OVERVIEW OF REDUNDANCY ANALYSIS AND PARTIAL LINEAR SQUARES AND THEIR EXTENSION TO THE FREQUENCY DOMAIN

by

Jinyi Liu

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE

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 "**Overview of Redundancy Analysis and Partial Linear Squares and Their Extension to the Frequency Domain**" by **Jinyi Liu** in partial fulfillment of the requirements for the degree of **Master of Science**.

Dated: 30 April 2011

Supervisor: ______________________________

Readers: ______________________________

____

DALHOUSIE UNIVERSITY

Date: **30 April 2011**

Author: **Jinyi Liu**

Title: **Overview of Redundancy Analysis and Partial Linear Squares and Their Extension to the Frequency Domain**

Department: **Department of Mathematics and Statistics**

Degree: **M.Sc.** Convocation: **May** Year: **2011**

Signature of Author

Table of Contents

List of Tables

List of Figures

Abstract

Applied statisticians are often faced with the problem of dealing with high dimensional data sets when attempting to describe the variability of a single set of variables, or trying to predict the variation of one set of variables from another. In this study, two data reduction methods are described: Redundancy Analysis and Partial Least Squares. A hybrid approach developed by Bougeard et al., (2007) and called Continuum Redundancy-Partial Least Squares, is described. All three methods are extended to the frequency domain in order to allow the lower dimensional subspace used to describe the variability to change with frequency. To illustrate and compare the three methods, and their frequency dependent generalizations, an idealized coupled atmosphere-ocean model is introduced in state space form. This model provides explicit expressions for the covariance and cross spectral matrices required by the various methods; this allows the strengths and weaknesses of the methods to be identified.

List of Symbols Used

$\boldsymbol{X}_1$	Random vector(Responses)	3
$\boldsymbol{X}_2$	Random vector(Predictors)	3
$\boldsymbol{Z}_1$	Latent vector(Responses)	21
$\boldsymbol{Z}_2$	Latent vector(Predictors)	21
$\boldsymbol{u}$	Loading vector for the responses	26
$\boldsymbol{v}$	Loading vector for the predictors	26
$\boldsymbol{A}$	Transition matrix	5
$\boldsymbol{M}\nu$	Noise vector	5
$\boldsymbol{\Sigma}_{ij}$	Covariance matrix	6
$\boldsymbol{h}_{ij}(\omega)$	Cross spectral matrix	7
λ_k	The kth eigenvalue	20
$\boldsymbol{R}_{ij}$	Correlation matrix	39
$\boldsymbol{W}_{ij}$	Coherence matrix	44

The parameters used in the idealized atmosphere-ocean coupling model are defined in Table 2.1.

Acknowledgements

I would like to dedicate this Master thesis to my parents, Benping Liu and LiRong Li, who have supported and encouraged me in every way since the beginning of my studies. This thesis is also dedicated to my girlfriend, Yi Zhang, who has always been a great source of motivation and inspiration.

I would like to especially thank my supervisor, Dr. Keith R. Thompson for his guidance, patience and encouragement. My appreciation also goes to Dr. Bruce Smith and Dr. Michael Dowd for their helpful suggestions and assistance.

Chapter 1

Introduction

Applied statisticians are often faced with the problem of dealing with high dimensional data sets in various fields such as environmental studies and bioinformatics. For example, in a gene expression microarray data set there could be tens or hundreds of dimensions, each of which corresponds to an experimental condition, and the corresponding multiple regression could have many possible predictors. The problem is compounded when attempting multivariate regression with many responses. The problem also appears when attempting to generalize the concept of correlation of scalar random variables to two high dimensional random vectors. All of the above problems lead to the need for effective ways of reducing the dimension of high dimensional random vectors. Such reduction techniques may lead not only to more robust estimates of unknown parameters but also a more parsimonious representation of the relationship between high dimensional data sets and random vectors, and a better understanding of the underlying relationships.

There are several well-known statistical dimension reduction methods including Canonical Correlation Analysis (CCA), Redundancy Analysis (RA) and Partial Least Squares (PLS). All of these methods are based on the search for a reduced set of *latent variables* which are taken to be linear combinations of the original variables. The differences amongst the methods arise from the different criteria used to select the latent variables.

Canonical Correlation Analysis was introduced by Hotelling (1936) as a method for identifying linear combinations of the two sets of variables in order to maximize their correlation. Although CCA is less popular than some other statistical techniques, it has been applied across various disciplines e.g., examining the relationship between adoption of outsourcing services and the characteristics of environments in which the services operate as discussed by Alpert (1975). CCA is conceptually attractive and simple to understand. It does however have some problems. For example it will be shown later that it is possible to obtain highly correlated canonical variates that explain well the relationship *between* but are not effective at explaining the variability *within* the random vectors.

Redundancy Analysis was introduced by Stewart and Love (1968) and aims to construct an asymmetrical measure of the dependence of one set of variables on the other, again using the concept of latent variables. Van den Wollenberg (1977) derived sets of latent variables which maximize the "redundancy" in each set, instead of maximizing the canonical correlation (as in CCA). The relationship between RA and CCA has been discussed by Muller (1981). Israels (1984) generalized redundancy analysis to qualitative variables and compared it with PCA. Overall RA is a powerful dimension reduction technique that performs well if the focus is prediction. It does however have problems as explained in the following paragraph.

Canonical Correlation and Redundancy Analysis share a common problem when the covariance matrices on which they are based are ill conditioned. For both approaches, the calculation of the latent vectors involves of the inversion of covariance matrices and their poor conditioning leads to unwanted sensitivity to small changes in the predictors and hence unstable results. Poor conditioning can be caused by small sample size, missing values and multicollinearity amongst predictors (leading to the risk of rejecting a theoretically sound predictor from a regression model).

Partial Least Squares was introduced by Wold and coworkers (Wold,1966) to overcome the above mentioned problem of ill conditioning. PLS refers to a wide class of methods for dimension reduction and also modeling the relationship between sets of variables (e.g., regression and classification). The underlying assumption of all PLS methods is that the original variables are driven by a small number of latent variables. In its general form, PLS creates latent vectors (sometimes called "score vectors") by maximizing the *covariance* between different sets of variables. In the present study we will consider only the application of PLS to two random vectors.

Based on the above discussion it is clear that there are two classes of problem requiring dimension reduction techniques. The "symmetrical" case treats the two sets of variables, stored in the vectors $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, on the same footing (e.g., CCA, PLS). We denote the symmetrical case by $\boldsymbol{X}_1 \leftrightarrow \boldsymbol{X}_2$. The asymmetrical case arises when we want to predict $\boldsymbol{X}_1$ from $\boldsymbol{X}_2$ (e.g., RA and regression based PLS). We denote the asymmetrical case by $\boldsymbol{X}_2 \rightarrow \boldsymbol{X}_1$. In this study the focus is on the asymmetrical case.

The similarities and differences among the above techniques have been studied extensively over recent decades. For example Van den Wollenberg (1977) described RA as an alternative method of CCA. Israels (1984) compared RA to PCA and CCA for qualitative variables. The equivalence between CCA and orthonormalized PLS has been studied in Sun Ji, (2008). In the present study, we note the good fitting ability of RA and the stability of PLS, and explore a hybrid approach proposed by Bougeard et al., (2007).

One of the novel aspects of the present study is the extension of RA and PLS to the frequency domain. In this case, the relationship between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ can change with frequency (e.g., a low dimensional relationship may only hold only a certain range of frequencies). This point is particularly important in many physically-based systems e.g., the coupled atmosphere-ocean system where it is well known that the interaction between the two fluids depends on time scale (Flato and Boer, 2000).

To illustrate RA and PLS, and their extension to the frequency domain, we have introduced an idealized atmosphere-ocean model in the form of a simple, linear state space model (Bakalian et al., 2009). The advantages of this model are (i) it is relevant for climate prediction and is thus practically relevant, (ii) the theoretical covariance matrices can be calculated explicitly in terms of covariance of forcing and the frequency dependent transfer functions that link the atmosphere and ocean. (There is thus no sampling variability associated with the covariances matrices used in the present study.)

In this study we will use the ocean state vector ($\boldsymbol{X}_2$) to predict the atmospheric state vector ($\boldsymbol{X}_1$) using the different dimension reduction approaches in both the time and frequency domains. The choice of optimal method would be straightforward if the covariance matrix of predictors was well conditioned. The focus of the present study is however the case of poorly conditioned covariance matrices.

The structure of the thesis is as follows. The linear state space model is introduced in Chapter 2 in both the time domain and frequency domains. We also introduce a more physically-based example based on an idealized, coupled atmosphere-ocean model which we will use to illustrate and evaluate the various dimension reduction methods. Chapter 3 provides a theoretical overview of CCA, PCA, RA and PLS (with a focus on two of its variants, PLS-SVD and PLS-W2A). In Chapter 4, the recent attempt of Bougeard et al., (2007) to develop a common theoretical framework for RA and PLS is described. Their "hybrid" approach is extended to PLS-SVD and PLS-W2A specifically. In Chapter 5, the above approaches are extended to the frequency domain. Chapter 6 illustrates the various methods, including some of their frequency dependent generalizations, using the idealized atmosphere-ocean model. The main results are summarized and discussed, and suggestions for future work are made, in Chapter 7.

Chapter 2

State Space and Coupled Atmosphere-Ocean Models

The overall goals of this study are to compare methods for relating high dimensional random vectors, using information contained in their covariance matrices, and extend them to the frequency domain. To make the discussion easier to understand, the linear state space model is introduced in this chapter. The model leads to analytic forms for both covariance and cross spectral matrices which are then used to compare the the dimension reduction methods in the time and frequency domains. Of particular interest is the case of covariance and cross spectral density matrices that are poorly conditioned.

The state space model is introduced in the first section and then described in the frequency domain in the next section. To interpret and specify the parameters of the state space model we follow Bakalian et al. (2010) and use the state space model to describe the coupled atmosphere-ocean system. This leads to the idealized coupled atmosphere-ocean model and it is described in the final section.

2.1 The Linear State Space Model

Let the state of a p dimensional random process at time t be defined by

$$\boldsymbol{X}_t = \begin{bmatrix} X_{1,t} \\ X_{2,t} \\ \vdots \\ X_{p,t} \end{bmatrix} \tag{2.1}$$

It is assumed that $\boldsymbol{X}_t$ evolves according to the following multivariate state equation:

$$\boldsymbol{X}_t = \boldsymbol{A}\boldsymbol{X}_{t-1} + \boldsymbol{M}\boldsymbol{\xi}_t \tag{2.2}$$

where $\boldsymbol{A}$ is a $p \times p$ transition matrix and $\boldsymbol{M}\boldsymbol{\xi}_t$ is the innovation or noise vector which is assumed uncorrelated with itself for all lags. It is assumed that $E(\boldsymbol{\xi}_t) = 0$ and the eigenvalues of $\boldsymbol{A}$ are all less than one in absolute value (to ensure $\boldsymbol{X}_t$ is a zero mean process that is asymptotically stationary to second order e.g., Priestley, 1982, p798-799). $\boldsymbol{M}$ is a $p \times q$ matrix that controls the covariance of the noise which is given by $\boldsymbol{M}\boldsymbol{\Sigma}_{\xi\xi}\boldsymbol{M}^T$.

It follows from (2.2) that the asymptotic covariance of $\boldsymbol{X}_t$, i.e., $E[\boldsymbol{X}_t\boldsymbol{X}_t^T]$ as $t \to \infty$, is given by

$$\boldsymbol{\Sigma}_{XX} \quad = \quad \boldsymbol{A}\boldsymbol{\Sigma}_{XX}\boldsymbol{A}^T + \boldsymbol{M}\boldsymbol{\Sigma}_{\xi\xi}\boldsymbol{M}^T$$

where $\boldsymbol{\Sigma}_{\xi\xi} = \mathrm{Cov}(\boldsymbol{\xi}_{\mathrm{t}})$.

An explicit expression for $\mathrm{Cov}(\boldsymbol{X}_{\mathrm{t}})$ in terms of $\boldsymbol{A}$, $\boldsymbol{M}$ and $\boldsymbol{\Sigma}_{XX}$ is given by (e.g., Harvey, 1982)

$$\mathrm{vec}\ \boldsymbol{\Sigma}_{XX} = (\boldsymbol{I} - \boldsymbol{A} \otimes \boldsymbol{A})^{-1}\mathrm{vec}\ (\boldsymbol{M}\boldsymbol{\Sigma}_{\boldsymbol{\xi}\boldsymbol{\xi}}\boldsymbol{M}^T) \tag{2.3}$$

where vec is an operator that converts a matrix to a vector by stacking its columns one upon the other, and ⊗ denotes a Kronecker product.

It is straightforward to show that the covariance between $\boldsymbol{X}_t$ and $\boldsymbol{X}_{t-k}$ satisfies the following recursive equation:

$$\boldsymbol{\Sigma}_{XX}(k) \quad = \quad \boldsymbol{A}\boldsymbol{\Sigma}_{XX}(k-1) \qquad k > 0$$

and so

$$\boldsymbol{\Sigma}_{XX}(k) = \boldsymbol{A}^k\boldsymbol{\Sigma}_{XX} \qquad \text{for any integer k} \tag{2.4}$$

The state state model also includes an equation that relates the state vector to the observation vector as shown below:

$$\boldsymbol{Y}_t = \boldsymbol{H}\boldsymbol{X}_t + \boldsymbol{\nu}_t$$

This is the so called observation equation. $\boldsymbol{H}$ is the observation operator matrix and $\boldsymbol{\nu}_t$ is the vector of observation noise. For simplicity, we will assume $\boldsymbol{\nu}_t = 0$ in this study.

2.2 Spectral Representation of the State Space Model

We now focus on the relationship between two random vectors in the frequency domain. According to the spectral representation of discrete parameter multivariate stationary processes, there exists an orthogonal process $\mathbf{X}(\omega)$ such that the state vector $\mathbf{X}_t$ can be written in the form

$$\boldsymbol{X}_t = \int_{-\pi}^{\pi} e^{it\omega}\, d\boldsymbol{X}(\omega) \tag{2.5}$$

for all t. Loosely speaking, $d\boldsymbol{X}(\omega)$ can be considered as the complex random amplitude of the sinusoidal function at frequency ω that makes up $\boldsymbol{X}_t$ (e.g., Priestley, 1982, p245)

Let $\boldsymbol{h}_{XX}(\omega)$ and $\boldsymbol{h}_{\xi\xi}(\omega)$ denote the cross spectral matrices of $\boldsymbol{X}_t$ and $\boldsymbol{\xi}_t$ respectively. The diagonal elements of these matrices define power spectral density at frequency ω; the off-diagonal elements defines the cross spectral densities. Note any cross spectral matrix is a positive semi-definite Hermitian matrix, i.e., $\boldsymbol{h}^*(\omega) = \boldsymbol{h}(\omega)$ where * denotes conjugate transpose (Priestley, 1982).

The random orthogonal increment process $d\boldsymbol{X}(\omega)$ has the following properties:

$$\begin{array}{rcll} E[d\boldsymbol{X}(\omega)] & = & 0 & \\ E[|d\mathbf{X}(\omega)|^2] & = & \boldsymbol{h}_{XX}(\omega)d\omega & \\ E[d\boldsymbol{X}(\omega)d\boldsymbol{X}(\omega')^*] & = & 0 & \qquad\qquad \omega \neq \omega' \end{array}$$

If the spectral representation of the multivariate random process $\boldsymbol{\xi}_t$ is given by

$$\boldsymbol{\xi}_t = \int_{-\pi}^{\pi} e^{it\omega}\, d\boldsymbol{\xi}(\omega) \tag{2.6}$$

it is possible to obtain from (2.2) the following frequency relationship between $d\boldsymbol{X}$ and $d\boldsymbol{\xi}$:

$$d\boldsymbol{X}(\omega) = e^{-i\omega}\boldsymbol{A}d\boldsymbol{X}(\omega) + \boldsymbol{M}d\boldsymbol{\xi}(\omega)$$

from which it follows that

$$d\boldsymbol{X}(\omega) = (\boldsymbol{I} - e^{-i\omega}\boldsymbol{A})^{-1}\boldsymbol{M}d\boldsymbol{\xi}(\omega) \tag{2.7}$$

Using (2.7) it follows that the cross spectral matrix of the state vector is given by

$$\boldsymbol{h}_{XX}(\omega) = \boldsymbol{Q}\boldsymbol{h}_{\xi\xi}(\omega)\boldsymbol{Q}^* \tag{2.8}$$

where

$$\boldsymbol{Q} = (\boldsymbol{I} - e^{-i\omega}\boldsymbol{A})^{-1}\boldsymbol{M}$$

Equation (2.8) is an elegant expression for the cross-spectral matrix of the state for each frequency in terms of $\mathbf{h}_{\xi\xi}(\omega)$ and transition matrix $\boldsymbol{A}$.

The relationship between the time and frequency domain descriptions of the state space model follows from the fact that $\boldsymbol{\Sigma}_{XX}(s)$ is the Fourier transform of $\boldsymbol{h}_{XX}(\omega)$:

$$\boldsymbol{\Sigma}_{XX}(k) = \int_{-\pi}^{\pi} e^{ik\omega}\boldsymbol{h}_{XX}(\omega)\, d\omega$$

and

$$\boldsymbol{h}_{XX}(\omega) = \frac{1}{2\pi}\sum_{-\infty}^{\infty}\boldsymbol{\Sigma}_{XX}(k)e^{-ik\omega}$$

Note that if $k = 0$ we obtain

$$\boldsymbol{\Sigma}_{XX}(0) = \int_{-\pi}^{\pi} \boldsymbol{h}_{XX}(\omega)\, d\omega$$

This shows that the variance and covariance of the elements of $\boldsymbol{X}$ can be expressed as integrals of the corresponding power and cross spectral densities.

Analysis in the time and frequency domains are thus equivalent but take different perspectives. They provide complementary infomation. For example the time domain representation is useful in terms of quantifying how quickly information is forgotten whereas analysis in the frequency domain can be very helpful in the exploration of the physical meaning of the relationship between random vectors. These points will be made more explicit in the next subsection which provides a physically-based application of the state space model.

2.3 An Idealized Coupled Atmosphere-Ocean Model

To provide a more physically-based form of the state space model, we follow Bakalian et al., (2009) and consider an annular atmosphere sitting above an annular ocean, each divided into N equally spaced sectors. Horizontal diffusion and horizontal advection link the states in adjacent sectors of each fluid, and vertical processes (e.g., latent and sensible heat exchange) allow the two fluids to communicate. A schematic of the model is shown in Figure 2.1.

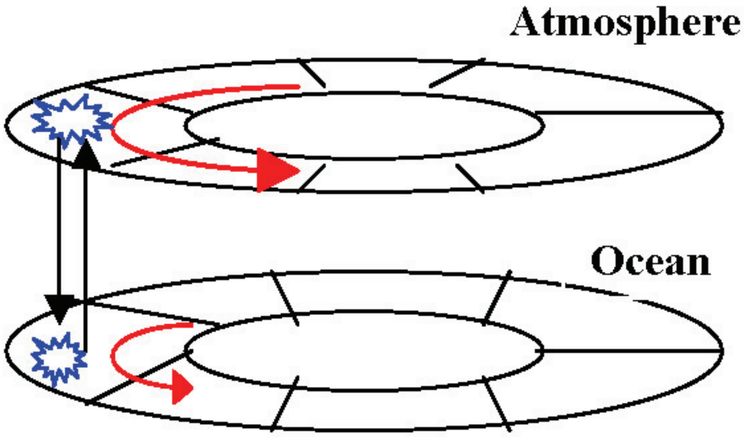

Figure 2.1: An idealized, coupled atmosphere-ocean model. The atmosphere and ocean are represented by two tori, each partitioned into N sectors. In the above schematic $N = 7$. The sectors within a specific fluid communicate with adjacent sectors through horizontal advection and diffusion (denoted by the red arrows). The two fluids also communicate through vertical processes that link changes in overlying sectors (denoted by the black arrows).

