\boldsymbol{\nu}^o|\boldsymbol{\nu}^n)}{f_{\boldsymbol{\theta}_j}(\boldsymbol{\nu}^o) q_{\boldsymbol{\theta}_j}(\boldsymbol{\nu}^n|\boldsymbol{\nu}^o)}, 1\right\}$$

8. For $s = 1, \ldots, w$, Generate $\boldsymbol{\tau}_s^*$ from the following Metropolis–Hastings algorithm with proposal distribution $q_{\boldsymbol{\tau}_s}(.|\boldsymbol{\tau}_{s,t-1})$

   (a) Generate $\boldsymbol{\nu} \sim q_{\boldsymbol{\tau}_s}(\boldsymbol{\nu}|\boldsymbol{\tau}_{s,t-1})$

   (b)

$$\boldsymbol{\tau}_s^* = \begin{cases} \boldsymbol{\nu} & \text{with probability} \rho_{\boldsymbol{\tau}_s}(\boldsymbol{\tau}_{s,t-1}, \boldsymbol{\nu}) \\ \boldsymbol{\tau}_{s,t-1} & \text{with probability} 1 - \rho_{\boldsymbol{\tau}_s}(\boldsymbol{\tau}_{s,t-1}, \boldsymbol{\nu}) \end{cases}$$

   where

$$\rho_{\boldsymbol{\tau}_s}(\boldsymbol{\nu}^o, \boldsymbol{\nu}^n) = \min\left\{\frac{f_{\boldsymbol{\tau}_s}(\boldsymbol{\nu}^n) q_{\boldsymbol{\tau}_s}(\boldsymbol{\nu}^o|\boldsymbol{\nu}^n)}{f_{\boldsymbol{\tau}_s}(\boldsymbol{\nu}^o) q_{\boldsymbol{\tau}_s}(\boldsymbol{\nu}^n|\boldsymbol{\nu}^o)}, 1\right\}$$

9. For $j = 1, \ldots, p$ sample $\Sigma_{\mathbf{z}_j}^*$ from

$$\mathcal{IW}^1(\Psi_{\mathbf{z}_j} + S(\mathbf{z}_{jk}, \boldsymbol{\lambda}_{j,t-1}^v), \delta_{\mathbf{z}_j} + n_j)$$

   where

$$S(\mathbf{z}_{jk}, \boldsymbol{\lambda}_{j,t-1}) = \sum_{k=1}^{n_j} (\phi(\mathbf{z}_{jk}) - \phi(\boldsymbol{\lambda}_{j,t-1}^v))(\phi(\mathbf{z}_{jk}) - \phi(\boldsymbol{\lambda}_{j,t-1}^v))'$$

10. For $j = 1, \ldots, p$ generate $\boldsymbol{\lambda}_j^{v*}$ from the following Metropolis–Hastings algorithm with proposal distribution $q_{\boldsymbol{\lambda}_j}(.|\boldsymbol{\lambda}_{j,t-1})$

    (a) Generate $\boldsymbol{\nu} \sim q_{\boldsymbol{\lambda}_j}(\boldsymbol{\nu}|\boldsymbol{\lambda}_{j,t-1})$

    (b)

$$\boldsymbol{\lambda}_j^{v*} = \begin{cases} \boldsymbol{\nu} & \text{with probability} \rho_{\boldsymbol{\lambda}_j^v}(\boldsymbol{\lambda}_{j,t-1}^v, \boldsymbol{\nu}) \\ \boldsymbol{\lambda}_{j,t-1}^v & \text{with probability} 1 - \rho_{\boldsymbol{\lambda}_j^v}(\boldsymbol{\lambda}_{j,t-1}^v, \boldsymbol{\nu}) \end{cases}$$

    where

$$\rho_{\boldsymbol{\lambda}_j^v}(\boldsymbol{\nu}^o, \boldsymbol{\nu}^n) = \min\left\{\frac{f_{\boldsymbol{\lambda}_j^v}(\boldsymbol{\nu}^n) q_{\boldsymbol{\lambda}_j^v}(\boldsymbol{\nu}^o|\boldsymbol{\nu}^n)}{f_{\boldsymbol{\lambda}_j^v}(\boldsymbol{\nu}^o) q_{\boldsymbol{\lambda}_j^v}(\boldsymbol{\nu}^n|\boldsymbol{\nu}^o)}, 1\right\}$$

Note $\boldsymbol{\lambda}_{p,t}^v = \boldsymbol{\lambda}_j^{v*}$, that is, we don't need to do the last one twice.

11. For $j = p-1, \ldots, 1$ generate $\boldsymbol{\lambda}^v_{j,t}$ from the following Metropolis–Hastings algorithm with proposal distribution $q_{\boldsymbol{\lambda}^v_j}(.|\boldsymbol{\lambda}^{v*}_j)$

    (a) Generate $\boldsymbol{\nu} \sim q_{\boldsymbol{\lambda}^v_j}(\boldsymbol{\nu}|\boldsymbol{\lambda}^{v*}_j)$

    (b)

$$\boldsymbol{\lambda}^v_{j,t} = \begin{cases} \boldsymbol{\nu} & \text{with probability} \rho_{\boldsymbol{\lambda}^v_j}(\boldsymbol{\lambda}^{v*}_j, \boldsymbol{\nu}) \\ \boldsymbol{\lambda}^{v*}_j & \text{with probability} 1 - \rho_{\boldsymbol{\lambda}^v_j}(\boldsymbol{\lambda}^{v*}_j, \boldsymbol{\nu}) \end{cases}$$

12. For $j = p, \ldots, 1$ sample $\Sigma_{\mathbf{z}_j,t}$ from

$$\mathcal{IW}^1(\Psi_{\mathbf{z}_j} + S(\mathbf{z}_{jk}, \boldsymbol{\lambda}^v_{j,t}), \delta_{\mathbf{z}_j} + n_j)$$

    where

$$S(\mathbf{z}_{jk}, \boldsymbol{\lambda}^v_{j,t}) = \sum_{i=1}^{n_j} (\phi(\mathbf{z}_{jk}) - \phi(\boldsymbol{\lambda}^v_{j,t}))(\phi(\mathbf{z}_{jk}) - \phi(\boldsymbol{\lambda}^v_{j,t}))'$$

13. For $s = w, \ldots, 1$, Generate $\boldsymbol{\tau}_{s,t}$ from the following Metropolis–Hastings algorithm with proposal distribution $q_{\boldsymbol{\tau}_s}(.|\boldsymbol{\tau}^*_s)$

    (a) Generate $\boldsymbol{\nu} \sim q_{\boldsymbol{\tau}_s}(\boldsymbol{\nu}|\boldsymbol{\tau}^*_s)$

    (b)

$$\boldsymbol{\tau}_{s,t} = \begin{cases} \boldsymbol{\nu} & \text{with probability} \rho_{\boldsymbol{\tau}_s}(\boldsymbol{\tau}^*_s, \boldsymbol{\nu}) \\ \boldsymbol{\tau}^*_s & \text{with probability} 1 - \rho_{\boldsymbol{\tau}_s}(\boldsymbol{\tau}^*_s, \boldsymbol{\nu}) \end{cases}$$

14. For $j = p, \ldots, 1$ generate $\boldsymbol{\theta}_{j,t}$ from the following Metropolis–Hastings algorithm with proposal distribution $q_{\boldsymbol{\theta}_j}(.|\boldsymbol{\theta}^*_j)$

    (a) Generate $\boldsymbol{\nu} \sim q_{\boldsymbol{\theta}_j}(\boldsymbol{\nu}|\boldsymbol{\theta}^*_j)$

    (b)

$$\boldsymbol{\theta}_{j,t} = \begin{cases} \boldsymbol{\nu} & \text{with probability} \rho_{\boldsymbol{\theta}_j}(\boldsymbol{\theta}^*_j, \boldsymbol{\nu}) \\ \boldsymbol{\theta}^*_j & \text{with probability} 1 - \rho_{\boldsymbol{\theta}_j}(\boldsymbol{\theta}^*_j, \boldsymbol{\nu}) \end{cases}$$

15. Generate $\boldsymbol{\theta}_{\mathrm{v},t}$ from the following Metropolis–Hastings algorithm with proposal distribution $q_{\boldsymbol{\theta}_{\mathrm{v}}}(.|\boldsymbol{\theta}_{\mathrm{v},t})$

    (a) Generate $\boldsymbol{\nu} \sim q_{\boldsymbol{\theta}_{\mathrm{v}}}(\boldsymbol{\nu}|\boldsymbol{\theta}_{\mathrm{v},t})$

(b)

$$\boldsymbol{\theta}_{\mathbf{v},t} = \begin{cases} \boldsymbol{\nu} & \text{with probability} \rho_{\boldsymbol{\theta}_{\mathbf{v}}}(\boldsymbol{\theta}_{\mathbf{v},t}, \boldsymbol{\nu}) \\ \boldsymbol{\theta}^*_{\mathbf{v}} & \text{with probability} 1 - \rho_{\boldsymbol{\theta}_{\mathbf{v}}}(\boldsymbol{\theta}_{\mathbf{v},t}, \boldsymbol{\nu}) \end{cases}$$

16. Generate $\boldsymbol{\theta}_{\mathbf{u},t}$ from the following Metropolis–Hastings algorithm with proposal distribution $q_{\boldsymbol{\theta}_{\mathbf{u}}}(.|\boldsymbol{\theta}_{\mathbf{u},t})$

    (a) Generate $\boldsymbol{\nu} \sim q_{\boldsymbol{\theta}_{\mathbf{u}}}(\boldsymbol{\nu}|\boldsymbol{\theta}_{\mathbf{u},t})$

    (b)

$$\boldsymbol{\theta}_{\mathbf{u},t} = \begin{cases} \boldsymbol{\nu} & \text{with probability} \rho_{\boldsymbol{\theta}_{\mathbf{u}}}(\boldsymbol{\theta}_{\mathbf{u},t}, \boldsymbol{\nu}) \\ \boldsymbol{\theta}^*_{\mathbf{u}} & \text{with probability} 1 - \rho_{\boldsymbol{\theta}_{\mathbf{u}}}(\boldsymbol{\theta}_{\mathbf{u},t}, \boldsymbol{\nu}) \end{cases}$$

17. Sample $\Sigma_{\mathbf{v},t}$ from

$$\mathcal{IW}^{a-1}(\Psi_{\mathbf{v}} + S(\mathbf{v}_m, \boldsymbol{\theta}_{\mathbf{v},t}), \delta_{\mathbf{v}} + M)$$

    where

$$S(\mathbf{v}_m, \boldsymbol{\theta}_{\mathbf{v},t}) = \sum_{m=1}^{M} (\phi(\mathbf{v}_m) - \phi(\boldsymbol{\theta}_{\mathbf{v},t}))(\phi(\mathbf{v}_m) - \phi(\boldsymbol{\theta}_{\mathbf{v},t}))'$$

18. Sample $\Sigma_{\mathbf{u},t}$ from

$$\mathcal{IW}^{a-1}(\Psi_{\mathbf{u}} + S(\mathbf{u}_l, \boldsymbol{\theta}_{\mathbf{u},t}), \delta_{\mathbf{u}} + L)$$

    where

$$S(\mathbf{u}_l, \boldsymbol{\theta}_{\mathbf{u},t}) = \sum_{l=1}^{L} (\phi(\mathbf{u}_l) - \phi(\boldsymbol{\theta}_{\mathbf{u},t}))(\phi(\mathbf{u}_l) - \phi(\boldsymbol{\theta}_{\mathbf{u},t}))'$$

19. For $j = p, \ldots, 1$ sample $\Sigma_{\mathbf{x}_j,t}$ from

$$\mathcal{IW}^{a-1}(\Psi_{\mathbf{x}_j} + S(\mathbf{x}_{jk}, \boldsymbol{\theta}_{j,t}), \delta_{\mathbf{x}_j} + n_j)$$

    where

$$S(\mathbf{x}_{jk}, \boldsymbol{\theta}_{j,t}) = \sum_{i=1}^{n_j} (\phi(\mathbf{x}_{jk}) - \phi(\boldsymbol{\theta}_{j,t}))(\phi(\mathbf{x}_{jk}) - \phi(\boldsymbol{\theta}_{j,t}))'$$

20. Sample $\Sigma_{\boldsymbol{\epsilon},t}$ from

$$\mathcal{IW}^{a-1}(\Psi_{\boldsymbol{\epsilon}} + S(\mathbf{y}_i, \boldsymbol{\mu}_{\mathbf{y}_i}), \delta_{\boldsymbol{\epsilon}} + n)$$

where

$$S(\mathbf{y}_i, \boldsymbol{\mu}_{\mathbf{y}_i}) = \sum_{i=1}^{n} (\phi(\mathbf{y}_i) - \phi(\boldsymbol{\mu}_{\mathbf{y}_i}))(\phi(\mathbf{y}_i) - \phi(\boldsymbol{\mu}_{\mathbf{y}_i}))'$$

and

$$\begin{aligned} \boldsymbol{\mu}_{\mathbf{y}_i} &= \boldsymbol{\Theta}_t \Gamma_t^i \oplus (\boldsymbol{\theta}_{\mathbf{v},t} \ominus \boldsymbol{\theta}_{\mathbf{u},t}) \\ \Gamma_t &= \boldsymbol{\phi}_c^{-1} \left(\boldsymbol{\phi}_c \left(T_t \right) \mathbf{W} \right) \oplus_c \boldsymbol{\lambda}_t, \end{aligned}$$

$\boldsymbol{\Theta}_{t-1} = [\boldsymbol{\theta}_{1,t-1}| \dots |\boldsymbol{\theta}_{p,t-1}]$, Γ^i means the ith column of the matrix Γ, $\mathbf{T}_t = [\boldsymbol{\tau}_{1,t}| \dots |\boldsymbol{\tau}_{w,t}]$ is an $p \times w$ matrix and $\boldsymbol{\lambda}_t = (\lambda_{1,t}, \dots, \lambda_{p,t})$ is a vector of length p.

