CROSS-VALIDATION ADJUSTMENT FOR MODEL SELECTION WITH CORRELATED DATA

by

Ebrahim Adeeb

Submitted in partial fulfilment of the requirements
for the degree of Master of Science

at

Dalhousie University
Halifax, Nova Scotia
August 2021

Table of Contents

List of Tables

List of Figures

Abstract

In the context of general linear models, often techniques are used with an independence assumption. Unfortunately, this assumption often does not hold in real data. Real data tends to have correlations in the errors which can take a variety of structures in the form of a covariance/correlation matrices. In this research we are primarily focused on the blocked correlation structures, phylogenetic tree structures. These correlation matrices arise in hierarchical models and come from phylogenetic modelling of trait evolution. Our research proposes an adjustment to cross-validation in the case of correlated data. We will produce a variety of candidate models and test how well our techniques do at selecting the true model from the set of candidate models. This research is focused on cross-validation techniques for model selection. Cross-validation techniques are focused on re-sampling data over K number of folds into training and testing samples. Historically, methods such as cross-validation account for the dependent data by transforming the data after the splitting of training and testing data. Our research looks at transforming our data with a square-root inverse covariance ($\mathbf{V^{-1/2}}$) matrix transformation that is applied prior to the sampling. We calculate a measure known as Expected Predictive Log Density (EPLD) and it is used to measure predictive accuracy across the folds. The loss function is applied on a variety of models. In the research we show the relationship between EPLD and square error loss, and argue that SSE can be used as the selection criterion for blocked models.

List of Abbreviations and Symbols Used

- K - Number of folds in a cross-validation split.
- $\mathbf{V}$ - Covariance matrix for data.
- V_{ij} - Covariance value from matrix at row i column j.
- $\mathbf{I}$ - Identity matrix as the covariance matrix.
- $\mathbf{V^{-1/2}}$ - Square root inverse of the covariance matrix.
- $\mathbf{Y}$ - Response data.
- $\tilde{\mathbf{Y}}$ - Transformed response data.
- $\mathbf{X}$ - Covariate data.
- $\tilde{\mathbf{X}}$ - Transformed covariate data.
- β - Regression coefficient.
- σ^2 - Variance of data.
- ρ - covariance value within a specific block.
- $p(y)$ - Probability density of data y.
- $p(y, \theta)$ - Parametric probability density of data y with parameters θ.
- $l(\theta)$ - Log likelihood.

Acknowledgements

I would like to acknowledge and thank all the people who supported and helped me along the way in conducting my research and receiving my (MSc.) Masters in Statistics from Dalhousie University. I will start by thanking my supervisors Edward Susko and Lam Ho for their insight and knowledge in the domain, and their empathy towards students, they helped expand my horizons and pushed me to be more detail oriented and helped facilitate this research. It would not have been possible without their support and guidance. I would also like to thank my family for their unwavering patience and support throughout my entire academic career. They've always believed in me and pushed me to take on new challenges and succeed. They have shown me that failure is a prerequisite for success and that sometimes difficult circumstances can lead to immense growth. I would like to acknowledge all the researchers and professors that have built the foundations for my research. Finally, I would like to thank the department and my peers for being a part of this journey.

Chapter 1

Introduction

In statistics, as in any field of science, background assumptions are required. Assumptions are often made to allow for simpler inferences. Although an assumption may help to simplify things mathematically, a faulty or untrue assumption may lead to heavily skewed and biased outcomes (1). The independence assumptions states that each observation has no effect on any of the other observations in the group. This can be interpreted as our observations are not correlated to each other. In statistical modelling, often an assumption of independence is needed to fit the model. This is problematic when it comes to selecting between the models of interest when data follows a non-independent correlation structure (2). For model selection in general linear models, many techniques are used, including coefficient of determination (R^2), Akaikes information criterion (AIC), and cross-validation (CV) (3). We are interested in the effect of the correlation structure in dependent data on these techniques and what adjustments can be made to the data prior to fitting the model. Our research is heavily focused on cross-validation, and adjustments to the CV process in regards to model selection. Our research examine the efficacy and use of predictive log density as a selection criterion. We showed a equivalence for model selection between square error lost and predictive log density for certain correlation structures. We proposed a corrected CV method, which involved transforming the data using a square-root inverse matrix of the correlation prior to splitting into a training and testing set. We examine cases where we used a blocked correlation matrix, and a