To apply the state space model to the coupled system the state vector and related quantities are partitioned according to their atmospheric (subscript 1) and oceanic (subscript 2) components:

$$
\begin{aligned}
\boldsymbol{X} &= [\boldsymbol{X}_1^T \ \boldsymbol{X}_2^T]^T \qquad \boldsymbol{\xi} = [\boldsymbol{\xi}_1^T \ \boldsymbol{\xi}_2^T]^T \\
\boldsymbol{A} &= \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix} \\
\boldsymbol{\Sigma}_{XX} &= \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \\
\boldsymbol{\Sigma}_{\xi\xi} &= \begin{bmatrix} \boldsymbol{\Sigma}_{\xi\xi 11} & \boldsymbol{\Sigma}_{\xi\xi 12} \\ \boldsymbol{\Sigma}_{\xi\xi 21} & \boldsymbol{\Sigma}_{\xi\xi 22} \end{bmatrix} \\
\boldsymbol{M} &= \begin{bmatrix} \boldsymbol{M}_1 & 0 \\ 0 & \boldsymbol{M}_2 \end{bmatrix}
\end{aligned}
$$

The state space model takes the form

$$
\begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix}_t = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix}_{t-1} + \begin{bmatrix} \boldsymbol{M}_1 \, \boldsymbol{\xi}_1 \\ \boldsymbol{M}_2 \, \boldsymbol{\xi}_2 \end{bmatrix}_t \tag{2.9}
$$

According to (2.9) the random vectors $\mathbf{X}_1$ and $\mathbf{X}_2$ describe the state of the atmosphere and the ocean respectively, and $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ correspond to the "forcing" vectors that drive the coupled system from a state of rest. The matrices $\boldsymbol{A}_{11}$ and $\boldsymbol{A}_{22}$ control the independent evolution of atmosphere and ocean, and $\boldsymbol{A}_{12}$ and $\boldsymbol{A}_{21}$ control their interaction.

To specify the parameters in (2.9) we follow Bakalian et al. (2010) and perform a simple discretization of the horizontal advection and diffusion processes that are known to operate in the ocean and atmosphere. This leads to (almost) tridiagonal $\boldsymbol{A}_{ii}$ matrices with terms of the form $1-2\alpha-\beta$ on the diagonal, α on the superdiagonal and $\alpha+\beta$ on the subdiagonal. The parameters α and β control the strength of diffusion and advection respectively in the atmosphere and ocean. Deviation from tridiagonal form is due to the addition of elements in the lower left and upper right of A_{ii} that result from the periodic nature of the system.

To specify the remaining parameters in (2.9) the vertical exchange processes are assumed proportional to the vertical difference in the state of the atmosphere and ocean. The result is that $\boldsymbol{A}_{12}$ is of the form $\eta_{12}\boldsymbol{I}$, where η_{12} is an (ocean to atmosphere) coupling coefficient. This means that η_{12} must be subtracted from the diagonal elements of $\boldsymbol{A}_{11}$. Similarly $\boldsymbol{A}_{21}$ is of the form $\eta_{21}\boldsymbol{I}$ and η_{21} is subtracted from the diagonal elements of $\boldsymbol{A}_{22}$. The exchange parameters η_{12} and η_{21} differ owing to the different heat capacities and densities of the two fluids. Finally a spatially varying coefficient $\eta_R\phi_i$ is subtracted from the diagonal elements of $\boldsymbol{A}_{11}$ to mimic radiative loss of heat to space from the atmosphere (where i denotes a dependence of the radiative transfer on sector).

The parameter values used in this thesis are given in Table 2.1.

Table 2.1: Parameters defining the coupled atmosphere-ocean model.

Parameter	Description	Value
N	Number of sectors in each fluid	15
α_1	Horizontal advection in atmosphere	0.3
β_1	Horizontal diffusion in atmosphere	0.01
η_{12}	Transfer from ocean to atmosphere	0.02
α_2	Horizontal advection in ocean	0.1
β_2	Horizontal diffusion in ocean	0.001
η_{21}	Transfer from atmosphere to ocean	0.002
η_R	Radiative transfer to space	0.04
ϕ_i	Spatial structure of transfer to space	$\phi_i = 1 \quad 7 \leq i \leq 13$ $\phi_i = 0 \quad$ Otherwise
σ_1	Standard deviation of atmosphere noise	0.6
σ_2	Standard deviation of ocean noise	0.1
m_1	Number of columns of $\boldsymbol{M}_1$	3
m_2	Number of columns of $\boldsymbol{M}_2$	2

The covariance matrix of $\boldsymbol{\xi}$ is also written in a partitioned form corresponding to the different forcing applied to the two fluids:

$$\boldsymbol{\Sigma}_{\xi\xi} = \begin{bmatrix} \sigma_1^2\boldsymbol{I} & 0 \\ 0 & \sigma_2^2\boldsymbol{I} \end{bmatrix} \tag{2.10}$$

It follows that the covariance matrix between the atmosphere forcing and ocean forcing is zero (i.e., it is assumed there is no interaction between the two forcing vectors) and given by

$$\mathrm{Cov}(\boldsymbol{M}\xi) = \begin{bmatrix} \sigma_1^2 \boldsymbol{M}_1 \boldsymbol{M}_1^T & 0 \\ 0 & \sigma_2^2 \boldsymbol{M}_2 \boldsymbol{M}_2^T \end{bmatrix} \tag{2.11}$$

To complete the specification of the state space model all that remains is the specification of the spatial structure of the atmospheric and oceanic forcing. For this study it is assumed that $\boldsymbol{M}_1$ is an $N \times 3$ matrix with i^{th} row of the form $[1/4 \ \sin(2\pi i/N) \ \cos(2\pi i/N)]$. $\boldsymbol{M}_2$ is taken to be an $N \times 2$ matrix with i^{th} row of the form $[\sin(2\pi i/N) \ \cos(2\pi i/N)]$.

It is important to note that the above forms for $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ imply that the covariance matrices of the atmospheric and oceanic forcing (i.e., $\sigma_1^2 \boldsymbol{M}_1 \boldsymbol{M}_1^T$ and $\sigma_2^2 \boldsymbol{M}_2 \boldsymbol{M}_2^T$) will be rank deficient (specifically they are of rank m_1 and m_2 respectively). It follows that the covariance matrices of the atmospheric and oceanic state ($\boldsymbol{\Sigma}_{X_N X_N} = \sum_{n=1}^{N} \boldsymbol{A}^{n-1} \boldsymbol{M} \boldsymbol{\Sigma}_{\xi\xi} \boldsymbol{M}^T \boldsymbol{A}^{n-1^T}$) will also be poorly conditioned. The reason that the possibility of poorly conditioned covariance matrices has been allowed in this study is that it will be shown in Chapter 4 that some of the methods for relating high dimensional random vectors perform poorly under such circumstances.

The response of atmosphere and ocean to an initial perturbation in the atmosphere assuming no noise ($\sigma_1 = \sigma_2 = 0$) is shown in Figure 2.2. The plots in the left panels show maps of the atmospheric and oceanic response. In the atmosphere, sectors 1 to 7 are non zero at the initial time as expected. The atmospheric state goes to zero as time passes due to the lose of heat to the ocean and space. The ocean sectors are initialized with no energy as shown in the figure. As time advances the ocean is warmed by the atmosphere and the heat is redistributed within the ocean by advection and diffusion. It is noticeable that the energy contained in the ocean is much less on average than the atmosphere and the natural period of variability of the ocean is longer than the atmosphere. In fact, recall the advection speed of atmosphere and ocean set in this model are 0.3 and 0.1 in (2.1). And then the corresponding natural periods are 50 and 150 time steps respectively. The right panels show the time varying response for sectors 1 and 10. The two atmospheric time series converge quickly and are close to zero after time step 150 due to the heat lost by vertical heat exchange. The ocean sectors converge more slowly and still contain a significant amount of energy at time

200, although the whole system will finally lose all of the heat to space.

The response of atmosphere and ocean to one realization of stochastic forcing (i.e., both σ_1 and σ_2 are non zero) is shown in Figure 2.3. The left panels show the atmosphere and ocean are dominated by quasi-periodic variations driven by the stochastic forcing. The natural period of variability is longer in the ocean compared to the atmosphere as expected (see above). The time varying response for sector 1 and 10 are shown in the right panels. The approach to asymptotic stationarity is evident in the way the variance builds from zero at the first time step to a constant value as time advances.

The standard deviation and correlation of the atmospheric and oceanic states are shown in Figure 2.4. The left panels show the standard deviation for the atmosphere and ocean as a function of position. The standard deviation of the atmosphere is higher than the ocean for each sector. The atmosphere in sectors 7 to 13 has relative low standard deviation because the atmospheric variability is more heavily damped in these regions of enhanced radiative loss to space (see ϕ in Table 2.1). By contrast, the standard deviations of the ocean sectors, which do not radiate heat to space, are relatively stable. The right panels show the correlation between the atmospheric state for sector 5 with the complete atmosphere (upper panel) and ocean (lower panel). The plot for the atmosphere shows a sinusoidal pattern with a maximum correlation of 1 for sector 5 as expected. The correlation of the atmosphere in sector 5 with the ocean (lower panel) also present a sinusoidal pattern but the correlations are much weaker.

Representative power, coherence and phase spectra are shown in Figure 2.5. The power spectra (left panels) show that the atmospheric energy is centered on a frequency corresponding to a period of 50 time steps which is just the natural advection period of the atmosphere. In the ocean the energy is found mostly around the frequency 0.007. It corresponds to a period which is equal to the natural advection period of the atmosphere. We can also note a peak in the power spectrum for the atmosphere due to the feedback of the heat from ocean. The right panels show the coherence and phase spectra for the atmosphere and underlying oceanic state for sector 1. The coherence is maximum at a frequency of 0.0071 where the atmospheric state and oceanic state have relatively high energy levels. The coherence drops as frequency increases and is almost zero as the frequency approaches 0.1. The bottom right panel shows the phase spectrum which is clearly quite complicated.

From the discussion above, it is clear that this idealized model supports quite complex covariance and spectral structures. This idealized atmosphere-ocean model is used in Chapter 6 to illustrate and evaluate the dimension reduction techniques described in the next chapter.

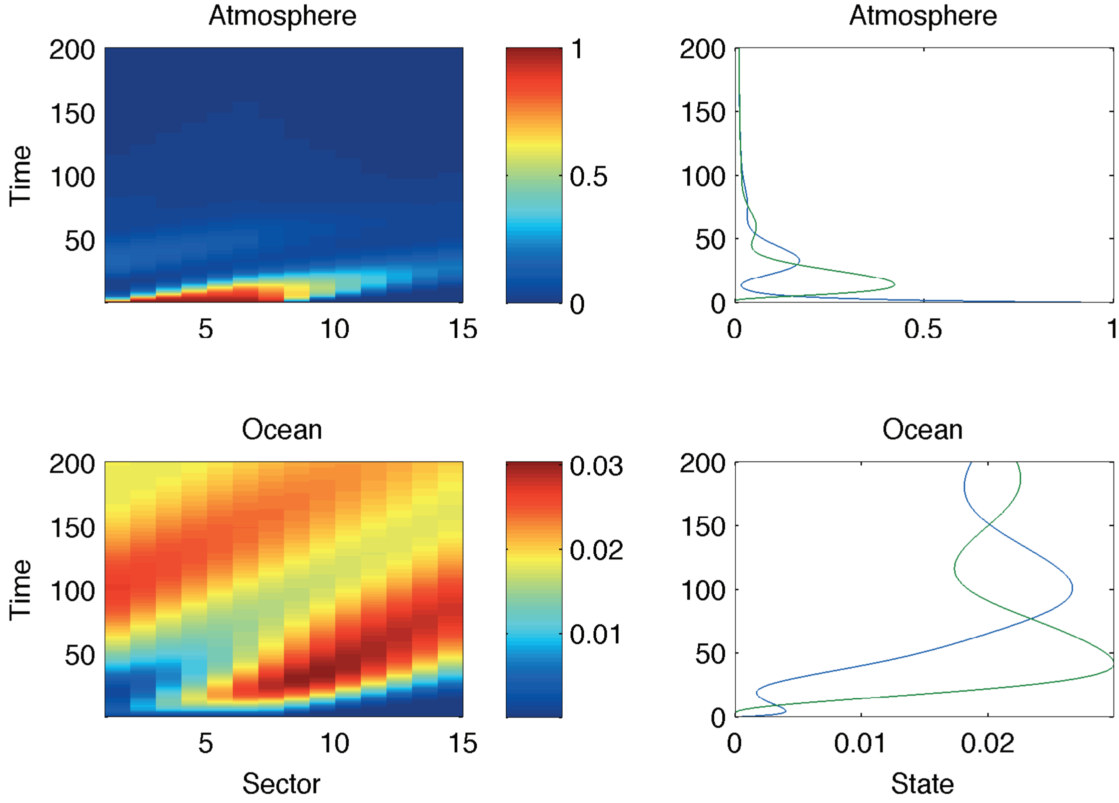

Figure 2.2: Response of atmosphere and ocean to an initial perturbation in the atmosphere. The upper and lower left panels show the atmospheric and oceanic response respectively. The right panels show the time varying response for sector 1 (blue line) and sector 10 (green line). The initial condition is that the atmospheric state is unity for sectors 1 through 7 inclusive, and zero elsewhere. The parameters are defined in Table 2.1 except that the noise is set to zero ($\sigma_1 = \sigma_2 = 0$).

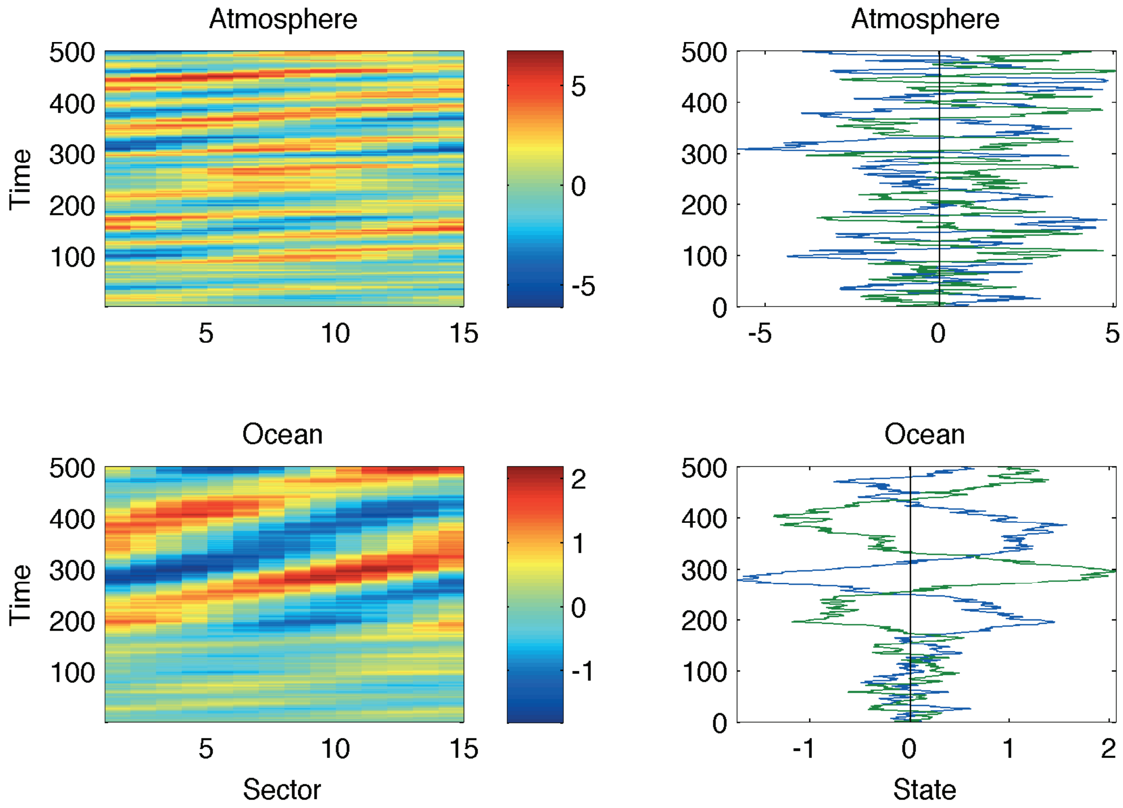

Figure 2.3: Response of atmosphere and ocean to one realization of stochastic forcing. Same format and parameters as Figure 2.2 except that the noise is not set to zero (see Table 2.1) and the model is integrated for 500 time steps.

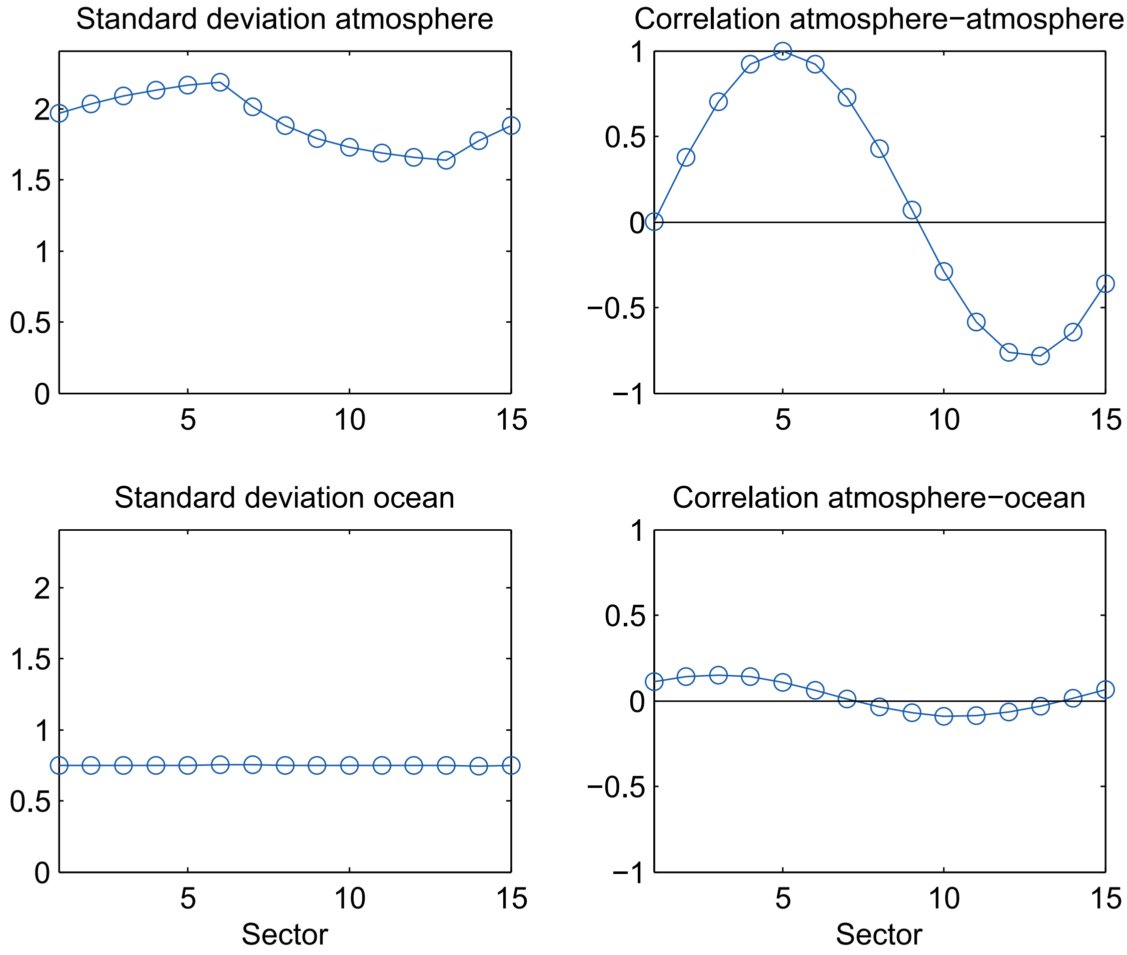

Figure 2.4: Standard deviation and autocorrelation of the state. The left panels show the standard deviation for the atmosphere (upper) and ocean (lower) as a function of sector. The right panels show the correlation at zero lag between the atmospheric state from sector 5 with the other atmospheric variables (upper) and ocean variables (lower) as a function of sector. The parameters are defined in Table 2.1.