21. Repeat steps 1-20 until convergence and increment t.

To complete the MCMC algorithm we need to specify the target densities, the proposal densities and the acceptance probabilities for each of the Metropolis–Hastings steps for each of $\boldsymbol{\theta}_{\mathbf{u}}$, $\boldsymbol{\theta}_{\mathbf{v}}$, $\boldsymbol{\theta}_j$, $\boldsymbol{\tau}_s$ and $\boldsymbol{\lambda}_j^v$. The target densities are the appropriate full Gibbs conditional distributions (see Figure B.1):

$$\begin{aligned} f_{\boldsymbol{\theta}_{\mathbf{u}}}(\boldsymbol{\nu}) &= \pi(\boldsymbol{\theta}_{\mathbf{u}} = \boldsymbol{\nu} | \boldsymbol{\mu}_{\boldsymbol{\theta}_{\mathbf{u}}}, \Sigma_{\boldsymbol{\theta}_{\mathbf{u}}}) \prod_{i=1}^{n} \pi(\mathbf{y}_i | \boldsymbol{\lambda}, \boldsymbol{\Theta}, \Sigma_{\boldsymbol{\epsilon}}, \boldsymbol{\theta}_{\mathbf{u}} = \boldsymbol{\nu}, \boldsymbol{\theta}_{\mathbf{v}}, \mathbf{T}) \\ & \prod_{l=1}^{L} \pi(\mathbf{u}_l | \boldsymbol{\theta}_{\mathbf{u}} = \boldsymbol{\nu}, \Sigma_{\mathbf{u}}) \end{aligned}$$

$$\begin{aligned} f_{\boldsymbol{\theta}_{\mathbf{v}}}(\boldsymbol{\nu}) &= \pi(\boldsymbol{\theta}_{\mathbf{v}} = \boldsymbol{\nu} | \boldsymbol{\mu}_{\boldsymbol{\theta}_{\mathbf{v}}}, \Sigma_{\boldsymbol{\theta}_{\mathbf{v}}}) \prod_{i=1}^{n} \pi(\mathbf{y}_i | \boldsymbol{\lambda}, \boldsymbol{\Theta}, \Sigma_{\boldsymbol{\epsilon}}, \boldsymbol{\theta}_{\mathbf{u}}, \boldsymbol{\theta}_{\mathbf{v}} = \boldsymbol{\nu}, \mathbf{T}) \\ & \prod_{m=1}^{M} \pi(\mathbf{v}_m | \boldsymbol{\theta}_{\mathbf{v}} = \boldsymbol{\nu}, \Sigma_{\mathbf{v}}) \end{aligned}$$

$$\begin{aligned} f_{\boldsymbol{\theta}_j}(\boldsymbol{\nu}) &= \pi(\boldsymbol{\theta}_j = \boldsymbol{\nu} | \boldsymbol{\mu}_{\boldsymbol{\theta}_j}, \Sigma_{\boldsymbol{\theta}_j}) \prod_{i=1}^{n} \pi(\mathbf{y}_i | \boldsymbol{\lambda}, \boldsymbol{\theta}_j = \boldsymbol{\nu}, \boldsymbol{\theta}_{-j}, \Sigma_{\boldsymbol{\epsilon}}, \boldsymbol{\theta}_{\mathbf{u}}, \boldsymbol{\theta}_{\mathbf{v}}, \mathbf{T}) \\ & \prod_{k=1}^{n_j} \pi(\mathbf{x}_k | \boldsymbol{\theta}_{\mathbf{x}_j}, \Sigma_{\mathbf{x}_j}) \end{aligned}$$

$$f_{\boldsymbol{\tau}_s}(\boldsymbol{\nu}) = \pi(\boldsymbol{\tau}_s = \boldsymbol{\nu}|\boldsymbol{\mu}_{\boldsymbol{\tau}_s}, \Sigma_{\boldsymbol{\tau}_s}) \prod_{i=1}^{n} \pi(\mathbf{y}_i|\boldsymbol{\lambda}, \boldsymbol{\Theta}, \Sigma_{\boldsymbol{\epsilon}}, \boldsymbol{\theta}_{\mathbf{u}}, \boldsymbol{\theta}_{\mathbf{v}}, \boldsymbol{\tau}_s = \boldsymbol{\nu}, \boldsymbol{\tau}_{-s})$$

$$\begin{aligned} f_{\boldsymbol{\lambda}_j^v}(\boldsymbol{\nu}) &= \pi(\boldsymbol{\lambda}_j = \boldsymbol{\nu}|\boldsymbol{\mu}_{\boldsymbol{\lambda}_j}, \Sigma_{\boldsymbol{\lambda}_j}) \prod_{i=1}^{n} \pi(\mathbf{y}_i|\boldsymbol{\lambda}_j^v = \boldsymbol{\nu}, \boldsymbol{\lambda}_{-j}, \boldsymbol{\Theta}, \Sigma_{\boldsymbol{\epsilon}}, \boldsymbol{\theta}_{\mathbf{u}}, \boldsymbol{\theta}_{\mathbf{v}}, \mathbf{T}) \\ & \prod_{k=1}^{n_j} \pi(\mathbf{z}_k|\boldsymbol{\theta}_{\mathbf{z}_j}, \Sigma_{\mathbf{z}_j}). \end{aligned}$$

The target density for $\boldsymbol{\tau}_s$ can be simplified depending on the form of the matrix $\mathbf{W}$.

The following logistic normal proposal distributions where used

$$\begin{aligned} q_{\boldsymbol{\theta}_{\mathbf{u}}}(\boldsymbol{\nu}^n|\boldsymbol{\nu}^o) &= \mathcal{L}^{a-1}(\phi(\boldsymbol{\nu}^o), \beta_{\boldsymbol{\theta}_{\mathbf{u}}}(\mathbf{I}+\mathbf{J})), \\ q_{\boldsymbol{\theta}_{\mathbf{v}}}(\boldsymbol{\nu}^n|\boldsymbol{\nu}^o) &= \mathcal{L}^{a-1}(\phi(\boldsymbol{\nu}^o), \beta_{\boldsymbol{\theta}_{\mathbf{v}}}(\mathbf{I}+\mathbf{J})), \\ q_{\boldsymbol{\theta}_j}(\boldsymbol{\nu}^n|\boldsymbol{\nu}^o) &= \mathcal{L}^{a-1}(\phi(\boldsymbol{\nu}^o), \beta_{\boldsymbol{\theta}_j}(\mathbf{I}+\mathbf{J})), \\ q_{\boldsymbol{\tau}_s}(\boldsymbol{\nu}^n|\boldsymbol{\nu}^o) &= \mathcal{L}^{p-1}(\phi(\boldsymbol{\nu}^o), \beta_{\boldsymbol{\tau}_s}(\mathbf{I}+\mathbf{J})), \\ q_{\boldsymbol{\lambda}_j^v}(\boldsymbol{\nu}^n|\boldsymbol{\nu}^o) &= \mathcal{L}^{1}(\phi(\boldsymbol{\nu}^o), \beta_{\boldsymbol{\lambda}_j}(\mathbf{I}+\mathbf{J})), \end{aligned}$$

where $\beta_{\boldsymbol{\theta}_{\mathbf{u}}}$, $\beta_{\boldsymbol{\theta}_{\mathbf{v}}}$,$\beta_{\boldsymbol{\theta}_j}$, $\beta_{\boldsymbol{\tau}_s}$ and $\beta_{\boldsymbol{\lambda}_j}$ are scale factors that control the acceptance rates for the Metropolis–Hastings algorithm.

The acceptance probabilities $\rho_{\boldsymbol{\theta}_j}(\boldsymbol{\nu}^o, \boldsymbol{\nu}^n)$, $\rho_{\boldsymbol{\theta}_j}(\boldsymbol{\nu}^o, \boldsymbol{\nu}^n)$, $\rho_{\boldsymbol{\theta}_j}(\boldsymbol{\nu}^o, \boldsymbol{\nu}^n)$ $\rho_{\boldsymbol{\tau}_s},(\boldsymbol{\nu}^o, \boldsymbol{\nu}^n)$ and $\rho_{\boldsymbol{\lambda}_j},(\boldsymbol{\nu}^o, \boldsymbol{\nu}^n)$ are given by

$$\begin{aligned} \rho_{\boldsymbol{\theta}_{\mathbf{u}}}(\boldsymbol{\nu}^o, \boldsymbol{\nu}^n) &= \min\left\{ \frac{f_{\boldsymbol{\theta}_{\mathbf{u}}}(\boldsymbol{\nu}^n) \prod_{j=1}^{p} [\boldsymbol{\nu}^n]_j}{f_{\boldsymbol{\theta}_{\mathbf{u}}}(\boldsymbol{\nu}^o) \prod_{j=1}^{p} [\boldsymbol{\nu}^o]_j}, 1 \right\} \\ \rho_{\boldsymbol{\theta}_{\mathbf{v}}}(\boldsymbol{\nu}^o, \boldsymbol{\nu}^n) &= \min\left\{ \frac{f_{\boldsymbol{\theta}_{\mathbf{v}}}(\boldsymbol{\nu}^n) \prod_{j=1}^{p} [\boldsymbol{\nu}^n]_j}{f_{\boldsymbol{\theta}_{\mathbf{v}}}(\boldsymbol{\nu}^o) \prod_{j=1}^{p} [\boldsymbol{\nu}^o]_j}, 1 \right\} \\ \rho_{\boldsymbol{\theta}_j}(\boldsymbol{\nu}^o, \boldsymbol{\nu}^n) &= \min\left\{ \frac{f_{\boldsymbol{\theta}_j}(\boldsymbol{\nu}^n) \prod_{j=1}^{p} [\boldsymbol{\nu}^n]_j}{f_{\boldsymbol{\theta}_j}(\boldsymbol{\nu}^o) \prod_{j=1}^{p} [\boldsymbol{\nu}^o]_j}, 1 \right\} \\ \rho_{\boldsymbol{\tau}_s}(\boldsymbol{\nu}^o, \boldsymbol{\nu}^n) &= \min\left\{ \frac{f_{\boldsymbol{\tau}_s}(\boldsymbol{\nu}^n) \prod_{j=1}^{p} [\boldsymbol{\nu}^n]_j}{f_{\boldsymbol{\tau}_s}(\boldsymbol{\nu}^o) \prod_{j=1}^{p} [\boldsymbol{\nu}^o]_j}, 1 \right\} \\ \rho_{\boldsymbol{\lambda}_j^v}(\boldsymbol{\nu}^o, \boldsymbol{\nu}^n) &= \min\left\{ \frac{f_{\boldsymbol{\lambda}_j}(\boldsymbol{\nu}^n) \prod_{j=1}^{p} [\boldsymbol{\nu}^n]_j}{f_{\boldsymbol{\lambda}_j}(\boldsymbol{\nu}^o) \prod_{j=1}^{p} [\boldsymbol{\nu}^o]_j}, 1 \right\} \end{aligned}$$

The acceptance probabilities are very similar to the random walk Metropolis–Hastings proposal, however, the logistic normal is not symmetric due to the Jacobin of the transformation $\prod_{i=1}^{a}[\mathbf{z}]_i^{-1}$.

B.2 Individual Diet Model

This section gives details for the reversible systematic scan Metropolis–within-Gibbs algorithm for the individual diet model.

$$\begin{aligned}
\underset{(a\times nr)}{\mathbf{Y}} &= \left[\left(\underset{(a\times p)}{\mathbf{\Theta}}\underset{(p\times n)}{\Gamma} \otimes \underset{(1\times r)}{\mathbf{U}}\right) \oplus_c \left(\underset{(a\times 1)}{\boldsymbol{\theta}_{\mathbf{v}}} \ominus \underset{(a\times 1)}{\boldsymbol{\theta}_{\mathbf{u}}}\right)\right] \oplus_c \underset{(a\times nr)}{\mathbf{E}}\,, \\
\underset{(p\times n)}{\Gamma} &= \boldsymbol{\phi}_c^{-1}\left(\boldsymbol{\phi}_c\left(\underset{(p\times w)}{\mathbf{T}}\right)\underset{(w\times n)}{\mathbf{W}} + \boldsymbol{\phi}_c\left(\underset{(p\times n)}{\Gamma^m}\right)\right) \oplus_c \underset{(p\times 1)}{\boldsymbol{\lambda}}\,, \\
\underset{(a\times n_j)}{\mathbf{X}_j} &= \underset{(a\times 1)}{\boldsymbol{\theta}_j}\underset{(1\times n_j)}{\mathbf{W}_{\mathbf{X}_j}} \oplus_c \underset{(a\times n_j)}{\mathbf{E}_{\mathbf{X}_j}}\,, \quad j = 1,\ldots,p, \\
\underset{(2\times n_j)}{\mathbf{Z}_j} &= \underset{(2\times 1)}{\boldsymbol{\lambda}_j^v}\underset{(1\times n_j)}{\mathbf{W}_{\mathbf{Z}_j}} \oplus_c \underset{(a\times n_j)}{\mathbf{E}_{\mathbf{Z}_j}}\,, \quad j = 1,\ldots,p, \\
\underset{(a\times L)}{\mathbf{U}} &= \underset{(a\times 1)}{\boldsymbol{\theta}_{\mathbf{u}}}\underset{(1\times L)}{\mathbf{W}_{\mathbf{U}}} \oplus_c \underset{(a\times L)}{\mathbf{E}_{\mathbf{U}}}\,, \\
\underset{(a\times M)}{\mathbf{V}} &= \underset{(a\times 1)}{\boldsymbol{\theta}_{\mathbf{v}}}\underset{(1\times M)}{\mathbf{W}_{\mathbf{V}}} \oplus_c \underset{(a\times M)}{\mathbf{E}_{\mathbf{V}}}
\end{aligned}$$