phylogentic tree based correlation matrix. The research found that in most situations our corrected method out performs the classical naive methods, although it is dependant on the structure of the correlation matrix. We show in our research that when correlation is a single block our corrected, naive, and AIC act equivalently for model selection. We show that when we have multiple blocks in the correlation matrix our corrected method out performs the naive method. We found that our corrected method mimics AIC in most instances in regards to model selection. We showed that our corrected method when the correlation is blocked describes a random effects model. We found that expected predictive log density does a good job as a selection criterion and that in the case of blocked correlation its equivalent to using sum of square error. The results begin to change significantly when we used the phylogenetic tree correlation structure. In this scenario the generating model and the underlying tree heavily effects the estimation of predictive log density, therefore heavily effects our model selection. Details of each simulation scenario are described in the research below.

Background

1.1 General Linear Models

The General Linear Model (GLM), is a statistical model that underlies many of the standard statistical techniques and tools. It is used to describe the relationship between response variables and a set of covariates. The relationship between the response and covariates is linear in the parameters, thus, it is a linear model (4). A GLM is defined as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \qquad \text{where} \quad \mathbf{e} \sim N(0, \mathbf{V}) \tag{1.1}$$

When the GLM is expanded for multiple linear equations, it takes the form

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + e_i, \tag{1.2}$$

where i is the i^{th} observation ($i = 1, \ldots, n$) and p is the number of covariates ($p = 1, \ldots, q$). One of the defining features of a GLM is that the residuals follow a normal distribution mean 0 and variance $\sigma^2\mathbf{V}$ or, $\mathbf{e} \sim N(0, \sigma^2\mathbf{V})$. Often, independence of our residuals is assumed, $\mathbf{e} \sim N(0, \mathbf{I})$. This assumption allows us to implement certain statistical techniques with more ease, but this assumption is often wrong. It is important to distinguish the difference between a General Linear Model (GLM) and a Generalized Linear Model (GLiM). In the latter, the response variables are allowed to be distributed according to the exponential family of distributions. Thus, the key difference is that the general linear model strictly assumes that the residuals will follow a conditionally normal distribution (5). For our work, we are only focused on GLM's with normally distributed response variables.

Our research looks at adjustments to transform our data into a space where the residuals are independent rather than just assuming the independence. In our research we examine a $\mathbf{V}$ matrix with a blocked and phylogenetic tree correlation structure. Below I define what a correlation structure is and define a block correlation structure.

1.2 Correlation Structures

Correlation refers to the statistical relationship between two entities, and in a data set it can be defined as the statistical relationship between all observations. That statistical relationship is mathematically represented by a correlation and/or covariance matrix. We refer to that matrix summarizing our correlation as our correlation structure. It is the standardized covariance

that exists between different units of observation in our data set. This is usually presented in the notation of a $\mathbf{V}$ matrix. It is a n by n matrix with 1 on the diagonal and the correlations (ρ_{ij}) on the off diagonal. The values of ρ are bounded between $-1 \leq \rho \leq 1$. For our research we only examine non-negative correlations, bounding our ρ between $0 \leq \rho \leq 1$. The general structure of a $\mathbf{V}$ matrix is symmetric:

$$\mathbf{V} = \begin{bmatrix} 1 & \rho_{12} & \dots & \rho_{1n} \\ \rho_{21} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho_{(n-1)n} \\ \rho_{n1} & \dots & \rho_{n(n-1)} & 1 \end{bmatrix} \qquad \text{where } i \text{ and } j = 1, \dots, n.$$