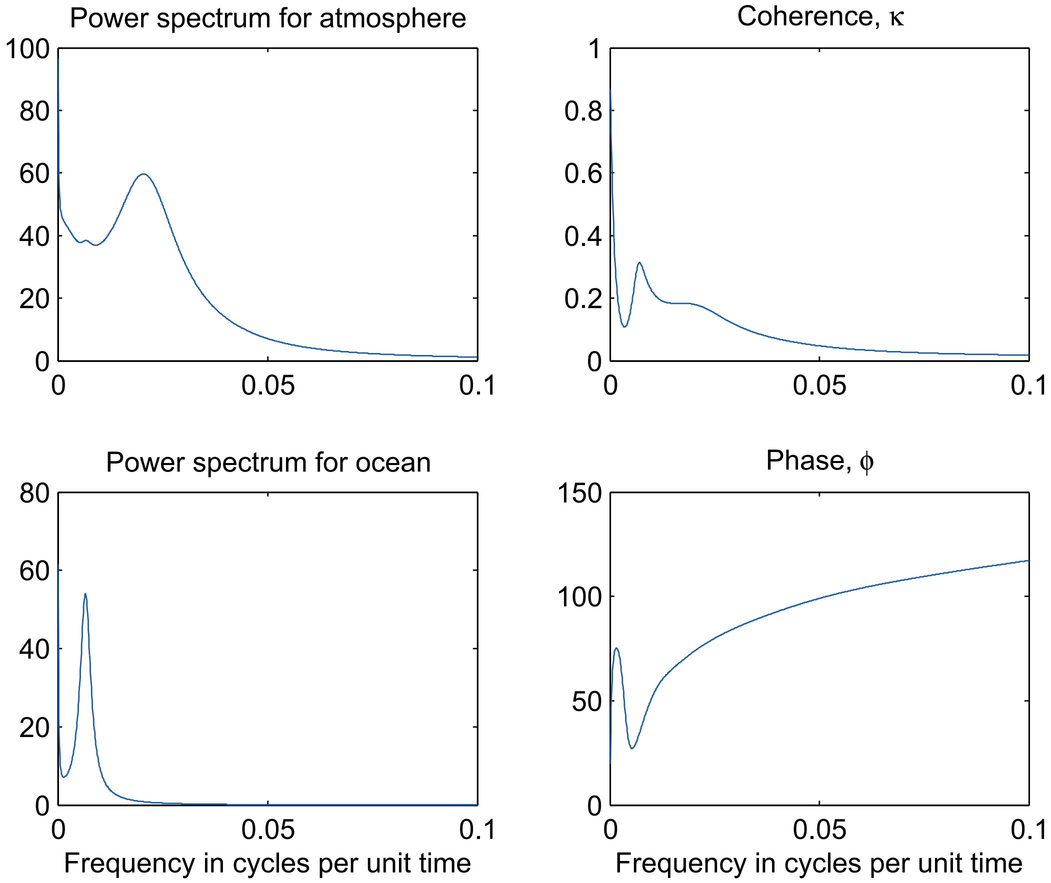

Figure 2.5: Power, coherence and phase spectra of the state. The left panels show the power spectral density for the atmosphere (upper) and ocean (lower) for sector 1. The right panels show the coherence (upper) and phase (lower) for the atmospheric and underlying oceanic state from sector 1. The parameters are defined in Table 2.1.

Chapter 3

Relating High Dimensional Random Vectors

Several statistical methods that provide a low dimensional representation of the relationship between two, high dimensional random vectors, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, are reviewed in this chapter. It is assumed that both vectors have zero mean and the stacked vector $\boldsymbol{X} = \left[\boldsymbol{X}_1^T \ \boldsymbol{X}_2^T\right]^T$ has the following covariance matrix:

$$\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}_{\mathrm{XX}} = \left[\begin{array}{cc} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array}\right] \tag{3.1}$$

where $\boldsymbol{\Sigma}_{11}$ is the $p_1 \times p_1$ covariance matrix of $\boldsymbol{X}_1$, $\boldsymbol{\Sigma}_{22}$ is the $p_2 \times p_2$ covariance matrix of $\boldsymbol{X}_2$, $\boldsymbol{\Sigma}_{12}$ is the $p_1 \times p_2$ covariance matrix of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ and $\boldsymbol{\Sigma}_{21} = \boldsymbol{\Sigma}_{12}^T$.

Two classes of method are discussed. The first class treats the two random vectors in a symmetric fashion and includes Canonical Correlation Analysis (CCA) and Principal Component Analysis (PCA). We denote such methods by $\boldsymbol{X}_1 \leftrightarrow \boldsymbol{X}_2$. The other class of methods treats $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ asymmetrically in the sense that one vector is treated as the predictor and the other as the response. This class of methods includes Multivariate Regression Analysis (MRA), Redundancy Analysis (RA) and Partial Least Squares (PLS) regression. Without loss of generality $\boldsymbol{X}_2$ is taken as the predictor and $\boldsymbol{X}_1$ as the response. The asymmetric methods are denoted by $\boldsymbol{X}_2 \to \boldsymbol{X}_1$.

The rest of this section reviews the above methods starting first with the symmetric methods.

3.1 Principal Component Analysis ($\boldsymbol{X}_1 \leftrightarrow \boldsymbol{X}_2$)

Principal Component Analysis is a well-known method of dimension reduction (e.g., Pearson, 1901 and Jolliffe, 1986). PCA can be used to provide a low dimensional representation of the covariation of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ by carrying out the analysis on the full covariance matrix $\boldsymbol{\Sigma}$ given in (3.1).

The basic idea of PCA is to find a set of "latent variables" that account for as much of the total variance of the p-dimensional random vector $\boldsymbol{X}$ (tr $\boldsymbol{\Sigma}$) as possible. The first principal component is defined by $Z_1 = \boldsymbol{e}_1^T \boldsymbol{X}$. The loading vector $\boldsymbol{e}_1$ is obtained by maximizing $\text{Var}(\boldsymbol{e}^{\text{T}}\boldsymbol{X})$ with respect to $\boldsymbol{e}$ subject to the constraint $|\boldsymbol{e}| = 1$. The second principal component $Z_2 = \boldsymbol{e}_2^T \boldsymbol{X}$ is obtained by maximizing $\text{Var}(\boldsymbol{e}^{\text{T}}\boldsymbol{X})$ with respect to $\boldsymbol{e}$ subject to the constraints $|\boldsymbol{e}| = 1$ and $\boldsymbol{e}^T \boldsymbol{e}_1 = 0$ and so on for higher order principal components.

The loading vectors are the eigenvectors of the covariance matrix of $\mathbf{X}$:

$$\boldsymbol{\Sigma}\boldsymbol{e}_i = \lambda_i \boldsymbol{e}_i, \qquad i = 1 \ldots r$$

where $\lambda_1 \geq \cdots \geq \lambda_r$ are the ordered, real eigenvalues of $\boldsymbol{\Sigma}$. In matrix notation the vector of principal components is given by

$$\boldsymbol{Z} = \boldsymbol{E}^T \boldsymbol{X}$$

where $\boldsymbol{Z} = [Z_1, \cdots, Z_p]^T$ is a $p \times 1$ random vector and $\boldsymbol{E}$ is a $p \times p$ orthogonal loading matrix with i^{th} column given by the ith eigenvector $\boldsymbol{e}_i$.

Principal Component Analysis is a very useful dimension reduction technique and is fundamental to Multivariate Analysis. In terms of the objectives of this thesis it does however have some limitations. First, the results of the PCA are strongly dependent on the units used to define $\boldsymbol{X}$. This can be problematic if the subvectors $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ correspond to different variables (e.g., how do you weight a temperature measurement relative to a pressure measurement?). Second, the loading vectors and corresponding principal components can be hard to interpret for $i > 1$. Third, PCA is not designed to predict one subvector from the other and can perform poorly in such situations.

3.2 Canonical Correlation Analysis ($\boldsymbol{X}_1 \leftrightarrow \boldsymbol{X}_2$)

Canonical Correlation Analysis is a way of measuring the strength of the linear relationship between two random vectors. The basic idea is to first determine the pair of linear combinations with the largest correlation among all possible pairs. The next pair is found by maximizing correlation subject to the constraint it is uncorrelated with the initially selected pair, and so on for higher order pairs. The pairs of linear combinations are called the canonical variables, and their correlations are called canonical correlations. The maximization aspect of the technique represents an attempt to concentrate a high-dimensional relationship between two sets of variables into a small number of canonical variables.

Mathematically CCA finds two bases, one for each variable, that are optimal with respect to correlation. More specifically CCA finds two bases for which the correlation matrix between the new variables is diagonal and the correlations on the diagonal are maximized. The dimensionality of the new bases is equal to, or less than, the smallest dimension of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. An important property of the canonical correlations is that they are invariant with respect to affine transformations of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. This is an most important difference between CCA and ordinary correlation analysis which depends strongly on the bases used to define the variables.

The canonical variables are defined by

$$\begin{aligned} \boldsymbol{Z}_1 &= \boldsymbol{A}^T\boldsymbol{X}_1 && (3.2)\\ \boldsymbol{Z}_2 &= \boldsymbol{B}^T\boldsymbol{X}_2 && (3.3) \end{aligned}$$

where $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ are $r \times 1$ random vectors, $\boldsymbol{A}$ is a $p_1 \times r$ loading matrix for $\boldsymbol{X}_1$ with i^{th} column $\boldsymbol{a}_i$, and $\boldsymbol{B}$ is a $p_2 \times r$ loading matrix for $\boldsymbol{X}_2$ with i^{th} column $\boldsymbol{b}_i$. $\boldsymbol{A}$ and $\boldsymbol{B}$ are called CCA coefficients. To find the CCA coefficients, consider the following singular value decomposition:

$$\boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-\frac{1}{2}} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are $p_1 \times p_1$ and $p_2 \times p_2$ orthogonal matrix respectively. It is straightforward to show (see Appendix B) that $\boldsymbol{a}_i$ is $\boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}\boldsymbol{u}_i$ and $\boldsymbol{b}_i$ is $\boldsymbol{\Sigma}_{22}^{-\frac{1}{2}}\boldsymbol{v}_i$.

Although CCA is straightforward to understand and implement, it has several disadvantages when used to provide a low dimensional representation of the relationship between two high dimensional random vectors. First, the method cannot be applied if $\boldsymbol{\Sigma}_{11}$ and $\boldsymbol{\Sigma}_{22}$ are singular. Second, the CCA coefficients can be hard to interpret because a high canonical correlation does not necessarily mean that the canonical variates account for a significant proportion of the total variance of $\boldsymbol{X}_1$ or $\boldsymbol{X}_2$. This is illustrated in Appendix A by means of a simple example.

3.3 Multivariate Regression Analysis ($\boldsymbol{X}_2 \to \boldsymbol{X}_1$)

Consider now the situation where $\boldsymbol{X}_2$ is treated as the predictor and $\boldsymbol{X}_1$ is the response. Assume both have zero means and predict $\boldsymbol{X}_1$ with a linear predictor of the form

$$\hat{\boldsymbol{X}}_1 = \boldsymbol{B}\boldsymbol{X}_2$$

The value of $\boldsymbol{B}$ that minimizes the trace of $\mathrm{Cov}(\boldsymbol{X}_1 - \boldsymbol{B}\boldsymbol{X}_2)$ is

$$\boldsymbol{B} = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}$$

and the associated prediction error is

$$\begin{aligned} \boldsymbol{R}_1 &= \boldsymbol{X}_1 - \hat{\boldsymbol{X}}_1 && (3.4) \\ &= \boldsymbol{X}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{X}_2 && (3.5) \end{aligned}$$

Note that $\mathrm{Cov}(\boldsymbol{R}_1, \hat{\boldsymbol{X}}_1) = 0$ and so the covariance of the response partitions into a part related to the predictor and an uncorrelated part associated with the prediction error:

$$\boldsymbol{\Sigma}_{11} = \boldsymbol{\Sigma}_{11\cdot 2} + \boldsymbol{\Sigma}_{RR}$$

where

$$\begin{aligned} \boldsymbol{\Sigma}_{11\cdot 2} &= \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \\ \boldsymbol{\Sigma}_{RR} &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \end{aligned}$$

Note that under the assumption of normality the predictor is consistent with the conditional mean of $\boldsymbol{X}_1$ given $\boldsymbol{X}_2 = \boldsymbol{x}_2$. Specifically $\boldsymbol{X}_1 | \boldsymbol{X}_2 = \boldsymbol{x}_2$ is distributed as $N_{p_1}(\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{x}_2,\ \boldsymbol{\Sigma}_{11\cdot 2})$.

Multivariate Regression is one of the most important concepts of Multivariate Analysis. It does not however provide a low dimensional representation of the relationship between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. This requires more advanced techniques including the two described next.

3.4 Redundancy Analysis $(\boldsymbol{X}_2 \rightarrow \boldsymbol{X}_1)$

Redundancy Analysis provides a low dimensional representation of the linear relationship between two random vectors (e.g.,Van den Wollenberg, 1977). The basic idea is to perform a principal component analysis on that part of the response $\boldsymbol{X}_1$ that is linearly related to $\boldsymbol{X}_2$, i.e., $\hat{\boldsymbol{X}}_1 = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{X}_2$.

Principal component analysis of $\mathrm{Cov}(\hat{\boldsymbol{X}}_1)$ yields an ordered set of principal components of the predictable part of $\boldsymbol{X}_1$ and a corresponding set loading vectors stored in the orthogonal matrix $\boldsymbol{U}$:

$$\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$$

where $\boldsymbol{\Lambda}$ denotes the diagonal matrix of eigenvalues, ordered from largest to smallest.

Principal Component Analysis provides the following optimal (in terms of $\mathrm{tr}[\mathrm{Cov}(\hat{\boldsymbol{X}}_1)]$) rank r representation of $\hat{\boldsymbol{X}}_1$:

$$\begin{aligned}\hat{\boldsymbol{X}}_1^r &= \boldsymbol{U}_r\boldsymbol{U}_r^T\hat{\boldsymbol{X}}_1 \\ &= \boldsymbol{U}_r\boldsymbol{U}_r^T\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{X}_2\end{aligned}$$

where the columns of $\boldsymbol{U}_r$ are the first r eigenvectors of $\mathrm{Cov}(\hat{\boldsymbol{X}}_1)$. According to this representation, the r dimensional random vector $\boldsymbol{U}_r^T\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{X}_2$ defines the associated r latent variables; multiplying these latent variables by $\boldsymbol{U}_r$ transforms them into reduced rank representation of $\hat{\boldsymbol{X}}_1$, and thus $\boldsymbol{X}_1$.

To assess the effectiveness of the dimension reduction it is possible to write (3.5) in the form

$$\boldsymbol{X}_1 = \boldsymbol{U}_r\boldsymbol{U}_r^T\hat{\boldsymbol{X}}_1 + (\boldsymbol{I} - \boldsymbol{U}_r\boldsymbol{U}_r^T)\hat{\boldsymbol{X}}_1 + \boldsymbol{R} \tag{3.6}$$

where $\boldsymbol{R}$ is the multivariate regression residual. According to this expression the response is expressed as the sum of (i) the rank r approximation of the predictable part of $\boldsymbol{X}_1$, (ii) the remainder of the predictable part, and (iii) the part of $\boldsymbol{X}_1$ that is not predictable by $\boldsymbol{X}_2$. Taking the trace of (3.6) gives the the following breakdown in the total variance of the response:

$$\text{tr } \boldsymbol{\Sigma}_{11} = (\lambda_1 + \ldots \lambda_{\mathrm{r}}) + (\lambda_{\mathrm{r}+1} + \ldots \lambda_{\mathrm{p}_1}) + \text{tr } \boldsymbol{\Sigma}_{\mathrm{RR}}$$

where λ_i is the i^{th} eigenvalue of $\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. Thus the proportion of the total variance of $\boldsymbol{X}_1$ accounted for by the first r latent variables is given by the so-called Redundancy Index:

$$R_{12}^2(r) = \frac{\sum_{i=1}^{r} \lambda_i}{\text{tr } \boldsymbol{\Sigma}_{11}} \tag{3.7}$$

It is straightforward to show that the Redundancy Index is constrained to lie between 0 and 1.

Based on the above description it is clear that RA makes the optimal prediction based on reduced number of latent variables derived from the predictors. One drawback of RA is that it is difficult to calculate the inverse of $\boldsymbol{\Sigma}_{22}$ when it is poorly conditioned. The technique described in the following section solves this problem.

3.5 Partial Least Squares Regression ($\boldsymbol{X}_2 \rightarrow \boldsymbol{X}_1$)

Partial Least Squares refers to a wide class of methods for dimension reduction. The underlying assumption of all PLS methods is that the original variables are generated by a system or process which is driven by a small number of latent variables. Projections of the original variables onto their latent structure by means of PLS was proposed and developed by Herman Wold and coworkers (see Chapter 1 for references).

PLS covers regression and classification as well as dimension reduction and modeling. The goal of partial least squares regression is to predict $\boldsymbol{X}_1$ from $\boldsymbol{X}_2$ and to describe the common structure underlying the two random vectors. Partial least squares regression allows for the identification of underlying factors (Talbot, 1997). Although similar to PCA and CCA, PLS is considered to be a better alternative to multiple linear regression and principal component-based regression because it provides more robust model parameters that do not change with new calibration samples from the population (Falk Miller, 1992, Geladi Kowalski, 1986). PLS is an improvement over PCA because it is constrained by the part of the covariance matrix that is directly related to the experimental manipulation or that relates to behavior (McIntosh, Chau, Protzner, 2004).

PLS can be applied to classification problems by encoding the class membership in an appropriate indicator matrix, although we mainly use it for prediction in this study. There is a close connection between PLS when used for classification and Fisher Discriminant Analysis.

Recently, PLS is also developed by connecting some statistical tools. For example, the powerful machinery of kernel-based learning can be applied to PLS. It is an elegant way to extend linear data analysis to nonlinear problems. The motivation of extending PLS in this way is that people sometimes prefer to set PLS latent variables as linear projections of the original variables is not adequate. The idea of the kernel PLS method is based on the mapping of the original random variables space into a high dimensional space.

Partial Least Squares is designed to cope with problems that result from small sample sizes, missing values and strong multicollinearity. By way of contrast, ordinary least squares regression can preform poorly when faced with such difficulties. Multicollinearity amongst predictors is a particularly important problem as it can increase the standard error of the estimated regression coefficients. Then it leads to theoretically predictors being omitted from the regression model because they are not statistically significant.

In its general form, PLS creates latent vectors (in some papers they are referred to as score vectors) by maximizing the covariance amongst different sets of variables. This study focuses on the application of PLS to only two random vectors, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. The first pair of latent variables, $\boldsymbol{u}^T\boldsymbol{X}_1$ and $\boldsymbol{v}^T\boldsymbol{X}_2$, are defined by maximizing their squared covariance as follows:

$$\max_{|\boldsymbol{u}|=|\boldsymbol{v}|=1} \mathrm{Cov}(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}_1, \boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}_2)^2$$

PLS can be readily extended to regression problems by treating the latent vectors as new predictor and response variables.

Partial Least Squares usually finds the latent vectors in an iterative fashion. (This is not true for PLS-SVD as explained below.) After the extraction of the first latent vectors of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, the matrices involved with $\boldsymbol{\Sigma}_{12}$ are "deflated" by subtracting their rank-one approximations generated by $\boldsymbol{u}$ and $\boldsymbol{v}$. The different forms of deflation define different variants of PLS. The description below focuses on two variants of PLS: the non iterative PLS-SVD method and the iterative PLS-W2A method.

3.5.1 PLS-SVD

The core part of PLS is the calculation of the loading vectors for $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ which are stored as the columns of the matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ respectively. The latent variables are then given by

$$\boldsymbol{Z}_1 = \boldsymbol{U}^T \boldsymbol{X}_1 \qquad \boldsymbol{Z}_2 = \boldsymbol{V}^T \boldsymbol{X}_2$$

By construction, the first pair of latent variables has the greatest covariance of any pair of linear combinations for which the latent vectors are of unit length, i.e., $|\boldsymbol{u}_1| = |\boldsymbol{v}| = 1$. In PLS-SVD the squared covariance of subsequent pairs of latent variables is maximized subject to the constraint that the next latent variable for $\boldsymbol{X}_1$ is uncorrelated with all previous latent variables for $\boldsymbol{X}_1$ and similarly for the next latent variables for $\boldsymbol{X}_2$. It is then straightforward to show that the latent vectors are the left and right singular vectors of $\boldsymbol{\Sigma}_{12}$, i.e.,

$$\boldsymbol{\Sigma}_{12} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T$$

where $\boldsymbol{U}$ is a $p_1 \times p_1$ orthonormal matrix, $\boldsymbol{V}$ is a $p_2 \times p_2$ orthonormal matrix and $\boldsymbol{\Lambda}$ is a $p_1 \times p_2$ diagonal matrix with ith diagonal element given by the ith singular value, λ_i.

Pre and postmultiplying $\boldsymbol{\Sigma}_{12}$ by $\boldsymbol{U}^T$ and $\boldsymbol{V}$ respectively gives the covariance of the latent vectors, $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$:

$$\mathrm{Cov}(\boldsymbol{Z}_1,\ \boldsymbol{Z}_2) = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Sigma}_{12}\boldsymbol{V} = \boldsymbol{\Lambda}$$

which, by construction, is a diagonal matrix. Note also that the latent vectors for $\boldsymbol{X}_1$ are orthogonal ($\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$) and similarly for the latent vectors of $\boldsymbol{X}_2$ ($\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$).