where $\underset{(a\times nr)}{\mathbf{Y}} = [\mathbf{y}_{11}|\ldots|\mathbf{y}_{1r}|\ldots,\mathbf{y}_{n1}|\ldots|\mathbf{y}_{nr}]$ is an $a \times nr$ matrix of observations on the predators, $\underset{(a\times p)}{\mathbf{\Theta}} = [\boldsymbol{\theta}_1|\ldots|\boldsymbol{\theta}_p]$ is an $a\times p$ matrix of predator(source) profiles, $\underset{(p\times n)}{\Gamma}$ is an $p\times n$ matrix of individual diet(mixing) compositions adjusted for fat content, $\otimes$ represents the Kronecker product of two matrices defined in the previous chapter, $\underset{(1\times r)}{\mathbf{U}}$ is an $1\times r$ matrix of ones which allows the predicted profile to be replicated r times to account for replicate measurements on the same predator, $\oplus_c$ is the perturbation operator applied column–wise for matrices of the same size (if the second argument is a vector, then the vector is replicated column–wise first then applied column wise), $\underset{(a\times 1)}{\boldsymbol{\theta}_{\mathbf{v}}}$ is an a–dimensional fatty acid profile of the calibration predator, $\underset{(a\times 1)}{\boldsymbol{\theta}_{\mathbf{u}}}$ is an a–dimensional fatty acid profile of the calibration prey, $\ominus$ is the inverse perturbation operator ($\mathbf{x} \ominus \mathbf{y} = \mathbf{x} \oplus \mathbf{y}^{-1}$), $\underset{(a\times nr)}{\mathbf{E}}$ is an $a\times nr$ dimensional matrix of compositional errors, $\boldsymbol{\phi}_c^{-1}$ is the logistic transformation applied column–wise, $\boldsymbol{\phi}_c$ is the log–ratio transformation, $\underset{(p\times w)}{\mathbf{T}} = [\boldsymbol{\mu}_{\boldsymbol{\tau}_1}|\ldots|\boldsymbol{\mu}_{\boldsymbol{\tau}_w}]$ is a $p\times w$ matrix of population diet compositions (table 6.2 gives the required matrices for the synthetic data), $\underset{(p\times n)}{\Gamma^m} = [\boldsymbol{\tau}_1^m|\ldots|\boldsymbol{\tau}_n^m]$ are samples from the multilevel distribution in our case the logistic normal with zero mean and covariance matrix $\Sigma_{\boldsymbol{\tau}}$, $\boldsymbol{\lambda}$ the p dimensional vector of fat contents for each prey type, $\underset{(a\times n_j)}{\mathbf{X}_j}$ is an $a \times n_j$ matrix of samples of the fatty acid profiles from the jth prey type, $\boldsymbol{\theta}_j$ is an a–dimensional vector consisting of the measure of location for the fatty acid profile of the jth prey type, $\underset{(1\times n_j)}{\mathbf{W}_{\mathbf{X}_j}}$ is a $1 \times n_j$ matrix of ones, $\underset{(a\times n_j)}{\mathbf{E}_{\mathbf{X}_j}}$ is an $a \times n_j$ dimensional matrix of compositional errors. We have similar definitions for the remaining parameters and observations for the model, however, note $\boldsymbol{\lambda}_j^v = (\lambda_j, 1-\lambda_j)'$ is the vector of fat and non–fat, similarly for the observations $\mathbf{Z}_j$. Let $\boldsymbol{\mu}_{\boldsymbol{\tau}_j} = \boldsymbol{\phi}(\boldsymbol{\tau}_j^p)$, or $\boldsymbol{\phi}_c(\mathbf{T}) = [\boldsymbol{\mu}_{\boldsymbol{\tau}_1}|\ldots|\boldsymbol{\mu}_{\boldsymbol{\tau}_w}]$, that is, $\boldsymbol{\mu}_{\boldsymbol{\tau}_j}$ are means of the logistic normal distributions of the individual populations. Note that, $\underset{(p\times n)}{\Gamma} = [\boldsymbol{\tau}_1|\ldots|\boldsymbol{\tau}_n$ consists samples from the mixing distribution with zero mean added to the appropriate population mean.

To complete the model specification we ass the following prior distributions for the location parameters

$$\begin{aligned}
\pi(\boldsymbol{\mu}_{\boldsymbol{\tau}_r}|\boldsymbol{\eta}, \Sigma_{\boldsymbol{\mu}_{\boldsymbol{\tau}_r}}) &\sim \mathcal{MN}^{p-1}(\boldsymbol{\eta}_r, \Sigma_{\boldsymbol{\mu}_{\boldsymbol{\tau}_r}}), r = 1, \ldots, w \\
\pi(\boldsymbol{\theta}_j|\boldsymbol{\mu}_{\theta_j}, \Sigma_{\theta_j}) &\sim \mathcal{L}^a(\boldsymbol{\mu}_{\theta_j}, \Sigma_{\theta_j}), \ \ j = 1, \ldots, p, \\
\pi(\boldsymbol{\lambda}_j^v|\mu_{\boldsymbol{\lambda}_j}, \Sigma_{\boldsymbol{\lambda}_j}) &\sim \mathcal{L}^2(\boldsymbol{\mu}_{\boldsymbol{\lambda}_j}, \Sigma_{\boldsymbol{\lambda}_j}), \ \ j = 1, \ldots, p, \\
\pi(\boldsymbol{\theta}_{\mathbf{u}}|\boldsymbol{\mu}_{\theta_{\mathbf{u}}}, \Sigma_{\theta_{\mathbf{u}}}) &\sim \mathcal{L}^a(\boldsymbol{\mu}_{\theta_{\mathbf{u}}}, \Sigma_{\theta_{\mathbf{u}}}), \\
\pi(\boldsymbol{\theta}_{\mathbf{v}}|\boldsymbol{\mu}_{\theta_{\mathbf{v}}}, \Sigma_{\theta_{\mathbf{v}}}) &\sim \mathcal{L}^a(\boldsymbol{\mu}_{\theta_{\mathbf{v}}}, \Sigma_{\theta_{\mathbf{v}}}),
\end{aligned}$$

and for the covariance matrices

$$\begin{aligned}
\pi(\Sigma_{\boldsymbol{\tau}}|\delta_{\boldsymbol{\tau}}, \Psi_{\boldsymbol{\tau}})) &\sim \mathcal{IW}^{p-1}(\delta_{\boldsymbol{\tau}}, \Psi_{\boldsymbol{\tau}}), \\
\pi(\Sigma_{\mathbf{x}_j}|\delta_{\mathbf{x}_j}, \Psi_{\mathbf{x}_j}) &\sim \mathcal{IW}^{a-1}(\delta_{\mathbf{x}_j}, \Psi_{\mathbf{x}_j}), \ \ j = 1, \ldots, p, \\
\pi(\Sigma_{\mathbf{z}_j}|\delta_{\mathbf{z}_j}, \Psi_{\mathbf{z}_j}) &\sim \mathcal{IW}^{1}(\delta_{\mathbf{z}_j}, \Psi_{\mathbf{z}_j}), \ \ j = 1, \ldots, p, \\
\pi(\Sigma_{\boldsymbol{\epsilon}}|\delta_{\boldsymbol{\epsilon}}, \Psi_{\boldsymbol{\epsilon}}) &\sim \mathcal{IW}^{a-1}(\delta_{\boldsymbol{\epsilon}}, \Psi_{\boldsymbol{\epsilon}}), \\
\pi(\Sigma_{\mathbf{u}}|\delta_{\mathbf{u}}, \Psi_{\mathbf{u}}) &\sim \mathcal{IW}^{a-1}(\delta_{\mathbf{u}}, \Psi_{\mathbf{u}}), \\
\pi(\Sigma_{\mathbf{v}}|\delta_{\mathbf{v}}, \Psi_{\mathbf{v}}) &\sim \mathcal{IW}^{a-1}(\delta_{\mathbf{v}}, \Psi_{\mathbf{v}}).
\end{aligned}$$

The sampling distributions are given by

$$\begin{aligned}
\pi(\mathbf{y}_{is}|\boldsymbol{\Theta},\boldsymbol{\tau}_i,\Sigma_{\boldsymbol{\epsilon}},\boldsymbol{\lambda},\boldsymbol{\theta}_{\mathbf{u}},\boldsymbol{\theta}_{\mathbf{v}}) &\sim \mathcal{L}^a\left(\boldsymbol{\phi}\left(\left[\underset{(a\times p)}{\boldsymbol{\Theta}}\underset{(p\times 1)}{\Gamma^i}\oplus_c\left(\underset{(a\times 1)}{\boldsymbol{\theta}_{\mathbf{v}}}\ominus\underset{(a\times 1)}{\boldsymbol{\theta}_{\mathbf{u}}}\right)\right]\right),\Sigma_{\boldsymbol{\epsilon}}\right), \quad \begin{matrix} i=1,\dots,n \\ s=1,\dots,r \end{matrix}\\
\pi(\mathbf{x}_{jk}|\boldsymbol{\theta}_j,\Sigma_{\mathbf{x}_j}) &\sim \mathcal{L}^a(\boldsymbol{\phi}(\boldsymbol{\theta}_j),\Sigma_{\mathbf{x}_j}),\ \ j=1,\dots,p;\ \ k=1,\dots,n_j,\\
\pi(\mathbf{z}_{jk}|\boldsymbol{\lambda}_j^v,\Sigma_{\mathbf{z}_j}) &\sim \mathcal{L}^2(\boldsymbol{\phi}(\boldsymbol{\lambda}_j^v),\Sigma_{\mathbf{z}_j}),\ \ j=1,\dots,p;\ \ k=1,\dots,n_j,\\
\pi(\mathbf{u}_l|\boldsymbol{\theta}_{\mathbf{u}},\Sigma_{\mathbf{u}}) &\sim \mathcal{L}^a(\boldsymbol{\phi}(\boldsymbol{\theta}_{\mathbf{u}}),\Sigma_{\mathbf{u}}),\ \ l=1,\dots,L,\\
\pi(\mathbf{v}_m|\boldsymbol{\theta}_{\mathbf{v}},\Sigma_{\mathbf{v}}) &\sim \mathcal{L}^a(\boldsymbol{\phi}(\boldsymbol{\theta}_{\mathbf{v}}),\Sigma_{\mathbf{v}}),\ \ m=1,\dots,M,
\end{aligned}$$

and the mixing distribution is given by

$$\pi(\boldsymbol{\tau}_i|\mathbf{0},\Sigma_{\boldsymbol{\tau}})\sim\mathcal{L}^p(\mathbf{0},\Sigma_{\boldsymbol{\tau}})$$

where $\mathbf{0}$ is a $p-1$ dimensional vector of zeros. We assume that the mixing distribution has the same covariance matrix in each population, although, this assumption could be easily adapted to allow a different mixing covariance per population. By comparison with more traditional analysis of variance models our individual diet model can be seen as a mixed model with the design matrix $\mathbf{W}$ giving the fixed effects and the covariance matrix $\Sigma_{\boldsymbol{\tau}}$ playing the role of controlling the amount of variability in the random effects.

Proposition B.2. *The posterior distribution for the individual diet model is proper.*

Proof. The proof is straightforward using the same bounding techniques used in the constant diet model. □

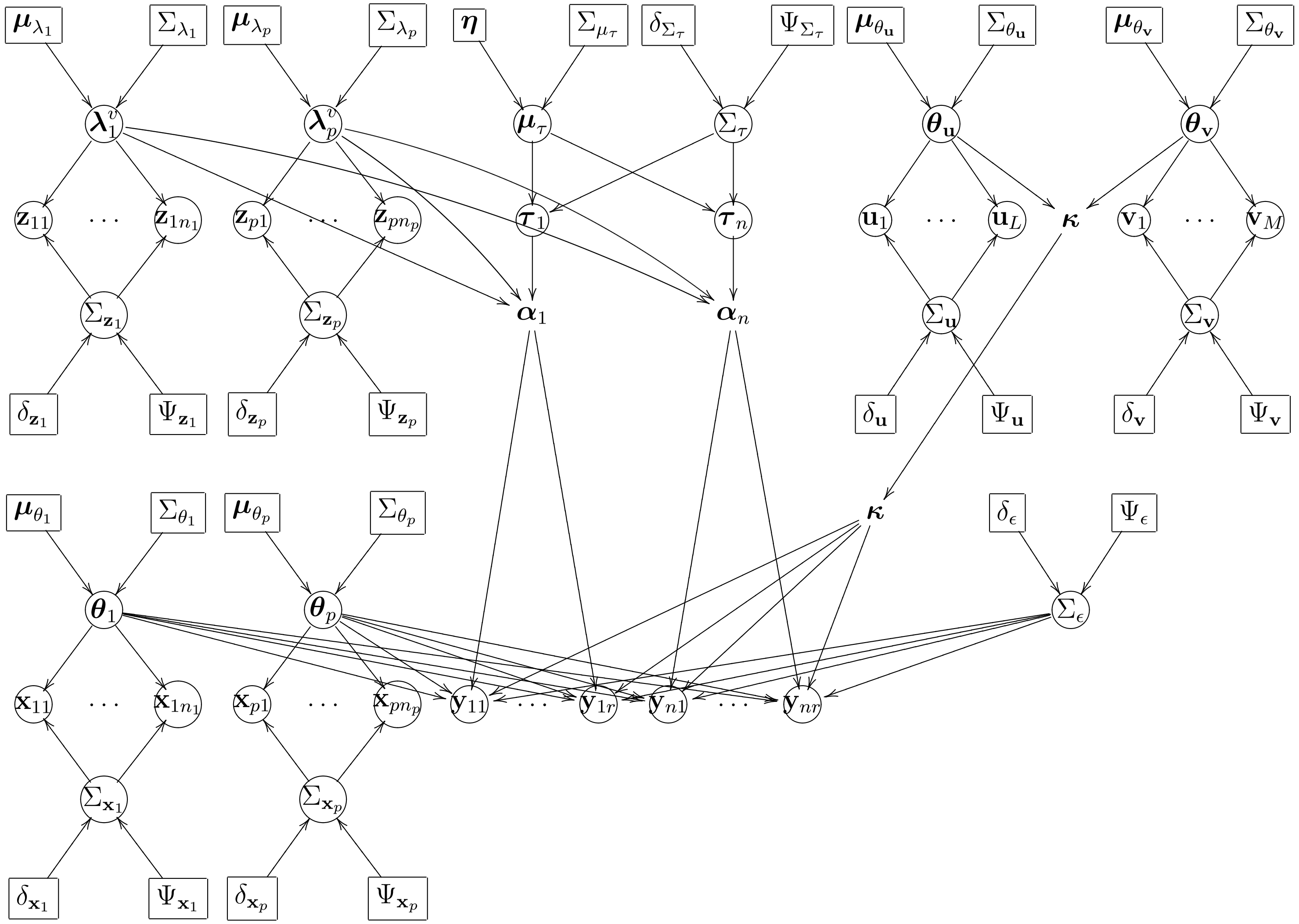

Figure B.2: DAG: Individual Diet Model with calibration and fat content. Note $\boldsymbol{\kappa} = \boldsymbol{\theta}_\mathbf{v} \oplus \boldsymbol{\theta}_\mathbf{u}$, $\boldsymbol{\alpha}_i = \boldsymbol{\tau}_i \oplus \boldsymbol{\lambda}$ and $\mathbf{Y} = [(\boldsymbol{\Theta}\Gamma \otimes \mathbf{U}) \oplus_c (\boldsymbol{\theta}_\mathbf{v} \ominus \boldsymbol{\theta}_\mathbf{u})] \oplus_c \mathbf{E}$. Nodes that are not contained in circles or squares are derived variables are used to simplify the graph.