Here the correlation exists in blocks of differing sizes and magnitude of ρ_{ij}. Blocked correlation has the following general form:

$$\mathbf{V_i} = \begin{bmatrix} 1 & \rho_i & \dots & \rho_i \\ \rho_i & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho_i \\ \rho_i & \dots & \rho_i & 1 \end{bmatrix} \qquad \text{where } i = 1, \dots, k,$$

$$\mathbf{V} = \begin{bmatrix} \mathbf{V_1} & 0 & \dots & 0 \\ 0 & \mathbf{V_2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \mathbf{V_k} \end{bmatrix}.$$

We will use our blocked $\mathbf{V}$ matrix to transform our data into a space where the errors are independent, $\mathbf{e} \sim N(0, \mathbf{I})$ in our GLM. Below I explain the process of accounting for the correlation in our data.

1.3 Square Root Inverse of Correlation Matrix

The Square Root Inverse (SRI) correlation matrix is the primary preconditioning technique used in our analysis to transform the data. In the case of correlated data, equation (1.1) is adjusted to:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \qquad \text{where } \mathbf{e} \sim N(0, \sigma^2\mathbf{V}). \tag{1.3}$$

The correlation matrix $\mathbf{V}$ is a symmetric $n \times n$ matrix that contains the structure of the correlation between our errors. We assume that there exists a matrix $\mathbf{Z} = \mathbf{V^{-1}}$. The inverse $\mathbf{V^{-1}}$ would not exist if one variable can be written as a linear combination of another. We assume that there exists a square root of the matrix $\mathbf{V^{-1/2}}$. We precondition our data by multiplying the $\mathbf{V^{-1/2}}$ through our response and covariates.

$$\mathbf{V^{-1/2}Y} = \mathbf{V^{-1/2}X}\boldsymbol{\beta} + \mathbf{V^{-1/2}e},$$

where $\mathbf{V^{-1/2}e} \sim N(0, \sigma^2\boldsymbol{V^{-1/2}VV^{-1/2}}) \sim N(0, \sigma^2\boldsymbol{I})$. This allows us to shift our data into a space where the errors are independent, after which many of the classical methods of model selection can begin to be applied. Moving forward in our research, we will refer to $\mathbf{V^{-1/2}}$ interchangeably as the square root inverse matrix.

1.4 Block Correlation and Random Effects Model

Recall that in (1.1) we fit a GLM with the form:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

where our $\mathbf{e}$ is the blocked correlation matrix with blocks that contain a value of ρ. We will form the relationship between our blocked correlation GLM and a random effects model (18). A random effects model has the form:

$$y = a + x'b + w.$$

where a is the random effect block mean, $x'b$ is the individuals variation, and w is the errors. A random effects model for the ith observation in the kth block has the following form:

$$y_{ik} = a_k + x'_{ik}b + w_{ik}, \tag{1.4}$$

the w_{ik} are independent and identically distributed, $iid \sim N(0, v_k^2)$. We assume that x'_{ik} contains a first entry equat to one, to account for the intercept, and that w_{ij} is independent of the a_k which are independent and identically distributed, $\sim N(0, u_k^2)$. If we set

$$v_k^2 = (1 - \rho_k^2)\sigma^2,$$

$$u_k^2 = \rho_k^2\sigma^2,$$

$$Var(y_{ik}) = Var(a_k) + Var(x'_{ik}b) + Var(w_{ik}),$$

$$Var(y_{ik}) = u_k^2 + 0 + v_k^2,$$

We can write this in the form of our model if we set $e_{ik} = a_j + w_{ik}$.

$$y_{ik} = x'_{ik}b + e_{ik}.$$

Because the w_{ik} and a_k are independent between blocks, the e_{ik} in different blocks are independent (just as in the blocks model). But because the a_k are the same for observations within the same block, the e_{ik} are dependent within blocks. What we get is

$$\begin{aligned} & Cov(e_{ik}, e_{i'k}) \\ = \; & Cov(a_k + w_{ik}, a_k + w_{i'k}) \\ = \; & Cov(a_k, a_k) + Cov(w_{ik}, a_k) + Cov(a_k, w_{i'k}) + Cov(w_{ik}, w_{i'k}) \\ = \; & Var(a_k) + 0 + 0 + 0 = u_k^2, \end{aligned}$$

the model is the same as the blocks model with

$$\begin{aligned} V_{ii} = v_k^2 + u_k^2 \qquad & \text{for } i \text{ in block } k, \\ \sigma^2 V_{ij} = u_k^2 \qquad & \text{for } i \text{ and } j \text{ in block } k, \end{aligned}$$