The latent vectors calculated by PLS-SVD have another useful property that follows the fact that if $\boldsymbol{M}^{(j)}$ denotes a $p_1 \times p_2$ matrix of rank less than or equal to j, the choice of $\boldsymbol{M}^{(j)}$ that minimizes

$$\mathrm{tr}\ (\boldsymbol{\Sigma}_{12} - \boldsymbol{M}^{(\mathrm{j})})(\boldsymbol{\Sigma}_{12} - \boldsymbol{M}^{(\mathrm{j})})^{\mathrm{T}} \tag{3.8}$$

is

$$\boldsymbol{M}^{(j)} = \boldsymbol{U}^{(j)}\boldsymbol{\Lambda}^{(j)}\boldsymbol{V}^{(j)^T} \tag{3.9}$$

where $\boldsymbol{U}^{(j)}$ and $\boldsymbol{V}^{(j)}$ hold the first j columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ respectively, and $\boldsymbol{\Lambda}^{(j)}$ is a $j \times j$ diagonal matrix holding the j largest singular values of $\boldsymbol{\Sigma}_{12}$. According to this result $\lambda_1 \boldsymbol{u}_1 \boldsymbol{v}_1^T$ is the best rank-one approximation of $\boldsymbol{\Sigma}_{12}$ in a least-squares sense. The goodness of fit given by (3.8) serves as a figure of merit for the overall PLS analysis.

One way to interpret (3.9) is to note that the vector $\boldsymbol{u}_1$ can be considered as the first sample principal component of the columns of $\boldsymbol{\Sigma}_{12}$. Thus $\boldsymbol{u}_1$ best fits the p_2 columns of covariance across the p_1 variables of $\boldsymbol{X}_1$. Similarly $\boldsymbol{v}_1$ best fits the p_1 rows of covariance across the p_2 variables of $\boldsymbol{X}_2$.

To interpret λ_1 note that it can be written as

$$\lambda_1 = \sum_{i=1}^{p_1} u_{1i} \mathrm{Cov}(\mathrm{x_i}, \boldsymbol{Z}_{21})$$

where u_{1i} is the i^{th} element of the vector $\boldsymbol{u}_1$. This implies u_{1i} is proportional to the covariances of the corresponding $\boldsymbol{X}_1$ variable with $\boldsymbol{Z}_{21}$, which is the first latent variable of $\boldsymbol{X}_2$. Following the same reason, v_{1j}, the j^{th} element of the vector $\boldsymbol{v}_1$, is proportional to the covariances of the corresponding $\boldsymbol{X}_2$ variable with $\boldsymbol{Z}_{11}$, the first latent variable of $\boldsymbol{X}_1$.

An attractive property of PLS-SVD from a theoretical perspective is that the loading vectors are found in a single step based on the singular decomposition on $\boldsymbol{\Sigma}_{12}$. It does not need to repeat the process of deflating $\boldsymbol{\Sigma}_{12}$ like other PLS algorithms such as PLS-W2A (see below). The PLS-SVD algorithm also generates the latent variables of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ symmetrically (in contrast to some variants of PLS such like PLS2).

An iterative form of PLS is described in the next section.

3.5.2 PLS-W2A

The goal of PLS-W2A is the same as that of PLS-SVD: to find a sequence of pairs of normalized loading vectors that together provide a low dimensional representation of the covariance of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. PLS-W2A differs from PLS-SVD in that it is an iterative method for calculating the loading vectors based on "deflating" the covariance matrix $\boldsymbol{\Sigma}_{12}$ each step. Details are given below.

<u>Step 1</u>: Initialize covariance matrix, random vectors and index

$$\boldsymbol{\Sigma}_{12}^{(1)} = \boldsymbol{\Sigma}_{12} \qquad \boldsymbol{X}_2^{(1)} = \boldsymbol{X}_2 \qquad \boldsymbol{X}_1^{(1)} = \mathbf{X}_1 \qquad r = 1$$

<u>Step 2:</u> Extract loading vector

Calculate $\boldsymbol{u}_1^{(r)}$ and $\boldsymbol{v}_1^{(r)}$, the first left and right singular vectors of $\boldsymbol{\Sigma}_{12}^{(r)}$, and $\lambda_1^{(r)}$, the corresponding singular value.

<u>Step 3:</u> Obtain the latent variables

$$\begin{aligned} \omega_r &= \boldsymbol{u}_1^{(r)T} \boldsymbol{X}_1^{(r)} \\ \xi_r &= \boldsymbol{v}_1^{(r)T} \boldsymbol{X}_2^{(r)} \end{aligned}$$

<u>Step 4:</u> Regress the deflated random vectors on the latent variables

$$\begin{aligned} \hat{\boldsymbol{X}}_1^{(r)} &= \frac{\mathrm{Cov}(\boldsymbol{X}_1^{(\mathrm{r})}, \omega_{\mathrm{r}})}{\mathrm{Var}(\omega_{\mathrm{r}})} \omega_r \\ \hat{\boldsymbol{X}}_2^{(r)} &= \frac{\mathrm{Cov}(\boldsymbol{X}_2^{(\mathrm{r})}, \xi_{\mathrm{r}})}{\mathrm{Var}(\xi_{\mathrm{r}})} \xi_r \end{aligned}$$

These equations can be written in the form

$$\begin{aligned} \hat{\boldsymbol{X}}_1^{(r)} &= \boldsymbol{Q}_1^{(r)} \boldsymbol{X}_1^{(r)} \qquad \boldsymbol{Q}_1^{(r)} = \frac{1}{\boldsymbol{u}_1^{(r)T} \boldsymbol{\Sigma}_{11}^{(r)} \boldsymbol{u}_1^{(r)}} \boldsymbol{\Sigma}_{11}^{(r)} \boldsymbol{u}_1^{(r)} \boldsymbol{u}_1^{(r)T} \\ \hat{\boldsymbol{X}}_2^{(r)} &= \boldsymbol{Q}_2^{(r)} \boldsymbol{X}_2^{(r)} \qquad \boldsymbol{Q}_2^{(r)} = \frac{1}{\boldsymbol{v}_1^{(r)T} \boldsymbol{\Sigma}_{22}^{(r)} \boldsymbol{v}_1^{(r)}} \boldsymbol{\Sigma}_{22}^{(r)} \boldsymbol{v}_1^{(r)} \boldsymbol{v}_1^{(r)T} \end{aligned}$$

<u>Step 5:</u> Update the random vectors

$$\begin{aligned} \boldsymbol{X}_2^{(r+1)} &= \boldsymbol{X}_2^{(r)} - \hat{\boldsymbol{X}}_2^{(r)} \\ \boldsymbol{X}_1^{(r+1)} &= \boldsymbol{X}_1^{(r)} - \hat{\boldsymbol{X}}_1^{(r)} \end{aligned}$$

STEP 6: More deflation required?

If yes, set $r = R$, exit.

If no, continue.

STEP 7: Deflate the covariance matrix and reset the index

$$\boldsymbol{\Sigma}_{12}^{(r+1)} = \left(\boldsymbol{I} - \boldsymbol{Q}_1^{(r)}\right) \boldsymbol{\Sigma}_{12}^{(r)} \left(\boldsymbol{I} - \boldsymbol{Q}_2^{(r)}\right)^T \tag{3.10}$$

Note that for the PLS-SVD method the deflation step takes the form

$$\boldsymbol{\Sigma}_{12}^{r+1} = \boldsymbol{\Sigma}_{12} - \sum_{s=1}^{r} \lambda_s \boldsymbol{U}_{.s} \boldsymbol{V}_{.s}^T$$

Set r to $r+1$ and go to STEP 2.

The main difference between PLS-SVD and PLS-W2A is that $\mathrm{Cov}(\hat{\boldsymbol{X}}_1^{(\mathrm{r})}, \hat{\boldsymbol{X}}_1^{(\mathrm{s})})$ and $\mathrm{Cov}(\hat{\boldsymbol{X}}_2^{(\mathrm{r})}, \hat{\boldsymbol{X}}_2^{(\mathrm{s})})$ are generally nonzero for PLA-W2A when $r \neq s$. The difference between the two methods is discussed in more detail in Appendix B.

Chapter 4

Comparison of Methods and Development of a Hybrid

Three methods (CCA, RA, and PLS) were introduced in Chapter 3 to provide a low dimensional representation of the relationship between two high dimensional vectors, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, based on the concept of latent variables.

Redundancy Analysis treats one vector as the predictor ($\boldsymbol{X}_2$) and the other ($\boldsymbol{X}_1$) as the response (i.e., the method treats the vectors asymmetrically). The choice of latent variables is based on PCA of that part of $\boldsymbol{X}_1$ that can be predicted by $\boldsymbol{X}_2$ using multivariate regression. PLS and CCA are similar in the sense they both treat $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ symmetrically. The methods differ in that PLS find pairs of latent variables with the largest covariance, while CCA find pairs of latent variables with the largest correlation.

Although the three methods are similar in principle, they differ fundamentally in their numerical properties. Redundancy Analysis is based on a PCA of $\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$ and requires the computation of inverse of $\boldsymbol{\Sigma}_{22}$. This can be problematic if $\boldsymbol{\Sigma}_{22}$ is singular or poorly conditioned. Canonical Correlation Analysis is based on the singular value decomposition of $\boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-\frac{1}{2}}$ and so, like RA, depends on the inversion of potentially poorly conditioned covariance matrices. Partial Least Squares differs from CCA and RA is that it does not require the inverse of $\boldsymbol{\Sigma}_{11}^{-1}$ or $\boldsymbol{\Sigma}_{22}^{-1}$ and is thus a more robust method.

There are many papers on the relationship between CCA and RA (e.g., Muller, 1981) and the relationship between CCA and PLS (e.g., Jacob, 2000). In this chapter, we focus on the relationship between RA and PLS, and pay particular attention to a hybrid method developed recently by Bougeard et al., (2007). A new, highly idealized model (simpler than the coupled atmosphere-ocean model) is introduced to better understand the differences between RA, PLS and the hybrid.

4.1 Comparison of PLS and RA Using an Idealized Example

A highly idealized model is now introduced to illustrate the differences between PLS and RA. Consider two random vectors that are related according to the following linear model:

$$\boldsymbol{X}_1 = \boldsymbol{E}_1\boldsymbol{S}\boldsymbol{E}_2^T\boldsymbol{X}_2 + \boldsymbol{n} \tag{4.1}$$

where $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$ are $p_1 \times r$ and $p_2 \times r$ matrices with orthonormal columns, $\boldsymbol{S}$ is an $r \times r$ diagonal "scale" matrix with positive diagonal elements (s_k), and $\boldsymbol{n}$ is an additive noise random vector. The columns of $\boldsymbol{E}_2$ are taken to be the r eigenvectors of $\boldsymbol{\Sigma}_{22}$ with the r largest eigenvalues.

It is straightforward to show

$$\boldsymbol{\Sigma}_{12} = \boldsymbol{E}_1\boldsymbol{S}\boldsymbol{\Lambda}_2\boldsymbol{E}_2^T \qquad \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} = \boldsymbol{E}_1\boldsymbol{S}\boldsymbol{E}_2^T \qquad \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} = \boldsymbol{E}_2\boldsymbol{S}\boldsymbol{\Lambda}_2\boldsymbol{S}\boldsymbol{E}_2^T \tag{4.2}$$

where $\boldsymbol{\Lambda}_2$ is a diagonal matrix holding the r largest eigenvalues of $\boldsymbol{\Sigma}_{22}$.

The matrix $\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}$ defines the regression coefficients for predicting $\boldsymbol{X}_1$ from $\boldsymbol{X}_2$. To predict $\boldsymbol{X}_1$, first project $\boldsymbol{X}_2$ onto its principal components, scale by the diagonal elements of $\boldsymbol{\Lambda}_2$ (the s_k), and then transform to the $\boldsymbol{X}_1$ prediction by multiplying by $\boldsymbol{E}_1$.

In PLS, the latent variables are based on the singular value decomposition of $\boldsymbol{\Sigma}_{12}$. From (4.2) it is clear that PLS-SVD will select $\boldsymbol{X}_2$ latent variables that are the principal components of $\boldsymbol{X}_2$ ordered by $s_k\lambda_k$. To interpret this result, note that the term $s_k\lambda_k$ can be written $(s_k\sqrt{\lambda_k})\sqrt{\lambda_k}$ which, roughly speaking, can be thought of as the product of the standard deviation of the predictable part of $\boldsymbol{X}_1$ and the standard deviation of $\boldsymbol{X}_2$.

In RA, the loading vectors are based on the eigenvalues and vectors of $\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. From (4.2) it is clear that RA will again select latent variables that are the principal components of $\boldsymbol{X}_2$ but now ordered by $s_k^2\lambda_k$. To interpret this result, note that the term $s_k^2\lambda_k$ which can be thought of as the variance of the predictable part of $\boldsymbol{X}_1$.

Thus according to this highly idealized model, the $\boldsymbol{X}_2$ latent variables for PLS-SVD and RA are simply the principal components of $\boldsymbol{X}_2$ but the choice of principal component depends on $s_k\lambda_k$ and $s_k^2\lambda_k$ respectively. This is illustrated in Figure 4.1 for a particular choice of s_k and λ_k. The left panels show the eigenvalues (λ_k) of $\boldsymbol{\Sigma}_{22}$ and the diagonal elements of the "scale" matrix (s_k). Note the s_k are assumed to increase with k meaning that the strength of the directed relationship from $\boldsymbol{X}_2$ to $\boldsymbol{X}_1$ is stronger for the higher principal components. The right panels show $s_k\lambda_k$ and $s_k^2\lambda_k$. For PLS, the first $\boldsymbol{X}_2$ latent variable will be the eleventh principal component of $\boldsymbol{X}_2$; for RA it will be the fourteenth principal component, reflecting the greater emphasis on prediction.

4.2 Combining the Strengths of RA and PLS: A Hybrid Approach

Given the good prediction ability of RA, and the robustness of PLS, we now follow Bougeard et al., (2007) and construct a blend of these two approaches using a hybrid approach. As shown in Chapter 3, the loading vectors ($\boldsymbol{v}$) for the $\boldsymbol{X}_2$ latent variables for PLS-SVD and RA satisfy the following equations:

$$\begin{aligned} &\text{PLS:} & \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{12}\boldsymbol{v} &= \lambda\boldsymbol{v} \\ &\text{RA:} & \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{12}\boldsymbol{v} &= \lambda\boldsymbol{v} \end{aligned}$$

Following Bougeard et al., (2007) we treat PLS and RA as the two endpoints of a hybrid approach to calculating the loading vector $\boldsymbol{v}$ which is assumed to satisfy

$$[\alpha\boldsymbol{I} + (1-\alpha)\boldsymbol{\Sigma}_{22}]^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{12}\boldsymbol{v} = \lambda\boldsymbol{v} \tag{4.3}$$

where the parameter α can be selected to be any real number between 0 and 1. Note that $\alpha = 0$ we recover RA, and if $\alpha = 1$ we recover PLS.

To calculate the loading vector $\boldsymbol{v}$, pre-multiply both sides of (4.3) by $\boldsymbol{P}^{\frac{1}{2}}$ to get

$$\boldsymbol{P}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{12}\boldsymbol{P}^{-\frac{1}{2}}\boldsymbol{P}^{\frac{1}{2}}\boldsymbol{v} = \lambda\boldsymbol{P}^{\frac{1}{2}}\boldsymbol{v} \tag{4.4}$$

where

$$\boldsymbol{P} = \alpha\boldsymbol{I} + (1-\alpha)\boldsymbol{\Sigma}_{22}$$

It is more convenient to rewrite (4.4) as a pair of equations:

$$\begin{aligned} \boldsymbol{v} &= \boldsymbol{P}^{-\frac{1}{2}}\boldsymbol{u} & (4.5) \\ \lambda\boldsymbol{u} &= \boldsymbol{P}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{12}\boldsymbol{P}^{-\frac{1}{2}}\boldsymbol{u} & (4.6) \end{aligned}$$

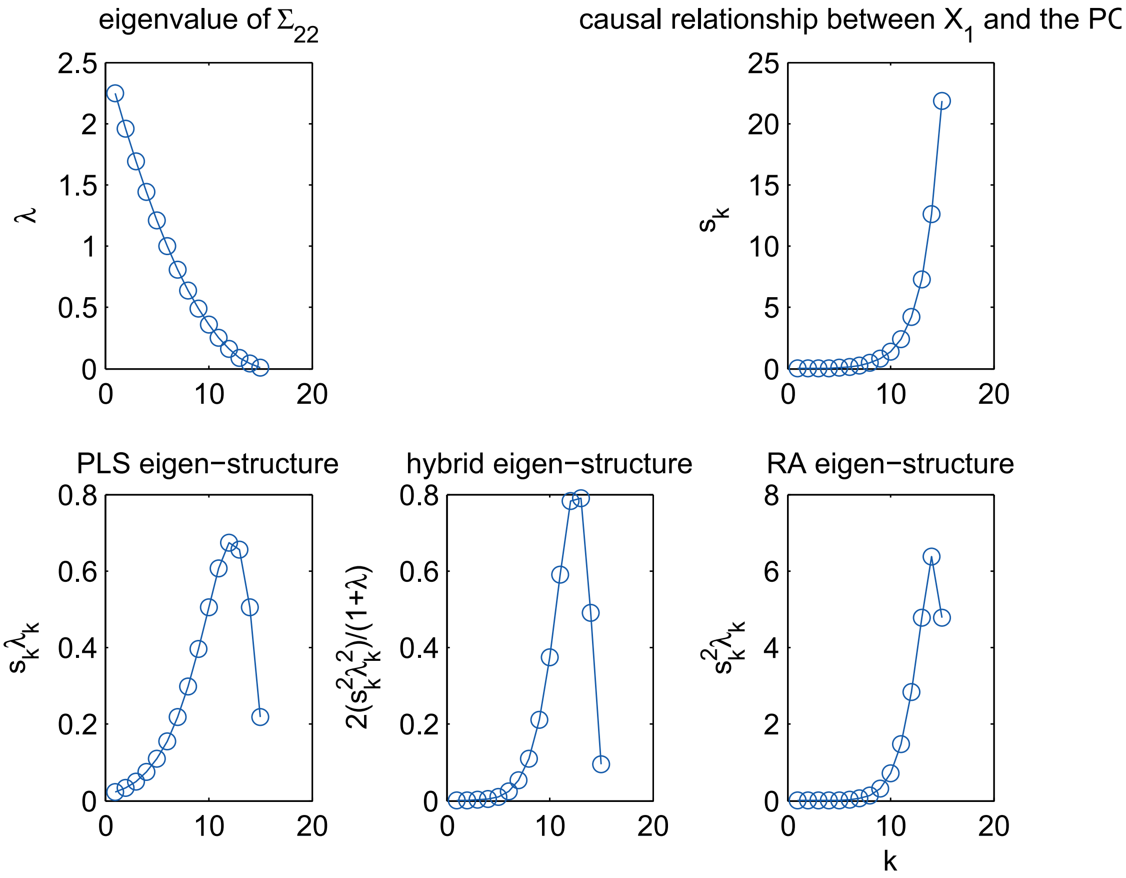

Figure 4.1: Selection of the latent variables of the highly idealized model for RA, PLS and CR-PLS. The upper panels show the eigenvalues of $\mathbf{\Sigma}_{22}$ (λ_k) and the diagonal elements of the "scale" matrix $\boldsymbol{S}$ (s_k). The lower left panel shows $s_k\lambda_k$ versus k. This is used to order the principal components of $\boldsymbol{X}_2$ when finding the latent variables for PLS. The lower middle panel shows $2(s_k^2\lambda_k)$ versus k. This is used to order the principal components when finding the latent variables for CR-PLS with $\alpha = 0.5$. The lower right panel shows $s_k^2\lambda_k$ versus k. This is used to order the principal components when finding the latent variables for RA. In this example, $\lambda_k = 0.01 \times (16 - k)^2$ and $s_k = 0.01 \times 3^{\frac{k-1}{2}}$.

From (4.6), it gives

$$\lambda = \boldsymbol{u}^T \boldsymbol{P}^{-\frac{1}{2}} \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{12} \boldsymbol{P}^{-\frac{1}{2}} \boldsymbol{u} \tag{4.7}$$

This means that for a specified α, we can get $\boldsymbol{v}$ by first calculating $\boldsymbol{u}$ as the first eigenvector of (4.6) and then calculating $\boldsymbol{v}$ from (4.5).

It is interesting to consider the behavour of CR-PLS as $\alpha \to 0$ in the case that $\boldsymbol{\Sigma}_{22}$ is singular. It is straightforward to show from (4.3) that in this limit the loading vector for $\boldsymbol{X}_2$ satisfies

$$\boldsymbol{\Sigma}_{22}^{+} \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{12} \boldsymbol{v} \quad = \quad \lambda \boldsymbol{v}$$

where $\boldsymbol{\Sigma}_{22}^{+}$ is the pseudo inverse of $\boldsymbol{\Sigma}_{22}$.

The hybrid method is a simple blend of PLS and RA. In prediction problems, the goal is to find an optimal value of α that simultaneously provides stable latent vectors and good fitting ability. Following Bougeard et al. (2007) we will refer to this approach as CR-PLS (Continuum Redundancy-Partial Least Squares).