The reversible systematic scan Metropolis–within–Gibbs sampler for the constant diet model is given below

0. Choose starting values for $\boldsymbol{\tau}_{i,0}, i = 1, \ldots, n$, $\boldsymbol{\theta}_{j,0}, j = 1, \ldots, p$, $\lambda_{j,0}, j = 1, \ldots, p$, $\boldsymbol{\theta}_{\mathbf{u},0}$ and $\boldsymbol{\theta}_{\mathbf{v},0}$

1. Sample $\Sigma^*_{\boldsymbol{\epsilon}}$ from

$$\mathcal{IW}^{a-1}\left(\Psi_{\boldsymbol{\epsilon}} + S\left(\mathbf{y}_{ir}, \boldsymbol{\mu}_{\mathbf{y}_i}\right), \delta_{\boldsymbol{\epsilon}} + nr\right)$$

   where

$$S\left(\mathbf{y}_{ir}, \boldsymbol{\mu}_{\mathbf{y}_i}\right) = \sum_{i=1}^{n}\sum_{r=1}^{r}\left(\phi\left(\mathbf{y}_{ir}\right) - \phi\left(\boldsymbol{\mu}_{\mathbf{y}_i}\right)\right)\left(\phi\left(\mathbf{y}_{ir}\right) - \phi\left(\boldsymbol{\mu}_{\mathbf{y}_i}\right)\right)',$$

$$\boldsymbol{\mu}_{\mathbf{y}_i} = \left(\boldsymbol{\Theta}_{t-1}\left(\boldsymbol{\tau}_{i,t-1} \oplus \boldsymbol{\lambda}_{t-1}\right)\right) \oplus \left(\boldsymbol{\theta}_{\mathbf{v},t-1} \ominus \boldsymbol{\theta}_{\mathbf{u},t-1}\right)$$

   $\boldsymbol{\Theta}_{t-1} = [\boldsymbol{\theta}_{1,t-1}|\ldots|\boldsymbol{\theta}_{p,t-1}]$ is a $a \times p$ matrix of prey fatty acid profiles and $\boldsymbol{\lambda}_{t-1} = (\lambda_{1,t-1}, \ldots, \lambda_{p,t-1})$ is a vector of length p of fat contents.

2. For $j = 1, \ldots, p$ sample $\Sigma^*_{\mathbf{x}_j}$ from

$$\mathcal{IW}^{a-1}\left(\Psi_{\mathbf{x}_j} + S\left(\mathbf{x}_{jk}, \boldsymbol{\theta}_{j,t-1}\right), \delta_{\mathbf{x}_j} + n_j\right)$$

   where

$$S\left(\mathbf{x}_{jk}, \boldsymbol{\theta}_{j,t-1}\right) = \sum_{k=1}^{n_j}\left(\phi\left(\mathbf{x}_{jk}\right) - \phi\left(\boldsymbol{\theta}_{j,t-1}\right)\right)\left(\phi\left(\mathbf{x}_{jk}\right) - \phi\left(\boldsymbol{\theta}_{j,t-1}\right)\right)'$$

3. Sample $\Sigma^*_{\mathbf{u}}$ from

$$\mathcal{IW}^{a-1}\left(\Psi_{\mathbf{u}} + S\left(\mathbf{u}_l, \boldsymbol{\theta}_{\mathbf{u},t-1}\right), \delta_{\mathbf{u}} + L\right)$$

   where

$$S\left(\mathbf{u}_l, \boldsymbol{\theta}_{\mathbf{u},t-1}\right) = \sum_{l=1}^{L}\left(\phi\left(\mathbf{u}_l\right) - \phi\left(\boldsymbol{\theta}_{\mathbf{u},t-1}\right)\right)\left(\phi\left(\mathbf{u}_l\right) - \phi\left(\boldsymbol{\theta}_{\mathbf{u},t-1}\right)\right)'$$

4. Sample $\Sigma^*_{\mathbf{v}}$ from

$$\mathcal{IW}^{a-1}\left(\Psi_{\mathbf{v}} + S\left(\mathbf{v}_m, \boldsymbol{\theta}_{\mathbf{v},t-1}\right), \delta_{\mathbf{v}} + M\right)$$

where

$$S\left(\mathbf{v}_m, \boldsymbol{\theta}_{\mathbf{v},t-1}\right) = \sum_{m=1}^{M} \left(\phi\left(\mathbf{v}_m\right) - \phi\left(\boldsymbol{\theta}_{\mathbf{v},t-1}\right)\right)\left(\phi\left(\mathbf{v}_m\right) - \phi\left(\boldsymbol{\theta}_{\mathbf{v},t-1}\right)\right)'$$

5. Generate $\boldsymbol{\theta}_{\mathbf{u}}^{*}$ from the following Metropolis–Hastings algorithm with proposal distribution $q_{\boldsymbol{\theta}_{\mathbf{u}}}\left(.|\boldsymbol{\theta}_{\mathbf{u},t-1}\right)$

   (a) Generate $\boldsymbol{\nu} \sim q_{\boldsymbol{\theta}_{\mathbf{u}}}\left(\boldsymbol{\nu}|\boldsymbol{\theta}_{\mathbf{u},t-1}\right)$

   (b)
$$\boldsymbol{\theta}_{\mathbf{u}}^{*} = \begin{cases} \boldsymbol{\nu} & \text{with probability} \rho_{\boldsymbol{\theta}_{\mathbf{u}}}\left(\boldsymbol{\theta}_{\mathbf{u},t-1}, \boldsymbol{\nu}\right) \\ \boldsymbol{\theta}_{\mathbf{u},t-1} & \text{with probability} 1 - \rho_{\boldsymbol{\theta}_{\mathbf{u}}}\left(\boldsymbol{\theta}_{\mathbf{u},t-1}, \boldsymbol{\nu}\right) \end{cases}$$

   where
$$\rho_{\boldsymbol{\theta}_{\mathbf{u}}}\left(\boldsymbol{\nu}^{o}, \boldsymbol{\nu}^{n}\right) = \min\left\{\frac{f_{\boldsymbol{\theta}_{\mathbf{u}}}\left(\boldsymbol{\nu}^{n}\right) q_{\boldsymbol{\theta}_{\mathbf{u}}}\left(\boldsymbol{\nu}^{o}|\boldsymbol{\nu}^{n}\right)}{f_{\boldsymbol{\theta}_{\mathbf{u}}}\left(\boldsymbol{\nu}^{o}\right) q_{\boldsymbol{\theta}_{\mathbf{u}}}\left(\boldsymbol{\nu}^{n}|\boldsymbol{\nu}^{o}\right)}, 1\right\}$$

6. Generate $\boldsymbol{\theta}_{\mathbf{v}}^{*}$ from the following Metropolis–Hastings algorithm with proposal distribution $q_{\boldsymbol{\theta}_{\mathbf{v}}}\left(.|\boldsymbol{\theta}_{\mathbf{v},t-1}\right)$

   (a) Generate $\boldsymbol{\nu} \sim q_{\boldsymbol{\theta}_{\mathbf{v}}}\left(\boldsymbol{\nu}|\boldsymbol{\theta}_{\mathbf{v},t-1}\right)$

   (b)
$$\boldsymbol{\theta}_{\mathbf{v}}^{*} = \begin{cases} \boldsymbol{\nu} & \text{with probability} \rho_{\boldsymbol{\theta}_{\mathbf{v}}}\left(\boldsymbol{\theta}_{\mathbf{v},t-1}, \boldsymbol{\nu}\right) \\ \boldsymbol{\theta}_{\mathbf{v},t-1} & \text{with probability} 1 - \rho_{\boldsymbol{\theta}_{\mathbf{v}}}\left(\boldsymbol{\theta}_{\mathbf{v},t-1}, \boldsymbol{\nu}\right) \end{cases}$$

   where
$$\rho_{\boldsymbol{\theta}_{\mathbf{v}}}\left(\boldsymbol{\nu}^{o}, \boldsymbol{\nu}^{n}\right) = \min\left\{\frac{f_{\boldsymbol{\theta}_{\mathbf{v}}}\left(\boldsymbol{\nu}^{n}\right) q_{\boldsymbol{\theta}_{\mathbf{v}}}\left(\boldsymbol{\nu}^{o}|\boldsymbol{\nu}^{n}\right)}{f_{\boldsymbol{\theta}_{\mathbf{v}}}\left(\boldsymbol{\nu}^{o}\right) q_{\boldsymbol{\theta}_{\mathbf{v}}}\left(\boldsymbol{\nu}^{n}|\boldsymbol{\nu}^{o}\right)}, 1\right\}$$

7. For $j = 1, \ldots, p$ generate $\boldsymbol{\theta}_{j}^{*}$ from the following Metropolis–Hastings algorithm with proposal distribution $q_{\boldsymbol{\theta}_{j}}\left(.|\boldsymbol{\theta}_{j,t-1}\right)$

   (a) Generate $\boldsymbol{\nu} \sim q_{\boldsymbol{\theta}_{j}}\left(\boldsymbol{\nu}|\boldsymbol{\theta}_{j,t-1}\right)$

   (b)
$$\boldsymbol{\theta}_{j}^{*} = \begin{cases} \boldsymbol{\nu} & \text{with probability} \rho_{\boldsymbol{\theta}_{j}}\left(\boldsymbol{\theta}_{j,t-1}, \boldsymbol{\nu}\right) \\ \boldsymbol{\theta}_{j,t-1} & \text{with probability} 1 - \rho_{\boldsymbol{\theta}_{j}}\left(\boldsymbol{\theta}_{j,t-1}, \boldsymbol{\nu}\right) \end{cases}$$

   where
$$\rho_{\boldsymbol{\theta}_{j}}\left(\boldsymbol{\nu}^{o}, \boldsymbol{\nu}^{n}\right) = \min\left\{\frac{f_{\boldsymbol{\theta}_{j}}\left(\boldsymbol{\nu}^{n}\right) q_{\boldsymbol{\theta}_{j}}\left(\boldsymbol{\nu}^{o}|\boldsymbol{\nu}^{n}\right)}{f_{\boldsymbol{\theta}_{j}}\left(\boldsymbol{\nu}^{o}\right) q_{\boldsymbol{\theta}_{j}}\left(\boldsymbol{\nu}^{n}|\boldsymbol{\nu}^{o}\right)}, 1\right\}$$

8. For $i = 1, \dots, n$, Generate $\boldsymbol{\tau}_i^*$ from the following Metropolis–Hastings algorithm with proposal distribution $q_{\boldsymbol{\tau}_i}(.|\boldsymbol{\tau}_{i,t-1})$

   (a) Generate $\boldsymbol{\nu} \sim q_{\boldsymbol{\tau}_i}(\boldsymbol{\nu}|\boldsymbol{\tau}_{i,t-1})$

   (b)
   $$\boldsymbol{\tau}_i^* = \begin{cases} \boldsymbol{\nu} & \text{with probability} \rho_{\boldsymbol{\tau}_i}(\boldsymbol{\tau}_{i,t-1}, \boldsymbol{\nu}) \\ \boldsymbol{\tau}_{i,t-1} & \text{with probability} 1 - \rho_{\boldsymbol{\tau}_i}(\boldsymbol{\tau}_{i,t-1}, \boldsymbol{\nu}) \end{cases}$$

   where
   $$\rho_{\boldsymbol{\tau}_i}(\boldsymbol{\nu}^o, \boldsymbol{\nu}^n) = \min\left\{ \frac{f_{\boldsymbol{\tau}_i}(\boldsymbol{\nu}^n) q_{\boldsymbol{\tau}_i}(\boldsymbol{\nu}^o|\boldsymbol{\nu}^n)}{f_{\boldsymbol{\tau}_i}(\boldsymbol{\nu}^o) q_{\boldsymbol{\tau}_i}(\boldsymbol{\nu}^n|\boldsymbol{\nu}^o)}, 1 \right\}$$

9. Sample $\mathbf{T}^*$ by generating a vector $\boldsymbol{\delta}$ from the following $(p-1)w$ dimensional normal distribution
   $$\mathcal{MN}^{(p-1)w}(\boldsymbol{\xi}, \Sigma)$$

   where
   $$\begin{aligned} \Sigma &= (\Delta + \Omega)^{-1}, \\ \boldsymbol{\xi} &= \Sigma(\Delta \boldsymbol{\eta}_v + \Omega \boldsymbol{\beta}). \end{aligned}$$

   And where
   $$\begin{aligned} \Delta &= I_{w \times w} \otimes \Psi_{\Sigma_{\boldsymbol{\tau}}}^{-1}, \\ \Omega &= (\mathbf{W}\mathbf{W}^{'}) \otimes \Sigma_{\boldsymbol{\tau},t-1}^{-1}, \\ \boldsymbol{\beta} &= \text{vec}\{\mathbf{G}\mathbf{W}^{'}(\mathbf{W}\mathbf{W}^{'})^{-1}\}, \end{aligned}$$

   $\boldsymbol{\eta}_v = \text{vec}([\boldsymbol{\eta}_1| \dots |\boldsymbol{\eta}_w])$ is a vector of length $(p-1)w$, $I_{w \times w}$ is an $w \times w$ identity matrix,$\mathbf{W}$ is the $w \times n$ design matrix, $\otimes$ is the Kronecker product, $\boldsymbol{\beta}$ is a vector of length $(p-1) \times w$, vec is the vector operator which turns a matrix into a vector column–wise and $\mathbf{G} = [\phi(\boldsymbol{\tau}_1^*)| \dots |\phi(\boldsymbol{\tau}_n^*)]$ is an $(p-1) \times n$ matrix.

   We then assign the first $p-1$ elements of $\boldsymbol{\delta}$ to $\boldsymbol{\mu}_{\boldsymbol{\tau}_1}^*$, the next $p-1$ elements of $\boldsymbol{\delta}$ to$\boldsymbol{\mu}_{\boldsymbol{\tau}_2}^*$ and the final $p-1$ elements of $\boldsymbol{\delta}$ to $\boldsymbol{\mu}_{\boldsymbol{\tau}_w}^*$. The $p \times w$ dimensional matrix $\mathbf{T}^*$ is then formed as follows
   $$\mathbf{T}^* = [\phi^{-1}(\boldsymbol{\mu}_{\boldsymbol{\tau}_1}^*)| \dots |\phi^{-1}(\boldsymbol{\mu}_{\boldsymbol{\tau}_w}^*)]$$

10. Sample $\Sigma^*_{\boldsymbol{\tau}}$ from

$$\mathcal{IW}^{p-1}\left(\Sigma_{\boldsymbol{\tau}} + S(\boldsymbol{\tau}^*_i, \phi_c(\mathbf{T})^*\mathbf{W}^i), \delta_{\boldsymbol{\tau}} + n\right)$$

where

$$S(\boldsymbol{\tau}^*_i, \phi_c(\mathbf{T}^*)\mathbf{W}^i) = \sum_{i=1}^{n}\sum_{r=1}^{r}(\phi(\boldsymbol{\tau}^*_i) - \phi_c(\mathbf{T}^*)\mathbf{W}^i)(\phi(\boldsymbol{\tau}^*_i) - \phi_c(\mathbf{T}^*)\mathbf{W}^i)'$$

and $\mathbf{W}^i$ is the i column of the $w \times n$ design matrix and $\mathbf{T} = [\boldsymbol{\tau}^{p*}_1|\ldots|\boldsymbol{\tau}_w{}^*]$ is an $(p-1) \times w$ matrix and recall that $\phi(\boldsymbol{\tau}^p_j) = \boldsymbol{\mu}_{\boldsymbol{\tau}_j}$.