Therefore,

$$\rho_k \;=\; u_k^2/(v_k^2 + u_k^2). \tag{1.5}$$

1.5 Model Selection

When data is modelled, several statistical models, which we will refer to as candidate models, are considered for a specific statistical problem. Model selection is a general term for a series of techniques used to choose from a set of candidate models (1). The model that yields the best predictive accuracy

for future data set is often desired. Many statistical methods such as likelihood maximization, least squares require the choice of a model. Many basic selection criterion exist but often ignore model complexity and over fitting. A common technique for model selection is penalization. Penalization consists of adding a penalty based on the complexity of the model (6). Model selection criterion and techniques that do account for model complexity include Akaike Information Criterion. These techniques often start with the assumption of independence in our observations, which can often be untrue in a real world context. In our research we assess the accuracy and efficiency of these techniques and discuss potential adjustments to account for dependency in our data.

1.6 Expected Predictive Log Density

Lets suppose we have a generating density for data $p(y)$, and that we are considering a model where the density is $p(y, \theta)$. We are able to use that data to fit a model and approximate our θ parameter, giving an estimated density $p(y, \hat{\theta}_y)$. To understand how well our approximation does, we take data z that comes from the same generating process, $z \sim p(z)$. Predictive log density is defined as $\log p(z, \hat{\theta}_y)$. Because we don't observe z we consider expected predictive log density (EPLD). We can then say that the Expected predictive log density is defined as $E_{z,y}[\log p(z, \hat{\theta}_y)]$. We try to choose models where EPLD is large, because if the fit on new data is in a high density zone we know our fitted model is appropriate. This approach is specifically important in our phylogenetic case study. A special case of interest is when the fitted model is a normal distribution. Then the model density is:

$$p(y, \hat{\theta}_y) = \frac{1}{(2\pi)\hat{\sigma}^n \left|\hat{\mathbf{V}}\right|^{n/2}} e^{[-(y-\hat{\mu})^T \hat{\mathbf{V}}^{-1}(y-\hat{\mu})/2\hat{\sigma^2}]} ,$$

here $\theta = \mu, \sigma, \mathbf{V}$. The PLD being:

$$\log(p(z, \hat{\theta}_y)) = -\frac{n}{2}\log(2\pi) - n\log(\hat{\sigma}) - \frac{n}{2}\log\left|\hat{\mathbf{V}}\right| - (z-\hat{\mu})^T\hat{\mathbf{V}}(z-\hat{\mu})/2\hat{\sigma^2}.$$

In the special case where $\mathbf{V} = \mathbf{I}$ and where σ^2 is known and equal to 1, the PLD is:

$$\log(p(z, \hat{\theta}_y)) = -n\log(2\pi) - \sum_i (z_i - \hat{\mu}_i)^2/2.$$

Therefore maximizing $E[\log(p(z, \hat{\theta}_y))]$ is equivalent to minimizing an approximation to the EPLD $E[\sum (z_i - \hat{\mu}_i)^2]$. This tells us that for model selection in this circumstance, the minimized square error loss is equivalent for selection as is the maximized EPLD. Now lets suppose that $y \sim N(\mu, \mathbf{V})$ with a fixed and known $\mathbf{V}$ Therefore, we can directly transform $\tilde{y} = \mathbf{V^{-1/2}}y$ and $\tilde{\hat{y}} = \mathbf{V^{-1/2}}\hat{y}$. Then we can define $\tilde{y} \sim N(\tilde{\mu}, \mathbf{I})$, if $\mathbf{V}$ is known, the result above still holds. The maximized EPLD is equivalent to the minimized square error loss. If we use the estimated correlation $\hat{\mathbf{V}}$, then our equivalency breaks down and square error loss ceases to be a sufficient statistic for model selection. We cover both cases in our research, for the blocked models we use the SSE as the $\mathbf{V}$ is fixed in this case. For the research done on data from a phylogenetic process, $\hat{\mathbf{V}}$ is the estimated correlation and we revert to using the EPLD directly for model selection.