The ratio of largest eigenvalue to the smallest eigenvalue of $\boldsymbol{P}$, denoted by β, reflect the stability of the model. A large β indicates serious multicolinearity of $\boldsymbol{\Sigma}_{22}$ and this will lead to a unstable model. In CR-PLS,

$$\beta = \frac{(1-\alpha)\lambda_1 + \alpha}{(1-\alpha)\lambda_p + \alpha} \tag{4.8}$$

where λ_1 and λ_p are the largest and smallest eigenvalues of $\boldsymbol{\Sigma}_{22}$. It is easy to prove that (4.8) is a decreasing function of α. Thus, as α increases, CR-PLS goes to the end point corresponding to PLS and the stability of model is enhanced.

We can take a closer look of the hybrid technique by applying it to the highly idealized example introduced earlier in this chapter. It is straightforward to show that the $\boldsymbol{X}_2$ loading vector satisfies

$$[\alpha \boldsymbol{I} + (1-\alpha)\boldsymbol{E}\boldsymbol{\Lambda}\boldsymbol{E}^T]^{-1} \boldsymbol{E}_2 \boldsymbol{\Sigma}_2 \boldsymbol{S} \boldsymbol{E}_1^T \boldsymbol{E}_1 \boldsymbol{S} \boldsymbol{\Lambda}_2 \boldsymbol{E}_2^T \boldsymbol{v} = \lambda \boldsymbol{v}$$

which simplifies to

$$\boldsymbol{E}_2 [\alpha \boldsymbol{I} + (1-\alpha)\boldsymbol{\Lambda}_2]^{-1} \boldsymbol{S}^2 \boldsymbol{\Lambda}_2^2 \boldsymbol{E}_2^T \boldsymbol{v} = \lambda \boldsymbol{v}$$

Thus for the highly idealized model, the latent variables are simply the principal components of $\boldsymbol{X}_2$ but ordered according to

$$\frac{s_k^2 \lambda_k^2}{\alpha + (1-\alpha)\lambda_k}$$

This again shows that CR-PLS is balancing the variance of $\boldsymbol{X}_2$ and the causal relationship between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ by varying α.

The implementation of CR-PLS in PLS-SVD and PLS-W2A is described in Appendix C.

Chapter 5

Allowing for Frequency Dependent Relationships

All of the covariance-based methods for dimension reduction discussed so far can be readily generalized to the frequency domain. The original random vectors and their covariance matrices are

$$\boldsymbol{X}_1,\ \boldsymbol{X}_2 : \qquad\qquad \boldsymbol{\Sigma}_{11} \qquad \boldsymbol{\Sigma}_{12} \qquad \boldsymbol{\Sigma}_{22}$$

Their counterparts in the frequency domain are

$$\boldsymbol{dX}_1(\omega),\ \boldsymbol{dX}_2(\omega) : \qquad\qquad \boldsymbol{h}_{11(\omega)} \quad \boldsymbol{h}_{12}(\omega) \qquad \boldsymbol{h}_{22}(\omega)$$

where $\boldsymbol{h}_{11}(\omega)$ is the power spectral matrix of $\boldsymbol{X}_1$, $\boldsymbol{h}_{22}(\omega)$ is the power spectral matrix of $\boldsymbol{X}_2$, and $\boldsymbol{h}_{12}(\omega) = \boldsymbol{h}_{21}(\omega)^*$ is the cross spectral matrix of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. As mentioned in Chapter 2, the $\boldsymbol{dX}(\omega)$ can be thought of as the complex amplitudes of the sinusoidal components that make up $\boldsymbol{X}_t$ (similar to a Fourier series expansion):

$$\boldsymbol{X}_t = \int_{-\pi}^{\pi} \mathrm{e}^{i\omega t}\, \boldsymbol{dX}(\omega)$$

The cross spectral and covariances matrices form a Fourier transform pair and thus essentially provide the same information but presented from differing perspectives (i.e., time and frequency domains). For example, the covariance between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ can be written

$$\boldsymbol{\Sigma}_{12} = \int_{-\pi}^{\pi} \boldsymbol{h}_{12}(\omega) d\omega$$

Thus the cross spectral matrix $\boldsymbol{h}_{12}(\omega)$ cab be thought of as the contribution to the covariance of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ at frequency ω. Note that the elements of $\boldsymbol{h}_{12}$ will, in general, be complex.

The extension of the covariance-based methods for dimension reduction to the frequency domain is straightforward in practice and simply involves replacing the covariance matrices by their spectral equivalents. This is illustrated below for principal component analysis and multivariate regression. The extension of the other methods, including the hybrid introduced in Chapter 4, is straightforward.

5.1 Principal Component Analysis in the Frequency Domain

Statisticians sometimes face the problem of dealing with redundant information, or the need for fewer "input" variables, in the frequency domain. Here we follow the development of Priestley (1973) and introduce principal component analysis in the frequency domain and show it is analogous to ordinary principal component analysis.

Ordinary PCA (see Chapter 3) searches for a latent variable

$$Z = \boldsymbol{e}^T \boldsymbol{X} \tag{5.1}$$

where $\boldsymbol{e}$ is selected to maximize $\mathrm{Var(Z)} = \boldsymbol{e}^{\mathrm{T}} \boldsymbol{\Sigma}_{\mathrm{xx}} \boldsymbol{e}$ subject to the constraint $\boldsymbol{e}^T \boldsymbol{e} = 1$.

Based on the spectral representation, (5.1) may be generalized to

$$dZ(\omega) = \boldsymbol{e}(\omega)^T d\boldsymbol{X}(\omega)$$

where the loading vector is now allowed to vary with frequency. It follows that

$$\mathrm{Var(Z)} = \int_{-\pi}^{\pi} \boldsymbol{e}(\omega)^{\mathrm{T}} \boldsymbol{h}_{\mathrm{xx}}(\omega) \boldsymbol{e}(\omega) \mathrm{d}\omega$$

This shows if we treat each frequency component separately, the loading vector is just the eigenvector of $\boldsymbol{h}_{xx}$ with the largest eigenvalue. Note that the loading vector will in general be complex but the eigenvalue will be real because $\boldsymbol{h}_{xx}$ is an Hermitian matrix. The extension to higher order loading vectors and principal components follows as an obvious generalization of the approach described in Chapter 3. The result is that the loading vector for the ith frequency dependent principal component is the ith eigenvector of $\boldsymbol{h}_{xx}$.

5.2 Multivariate Regression in the Frequency Domain

The ordinary multivariate regression model is of the form

$$\boldsymbol{X}_1 = \boldsymbol{B}\boldsymbol{X}_2 + \boldsymbol{\epsilon}$$

where $\boldsymbol{X}_1$ is the response, $\boldsymbol{X}_2$ is the predictor and $\boldsymbol{\epsilon}$ is a noise vector assumed uncorrelated with $\boldsymbol{X}_2$. The $\boldsymbol{B}$ matrix transforms the predictor into the response. We now generalize this model to allow $\boldsymbol{B}$ to vary with frequency according to the following equation:

$$d\boldsymbol{X}_1(\omega) = \boldsymbol{B}(\omega)d\boldsymbol{X}_2(\omega) + d\boldsymbol{E}(\omega)$$

where $d\boldsymbol{X}_1$, $d\boldsymbol{X}_2$ and $d\boldsymbol{E}$ come from the spectral representations of $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ and $\boldsymbol{\epsilon}$.

If we assume $d\boldsymbol{X}_2(\omega)$ and $d\boldsymbol{E}(\omega)$ are uncorrelated, and multiply both sides of the above equation by $d\boldsymbol{X}_2(\omega)^*$ and take expectations, we obtain

$$\boldsymbol{B}(\omega) = \boldsymbol{h}_{12}\boldsymbol{h}_{22}^{-1}$$

It is straightforward to show that this expression for $\boldsymbol{B}(\omega)$ minimizes the trace of the power spectral matrix of prediction errors at each frequency.

In general, the elements of $\boldsymbol{B}(\omega)$ are complex and the i, jth element can be written in the form

$$B_{ij}(\omega) = |B_{ij}(\omega)|\mathrm{e}^{i\phi(\omega)} \tag{5.2}$$

According to the language of filtering theory, B_{ij} is called the "gain" of the transfer function linking the j element of $d\boldsymbol{X}_2$ with the ith element of $d\boldsymbol{X}_1$. The quantity $\phi_{ij}(\omega)$ is the corresponding "phase". Both vary with frequency and so we will refer to the gain spectrum and phase spectrum.

To interpret the gain and phase we refer to (Priestley, 1982) and note that the ith element of $\boldsymbol{X}_1$ can be written

$$dX_{1i}(\omega) = \sum_j |B_{ij}(\omega)|\mathrm{e}^{i\phi(\omega)}dX_{2j}(\omega)$$

Thus in frequency domain, the ith element of $d\boldsymbol{X}_1(\omega)$ is obtained by scaling the amplitude of the jth element of $d\boldsymbol{X}_2$ by the gain $|B_{ij}(\omega)|$ and shifting the phase by $\phi(\omega)$.

Chapter 6

Comparing Approaches Using the Coupled Model

We now illustrate the dimension reduction methods using covariances calculated from the coupled atmosphere-ocean model introduced in Chapter 2. This model is more complicated than the highly idealized model introduced in Chapter 4 (see (4.1)), is physically realistic and relevant, and provides explicit covariance and cross spectral matrices (see (2.3) and (2.8)) . It follows that all of the results shown in this chapter are not subject to sampling variability.

The covariance structure of the noise, $\boldsymbol{M}\boldsymbol{\xi}$, is controlled by the variance of $\boldsymbol{\xi}$ (denoted by $\boldsymbol{\Sigma}_{\xi\xi}$) and the spatial structure of the noise (determined by the columns of $\boldsymbol{M}$). If the ocean component of the noise has a covariance matrix that is full rank, and the ocean noise is large compared to the atmospheric noise, the covariance matrix of the ocean state, $\boldsymbol{\Sigma}_{22}$, will in general be well conditioned. In this study, however, we are interested in the opposite situation where $\boldsymbol{\Sigma}_{22}$ is poorly conditioned. This can be caused by weak ocean noise or a small number of columns for the ocean part of $\boldsymbol{M}$. The variances of the atmosphere and ocean components of $\boldsymbol{\xi}$ are given in Table 2.1 and the spatial structure of the noise is controlled by the $\boldsymbol{M}$ matrix; both were chosen to ensure $\boldsymbol{\Sigma}_{22}$ is poorly conditioned.

In the remainder of this chapter, covariance and cross spectral matrices from the coupled atmosphere-ocean model are used to illustrate and evaluate PCA, CCA, RA , PLS and finally the hybrid approach. All of the covariance matrices, and their frequency dependent generalizations (e.g., $\boldsymbol{h}_{22}$) are identical to those introduced in Chapter 2. We also use the theoretical correlation ($\boldsymbol{R}$) and coherency ($\boldsymbol{W}$) matrices of the state vector $\boldsymbol{X}$. The correlation matrix is shown below in partitioned form:

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}_{11} & \boldsymbol{R}_{12} \\ \boldsymbol{R}_{21} & \boldsymbol{R}_{22} \end{bmatrix}$$

where $\boldsymbol{R}_{11}$ and $\boldsymbol{R}_{22}$ are the correlation matrices of atmosphere and ocean state respectively, and $\boldsymbol{R}_{12}$ is the correlation matrix between atmosphere and ocean state. The coherency matrix takes on a similar form but the off diagonal elements are complex, the matrix is Hermitian, and it depends of frequency.

6.1 Principal Component Analysis

A principal component analysis was performed on the correlation matrix $\boldsymbol{R}$ (corresponding to an analysis of the standardized state vector, $\boldsymbol{X}$). The reason the correlation matrix is used is that the variability in the atmosphere is much greater than the variability in the ocean (see Figure 2.4).

The "scree" plot (Figure 6.1) shows that almost all of the total variance of the standardized state (equal to $2N$) is captured by the first six principal components; the rest of the principal components have relative small variances (i.e., the eigenvalues of $\boldsymbol{R}$). This implies $\boldsymbol{\Sigma}$ is ill-conditioned as expected.

The loading vectors for the principal components are given by the eigenvectors of $\boldsymbol{R}$. The first eigenvector, $\boldsymbol{e}_1$, is shown in the left panels of Figure 6.2. The upper panel shows the loadings for the atmosphere, and the lower panel shows the loadings for the ocean. The loadings for the ocean and atmosphere each correspond to a wave-like pattern with a wavelength that is equal to the circumference of the torus. We will call such spatial structures a mode 1 pattern. It can be seen that $\boldsymbol{e}_3$ has the same mode 1 structure as $\boldsymbol{e}_1$ but it is in quadrature. The second eigenvector, $\boldsymbol{e}_2$, is similar to $\boldsymbol{e}_1$ except the mode 1 variations in the ocean are shifted by half a cycle. The fourth eigenvector, $\boldsymbol{e}_4$, is similar to $\boldsymbol{e}_2$ but in quadrature. The fifth and sixth eigenvectors describe changes in the mean temperature of the atmosphere and ocean. Eigenvectors $\boldsymbol{e}_7$ and $\boldsymbol{e}_8$ are similar to $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ except they describe mode 2 variations (i.e., variations with two peaks during one circuit of the torus). Taken together the loading vectors are similar to a Fourier decomposition of the spatial variations in the atmosphere and ocean.

6.2 Canonical Correlation Analysis

The canonical correlations are shown in Figure 6.3. The first three correlations are very close to one; the last pair of canonical variates have almost zero correlation.

The loading vectors used to define the first three pairs of canonical variates (see (3.3)) are plotted in Figure 6.4. The first pair are plotted in the leftmost column of panels. Unlike the simple mode 1 patterns of the eigenvectors of $\mathbf{\Sigma}_{22}$, the spatial pattern of the canonical coefficients are highly irregular and difficult to interpret physically. The first three canonical variates also do not describe well the variability in $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ as shown by Table 6.1.

Table 6.1: Proportion of total variance of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ that can be predicted by the first canonical variate (column 2), the first two canonical variates (column 3) and the first three canonical variates (column 3).

Vector	Z_1	Z_1, Z_2	Z_1, Z_2, Z_3
$\boldsymbol{X}_1$	0.12	0.45	0.46
$\boldsymbol{X}_2$	0.01	0.01	0.02

The results based on the coupled atmosphere-ocean model clearly demonstrate the problems that can occur with CCA: (i) loading vectors are unstable and difficult to interpret (ii) canonical variables may not able to represent well the trace of $\mathbf{\Sigma}_{11}$ or $\mathbf{\Sigma}_{22}$.

6.3 Redundancy Analysis

Proportion of the total variance of $\boldsymbol{X}_1$ predicted by the latent variables of $\boldsymbol{X}_2$ from RA. The blue line shows the proportion of total variance account for by the kth latent variable. The green line shows the cumulative proportion of variance account for by the first k latent variables and red line plot gives the proportion of the total variance of $\boldsymbol{X}_1$ accounted for by multivariate regression. As discussed in Chapter 3, the proportion of variance explained by multivariate regression is the largest scale of any statistical technique can reach. In this example, the variance accounted by multivariate regression is

$$R^2 = \frac{\mathrm{tr}(\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}^{\mathrm{T}})}{\mathrm{tr}(\boldsymbol{\Sigma}_{11})} = 0.65$$

It is noted that the redundancy index based on the first 3 RA latent variables is pretty much the same as 0.65. This indicates RA provides a very good prediction and accounts for as much total variance of $\boldsymbol{X}_1$ as possible in this example.

The loading vectors for RA are shown in Figure 6.6 as a function of sector position. Note that although the loading vectors for the atmosphere have a simple mode 1 shape, the corresponding loading vectors for the ocean have a very irregular pattern which is caused by the poor condition of $\boldsymbol{\Sigma}_{22}$. This implies that the ocean latent variables generated by RA are not stable with respect to small perturbations in $\boldsymbol{\Sigma}_{22}$.

6.4 Partial Least Squares

In this section we focus on PLS-SVD which is more amenable to theoretical analysis.

The loading vectors for the first three latent variables for the atmosphere and ocean, calculated from PLS-SVD of $\boldsymbol{\Sigma}_{12}$, are shown in the left panels of Figure 6.7. Note that for both atmosphere and ocean, the loading vectors show a simple mode 1 pattern with similar amplitudes. This implies that PLS-SVD regression based on the ocean latent variables will be robust. The right panels show the proportion of the trace of $\boldsymbol{\Sigma}_{11}$ and $\boldsymbol{\Sigma}_{22}$ that can be accounted for regression of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ on an increasing number of ocean latent variables. In the right upper panel, it can be seen that the first latent variable is a poor predictor of the atmospheric state and the increase in prediction skill is quite slow as more latent variables are added to the regression model. The right lower panel, on the other hand, shows the proportion of total variance of the ocean state vector $\boldsymbol{X}_2$ that can be accounted for regression on the ocean latent variables. PLS performs very well from this perspective, and almost all of the total variance of $\boldsymbol{X}_2$ is accounted for by the first three latent variables.

Differences in the performance of PLS-based regression and RA are further highlighted by Figures 6.7 and Figure 6.8. Both figures have the same format to aid comparison. As expected, RA clearly places more emphasis on prediction whereas

PLS balances prediction against an ability to recover the predictor variability from the selected latent variables.

6.5 Hybrid Approach

The proportions of the total variance of $\boldsymbol{\Sigma}_{11}$ and $\boldsymbol{\Sigma}_{22}$ accounted for by regression on the predictor latent variables from the hybrid approach are shown in Figure 6.9. The proportions of explained variance are plotted as a function of α in separate panels. The lines in each panel correspond to the number of predictor latent variables included in the regression. The upper panel shows that as α increases from 0 to 1, corresponding to a gradual shift from RA to PLS, the proportion of explained tr $\boldsymbol{\Sigma}_{11}$ decreases; the lower panel shows the opposite behavior with the proportion of explained tr $\boldsymbol{\Sigma}_{22}$ increasing with α. These results are in accord with the discussion in Chapter 4 and clearly illustrate the trade off between predictive skill and stability of the predictors that can be controlled with α.

The sensitivity of the loading vectors, and the proportion of explained variance, is shown in Figure 6.10. Each row corresponds to a different value of α. The first column of panels shows the first two loading vectors for $\boldsymbol{X}_1$ calculated by PLS-SVD as a function of sector. The second column shows the corresponding first two loading vectors for $\boldsymbol{X}_2$. The third and fourth columns show the proportion of total variance of tr $\boldsymbol{\Sigma}_{11}$ and tr $\boldsymbol{\Sigma}_{22}$ accounted for by k predictor latent variables as a function of k. It is clear that as α increases, the predictor loading vectors become smoother and eventually take on a mode 1 pattern. This shows that CR-PLS becomes more stable with increasing α. The third and fourth columns show how the proportion of explained variance for $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ increase with the number of predictor latent variables.

6.6 Extension in Frequency Domain

The extension of several of the dimension reduction methods to the frequency domain is illustrated in this section. As discussed in Chapter 5, the computation of latent variables can be readily applied in the frequency domain by replacing the covariance matrices by the corresponding cross spectral matrices.

Principal Component Analysis: The proportion of variance accounted for by the first principal component is shown as a function of frequency in Figure 6.11. The analysis is based on the coherency matrix, $\boldsymbol{W}$. This principal component accounts for a relative high proportion of the power spectral density (almost 0.9) at low frequencies. The proportion reaches a minimum at a frequency of about 1.5×10^{-3} cycles per unit time, corresponding a minimum in the power spectral density of the ocean and atmosphere (see Figure 6.11).

The scree plot for a frequency of 0.01 (corresponding to a period of 100 time steps) shows that almost all of the power is captured by the first three principal components (Figure 6.12). The remaining principal components have relative small variances. This implies $\boldsymbol{W}$ is ill-conditioned as expected.

The first four loading vectors computed by PCA (i.e. the eigenvectors of $\boldsymbol{W}$, $\boldsymbol{e}_1$ to $\boldsymbol{e}_4$) at the same frequency of 0.01 are plotted in Figure 6.13. The real parts and imaginary parts are plotted separately. Each column of panels corresponds to a particular eigenvector; the atmosphere and ocean subvectors are plotted in the upper and lower panels respectively. For $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ the atmosphere atmosphere and ocean loadings each have simple mode 1 pattern with similar amplitude. The situation is similar for $\boldsymbol{e}_3$ and $\boldsymbol{e}_4$ except that the ocean amplitudes are small and also have a relatively strong contribution from the mean. Higher order eigenvectors (not shown) exhibit more spatial variability (e.g., mode 2 patterns).

Partial Least Squares: The total power of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ (i.e., the trace of $\boldsymbol{h}_{11}$ and $\boldsymbol{h}_{22}$) as a function of frequency are shown by the black lines in the upper and lower panels of Figure 6.14. The peaks in the power spectra are in agreement with the spectra shown and discussed in Chapter 2. The remaining lines in each panel show the prediction of the total power using the predictor latent variables from frequency dependent PLS-SVD. Note that as the number of predictors increases so does the proportion of explained power. When all N predictors are used, all of the power is recovered (i.e., multiple squared coherence is 1). It can also be seen that $\mathrm{tr}\boldsymbol{h}_{22}$ is accounted for with far fewer predictors than $\mathrm{tr}\boldsymbol{h}_{11}$. This is in accord with the above covariance-based discussion of the way PLS balances prediction skill and ability to explain variability in the predictors.