11. For $j = 1, \ldots, p$ sample $\Sigma^*_{\mathbf{z}_j}$ from

$$\mathcal{IW}^{1}\left(\Sigma_{\mathbf{z}_j} + S(\mathbf{z}_{jk}, \boldsymbol{\lambda}^v_{j,t-1}), \delta_{\mathbf{z}_j} + n_j\right)$$

where

$$S(\mathbf{z}_{jk}, \boldsymbol{\lambda}^v_{j,t-1}) = \sum_{k=1}^{n_j}(\phi(\mathbf{z}_{jk}) - \phi(\boldsymbol{\lambda}^v_{j,t-1}))(\phi(\mathbf{z}_{jk}) - \phi(\boldsymbol{\lambda}^v_{j,t-1}))'$$

12. For $j = 1, \ldots, p$ generate $\boldsymbol{\lambda}^{v*}_j$ from the following Metropolis–Hastings algorithm with proposal distribution $q_{\boldsymbol{\lambda}^v_j}(.|\boldsymbol{\lambda}^v_{j,t-1})$

    (a) Generate $\boldsymbol{\nu} \sim q_{\boldsymbol{\lambda}^v_j}(\boldsymbol{\nu}|\boldsymbol{\lambda}^v_{j,t-1})$

    (b)
    $$\boldsymbol{\lambda}^{v*}_j = \begin{cases} \boldsymbol{\nu} & \text{with probability}\rho_{\boldsymbol{\lambda}^v_j}(\boldsymbol{\lambda}^v_{j,t-1}, \boldsymbol{\nu}) \\ \boldsymbol{\lambda}^v_{j,t-1} & \text{with probability}1 - \rho_{\boldsymbol{\lambda}^v_j}(\boldsymbol{\lambda}^v_{j,t-1}, \boldsymbol{\nu}) \end{cases}$$

    where
    $$\rho_{\boldsymbol{\lambda}^v_j}(\boldsymbol{\nu}^o, \boldsymbol{\nu}^n) = \min\left\{\frac{f_{\boldsymbol{\lambda}^v_j}(\boldsymbol{\nu}^n)q_{\boldsymbol{\lambda}^v_j}(\boldsymbol{\nu}^o|\boldsymbol{\nu}^n)}{f_{\boldsymbol{\lambda}^v_j}(\boldsymbol{\nu}^o)q_{\boldsymbol{\lambda}^v_j}(\boldsymbol{\nu}^n|\boldsymbol{\nu}^o)}, 1\right\}$$

    Note $\boldsymbol{\lambda}^v_{p,t} = \boldsymbol{\lambda}^{v*}_j$, that is, we don't need to do the last one twice.

13. For $j = p-1, \ldots, 1$ generate $\boldsymbol{\lambda}^v_{j,t}$ from the following Metropolis–Hastings algorithm with proposal distribution $q_{\boldsymbol{\lambda}^v_j}(.|\boldsymbol{\lambda}^{v*}_j)$

    (a) Generate $\boldsymbol{\nu} \sim q_{\boldsymbol{\lambda}^v_j}(\boldsymbol{\nu}|\boldsymbol{\lambda}^{v*}_j)$

(b)

$$\boldsymbol{\lambda}_{j,t}^{v} = \begin{cases} \boldsymbol{\nu} & \text{with probability} \rho_{\boldsymbol{\lambda}_j^v}(\boldsymbol{\lambda}_j^{v*}, \boldsymbol{\nu}) \\ \boldsymbol{\lambda}_j^{v*} & \text{with probability} 1 - \rho_{\boldsymbol{\lambda}_j^v}(\boldsymbol{\lambda}_j^{v*}, \boldsymbol{\nu}) \end{cases}$$

14. For $j = p, \ldots, 1$ sample $\Sigma_{\mathbf{z}_j,t}$ from p

$$\mathcal{IW}^1\left(\Sigma_{\mathbf{z}_j} + S(\mathbf{z}_{jk}, \boldsymbol{\lambda}_{j,t}^v), \delta_{\mathbf{z}_j} + n_j\right)$$

where

$$S(\mathbf{z}_{jk}, \boldsymbol{\lambda}_{j,t}^v) = \sum_{i=1}^{n_j} (\phi(\mathbf{z}_{jk}) - \phi(\boldsymbol{\lambda}_{j,t}^v))(\phi(\mathbf{z}_{jk}) - \phi(\boldsymbol{\lambda}_{j,t}^v))'$$

15. Sample $\Sigma_{\boldsymbol{\tau},t}$ from

$$\mathcal{IW}^{p-1}\left(\Sigma_{\boldsymbol{\tau}} + S(\boldsymbol{\tau}_i^*, \phi_c(\mathbf{T}^*)\mathbf{W}^i), \delta_{\boldsymbol{\tau}} + n\right)$$

where

$$S(\boldsymbol{\tau}_i^*, \phi_c(\mathbf{T}^*)\mathbf{W}^i) = \sum_{i=1}^{n} \sum_{r=1}^{r} (\phi(\boldsymbol{\tau}_i^*) - \phi_c(\mathbf{T}^*)\mathbf{W}^i)(\phi(\boldsymbol{\tau}_i^*) - \phi_c(\mathbf{T}^*)\mathbf{W}^i)'$$

and $\mathbf{W}$ is an $w \times n$ design matrix and $\mathbf{T}^* = [\boldsymbol{\mu}_{\boldsymbol{\tau}_1}^* | \ldots | \boldsymbol{\mu}_{\boldsymbol{\tau}_w}^*]$ is an $(p-1) \times w$ matrix and recall that $\phi(\boldsymbol{\tau}_j^p) = \boldsymbol{\mu}_{\boldsymbol{\tau}_j}$.

16. Sample $\mathbf{T}_t$ by generating a vector $\boldsymbol{\delta}$ from the following $(p-1)w$ dimensional normal distribution

$$\mathcal{MN}^{(p-1)w}(\boldsymbol{\xi}, \Sigma)$$

where

$$\begin{aligned} \Sigma &= (\Delta + \Omega)^{-1}, \\ \boldsymbol{\xi} &= (\Delta \boldsymbol{\eta}_v + \Omega \boldsymbol{\beta}). \end{aligned}$$

And where

$$\begin{aligned}\Delta &= I_{w\times w}\otimes\Psi_{\Sigma_{\boldsymbol{\tau}}}^{-1},\\ \Omega &= (\mathbf{W}\mathbf{W}^{'})\otimes\Sigma_{\boldsymbol{\tau},t}^{-1},\\ \boldsymbol{\beta} &= \text{vec}\{\mathbf{G}\mathbf{W}^{'}(\mathbf{W}\mathbf{W}^{'})^{-1}\},\end{aligned}$$

$\boldsymbol{\eta}_v = \text{vec}([\boldsymbol{\eta}_1|\ldots|\boldsymbol{\eta}_w])$ is a vector of length $(p-1)w$, $I_{w\times w}$ is an $w\times w$ identity matrix,$\mathbf{W}$ is the $w\times n$ design matrix, $\otimes$ is the Kronecker product, $\boldsymbol{\beta}$ is a vector of length $(p-1)\times w$, vec is the vector operator which turns a matrix into a vector column–wise and $\mathbf{G} = [\phi(\boldsymbol{\tau}_1^*)|\ldots|\phi(\boldsymbol{\tau}_n^*)]$ is an $(p-1)\times n$ matrix.

We then assign the first $p-1$ elements of $\boldsymbol{\delta}$ to $\boldsymbol{\mu}_{\boldsymbol{\tau}_1}^*$, the next $p-1$ elements of $\boldsymbol{\delta}$ to$\boldsymbol{\mu}_{\boldsymbol{\tau}_2}^*$ and the final $p-1$ elements of $\boldsymbol{\delta}$ to $\boldsymbol{\mu}_{\boldsymbol{\tau}_w}^*$. The $p\times w$ dimensional matrix $\mathbf{T}_t$ is then formed as follows

$$\mathbf{T}_t = [\phi^{-1}(\boldsymbol{\mu}_{\boldsymbol{\tau}_1}^*)|\ldots|\phi^{-1}(\boldsymbol{\mu}_{\boldsymbol{\tau}_w}^*)]$$

17. For $i = n,\ldots,1$, Generate $\boldsymbol{\tau}_{i,t}$ from the following Metropolis–Hastings algorithm with proposal distribution $q_{\boldsymbol{\tau}_i}(.|\boldsymbol{\tau}_i^*)$

    (a) Generate $\boldsymbol{\nu}\sim q_{\boldsymbol{\tau}_i}(\boldsymbol{\nu}|\boldsymbol{\tau}_i^*)$

    (b)
    $$\boldsymbol{\tau}_{i,t} = \begin{cases}\boldsymbol{\nu} & \text{with probability}\rho_{\boldsymbol{\tau}_i}(\boldsymbol{\tau}_i^*,\boldsymbol{\nu})\\ \boldsymbol{\tau}_i^* & \text{with probability}1-\rho_{\boldsymbol{\tau}_i}(\boldsymbol{\tau}_i^*,\boldsymbol{\nu})\end{cases}$$

18. For $j = p,\ldots,1$ generate $\boldsymbol{\theta}_{j,t}$ from the following Metropolis–Hastings algorithm with proposal distribution $q_{\boldsymbol{\theta}_j}(.|\boldsymbol{\theta}_j^*)$

    (a) Generate $\boldsymbol{\nu}\sim q_{\boldsymbol{\theta}_j}(\boldsymbol{\nu}|\boldsymbol{\theta}_j^*)$

    (b)
    $$\boldsymbol{\theta}_{j,t} = \begin{cases}\boldsymbol{\nu} & \text{with probability}\rho_{\boldsymbol{\theta}_j}(\boldsymbol{\theta}_j^*,\boldsymbol{\nu})\\ \boldsymbol{\theta}_j^* & \text{with probability}1-\rho_{\boldsymbol{\theta}_j}(\boldsymbol{\theta}_j^*,\boldsymbol{\nu})\end{cases}$$

19. Generate $\boldsymbol{\theta}_{\text{v},t}$ from the following Metropolis–Hastings algorithm with proposal distribution $q_{\boldsymbol{\theta}_{\text{v}}}(.|\boldsymbol{\theta}_{\text{v},t})$

(a) Generate $\boldsymbol{\nu} \sim q_{\boldsymbol{\theta}_{\mathbf{v}}}(\boldsymbol{\nu}|\boldsymbol{\theta}_{\mathbf{v},t})$

(b)

$$\boldsymbol{\theta}_{\mathbf{v},t} = \begin{cases} \boldsymbol{\nu} & \text{with probability} \rho_{\boldsymbol{\theta}_{\mathbf{v}}}(\boldsymbol{\theta}_{\mathbf{v},t}, \boldsymbol{\nu}) \\ \boldsymbol{\theta}^*_{\mathbf{v}} & \text{with probability} 1 - \rho_{\boldsymbol{\theta}_{\mathbf{v}}}(\boldsymbol{\theta}_{\mathbf{v},t}, \boldsymbol{\nu}) \end{cases}$$

20. Generate $\boldsymbol{\theta}_{\mathbf{u},t}$ from the following Metropolis–Hastings algorithm with proposal distribution $q_{\boldsymbol{\theta}_{\mathbf{u}}}(.|\boldsymbol{\theta}_{\mathbf{u},t})$

(a) Generate $\boldsymbol{\nu} \sim q_{\boldsymbol{\theta}_{\mathbf{u}}}(\boldsymbol{\nu}|\boldsymbol{\theta}_{\mathbf{u},t})$

(b)

$$\boldsymbol{\theta}_{\mathbf{u},t} = \begin{cases} \boldsymbol{\nu} & \text{with probability} \rho_{\boldsymbol{\theta}_{\mathbf{u}}}(\boldsymbol{\theta}_{\mathbf{u},t}, \boldsymbol{\nu}) \\ \boldsymbol{\theta}^*_{\mathbf{u}} & \text{with probability} 1 - \rho_{\boldsymbol{\theta}_{\mathbf{u}}}(\boldsymbol{\theta}_{\mathbf{u},t}, \boldsymbol{\nu}) \end{cases}$$

21. Sample $\Sigma_{\mathbf{v},t}$ from

$$\mathcal{IW}^{a-1}\left(\Sigma_{\mathbf{v}} + S(\mathbf{v}_m, \boldsymbol{\theta}_{\mathbf{v},t}), \delta_{\mathbf{v}} + M\right)$$

where

$$S(\mathbf{v}_m, \boldsymbol{\theta}_{\mathbf{v},t}) = \sum_{m=1}^{M} (\phi(\mathbf{v}_m) - \phi(\boldsymbol{\theta}_{\mathbf{v},t}))(\phi(\mathbf{v}_m) - \phi(\boldsymbol{\theta}_{\mathbf{v},t}))'$$

22. Sample $\Sigma_{\mathbf{u},t}$ from

$$\mathcal{W}^{a-1}\left(\Sigma_{\mathbf{u}} + S(\mathbf{u}_l, \boldsymbol{\theta}_{\mathbf{u},t}), \delta_{\mathbf{u}} + L\right)$$

where

$$S(\mathbf{u}_l, \boldsymbol{\theta}_{\mathbf{u},t}) = \sum_{l=1}^{L} (\phi(\mathbf{u}_l) - \phi(\boldsymbol{\theta}_{\mathbf{u},t}))(\phi(\mathbf{u}_l) - \phi(\boldsymbol{\theta}_{\mathbf{u},t}))'$$

23. For $j = p, \ldots, 1$ sample $\Sigma_{\mathbf{x}_j,t}$ from

$$\mathcal{W}^{a-1}\left(\Sigma_{\mathbf{x}_j} + S(\mathbf{x}_{jk}, \boldsymbol{\theta}_{j,t}), \delta_{\mathbf{x}_j} + n_j\right)$$

where

$$S(\mathbf{x}_{jk}, \boldsymbol{\theta}_{j,t}) = \sum_{i=1}^{n_j} (\phi(\mathbf{x}_{jk}) - \phi(\boldsymbol{\theta}_{j,t}))(\phi(\mathbf{x}_{jk}) - \phi(\boldsymbol{\theta}_{j,t}))'$$

24. Sample $\Sigma_{\boldsymbol{\epsilon},t}$ from

$$\mathcal{W}^{a-1}\left(\Sigma_{\boldsymbol{\epsilon}} + S(\mathbf{y}_i, \boldsymbol{\mu}_{\mathbf{y}_i}), \delta_{\boldsymbol{\epsilon}} + nr\right)$$

where

$$S(\mathbf{y}_i, \boldsymbol{\mu}_{\mathbf{y}_i}) = \sum_{i=1}^{n} (\phi(\mathbf{y}_i) - \phi(\boldsymbol{\mu}_{\mathbf{y}_i}))(\phi(\mathbf{y}_i) - \phi(\boldsymbol{\mu}_{\mathbf{y}_i}))',$$

$$\boldsymbol{\mu}_{\mathbf{y}_i} = (\boldsymbol{\Theta}_t \left(\boldsymbol{\tau}_{i,t} \oplus \boldsymbol{\lambda}_t\right)) \oplus \left(\boldsymbol{\theta}_{\mathbf{v},t} \ominus \boldsymbol{\theta}_{\mathbf{u},t}\right),$$

$\boldsymbol{\Theta}_t = [\boldsymbol{\theta}_{1,t}| \dots |\boldsymbol{\theta}_{p,t}]$ is a $a \times p$ matrix and $\boldsymbol{\lambda}_t = (\lambda_{1,t}, \dots, \lambda_{p,t})$ is a vector of length p.

25. Repeat steps 1-24 until convergence and increment t.

To complete the MCMC algorithm we need to specify the target densities, the proposal densities and the acceptance probabilities for each of the Metropolis–Hastings steps for each of $\boldsymbol{\theta}_{\mathbf{u}}$, $\boldsymbol{\theta}_{\mathbf{v}}$, $\boldsymbol{\theta}_j$, $\boldsymbol{\tau}_i$ and $\boldsymbol{\lambda}_j$. The target densities are the full Gibbs conditional distributions (see Figure B.2)

$$\begin{aligned} f_{\boldsymbol{\theta}_{\mathbf{u}}}(\boldsymbol{\nu}) &= \pi(\boldsymbol{\theta}_{\mathbf{u}} = \boldsymbol{\nu} | \boldsymbol{\mu}_{\boldsymbol{\theta}_{\mathbf{u}}}, \Sigma_{\boldsymbol{\theta}_{\mathbf{u}}}) \prod_{i=1}^{n} \prod_{i=1}^{r} \pi(\mathbf{y}_{ir} | \boldsymbol{\lambda}, \boldsymbol{\Theta}, \Sigma_{\boldsymbol{\epsilon}}, \boldsymbol{\theta}_{\mathbf{u}} = \boldsymbol{\nu}, \boldsymbol{\theta}_{\mathbf{v}}, \boldsymbol{\tau}_i) \\ & \prod_{l=1}^{L} \pi(\mathbf{u}_l | \boldsymbol{\theta}_{\mathbf{u}} = \boldsymbol{\nu}, \Sigma_{\mathbf{u}}) \end{aligned}$$

$$\begin{aligned} f_{\boldsymbol{\theta}_{\mathbf{v}}}(\boldsymbol{\nu}) &= \pi(\boldsymbol{\theta}_{\mathbf{v}} = \boldsymbol{\nu} | \boldsymbol{\mu}_{\boldsymbol{\theta}_{\mathbf{v}}}, \Sigma_{\boldsymbol{\theta}_{\mathbf{v}}}) \prod_{i=1}^{n} \prod_{i=1}^{r} \pi(\mathbf{y}_{ir} | \boldsymbol{\lambda}, \boldsymbol{\Theta}, \Sigma_{\boldsymbol{\epsilon}}, \boldsymbol{\theta}_{\mathbf{u}}, \boldsymbol{\theta}_{\mathbf{v}} = \boldsymbol{\nu}, \boldsymbol{\tau}_i) \\ & \prod_{m=1}^{M} \pi(\mathbf{v}_m | \boldsymbol{\theta}_{\mathbf{v}} = \boldsymbol{\nu}, \Sigma_{\mathbf{v}}) \end{aligned}$$

$$\begin{aligned} f_{\boldsymbol{\theta}_j}(\boldsymbol{\nu}) &= \pi(\boldsymbol{\theta}_j = \boldsymbol{\nu} | \boldsymbol{\mu}_{\boldsymbol{\theta}_j}, \Sigma_{\boldsymbol{\theta}_j}) \prod_{i=1}^{n} \prod_{i=1}^{r} \pi(\mathbf{y}_{ir} | \boldsymbol{\lambda}, \boldsymbol{\theta}_j = \boldsymbol{\nu}, \boldsymbol{\theta}_{-j}, \Sigma_{\boldsymbol{\epsilon}}, \boldsymbol{\theta}_{\mathbf{u}}, \boldsymbol{\theta}_{\mathbf{v}}, \boldsymbol{\tau}_i) \\ & \prod_{k=1}^{n_j} \pi(\mathbf{x}_k | \boldsymbol{\theta}_{\mathbf{x}_j}, \Sigma_{\mathbf{x}_j}) \end{aligned}$$

$$f_{\boldsymbol{\tau}_i}(\boldsymbol{\nu}) = \pi(\boldsymbol{\tau}_i = \boldsymbol{\nu} | [\mathbf{W}]_i \mathbf{S}, \Sigma_{\boldsymbol{\tau}}) \prod_{r=1}^{r} \pi(\mathbf{y}_{ir} | \boldsymbol{\lambda}, \boldsymbol{\Theta}, \Sigma_{\boldsymbol{\epsilon}}, \boldsymbol{\theta}_{\mathbf{u}}, \boldsymbol{\theta}_{\mathbf{v}}, \boldsymbol{\tau}_i = \boldsymbol{\nu})$$

where $[\mathbf{W}]_i$ means the ith row of the $n \times w$ design matrix $\mathbf{W}$ and $\mathbf{S} = [\boldsymbol{\mu}_{\boldsymbol{\tau}_1}; \ldots; \boldsymbol{\mu}_{\boldsymbol{\tau}_w}]$ is a $w \times (p-1)$ matrix.

$$\begin{aligned} f_{\boldsymbol{\lambda}_j}(\boldsymbol{\nu}) &= \pi(\boldsymbol{\lambda}_j = \boldsymbol{\nu} | \boldsymbol{\mu}_{\boldsymbol{\lambda}_j}, \Sigma_{\boldsymbol{\lambda}_j}) \prod_{i=1}^{n} \prod_{i=1}^{r} \pi(\mathbf{y}_{ir} | \boldsymbol{\lambda}_j = \boldsymbol{\nu}, \boldsymbol{\lambda}_{-j}, \boldsymbol{\Theta}, \Sigma_{\boldsymbol{\epsilon}}, \boldsymbol{\theta}_{\mathbf{u}}, \boldsymbol{\theta}_{\mathbf{v}}, \boldsymbol{\tau}_i) \\ & \prod_{k=1}^{n_j} \pi(\mathbf{z}_k | \boldsymbol{\theta}_{\mathbf{z}_j}, \Sigma_{\mathbf{z}_j}). \end{aligned}$$

The following logistic normal proposal distributions where used

$$\begin{aligned} q_{\boldsymbol{\theta}_{\mathbf{u}}}(\boldsymbol{\nu}^n | \boldsymbol{\nu}^o) &= \mathcal{L}^{a-1}(\phi(\boldsymbol{\nu}^o), \beta_{\boldsymbol{\theta}_{\mathbf{u}}}(\mathbf{I} + \mathbf{J})), \\ q_{\boldsymbol{\theta}_{\mathbf{v}}}(\boldsymbol{\nu}^n | \boldsymbol{\nu}^o) &= \mathcal{L}^{a-1}(\phi(\boldsymbol{\nu}^o), \beta_{\boldsymbol{\theta}_{\mathbf{v}}}(\mathbf{I} + \mathbf{J})), \\ q_{\boldsymbol{\theta}_j}(\boldsymbol{\nu}^n | \boldsymbol{\nu}^o) &= \mathcal{L}^{a-1}(\phi(\boldsymbol{\nu}^o), \beta_{\boldsymbol{\theta}_j}(\mathbf{I} + \mathbf{J})), \\ q_{\boldsymbol{\tau}_i}(\boldsymbol{\nu}^n | \boldsymbol{\nu}^o) &= \mathcal{L}^{p-1}(\phi(\boldsymbol{\nu}^o), \beta_{\boldsymbol{\tau}_i}(\mathbf{I} + \mathbf{J})), \\ q_{\boldsymbol{\lambda}_j}(\boldsymbol{\nu}^n | \boldsymbol{\nu}^o) &= \mathcal{L}^{1}(\phi(\boldsymbol{\nu}^o), \beta_{\boldsymbol{\lambda}_j}(\mathbf{I} + \mathbf{J})), \end{aligned}$$

where $\beta_{\boldsymbol{\theta}_{\mathbf{u}}}$, $\beta_{\boldsymbol{\theta}_{\mathbf{v}}}$,$\beta_{\boldsymbol{\theta}_j}$, $\beta_{\boldsymbol{\tau}_i}$ and $\beta_{\boldsymbol{\lambda}_j}$ are scale factors that control the acceptance rates for the Metropolis–Hastings algorithm.

The acceptance probabilities $\rho_{\boldsymbol{\theta}_j}(\boldsymbol{\nu}^o, \boldsymbol{\nu}^n)$, $\rho_{\boldsymbol{\theta}_j}(\boldsymbol{\nu}^o, \boldsymbol{\nu}^n)$, $\rho_{\boldsymbol{\theta}_j}(\boldsymbol{\nu}^o, \boldsymbol{\nu}^n)$ $\rho_{\boldsymbol{\tau}_i}, (\boldsymbol{\nu}^o, \boldsymbol{\nu}^n)$ and $\rho_{\boldsymbol{\lambda}_j}, (\boldsymbol{\nu}^o, \boldsymbol{\nu}^n)$ are given by

$$\begin{aligned} \rho_{\boldsymbol{\theta}_{\mathbf{u}}}(\boldsymbol{\nu}^o, \boldsymbol{\nu}^n) &= \min \left\{ \frac{f_{\boldsymbol{\theta}_{\mathbf{u}}}(\boldsymbol{\nu}^n) \prod_{j=1}^{p} [\boldsymbol{\nu}^n]_j}{f_{\boldsymbol{\theta}_{\mathbf{u}}}(\boldsymbol{\nu}^o) \prod_{j=1}^{p} [\boldsymbol{\nu}^o]_j}, 1 \right\} \\ \rho_{\boldsymbol{\theta}_{\mathbf{v}}}(\boldsymbol{\nu}^o, \boldsymbol{\nu}^n) &= \min \left\{ \frac{f_{\boldsymbol{\theta}_{\mathbf{v}}}(\boldsymbol{\nu}^n) \prod_{j=1}^{p} [\boldsymbol{\nu}^n]_j}{f_{\boldsymbol{\theta}_{\mathbf{v}}}(\boldsymbol{\nu}^o) \prod_{j=1}^{p} [\boldsymbol{\nu}^o]_j}, 1 \right\} \\ \rho_{\boldsymbol{\theta}_j}(\boldsymbol{\nu}^o, \boldsymbol{\nu}^n) &= \min \left\{ \frac{f_{\boldsymbol{\theta}_j}(\boldsymbol{\nu}^n) \prod_{j=1}^{p} [\boldsymbol{\nu}^n]_j}{f_{\boldsymbol{\theta}_j}(\boldsymbol{\nu}^o) \prod_{j=1}^{p} [\boldsymbol{\nu}^o]_j}, 1 \right\} \\ \rho_{\boldsymbol{\tau}_i}(\boldsymbol{\nu}^o, \boldsymbol{\nu}^n) &= \min \left\{ \frac{f_{\boldsymbol{\tau}_i}(\boldsymbol{\nu}^n) \prod_{j=1}^{p} [\boldsymbol{\nu}^n]_j}{f_{\boldsymbol{\tau}_i}(\boldsymbol{\nu}^o) \prod_{j=1}^{p} [\boldsymbol{\nu}^o]_j}, 1 \right\} \\ \rho\boldsymbol{\lambda}_j(\boldsymbol{\nu}^o, \boldsymbol{\nu}^n) &= \min \left\{ \frac{f_{\boldsymbol{\lambda}_j}(\boldsymbol{\nu}^n) \prod_{j=1}^{p} [\boldsymbol{\nu}^n]_j}{f_{\boldsymbol{\lambda}_j}(\boldsymbol{\nu}^o) \prod_{j=1}^{p} [\boldsymbol{\nu}^o]_j}, 1 \right\} \end{aligned}$$

APPENDIX C

DISTRIBUTIONAL FUNCTIONAL FORMS

- Normal distribution with location $-\infty < \mu < \infty$ and variance $\sigma^2 > 0$

  $$\pi(y|\mu, \sigma^2) = \mathcal{N}(\mu, \sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}, \quad -\infty < y < \infty.$$

  $E(y) = \mu$ and Var$(y) = \sigma^2$.

- t-distribution distribution with location $-\infty < \mu < \infty$, scale $\sigma^2 > 0$ and degrees of freedom $n > 0$

  $$\pi(y|\mu, \sigma^2, n) = \mathcal{T}(\mu, \sigma^2, n) = \frac{\Gamma((n+1)/2)}{(n\pi\sigma^2)^{1/2}\Gamma(n/2)} \left[1 + \frac{1}{n\sigma^2}(y-\mu)^2\right]^{-(n+1)/2}$$

  If $n > 2$ then $E(y) = \mu$ and Var$(y) = \frac{n}{n-2}\sigma^2$.

- Multivariate Normal with location vector $-\boldsymbol{\infty} < \boldsymbol{\mu} < \boldsymbol{\infty}$ and positive definite covariance matrix Σ

  $$\pi(\mathbf{y}|\boldsymbol{\mu}, \Sigma) = \mathcal{N}^d(\boldsymbol{\mu}, \Sigma) = (2\pi)^{d/2}|\Sigma|^{-1/2} e^{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{y}-\boldsymbol{\mu})}$$

  $E(\mathbf{y}) = \mu$ and Var$(\mathbf{y}) = \Sigma$.

- Multivariate Log–Normal with location vector $-\boldsymbol{\infty} < \boldsymbol{\mu} < \boldsymbol{\infty}$ and positive definite covariance matrix Σ

$$\pi(\mathbf{y}|\boldsymbol{\mu},\Sigma) = \mathcal{LN}^d(\boldsymbol{\mu},\Sigma) = (2\pi)^{d/2}|\Sigma|^{-1/2}\left(\prod_{i=1}^{d}[\mathbf{y}]_i^{-1}\right) e^{-\frac{1}{2}(\log(\mathbf{y})-\boldsymbol{\mu})'\Sigma^{-1}(\log(\mathbf{y})-\boldsymbol{\mu})}$$

$E(\log(\mathbf{y})) = \mu$ and $\mathrm{Var}(\log(\mathbf{y})) = \Sigma$.