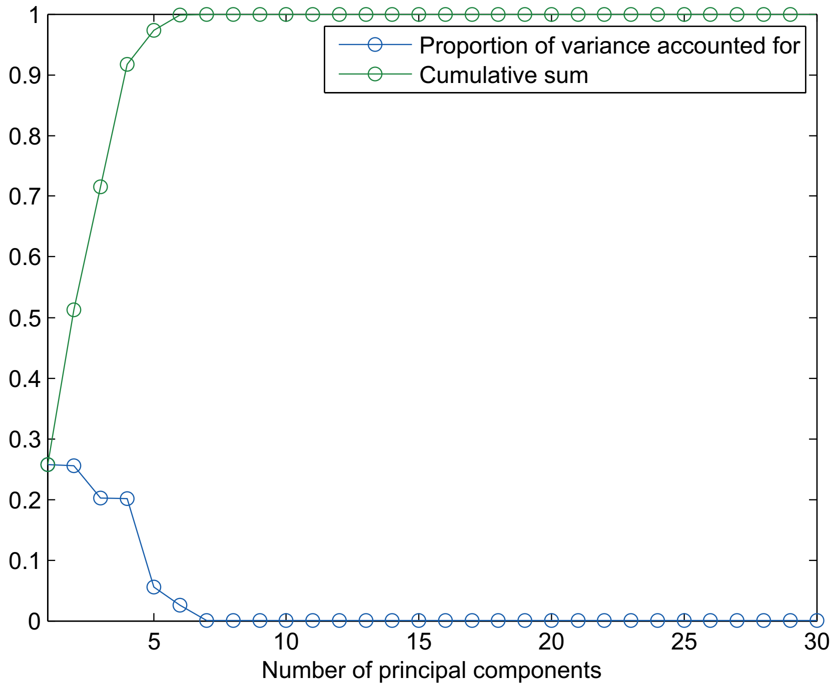

Figure 6.1: A “scree plot” for the principal component analysis of the atmosphere-ocean state vector based on the correlation matrix, $\boldsymbol{R}$. The blue line shows the proportion of total variance account for by the kth principal component. The green line shows the cumulative proportion of variance account for by the first k principal components.

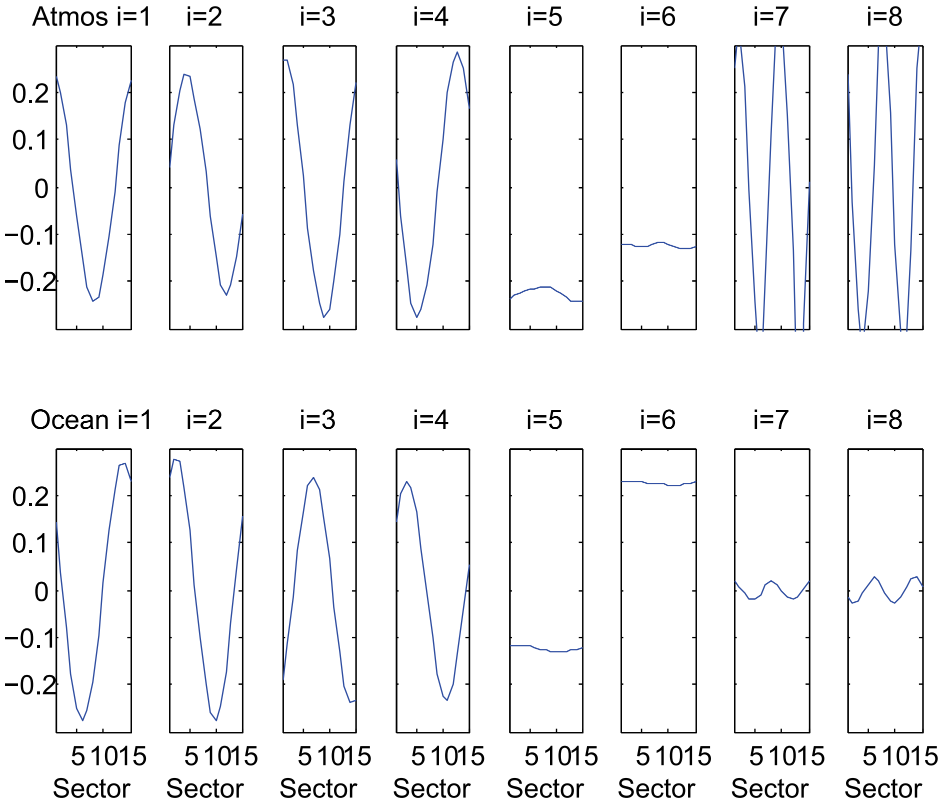

Figure 6.2: Loading vectors for the first eight principal components from a PCA of the atmosphere-ocean state vector based on the correlation matrix, $\boldsymbol{R}$. For each panel the x-axis corresponds to sector (i.e., position). The upper panels are the loading subvectors for the atmosphere and the lower panels are the loading subvectors for the ocean. The ith principal component corresponds to the the ith column of panels.

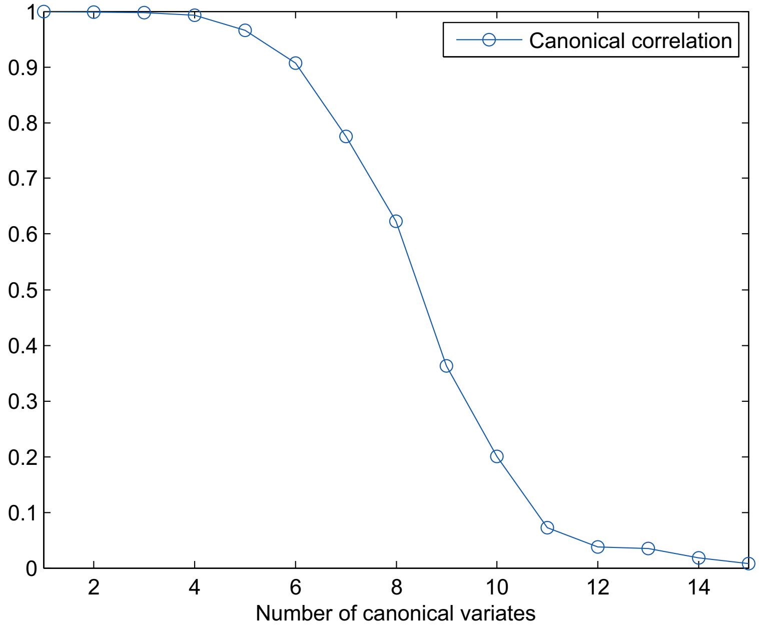

Figure 6.3: Canonical correlations based on a CCA of the state vector from the coupled atmosphere-ocean model.

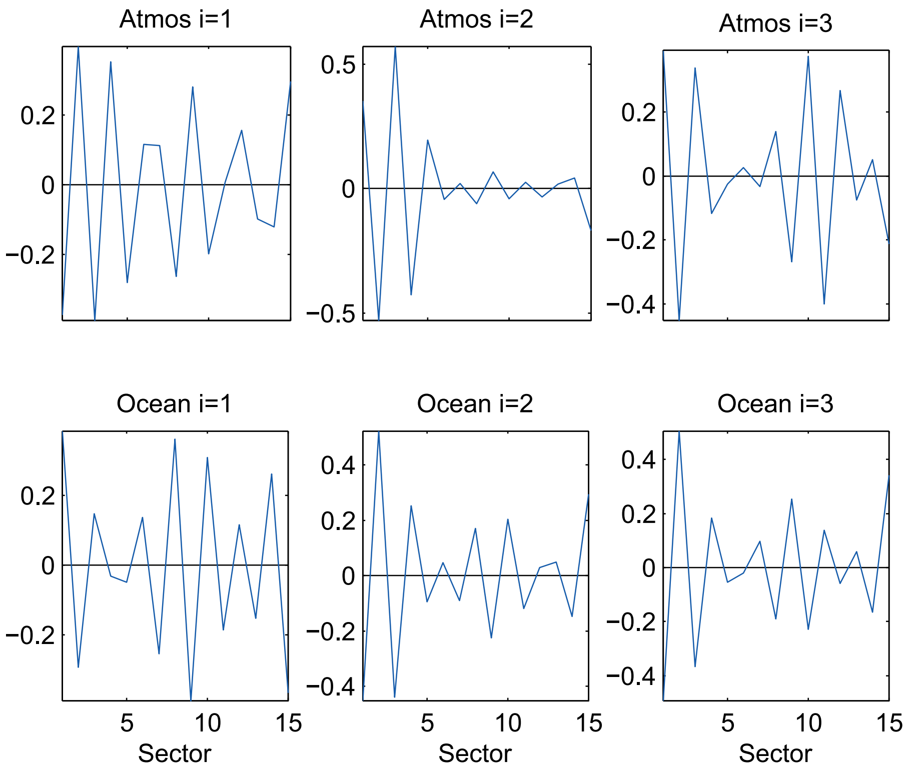

Figure 6.4: Loading vectors for the first three canonical variates from a CCA of the state vector from the coupled atmosphere-ocean model. For each panel the x-axis corresponds to sector (i.e., position). The upper panels are the loading subvectors for the atmosphere and the lower panels are the loading subvectors for the ocean. The ith loading vectors corresponds to the the ith column of panels.

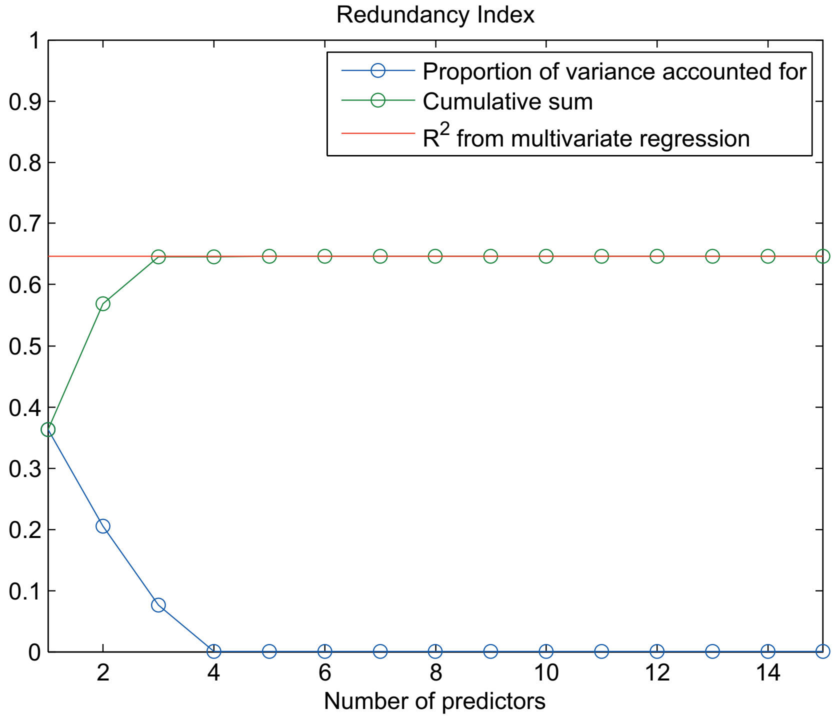

Figure 6.5: Proportion of the total variance of $\boldsymbol{X}_1$ predicted by the latent variables of $\boldsymbol{X}_2$ from RA. The blue line shows the proportion of total variance account for by the kth latent variable. The green line shows the cumulative proportion of variance account for by the first k latent variables (i.e., the Redundancy Index, see (3.7)).

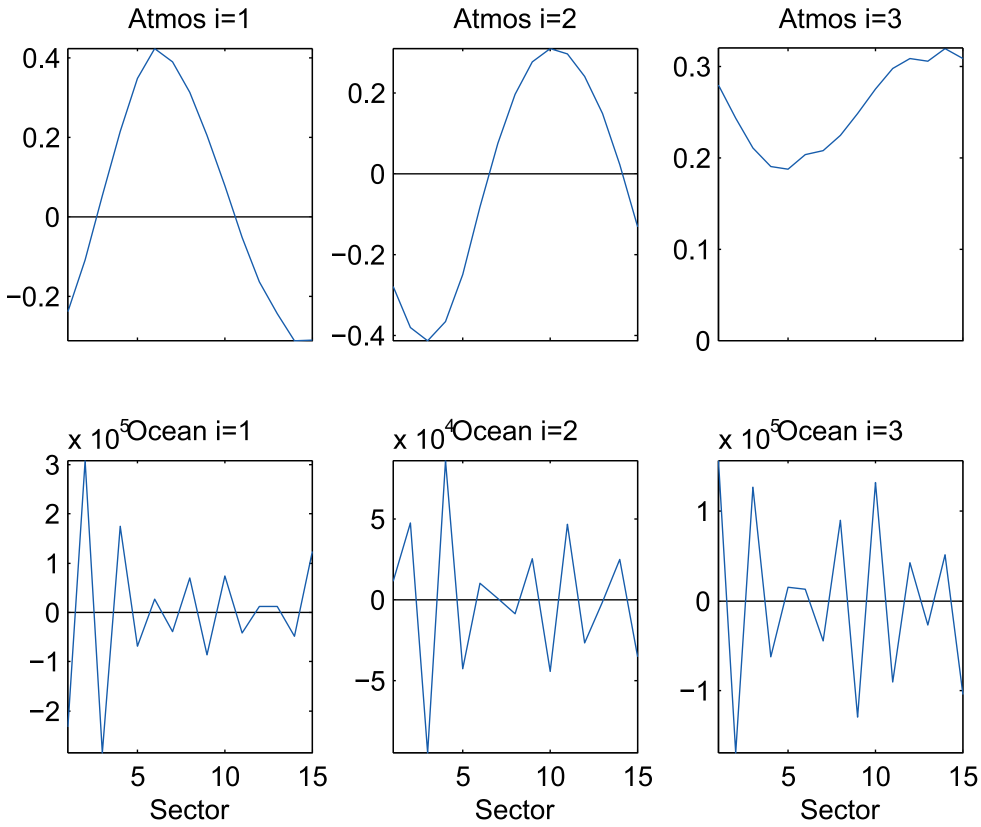

Figure 6.6: The loading vectors for the first three latent variables from a RA of the atmosphere-ocean state vector. For each panel the x axis corresponds to position (i.e., sector). The upper panels are the loading vectors for the atmosphere and the lower panels are the loading vectors for the ocean. Each column of panels corresponds to a specific principal component (denoted by i in the title of each panel).

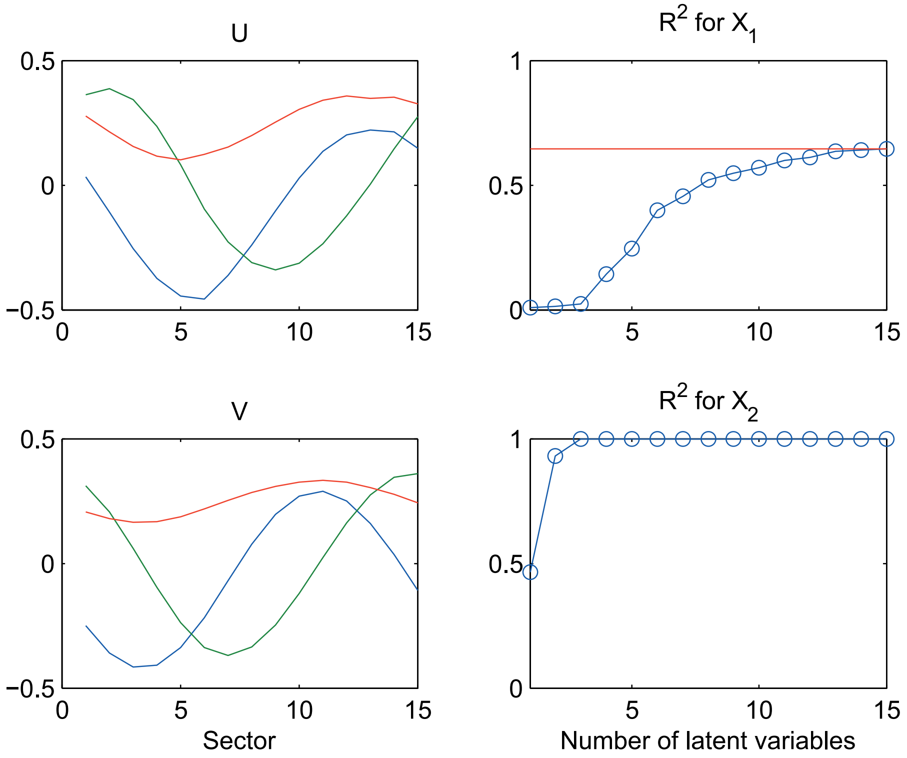

Figure 6.7: Loading vectors and proportion of variance explained by latent variables from a PLS-SVD analysis of the state vector from the coupled atmosphere-ocean model. The left panels show the first three loading vectors for the atmosphere (upper panel) and ocean (lower panel) as a function of sector (i.e., position). The right panels show the proportion of the total variance of the atmosphere (upper) and ocean (lower) explained by the first k predictor latent variables as a function of k.

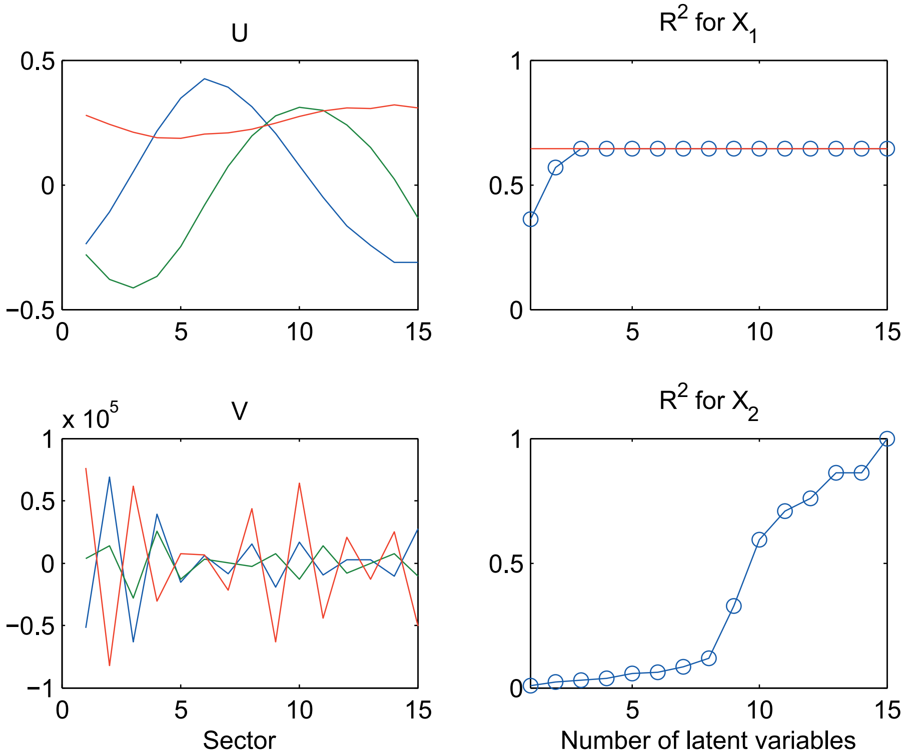

Figure 6.8: Loading vectors and proportion of variance explained by latent variables from an RA of the state vector from the coupled atmosphere-ocean model. Same format as Figure 6.7.

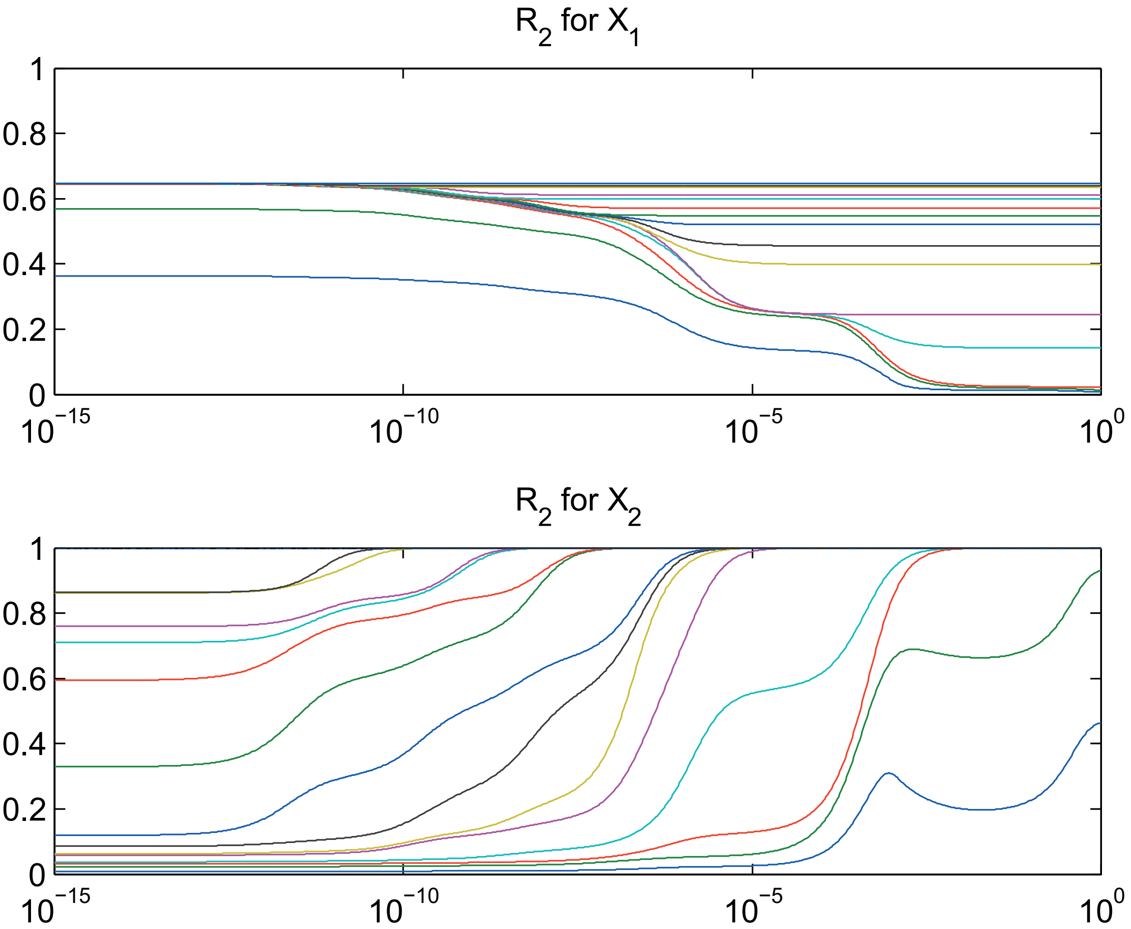

Figure 6.9: Proportion of the total variance of the atmosphere (upper panel) and ocean (lower panel) accounted for by CR-PLS, the hybrid of RA and PLS. The x-axis corresponds to α which ranges from 0 (RA) to 1 (PLS). The lines in each panel correspond to different numbers of predictor latent variables.

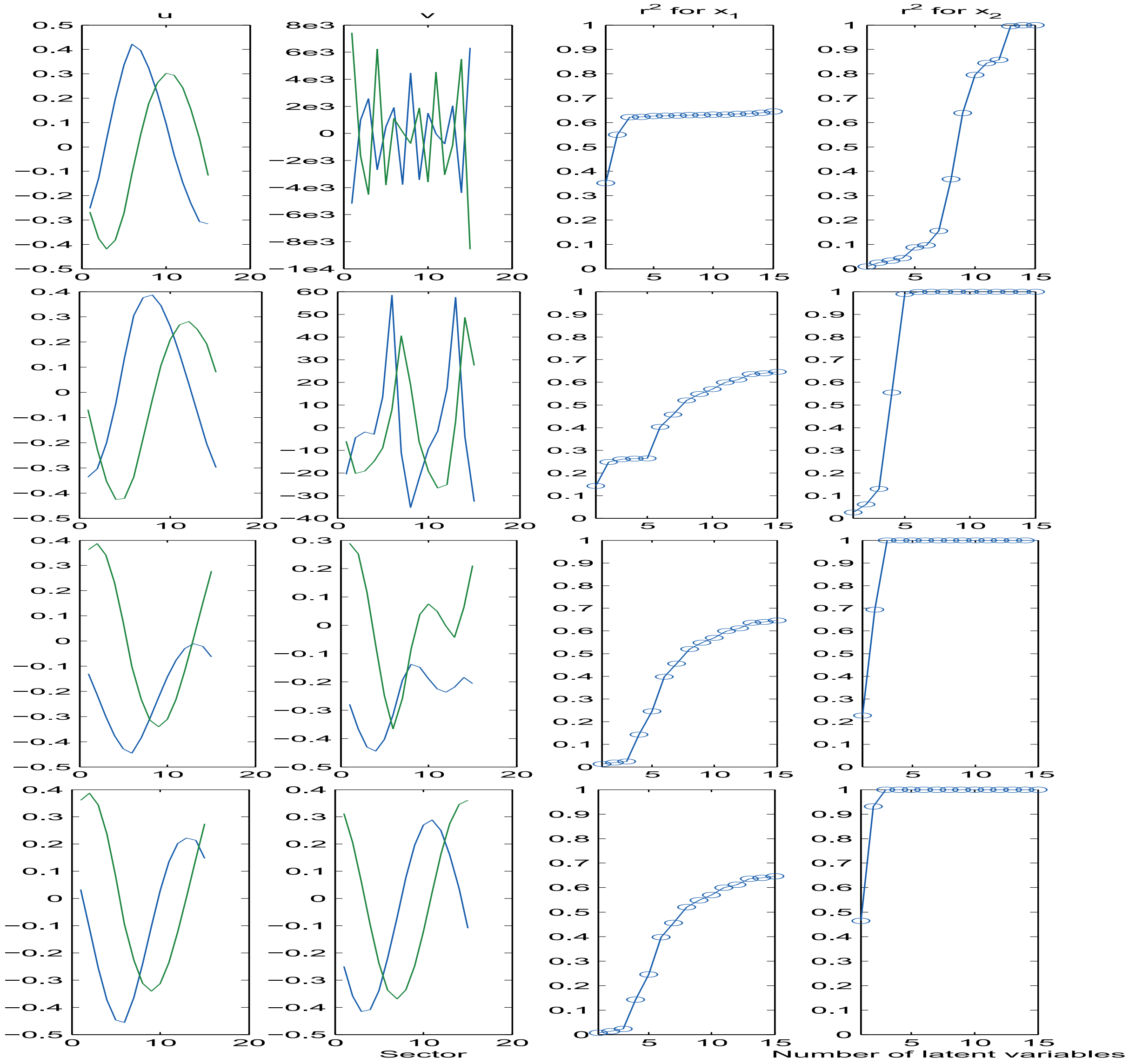

Figure 6.10: The sensitivity of the loading vectors, and the proportion of explained variance. Each row corresponds to a different value of α ($\alpha = 10^{-10}, 10^{-5}, 10^{-1}$ and 1 respectively). The first column of panels shows the first two loading vectors for $\boldsymbol{X}_1$ calculated by PLS-SVD as a function of sector. The second column shows the corresponding first two loading vectors for $\boldsymbol{X}_2$. The third and fourth columns show the proportion of total variance of tr $\boldsymbol{\Sigma}_{11}$ and tr $\boldsymbol{\Sigma}_{22}$ accounted for by k predictor latent variables as a function of k.

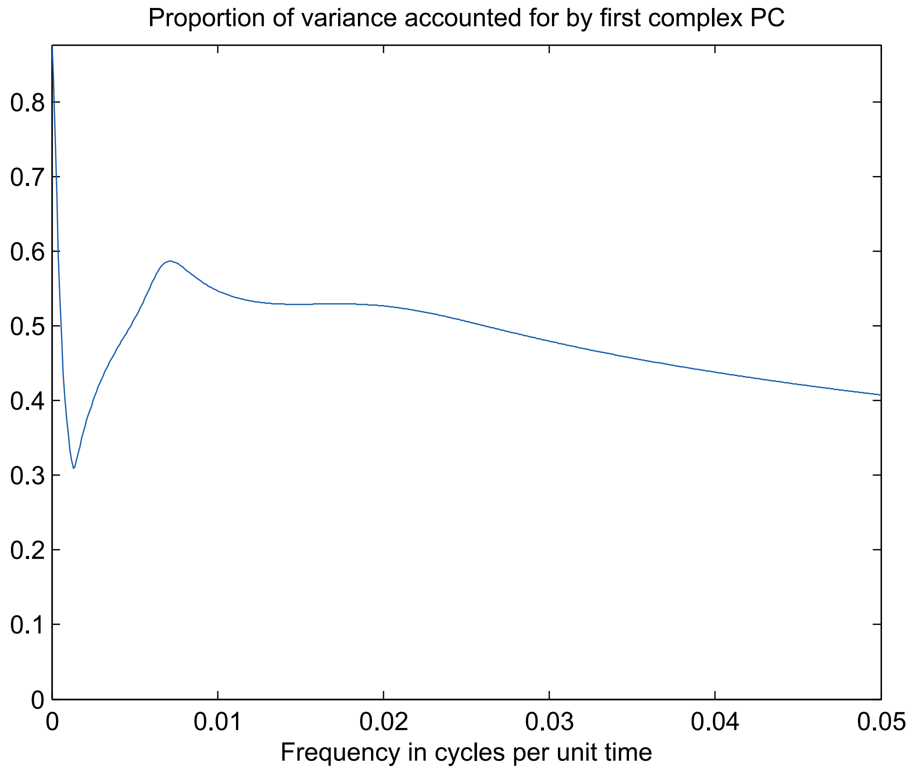

Figure 6.11: Proportion of total power spectral density of the state vector from the coupled atmosphere-ocean model accounted for by first frequency dependent principal component. The coherency matrix ($\boldsymbol{W}$, see Chapter 2) was used.

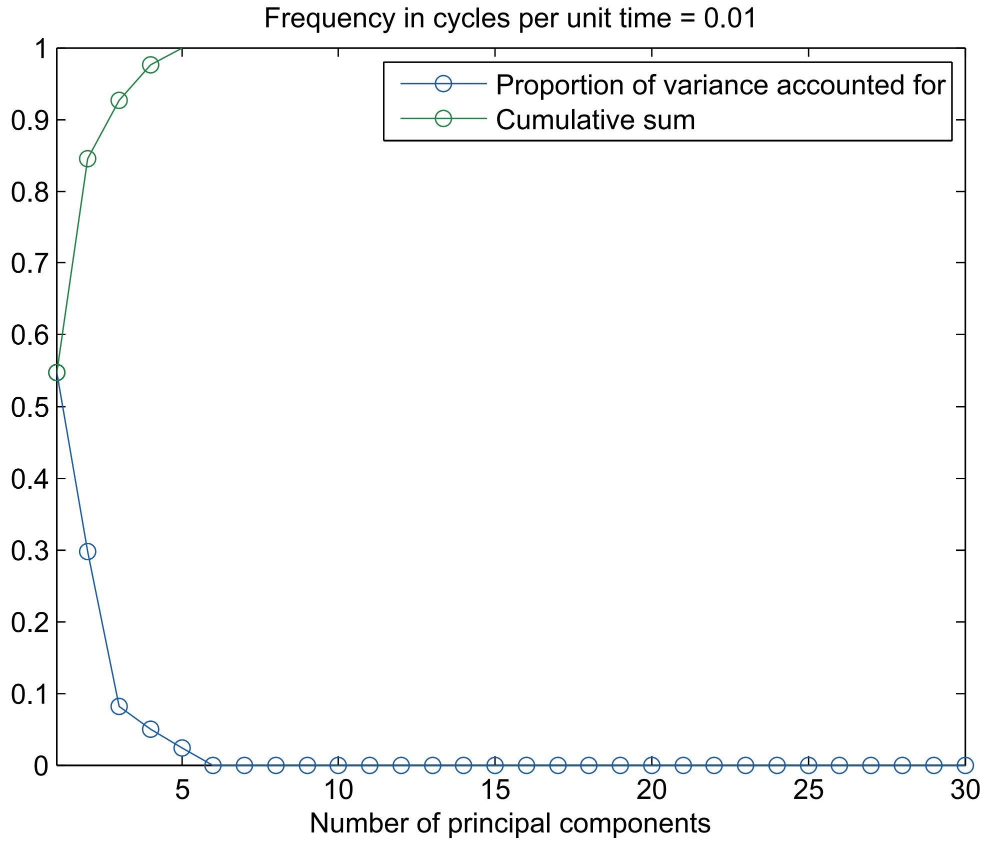

Figure 6.12: Scree plot for the frequency-dependent principal component analysis of the atmosphere-ocean state for a frequency of 0.01 cycles per unit time. The coherency matrix, $\boldsymbol{W}$, was used. The format is the same as Figure 6.1.

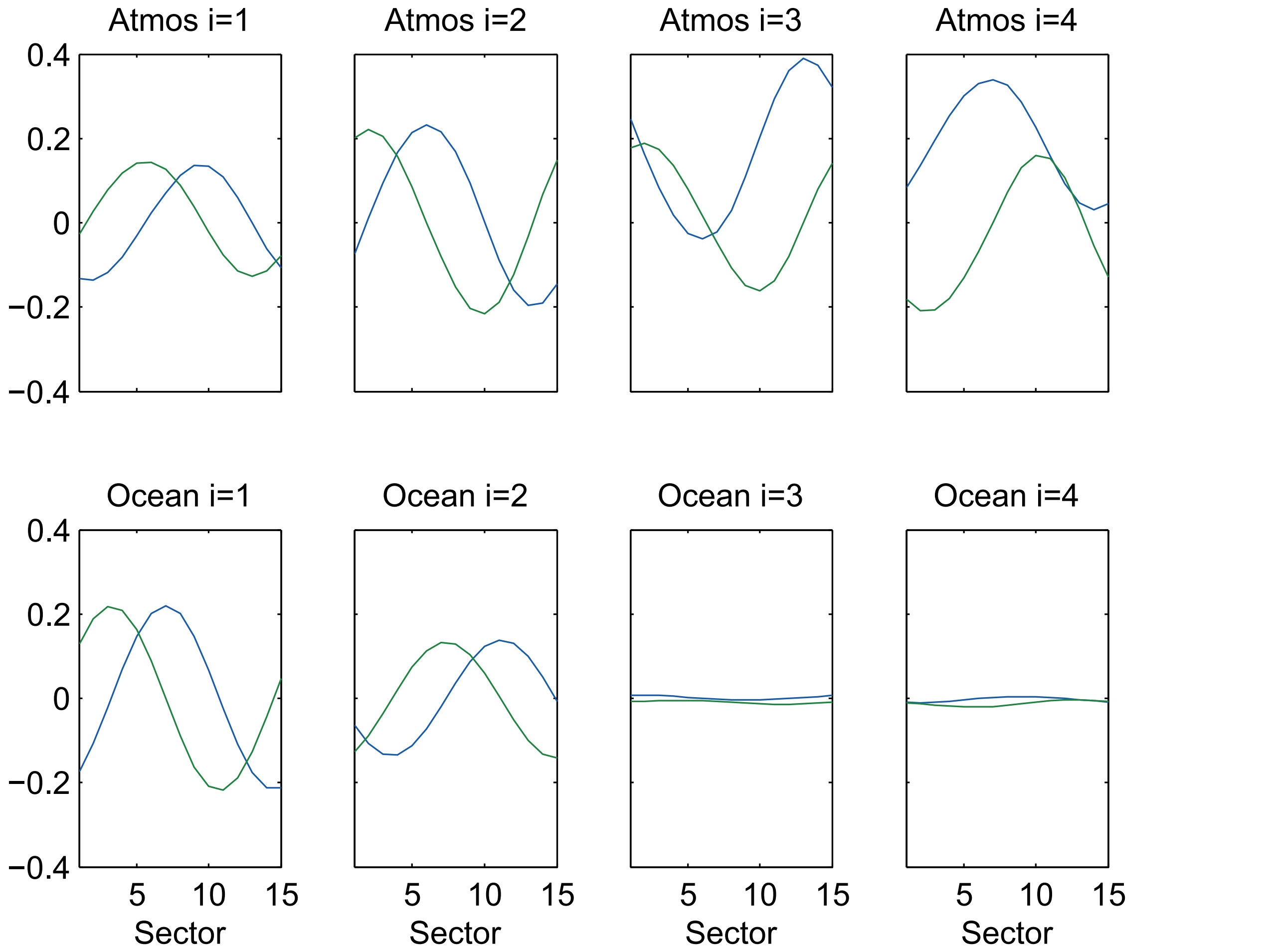

Figure 6.13: Loading vectors for the first four principal components from a frequency-dependent PCA of the atmosphere-ocean state vector. The coherency matrix, $\boldsymbol{W}$, was used. The format is the same as Figure 6.2 expect that there are two curves for each panel corresponding the real (blue) and imaginary (green) parts of each eigenvector.

ht

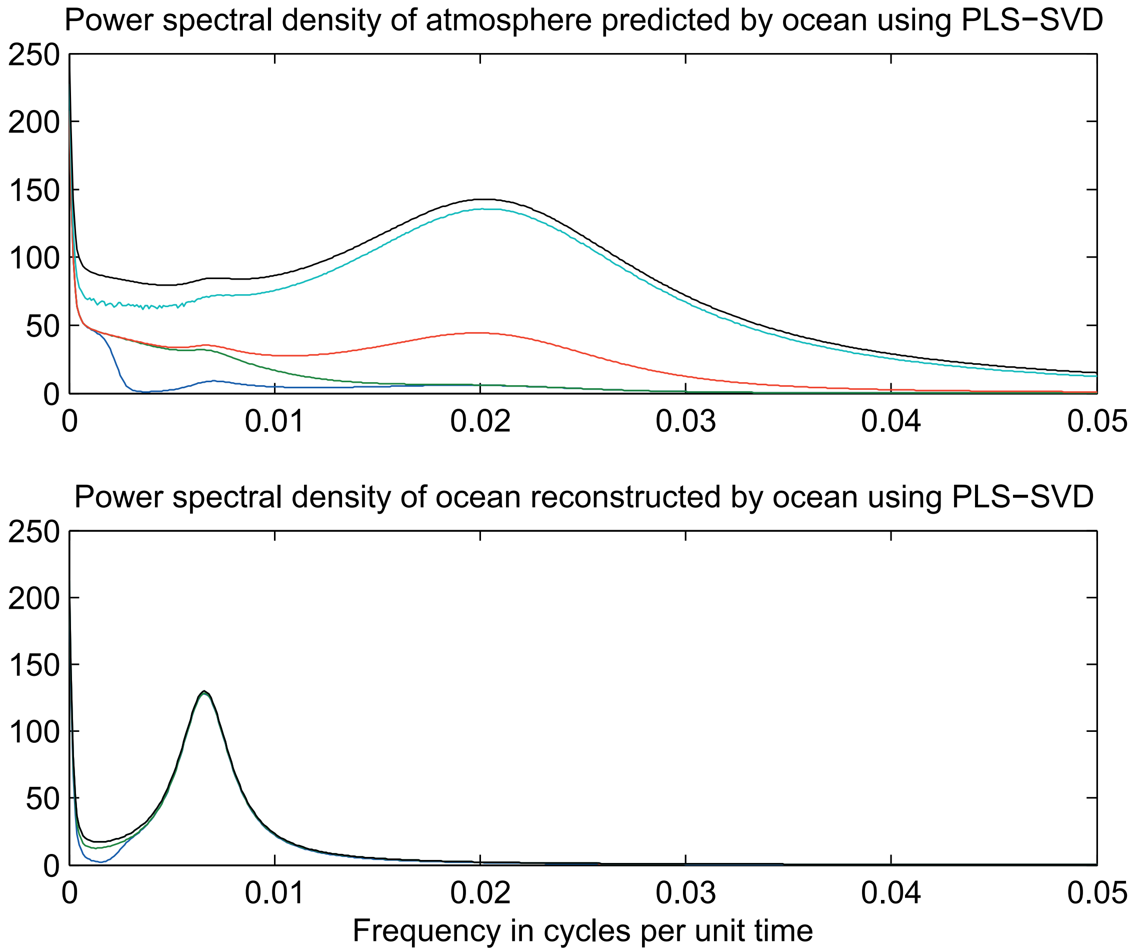

Figure 6.14: The spectral power density of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ as a function of frequency. The black lines in the upper and lower panels give the total power of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ respectively. The remaining lines in each panel show the prediction of the total power using the predictor latent variables from frequency dependent PLS-SVD.

Chapter 7

Summary and Discussion

Several dimension reduction methods have been reviewed in this study including principal component analysis, canonical correlation analysis, redundancy analysis and partial least squares. The focus of this study is the prediction of a random response vector ($\boldsymbol{X}_1$) from a random predictor vector ($\boldsymbol{X}_2$). Two variants of partial least squares (PLS-SVD and PLS-W2A), and also a hybrid method that blends redundancy analysis and partial least squares, were also reviewed.