- Logistic Normal with location vector $-\infty < \boldsymbol{\mu} < \infty$ and positive definite covariance matrix Σ

$$\pi(\mathbf{z}|\boldsymbol{\mu},\Sigma) = \mathcal{L}^d(\boldsymbol{\mu},\Sigma) = (2\pi)^{d/2}\left(\prod_{i=1}^{D}[\mathbf{z}]_i^{-1}\right)|\Sigma|^{-1/2} e^{-\frac{1}{2}(\phi(\mathbf{z})-\boldsymbol{\mu})'\Sigma^{-1}(\phi(\mathbf{z})-\boldsymbol{\mu})}$$

where $\mathbf{z}$ is of dimension $D = d+1$, $\boldsymbol{\mu}$ is of dimension d and Σ is of dimension $d \times d$. $E(\phi(\mathbf{z})) = \boldsymbol{\mu}$ and $\mathrm{Var}(\phi(\mathbf{z})) = \Sigma$.

- Multivariate T-distribution with location vector $-\infty < \boldsymbol{\mu} < \infty$, positive definite covariance matrix Σ and degrees of freedom $n > 0$

$$\pi(\mathbf{y}|\boldsymbol{\mu},\Sigma,n) = \mathcal{T}^d(\boldsymbol{\mu},\Sigma,n) = \frac{\Gamma((n+d)/2)}{|\Sigma|^{1/2}\Gamma(n/2)(n\pi)^{d/2}}\left[1+\frac{1}{n}(\mathbf{y}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{y}-\boldsymbol{\mu})\right]^{-(n+d)/2}$$

If $n > 2$ then $E(\mathbf{y}) = \boldsymbol{\mu}$ and $\mathrm{Var}(\mathbf{y}) = \frac{n}{n-2}\Sigma$.

- Logistic T-distribution with location vector $-\infty < \boldsymbol{\mu} < \infty$, positive definite covariance matrix Σ and degrees of freedom $n > 0$

$$\begin{aligned}\pi(\mathbf{z}|\boldsymbol{\mu},\Sigma,n) = \mathcal{LT}^d(\boldsymbol{\mu},\Sigma,n) &= \frac{\Gamma((n+d)/2)}{\Gamma(n/2)(n\pi)^{d/2}|\Sigma|^{1/2}}\left(\prod_{i=1}^{D}[\mathbf{z}]_i^{-1}\right)\\ &\quad\times\left[1+\frac{1}{n}(\phi(\mathbf{z})-\boldsymbol{\mu})'\Sigma^{-1}(\phi(\mathbf{z})-\boldsymbol{\mu})\right]^{-(n+d)/2}\end{aligned}$$

If $n > 2$ then $E(\phi(\mathbf{z})) = \boldsymbol{\mu}$ and $\mathrm{Var}(\phi(\mathbf{z})) = \frac{n}{n-2}\Sigma$.

- Skew Normal with mean vector $-\infty < \boldsymbol{\mu} < \infty$, positive definite covariance matrix

Σ and skew vector $-\infty < \boldsymbol{\beta} < \infty$.

$$\pi(\mathbf{y}|\boldsymbol{\mu}, \Sigma, \boldsymbol{\beta}) = \mathcal{SN}^d(\boldsymbol{\mu}, \Sigma, \boldsymbol{\beta}) =$$
$$2(2\pi)^{-d/2}|\Sigma|^{-1/2}e^{\left[-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{y}-\boldsymbol{\mu})\right]}\Phi(\boldsymbol{\beta}'\Omega^{-1}(\mathbf{y}-\boldsymbol{\mu})),$$

where $\Phi(.)$ is the standard normal distribution function and Ω is the square root of diag(Σ) and diag(Σ) is the diagonal matrix of Σ. The vector $\boldsymbol{\beta}$ controls the shape of the distribution and determines the direction of maximum skewness. If $\boldsymbol{\beta} = \mathbf{0}$ the skew–normal reduces to the Multivariate normal distribution.

- Additive skew Logistic Normal with mean vector $-\infty < \boldsymbol{\mu} < \infty$, positive definite covariance matrix Σ and skew vector $-\infty < \boldsymbol{\beta} < \infty$

$$\pi(\mathbf{z}|\boldsymbol{\mu}, \Sigma, \boldsymbol{\beta}) = \mathcal{SL}^d(\boldsymbol{\mu}, \Sigma, \boldsymbol{\beta}) =$$
$$2(2\pi)^{-d/2}|\Sigma|^{-1/2}\left(\prod_{i=1}^{D}[\mathbf{z}]_i^{-1}\right)e^{\left[-\frac{1}{2}(\phi(\mathbf{z})-\boldsymbol{\mu})'\Sigma^{-1}(\phi(\mathbf{z})-\boldsymbol{\mu})\right]}\Phi(\boldsymbol{\beta}'\Omega^{-1}(\phi(\mathbf{z})-\boldsymbol{\mu}))$$

where $\Phi(.)$ is the standard normal distribution function and Ω is the square root of diag(Ψ^{-1}) and diag(Ψ^{-1}) is the diagonal matrix of Ψ^{-1}. The vector $\boldsymbol{\beta}$ controls the shape of the distribution and determines the direction of maximum skewness. If $\boldsymbol{\beta} = \mathbf{0}$ the skew–normal reduces to the Multivariate normal distribution.

- Wishart

$$\pi(\Psi|n, M) = \mathcal{W}^d(\Psi|M, n) = \frac{|M|^{-n/2}|\Psi|^{(n-d-1)/2}}{\left[2^{nd/2}\pi^{d(d-1)/4}\prod_{j=1}^{d}\Gamma\left(\frac{n+1-j}{2}\right)\right]}e^{-\frac{1}{2}tr(M^{-1}\Psi)}$$

where M is a $d \times d$ positive definite matrix and n is the degrees of freedom $n > d-1$. The support of this distribution is the set of positive definite matrices Ψ.

- Inverse–Wishart

$$\pi(\Sigma|n, M) = \mathcal{W}^d(\Sigma|M, n) = \frac{|M|^{n/2}|\Sigma|^{-(n+d+1)/2}}{\left[2^{nd/2}\pi^{d(d-1)/4}\prod_{j=1}^{d}\Gamma\left(\frac{n+1-j}{2}\right)\right]}e^{-\frac{1}{2}tr(M\Sigma^{-1})}$$

where M is a $d \times d$ positive definite matrix and n is the degrees of freedom $n > d-1$.

The support of this distribution is the set of positive definite matrices Σ.

BIBLIOGRAPHY

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society, Series B (Methodological) 44*(2), 139–177.

Aitchison, J. (1992). On criteria for measures of compositional difference. *Mathematical Geology 24*(4), 365–379.

Aitchison, J. (2001). Simplicial inference. In M. A. G. Vianna and D. S. P. Richards (Eds.), *Algebraic Structures in Statistics and Probability*, pp. 1–22. Providence, Rhode Island, USA: American Mathematical Society.

Aitchison, J. (2003). *The Statistical Analysis of Compositional Data* (reprint ed.). Caldwell, New Jersey, USA: The Blackburn Press.

Aitchison, J. and J. Bacon-Shone (1999). Convex linear combinations of compositions. *Biometrika 86*(2), 351–364.

Aldershof, B. and D. Ruppert (1987). A statistical analysis of wood–stove PAH emissions and source apportionment of ambient air samples. *Unpublished EPA report: Research Triangle Park*.

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis* (2nd ed.). John Wiley & Sons.

Andrieu, C. and C. P. Robert (2001). Controlled MCMC for optimal sampling. *Technical Rep. 0125, Cahiers de Mathématiques du Ceremade, Université Paris-Dauphine*, 1–37.

Andrieu, C. and J. Thoms (2008). A tutorial on adaptive MCMC. *Statistics and Computing 18*, 343–373.

Azzalini, A. and A. Capitanio (1999). Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society, Series B 61*(3), 579–602.

Azzalini, A. and A. DallaValle (1996). The multivariate skew–normal distribution. *Biometrika 83*(4), 715–726.

Bandeen-Roche, K. (1994). Resolution of additive mixtures into source components and contributions: A compositional approach. *Journal of the American Statistical Association 89*(428), 1450–1458.

Bandeen-Roche, K. and D. Ruppert (1991). Source apportionment with one source unknown. *Chemometrics and Intelligent Laboratory Systems 10*, 169–184.

Ben-David, M. and D. M. Schell (2001). Mixing models in analysis of diet using multiple stable isotopes: A response. *Oecologia 127*(2), 180–184.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis* (2nd ed.). New York, New York, USA: Springer–Verlag.

Besag, J. (1994). Discussion of ”Markov chains for exploring posterior distributions”. *Annals of Statistics 22*, 1734–1741.

Bickel, P. J. and K. A. Doksum (2000). *Mathematical Statistics, Basic Ideas and Selected Topics, Vol 1* (2nd ed.). New Jersey, USA: Prentice Hall.

Billheimer, D. (2001). Compositional receptor modeling. *Environmetrics 12*, 451–467.

Billheimer, D., P. Guttorp, and W. F. Fagan (2001). Statistical interpretation of species composition. *Journal of the American Statistical Association 96*(456), 1205–1214.

Birnbaum, A. (1962). On the foundations of statistical inference (with discussion). *Journal of the American Statistical Association 57*, 269–326.

Bloomfield, P. (2000). *Fourier Analysis of Time Series: An Introduction, 2nd edition*. Wiley.

Bowen, W. D., D. Tully, D. J. Boness, B. Bulhier, and G. Marshall (2002). Prey–dependent foraging tactics and prey profitability in a marine mammal. *Ecological Applications 244*, 235–245.

Brillinger, D. R. (1981). *Time Series: Data Analysis and Theory*. Holden–Day Inc.

Brooks, S. P. and G. O. Roberts (1998). Convergence assessment techniques for Markov chain Monte Carlo. *Statistics and Computing 8*, 319–335.

Budge, S. M., S. J. Iverson, and H. N. Koopman (2006). Studying trophic ecology in marine ecosystems using fatty acids: A primer on analysis and interpretation. *Marine Mammal Science 22*(4), 759–801.

Celeux, G., M. Hurn, and C. Robert (2000). Computational and inferential difficulties with mixtures posterior distribution. *Journal of the American Statistical Association 95*(3), 957–979.

Chakraborty, B. (2001). On affine equivariant multivariate quantiles. *Annals of the Institute of Statistical Mathematics 53*(2), 380–403.

Chen, M.-H. and B. Schmeiser (1998). Toward black–box sampling: A random–direction interior–point Markov chain approach. *Journal of Computational and Graphical Statistics 7*(1), 1–22.

Chib, S. and E. Greenberg (1994). Bayes inference for regression models with arma(p,q) errors. *Journal of Econmetrics 64*, 183–206.

Chib, S. and E. Greenberg (1995). Understanding the Metropolis–Hastings algorithm. *The American Statistician 49*(4), 327–335.

Christensen, W. F. and S. R. Sain (2002). Accounting for dependence in a flexible multivariate receptor model. *Technometrics 44*(4), 328–337.

Consonni, G. and V. Leucari (2001). Model determination for directed acyclic graphs. *The Statistician 50*(3), 243–256.

Cook, H. W. (1991). Fatty acid desaturation and chain elongation in eukaryotes. In D. E. Vance and J. E. Vance (Eds.), *Biochemistry of Lipids and Membranes*, pp. 141–169. The Netherlands: Elzevier Science.

Cowell, R. G., A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter (1999). *Probabilistic Networks and Expert Systems*. New York, New York, USA: Springer–Verlag.

Cowles, M. K. and B. P. Carlin (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of American Statistical Association 91*(434), 883–904.

Cox, D. R. (1958). Some problems connected with statistical inference. *Annals of Mathematical Statistics 29*, 357–425.

Cox, R. T. (1946). Probability, frequency and reasonable expectation. *American Journal of Physics 14*(1), 1–13.

Cox, R. T. (1962). *The Algebra of Probable Inference*. Baltimore, MD: Johns Hopkins Press.

Dawid, A. P. (1979a). Conditional independence in statistical theory. *Journal of the Royal Statistical Society, Series B 41*(1), 1–31.

Dawid, A. P. (1979b). Conditional independence in statistical theory (with discussion). *Journal of the Royal Statistical Society, Series B 41*(1), 1–31.

de Finetti, B. (1990a). *Theory of Probability, A critical introductory treatment (volume 1)* (Wiley Classics Library Edition ed.). Chichester, England: John Wiley & Sons.

de Finetti, B. (1990b). *Theory of Probability, A critical introductory treatment (volume 2)* (Wiley Classics Library Edition ed.). Chichester, England: John Wiley & Sons.

Diebolt, J. and C. Robert (1994). Estimation of finite mixture distributions by Bayesian sampling. *Journal of Royal Statistical Society, Series B 56*, 363–375.

Diggle, P., P. Heagerty, K.-Y. Liang, and S. Zeger (2002). *Analysis of Longitudinal Data* (second ed.). New York, USA: Oxford University Press.

Egozcue, J., V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology 35*(3), 279–300.

Encyclopedia of Statistics (1983). The Kulback distance. In S. Kotz and N. L. Johnson (Eds.), *Encyclopedia of Statistics*, pp. 421–425. New York, New York, USA: John Wiley and Sons.

Evans, M. and T. Swartz (2000). *Approximating Integrals via Monte Carlo and Deterministic Methods*. Oxford University Press.

Gaston, A. J. and D. G. Noble (1985). The diet of thick–billed murres (*Uria lomvia)* in west hudson straight and northeast hudson bay. *Canadian Journal of Zoology 93*, 1148–1160.

Gelfand, A. E. and J. K. Sahu (1994). On Markov chain Monte Carlo acceleration. *Journal of Computational and Graphical Statistics 3*(3), 261–276.

Gelfand, A. E., S. K. Sahu, and B. P. Carlin (1995). Efficient parametrizations for normal linear mixed models. *Biometrika 82*(3), 479–488.

Gelfand, A. E. and A. F. M. Smith (1990). Sampling–based approaches to calculating marginal densities. *Journal of the American Statistical Association 85*(410), 398–409.

Gelman, A. (2004). Parametrization and Bayesian modeling. *Journal of the American Statistical Association 99*(466), 537–545.

Gelman, A., J. B. C. an d Hal S. Stern, and D. B. Rubin (2003). *Bayesian Data Analysis*. New York, New York, USA: Chapman and Hall.