One of the novel features of this study is the extension of some of the dimension reduction techniques to the frequency domain based on the spectral representation of stationary random processes. The corresponding results are explicit and similar to those obtained with standard techniques; the frequency dependent generalizations were obtained by replacing covariance matrices by cross spectral matrices. By way of illustration, it was shown in Chapter 5 that the extension of PCA to the frequency domain lead to new insights into the covariation of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$; not only was it shown that the proportion of variance accounted for by the first principal component changed with frequency but so did the loading vector that defined the first principal componnet.

Although CCA, PCA and PLS are easy to understand and powerful in terms of defining latent variables, they do not perform efficiently in terms of prediction. Redundancy Analysis on the other hand is derived directly from multivariate regression and naturally has a better ability to predict. It was shown however that the performance of RA is seriously impaired by poor conditioning of $\boldsymbol{\Sigma}_{22}$ and that under such circumstances RA is not robust. Noting that PLS is not affected by ill-conditioned $\boldsymbol{\Sigma}_{22}$, we followed S. Bougeard et al. (2007) and constructed a blend of RA and PLS that trades off prediction skill against robustness of the prediction model.

Another feature of the present study is the use of idealized models to illustrate the strengths and weaknesses of the various methods. The main advantage of the use of such models is that explicit covariance and cross spectral matrices can be defined leading to results that are not subject to sampling variability. Extensive use was made in Chapter 6 of an idealized coupled atmosphere-ocean model. It is important to note however that the methods described in this study have applicability beyond just climate prediction. The results shown in Chapter 6 are generally in accord with the corresponding theoretical conclusion e.g., the loading vectors from PCA, CCA and RA show irregular patterns when $\mathbf{\Sigma}_{22}$ is poorly conditioned, leading to results that are unstable and difficult to interpret. PLS on the other hand was shown to give stable results but poor prediction skill using a small number of latent variables. It was also shown that the hybrid approach provided the required trade off between prediction skill and robustness, and that the generalization to frequency dependence of PCA was useful.

As to future work, we note that the hybrid method is determined by its two end members; in the present case they are RA ($\alpha = 0$) and PLS ($\alpha = 1$). The choice of RA as an end member is quite reasonable because it is based on maximizing prediction skill. The choice of PLS as the other end member is less clear and one could argue that it is somewhat arbitrary in terms of maximizing robustness and ability to identify physically meaningful latent variables. A useful avenue for future work would be to construct a hybrid based on a Bayesian perspective and the introduction of additional prior information on the structure of the loading vectors.

Appendix A

Canonical Correlation Analysis

Proof. For any pair of a weight vectors, say $\mathbf{a}$ and $\mathbf{b}$, the correlation between the latent variables $\boldsymbol{a}^T\boldsymbol{X}_1$ and $\boldsymbol{b}^T\boldsymbol{X}_2$ is

$$\mathrm{Corr}(\boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}_1, \boldsymbol{b}^{\mathrm{T}}\boldsymbol{X}_2) = \frac{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{\Sigma}_{12}\boldsymbol{b}}{\sqrt{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{\Sigma}_{11}\boldsymbol{a}}\sqrt{\boldsymbol{b}^{\mathrm{T}}\boldsymbol{\Sigma}_{22}\boldsymbol{b}}} \tag{A.1}$$

If we set $\mathbf{c} = \boldsymbol{\Sigma}_{11}^{\frac{1}{2}}\mathbf{a}$ and $\mathbf{d} = \boldsymbol{\Sigma}_{22}^{\frac{1}{2}}\mathbf{b}$, then (A.1) becomes

$$\mathrm{Corr}(\mathbf{a}^{\mathrm{T}}\boldsymbol{X}_1, \mathbf{b}^{\mathrm{T}}\boldsymbol{X}_2) = \frac{\mathbf{c}^{\mathrm{T}}\boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-\frac{1}{2}}\mathbf{d}}{\sqrt{\mathbf{c}^{\mathrm{T}}\mathbf{c}}\sqrt{\mathbf{d}^{\mathrm{T}}\mathbf{d}}}$$

Using the extended Cauchy-Schwarz inequality, it can be shown

$$\begin{aligned}\mathrm{Corr}(\mathbf{a}^{\mathrm{T}}\boldsymbol{X}_1, \mathbf{b}^{\mathrm{T}}\boldsymbol{X}_2) &\leq \frac{\mathbf{c}^{T}\boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}\mathbf{c}}{\sqrt{\mathbf{c}^{T}\mathbf{c}}} \\ &\leq \sqrt{\lambda_1}\end{aligned}$$

where λ_1 is the largest eigenvalue of $\boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}$. Equivalently, $\sqrt{\lambda_1}$ is the largest singular value of $\boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-\frac{1}{2}}$ and $\mathbf{c}$ and $\mathbf{d}$ are selected as the first pair of eigenvectors of $\boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-\frac{1}{2}}$, say $\boldsymbol{u}_1$ and $\boldsymbol{v}_1$. If we take $\mathbf{a} = \boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}\boldsymbol{u}_1$ and $\mathbf{b} = \boldsymbol{\Sigma}_{22}^{-\frac{1}{2}}\boldsymbol{v}_1$, then (A.1) reaches the maximum value.

Similarly, the k^{th} pair of weight vectors, whose corresponding latent variables are uncorrelated with preceding latent variables, are given by

$$\mathbf{a}_k = \boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}\boldsymbol{u}_k, \qquad \mathbf{b}_k = \boldsymbol{\Sigma}_{22}^{-\frac{1}{2}}\boldsymbol{v}_k$$

where $\boldsymbol{u}_k$ and $\boldsymbol{v}_k$ are the k^{th} pair of eigenvectors of $\boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-\frac{1}{2}}$.

Appendix B

Comparison of PLS-SVD and PLS-W2A

As introduced in Chapter 3, there is no difference between PLS-SVD and PLS-W2A in the process of calculating the first pair of latent variables. Two methods give the same patterns of $\boldsymbol{u}_1$ and $\boldsymbol{v}_1$. For the subsequent latent variables, PLS-SVD computes loading vectors as the corresponding singular vectors of $\boldsymbol{\Sigma}_{12}$. On the other hand, this is not true for PLS-W2A in general.

In some particular situation, however, the outcomes generated by the PLS-SVD algorithm are equivalent to those from the PLS-W2A algorithm. Consider the case where the first eigen-vector of $\boldsymbol{\Sigma}_{11}^{(r)}$ is proportional to $\boldsymbol{u}_1^{(r)}$ and the first eigen-vector of $\boldsymbol{\Sigma}_{22}^{(r)}$ is proportional to $\boldsymbol{v}_1^{(r)}$. According to PLS-W2A, $\boldsymbol{\Sigma}_{12}^{(r)}$ is simplified into:

$$\lambda_1^{(r)}\left[\boldsymbol{Q}_1^{(r)}\boldsymbol{u}_1^{(r)}\boldsymbol{v}_1^{(r)T}+\boldsymbol{u}_1^{(r)}\boldsymbol{v}_1^{(r)T}\boldsymbol{Q}_2^{(r)T}-\boldsymbol{Q}_1^{(r)}\boldsymbol{u}_1^{(r)}\boldsymbol{v}_1^{(r)T}\boldsymbol{Q}_2^{(r)T}\right]=\lambda_1^{(r)}\boldsymbol{u}_1^{(r)}\boldsymbol{v}_1^{(r)T}$$

which has the same form of deflation with the one in PLS-SVD.

The difference and similarity between PLS-SVD and PLS-W2A can be also studied from the geometrical angle. For simplicity, we focus on a two dimensional example. The spectral representation of $\boldsymbol{\Sigma}_{22}$ is

$$\boldsymbol{\Sigma}_{22}=\mu_1\boldsymbol{e}_1\boldsymbol{e}_1^T+\mu_2\boldsymbol{e}_2\boldsymbol{e}_2^T,$$

where μ_1 is the largest eigenvalue of $\boldsymbol{\Sigma}_{22}$, μ_2 is the smallest one. $\boldsymbol{e}_1$, $\boldsymbol{e}_2$ are the corresponding eigenvectors. And rewrite the fifth step in the algorithm of PLS-W2A by replacing $\boldsymbol{\Sigma}_{22}$ with its spectral representation form. The equation becomes

$$\hat{\boldsymbol{X}}_2=\frac{1}{\mu_1\cos(\theta_1)^2+\mu_2\cos(\theta_2)^2}(\mu_1\cos(\theta_1)\boldsymbol{e}_1+\mu_2\cos(\theta_2)\boldsymbol{e}_2)\boldsymbol{v}_1^T\boldsymbol{X}_2, \tag{B.1}$$

where θ_1 and θ_2 are the angles between $\boldsymbol{e}_1$ and $\boldsymbol{v}_1$, $\boldsymbol{e}_2$ and $\boldsymbol{v}_1$ respectively.

Geometrically, (B.1) means that random vector $\boldsymbol{X}_2$ is projected down onto $\boldsymbol{v}_1$ and then re-expanded back through the vector $\mathbf{w}_1$, where

$$\mathbf{w}_1 = \frac{1}{\mu_1 \cos(\theta_1)^2 + \mu_2 \cos(\theta_2)^2}[\mu_1 \cos(\theta_1)\boldsymbol{e}_1 + \mu_2 \cos(\theta_2)\boldsymbol{e}_2].$$

From the equation above, we can see $\mathbf{w}_1$ is determined by the eigen-structure of $\boldsymbol{\Sigma}_{22}$.

It is easy to prove that for fixed vector $\hat{\boldsymbol{X}}_2$ and $\boldsymbol{v}_1 = [1, 0]^T$, the polar coordinates for $\mathbf{w}_1$ according to different 'shapes' of $\boldsymbol{\Sigma}_{22}$, are given explicitly by

$$\left[\frac{a}{\mu_1 \cos(\theta_1)^2 + \mu_2 \cos(\theta_2)^2}(\mu_1^2 \cos(\theta_1)^2 + \mu_2^2 \cos(\theta_2)^2)^{\frac{1}{2}}, \quad \arctan(\frac{\mu_2}{\mu_1} \tan \theta_1)\right] \tag{B.2}$$

where $a = |\boldsymbol{X}_2| \boldsymbol{v}_1^T \boldsymbol{X}_2$.

On the other hand, if we just re-expand $\boldsymbol{v}_1^T \boldsymbol{X}_2$ back across $\boldsymbol{v}_1$ without the influence of $\boldsymbol{\Sigma}_{22}$, the approximation of $\hat{\boldsymbol{X}}_2$ becomes

$$\hat{\boldsymbol{X}}_2 = \boldsymbol{v}_1 \boldsymbol{v}_1^T \boldsymbol{X}_2. \tag{B.3}$$

This result is equivalent to re-expanding $\boldsymbol{v}_1^T \boldsymbol{X}_2$ using the PLS-SVD algorithm.

In summary, in order to compute the one dimensional approximation of $\boldsymbol{X}_2$, both PLS-SVD and PLS-W2A project $\boldsymbol{X}_2$ on the loading vector $\boldsymbol{v}_1$ at first. The difference appears in the process of re-expanding it back: with PLS-SVD, $\boldsymbol{v}_1^T \boldsymbol{X}_2$ is re-expanded through the direction across $\boldsymbol{v}_1$; by PLS-W2A, it is re-expanded back through the direction across $\mathbf{w}_1$ which is determined by the eigen-structure of $\boldsymbol{\Sigma}_{22}$. Note, when $\boldsymbol{v}_1$ and $\boldsymbol{e}_1$ are parallel, that is $\theta_1 = 0$ or $\theta_1 = \pi$, by $(B.2)$ the angle between $\hat{\boldsymbol{X}}_2$ and $\boldsymbol{v}_1$ is zero. It means that PLS-W2A method also re-expand $\boldsymbol{v}_1^T \boldsymbol{X}_2$ across $\boldsymbol{v}_1$ direction. Thus, PLS-SVD and PLS-W2A give the exactly same one dimensional approximation of $\boldsymbol{X}_2$ when $\boldsymbol{e}_1 = \pm \boldsymbol{v}_1$. Recall the situation discussed at the beginning part in this section; we come to the same conclusion geometrically.

PLS only uses one step of singular decomposition to find all the latent variables. This makes mathematical analysis much simpler to understand and express. This property also allows PLS-SVD to be more time saving when dealing with large data sets. On the other hand, PLS-W2A has to perform a singular decomposition of updated covariance matrix for every iteration of the loop.

It is important to note that the property of full rank in PLS-SVD regression is requiring the number of selected latent variables is smaller or equivalent to the rank of $\mathbf{\Sigma}_{12}$. If this is not true, PLS-SVD regression will perform poorly. According to PLS-SVD, if we select more pairs of latent variables than the rank of $\mathbf{\Sigma}_{12}$, their covariance will be zero. That means those extra latent variables would contribute nothing to reconstruction of Σ_{12}.

To show this, assume $\mathbf{\Sigma}_{12}$ has rank one and $\mathbf{\Sigma}_{22}$ has full rank. PLS-SVD latent variables are selected essentially by the eigenvector of $\mathbf{\Sigma}_{21}\mathbf{\Sigma}_{12}$ whose rank is one as well. Thus, there is no useful information about covariance of the remaining latent variables. The regression of $\boldsymbol{X}_1$ on $\boldsymbol{X}_2$ based on one pair of latent variables is

$$\hat{\boldsymbol{X}}_1^{(1)} = \lambda_1 \boldsymbol{u}_1 \frac{1}{a} \boldsymbol{v}_1^T \boldsymbol{X}_2,$$

and the regression of $\boldsymbol{X}_1$ the two pairs of latent variables is

$$\hat{\boldsymbol{X}_1}^{(2)} = \begin{bmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & 0 \end{bmatrix} [\begin{bmatrix} \boldsymbol{v}_1^T \\ \boldsymbol{v}_2^T \end{bmatrix} \mathbf{\Sigma}_{22} \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 \end{bmatrix}]^{-1} \begin{bmatrix} \boldsymbol{v}_1^T \\ \boldsymbol{v}_2^T \end{bmatrix} \boldsymbol{X}_2.$$

This equation can be simplified to give

$$\hat{\boldsymbol{X}_1}^{(2)} = \lambda_1 \boldsymbol{u}_1 [\frac{1}{ad - b^2} (d\boldsymbol{v}_1 - b\boldsymbol{v}_2)^T] \boldsymbol{X}_2. \tag{B.4}$$

where $a = \boldsymbol{v}_1^T \mathbf{\Sigma}_{22} \boldsymbol{v}_1$, $b = \boldsymbol{v}_1^T \mathbf{\Sigma}_{22} \boldsymbol{v}_2 = \boldsymbol{v}_2^T \mathbf{\Sigma}_{22} \boldsymbol{v}_1$ and $d = \boldsymbol{v}_2^T \mathbf{\Sigma}_{22} \boldsymbol{v}_2$.

In this example, when we choose two pairs of latent variables, the expression for $\hat{\boldsymbol{X}}_1$ finally degenerates to using one pair of latent variables. The reason is rank-deficient $\mathbf{\Sigma}_{12}$ leads to a diagonal matrix Λ which is rank-deficient and vanishes the loading vector of the second pair of latent variables. (B.4) basically chooses one pair of variables but not with the largest covariance. Thus, this example illustrates PLS-regression's drawback under the condition that the number of selected pairs of latent variables is larger than the rank of $\mathbf{\Sigma}_{12}$. But in PLS-W2A, it is possible to have nonzero covariance for all the latent variables, even when $\mathbf{\Sigma}_{12}$ has a low rank.

Appendix C

Implementing CR-PLS

A hybrid method, Continuum Redundancy-Partial Least Squares, has been introduced in chapter 4 as a mean of relating PLS and RA based on the eigen-structure of various covariance matrices. Now, we will implement this approach specifically on those two specific variants of PLS (PLS-SVD and PLS-W2A) respectively.

As discussed in the latter part in Chapter 3, the essential difference between PLS-W2A and PLS-SVD is the way they use to deflate the covariance matrix. The PLS-SVD approach deflates $\mathbf{\Sigma}_{12}$ once and for all by calculating the singular decomposition of the original covariance matrix instead of iterative computation. In the hybrid case, where PLS-SVD and RA are combined together, this attractive property has been retained.

Combining RA and PLS-SVD:

For convenience, we review the algorithm of PLS-SVD and RA:

$$\begin{array}{lrcl} \text{PLS:} & \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{12}\boldsymbol{v} & = & \lambda\boldsymbol{v} \\ \text{RA:} & \mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}\mathbf{\Sigma}_{12}\boldsymbol{v} & = & \lambda\boldsymbol{v} \end{array}$$

From (4.5),CR-PLS combines RA and PLS-SVD and is based on the following equation:

$$\boldsymbol{V} \;=\; [\alpha\boldsymbol{I} + (1-\alpha)\mathbf{\Sigma}_{22}]^{-\frac{1}{2}}\boldsymbol{U} \tag{C.1}$$

where $\boldsymbol{V}_{CR-PLSSVD}$ is a $p_2 \times r$ matrix and $\boldsymbol{U}$ is the matrix of first r eigenvectors of $\boldsymbol{P}^{-\frac{1}{2}}\mathbf{\Sigma}_{21}\mathbf{\Sigma}_{12}\boldsymbol{P}^{-\frac{1}{2}}$.

Combining RA and PLS-SVD:

The iterative process of deflating covariance matrices is still needed to combine RA and PLS-W2A. Specifically, for a selected α, let $\boldsymbol{u}_1^{(r)}$ be the first eigen-vector of

$\mathbf{\Sigma}_{12}^{(r)}[\alpha I + (1-\alpha)\mathbf{\Sigma}_{22}^{(r)}]^{-1}\mathbf{\Sigma}_{21}^{(r)}$. We compute $\boldsymbol{v}_1^{(r)}$ as follows:

$$\boldsymbol{v}_1^{(r)} = [\alpha \boldsymbol{I} + (1-\alpha)\mathbf{\Sigma}_{22}^{(r)}]^{-1}\mathbf{\Sigma}_{21}^{(r)}\boldsymbol{u}_1^{(r)}$$

Bibliography

[1] Abby Z. Israels *Redundancy Analysis for Qualitative Variables* Psychometrika–Vol. 49, No.3, 331-346, September 1984.

[2] Alexander Graham *Kronecker Products and Matrix Calculus With Applications* Ellis Horwood Limited, 1981

[3] Faez Bakalian, Harold Ritchie, Keith Thompson, William MerryField *Exploring Atmosphere-Ocean Coupling Using Principal Component and Redundancy Analysis* AMS Journals 2010; Vol. 23, 4926-4923

[4] Fred L. Bookstein *Overview of Partial Least Squares from The Viewpoint of The Natural Sciences*

[5] Fred L. Bookstein, Paul D. Sampson, Ann P. Streissguth and Helen M. Barr *Exploiting Redundant Measurement of Dose and Developmental Outcom: New Methods From the Behavioral Teratology of Alcohol* Developmental Psychology 1996, Vol.32, No.3, 404-415.

[6] Gang Li, S. Joe Qin, and Donghua Zhou, *Geometric properties of partial least squares for process monitoring* Automatica 46 (2010) 204-210.

[7] S. Bougeard, M. Hanafi, and E.M. Qannari, *Continuum redundancy-PLS regression: A simple continuum approach* Computational Statistics and Data Analysis. 52 (2008) 3688-3696.

[8] Inge S. Helland *Partial Least Squares Regression and Statistical Models* Scandinavian Journal of Statistics, Vol. 17, No. 2 (1990), pp. 97-114

[9] Jacob A. Wegelin *A Survey of Partial Least Squares (PLS) Methods, with Emphasis on the Two-Block Case* Technical Report No. 371, University of Washington, Department of Statistics.

[10] Matthew Barker, William Rayens *Partial Least Squares for Discrimination* Journal of Chemometrics 2003; 17: 166-173

[11] Roman Rosipal, and Nicole Kramer *Overview and Recent Advances in Partial Least Squares* SLSFS 2005, LNCS 3940, pp. 34-51, 2006.

[12] M. B. Priestley *Spectral Analysis and Times Series* Academic Press, 1996

[13] Ozgur Yeniay, Atilla Goktas *A Comparison of Partial Least Squares Regression with Other Prediction Methods* Hacettepe Journal of Mathematics and Statistics Volume 31 (2002), 99-111

[14] Paul Geladi, Bruce R. Kowalski *Partial Least Squares Regression: A Tutorial* Analytica Chimica Acta, 185 (1986) 1-17

[15] Peter D. Wentzell, Lorenzo Vega Montoto *Comparison of principal components regression and partial least squares regression through generic simulations of complex mixtures* Chemometrics and Intelligent Laboratory Systems 65 (2003) 257C279

[16] Richard A. Johnson, Dean W. Wichern *Applied Multivariate Statistical Analysis Fifth Edition* Prentice-Hall, 2002

[17] Rolf Ergon, Kim H. Esbensen *A Didactically Motivated PLS Prediction Algorithm*

[18] Roman Rosipal *Kernel Partial Least Squares for Nonlinear Regression and Discrimination*