Gelman, A. and D. B. Rubin (1992). Inference from iterative simulation using multiple sequences. *Statistical Science 7*(4), 457–472.

Gelman, A., D. A. van Dyk, Z. Huang, and W. J. Boscardin (2008). Using redundant parametrizations to fit hierarchical models. *Journal of Computational and Graphical Statistics 17*(1), 95–122.

Geman, S. and D. Geman (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEE Transactions in Pattern Analysis and Machine Intelligence 6*, 721–741.

Gilks, W. R., G. O. Roberts, and E. I. George (1994). Adaptive direction sampling. *The Statistician 43*(1), 179–189.

Gilks, W. R., G. O. Roberts, and S. K. Sahu (1998). Adaptive Markov chain Monte Carlo through regeneration. *Journal of the American Statistical Association 93*(443), 1045–1054.

Gilks, W. R. and P. Wild (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics 41*(2), 337–348.

Gilmore, I., M. A. Johnston, C. T. Pillinger, C. M. Pond, C. A. Mattacks, and P. Prestrud (1995). The carbon isotopic composition of individual fatty acids as indicators of dietary history in arctic foxes on svalbard. *Philosophical Transactions of the Royal Society of London B 349*, 135–142.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika 82*(4), 711–732.

Haario, H., E. Saksman, and J. Tamminen (1999). Adaptive proposal distribution for random walk metropolis algorithm. *Computational Statistics 14*(3), 375–395.

Haario, H., E. Saksman, and J. Tamminen (2001). An adaptive metropolis algorithm. *Bernoulli 7*(2), 223–242.

Hamilton, D. D. (1987). Sometimes $r^2 > r^2_{yx_1} + r^2_{yx_2}$: Correlated variables are not always redundant. *Journal of American Statistician 41*(2), 129–132.

Hammersley, J. M. and D. C. Handscomb (1964). *Monte Carlo Methods*. New York, New York, USA: John Wiley and Sons.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika 57*(1), 97–109.

Henry, R. C., C. W. Lewis, P. K. Hopke, and H. J. Williamson (1984). Review of receptor model fundamentals. *Atmospheric Enivronment 18*(8), 1507–1515.

Hills, S. E. and A. F. M. Smith (1993). Diagnostic plots of improved parameterization in Bayesian inference. *Biometrika 80*(1), 61–74.

Hobert, J. P., C. P. Robert, and C. Goutis (1997). Connectedness conditions for the convergence of the Gibbs sampler. *Statistics and Probability Letters 33*, 235–240.

Hobson, K. A. (1993). Trophic relationships among high arctic seabirds: Insights from tissue–dependent stable–isotope models. *Marine Ecology Progress Series 8*, 509–516.

Iverson, S., C. Field, D. Bowen, and W. Blanchard (2004). Quantitative fatty acid signature analysis: A new method of estimating predator diets. *Ecological Monographs 72*(2), 211–235.

Iverson, S. J., A. M. Springer, and A. S. Kitaysky (2007). Seabirds as indicators of food web structure and ecosystem variability: Qualitative and quantitative diet analysis using fatty acids. *Marine Ecology Progress Series 352*, 235–244.

Jackson, A. L., R. Inger, S. Bearhop, and A. Parnell (2009). Erroneous behaviour of MixSIR, a recently published Bayesian isotope mixing model: A discussion of moore and semmens (2008). *Ecology Letters 12*, E1–E5.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge, UK: Cambridge University Press.

Jeffreys, H. (1961). *Theory of Probability* (3nd ed.). London, UK: Cambridge University Press.

Jenkins, G. M. and M. B. Priestley (1957). The spectral analysis of time–series. *Journal of the Royal Statistical Society. Series B (Methodological) 19*(1), 1–12.

Kashiwagi, N. (2004). Chemical mass balance when an unknown source exists. *Environmetrics 15*, 777–796.

Kiefer, J. and J. Wolfowitz (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics 27*(4), 887–906.

Knuth, D. E. (1964). *The Art of Computer Programming: Volume 2: Seminumerical Algorithms* (2nd ed.). Reading, Massachusetts, USA: Addison–Wesley.

Koch, P. L., J. Heisinger, C. Moss, R. W. Carlson, M. L. Fogel, and A. K. Behrensmeyer (1995). Isotropic tracking of change in diet and habitat use in African elephants. *Science 267*, 1340–1343.

Lauritzen, S. L. (1996). *Graphical Models.* New York, New York, USA: Oxford University Press.

Lauritzen, S. L., A. P. Dawid, B. N. Larsen, and H. G. Leimer (1990). Independence properties of directed Markov fields. *Networks 20*, 491–505.

Lee, S.-Y. (2007). *Structural Equation Modeling: A Bayesian Approach.* Sussex, England: John Wiley & Sons.

Lehmann, E. L. (1983). *Theory of Point Estimation.* John Wiley & Sons.

Lindley, D. V. (1965a). *Introduction to Probability and Statistics: From a Bayesian Viewpoin: Part 1 Probability*. Cambridge, England: Cambridge University Press.

Lindley, D. V. (1965b). *Introduction to Probability and Statistics: From a Bayesian Viewpoin: Part 2 Inference*. Cambridge, England: Cambridge University Press.

Lindley, D. V. and L. D. Phillips (1976). Inference for a Bernoulli process (a Bayesian view). *The American Statistician 30*(1), 112–119.

Liu, F., M. J. Bayarri, and J. O. Berger (2009). Modularization in Bayesian analysis, with emphasis on analysis of computer models. *Bayesian Analysis 4*(1), 119–150.

Liu, J., W. Wong, and A. Kong (1995). Correlation structure and convergence rate of the Gibbs sampler with various scans. *Journal of the Royal Statistical Society Series B 57*, 157–169.

Lubetkin, S. C. and C. A. Simenstad (2004). Multi–source mixing models to quantify food web sources and pathways. *Journal of Applied Ecology 41*, 996–1008.

Lunn, D. J., A. Thomas, N. Best, and D. Spiegelhalter (2000). WinBUGS – a Bayesian modeling framework: Concepts, structure and extensibility. *Statistics and Computing 10*, 325–337.

Martin-Fernández, J. A., C. Barceló-Vidal, and V. Pawlowsky-Glahn (2003). Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology 35*(3), 253–278.

Martín-Fernández, J. A., C. Barceló-Vidal, and V. Pawlowsky-Glahn (2003). Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology 35*(3), 253–278.

Mateu-Figueras, G. and V. Pawlowsky-Glahn (2007). The skew–normal distribution on the simplex. *Communications in Statistics – Theory and Methods 36*, 1787–1802.

Mateu-Figueras, G., V. Pawlowsky-Glahn, and C. Barceló-Vidal (2005). The additive logistic skew–normal distribution on the simplex. *Stochastic Environmental Risk Assessment 19*, 205–214.

Mengersen, K. L., C. P. Robert, and C. Guihenneuc-Jouyaux (1999). MCMC convergence diagnostics: a ”reviewww”. In J. Berger, J. Bernardo, A. Dawid, and A. Smith (Eds.), *Bayesian Statistics 6*, pp. 415–440. Oxford Sciences Publications.

Mengersen, K. L. and R. L. Tweedie (1996). Rates of convergence of the hasting and metropolis algorithms. *The Annals of Statistics 24*(1), 101–121.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics 21*, 1087–1091.

Meyn, S. and R. L. Tweedie (2009). *Markov Chains and Stochastic Stability (Second Edition)* (2nd ed.). Cambridge University Press.

Moore, J. W. and B. X. Semmens (2008). Incorporating uncertainty and prior information into stable isotope mixing models. *Ecology Letters 11*, 470–480.

Morrison, D. F. (1976). *Multivariate Statistical Methods* (2nd ed.). McGraw–Hill.

Müller, P. (1991). A generic approach to posterior integration and Gibbs sampling. *Technical Report # 91-09,Perdue University ,West Lafayette , Indiana*.

Müller, P. (1993). Alternatives to the Gibbs sampling scheme. *Technical Report ,Institute of Statistics and Decision Sciences ,Duke University*.

Neymann, J. and E. L. Scott (1948). Consistent estimates based on partially consistent observations. *Econometrica 16*, 1–32.

Nummelin, E. (1984). *General Irreducible Markov Chains and Non-Negative Operators*. Cambridge,UK: Cambridge University Press.

Nummelin, E. (2002). MC’s for MCMC’ists. *International Statistical Review 70*(2), 215–240.

Palarea-Albaladejo, J., J. A. Martin-Fernández, and J. Gómez-Garcia (2007). A parametric approach for dealing with compositional rounded zeros. *Mathematical Geology 39*, 629–645.

Park, E. S., P. Guttorp, and R. C. Henry (2001). Multivariate receptor modeling for temporally correlated data using MCMC. *Journal of the American Statistical Association 96*(456), 1171–1183.

Park, E. S., R. C. Henry, and C. H. Spiegelman (2000). Estimating the number of factors to include in a high–dimensional multivariate bilinear model. *Communications in Statistics – Simulation 29*(3), 723–746.

Park, E. S., C. H. Spiegelman, and R. C. Henry (2002). Bilinear estimation of pollution source profiles and amounts by using multivariate receptor models. *Envionmetrics 13*, 775–798.

Pawlowsky-Glahn, V. and J. Egozcue (2001). Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Risk Assessment 15*, 384–398.

Pawlowsky-Glahn, V. and J. J. Egozcue (2002). BLU estimators and compositional data. *Mathematical Geology 34*(3), 259–274.

Pearl, J. (1987). Evidential reasoning using stochastic simulation of causal models. *Artificial Intelligence 32*, 245–257.

Peirce, G. J. and P. R. Boyle (1991). A review of methods for diet analysis in piscivorous marine mammals. *Oceanography and Marine Biology 29*, 409–486.

Phillips, D. L. (2001). Mixing models in the analysis of diet using multiple stable isotopes: A critique. *Oecologia 127*(2), 166–170.

Phillips, D. L. and J. W. Gregg (2001). Uncertainty in source partitioning using stable isotopes. *Oecologia 127*(2), 171–179.

Phillips, D. L. and J. W. Gregg (2003). Source partitioning using stable isotopes: Coping with too many sources. *Oecologia 136*(2), 261–269.

Phillips, D. L. and P. L. Koch (2001). Incorporating concentration dependence in stable isotope mixing models. *Oecologia 130*(1), 114–125.

Phillips, D. L., S. D. Newsome, and J. W. Gregg (2005). Combining sources in stable isotope mixing models: Alternative methods. *Oecologia 144*(4), 520–527.

Plummer, M., N. Best, K. Cowles, and K. Vines (2006, March). CODA: Convergence diagnosis and output analysis for MCMC. *R News 6*(1), 7–11.

Priestley, M. B. (1965). Evolutionary spectra and non–stationary processes. *Journal of the Royal Statistical Society. Series B (Methodological) 27*(2), 204–237.

Priestley, M. B. (1966). Design relations for non–stationary processes. *Journal of the Royal Statistical Society. Series B (Methodological) 28*(1), 228–240.

Priestley, M. B. (1981). *Spectral Analysis and Time Series*. Academic Press.

Priestley, M. B. (1988). *Non–linear and Non–stationary Time Series Analysis*. Academic Press.

Priestley, M. B. (1996). Wavelets and time–dependent spectral analysis. *Journal of Time Series Analysis 17*(1), 85–103.

Priestley, M. B. and T. Subba Rao (1969). A test for non–stationarity of time series. *Journal of the Royal Statistical Society, Series B (Methodological) 31*(1), 140–149.

Priestley, M. B. and H. Tong (1973). On the analysis of bivariate non-stationary processes. *Journal of the Royal Statistical Society, Series B (Methodological) 35*(2), 153–166.

Renner, R. M. (1993). The resolution of a compositional data set into mixtures of fixed source compositions. *Applied Statistics 42*(4), 615–631.

Ripley, B. D. (1987). *Stochastic Simulation*. New York, New York, USA: John Wiley and Sons.

Robert, C. P. (2004). *The Bayesian Choice: From Decision–Theoretic Foundations to Computational Implementation* (2nd ed.). New York, New York, USA: Springer.

Robert, C. P. and G. Casella (2004). *Monte Carlo Statistical Methods*. Springer.

Roberts, G. O., A. Gelman, and W. R. Gilks (1997). Weak convergence and optimal scale of random walk Metropolis algorithms. *Annals of Applied Probability 7*, 110–120.

Roberts, G. O. and J. S. Rosenthal (2001). Optimal scaling for various Metropolis–Hastings algorithms. *Statistical Science 16*(4), 351–367.

Roberts, G. O. and J. S. Rosenthal (2006). Harris recurrence of Metropolis–within–Gibbs and trans–dimensional Markov chains. *The Annals of Applied Probability 16*(4), 2123–2139.

Roberts, G. O. and J. S. Rosenthal (2009). Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics 18*(2), 349–367.

Roberts, G. O. and S. K. Sahu (1997). Updating schemes, correlation structure, blocking and parameterization for the gibbs sampler. *Journal of the Royal Statistical Society. Series B (Methodological) 59*(2), 291–317.

Rowe, D. B. (2003). *Multivariate Bayesian Statistics: Models for Source Separation and Signal Unmixing*. Baca Raton, FL, USA: CRC Press.

Savage, L. J. (1954). *The Foundations of Statistics*. New York, USA: John Wiley & Sons.

Semmens, B. X., J. W. Moore, and E. J. Ward (2009). Improving Bayesian isotope mixing models: A response to Jackson *et al.* (2009). *Ecology Letters 12*, E6–E8.

Spiegelhalter, D. J. (1998). Bayesian graphical modeling: A case–study in monitoring health outcomes. *Applied Statistics 47*(1), 115–133.

Spiegelhalter, D. J. and S. L. Lauritzen (1990). Sequential updating of conditional probabilities on directed graphical structures. *Networks 20*, 579–605.

Srivastava, M. S. and C. G. Khatri (1979). *An Introduction to Multivariate Statistics*. North–Holland.

Stewart, C. (2005). *Inference On The Diet of Predators Using Fatty Acid Signatures*. Ph. D. thesis, Dalhousie University.

Subba Rao, T. and H. Tong (1972). A test for time–dependence of linear open–loop systems. *Journal of the Royal Statistical Society, Series B (Methodological) 34*(2), 235–250.

Tanner, M. A. and W. H. Wong (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association 82*(398), 528–540.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics 22*(4), 1701–1762.

Titterington, D. M., A. F. M. Smith, and U. E. Makov (1985). *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley.

Wolbers, M. and W. Stahel (2005). Linear unmixing of multivariate observations: A structural model. *Journal of the American Statistical Association 100*(472), 1328–1342.