1.7 Cross Validation (CV)

When the main purpose of a statistical model is prediction, the model should hold for future data. Cross-validation (CV) is a sampling technique that splits data into a hold-out testing set and a fitting training set (8). The training set is used to fit a model, and the testing set is used to predict, and this process is repeated many times. This is in line with our process of calculating

EPLD from Section 1.6. We are repeatedly fitting a model from data $y = y_{(-n)}$ and predicting on data z. The z_i are independent and the PLD in this case is defined as $\log p(z, \hat{\theta}_y)$. As an approximation to the expectation of this you use $\sum_i \log p(y_i, \hat{\theta}_{y_{(-ii)}})$. Therefore, when this process is repeated we can generate the EPLD. We've also shown above in Section 1.6 that when data comes from a normal distribution, and the covariance matrix $\mathbf{V}$ is fixed, we are able to use a minimized square error loss as it is equivalent to maximizing EPLD. We used square error loss as our model selection criterion for the blocked modelling simulations, this is a consequence of an equivalence between square error loss and expected predictive log density (EPLD), as discussed in Section 1.6. This equivalence broke down when we simulated from phylogenetic tree correlation structures. Further details are provided in Section 3.2.1.

The CV process of sampling is repeated a set number of times and the sum of our validation error is used as a selection criterion. The sum is known as the cross-validation error and the model that minimizes our CV error is selected as the most efficient model at prediction from our candidate models. The most common form of CV is known as K-fold CV (9). This is where our data is partitioned into K equal sized prediction samples. The sample is withheld for validation while the model is fit on the remaining data. this process is then repeated K times with each sub sample used once for validation. A form of K-fold CV known as leave-one-out CV is what we've focused our research on. Leave-one-out (LOO) is a form of cross validation where a single observation is withheld for prediction (10). This requires our model to be fit with $n - 1$ observations, and this process is repeated $K = n$ times. The prediction is conducted on that single observation that was removed. It is then compared to the true value and a square error is calculated (11). The square error is summed up to generate my loss function. As stated previously, square error loss is used to estimate EPLD when $\mathbf{V}$ is known, the blocked correlation case. CV can be conducted in a variety of ways where $K = 1...n$ samples

are withheld. with smaller K values leading to a less exhaustive process. For our research purposes we will focus primarily on leave-one-out cross validation where $K = n$.

1.8 Akaike information Criterion (AIC)

AIC is a method for model selection that will be examined alongside our LOO-CV. Developed in 1973 by Hirotugu Akaike (7). AIC is widely used in model selection tasks. It incorporates an estimated in-sample log likelihood and the number of parameters included in the model. This allows for a trade-off between goodness of fit and model complexity. We defined AIC as:

$$AIC = l(\hat{\theta}) - p,$$

which is a constant multiple of -1/2 from the original definition of AIC. Here p is the number of parameters and $l(\hat{\theta})$ is the log likelihood. The results of Akaike (7), imply that AIC as defined above estimates the expected predictive log density for the model.

Our research examined two aspects of AIC. The first being how well AIC performed compared to our corrected cross-validation and naive cross validation in regards to model selection. The second being how well AIC performed at estimating the true EPLD.

1.9 Relationship Between AIC and Cross-Validation

It is proved in the paper by Stone (8), that there exists and asymptotic equivalence between AIC and Cross-Validation methods. In the paper AIC is defined as $l(\alpha, \hat{\theta}_\alpha) - p_\alpha$ where:

$$l() - \text{log likelihood},$$

$$\alpha - \text{model selected},$$

$$\hat{\theta}_\alpha - \text{Maximum Likelihood Estimator for model } \alpha,$$

$$p_\alpha - \text{ number of parameters in model } \alpha.$$

Stone states that AIC stemmed from the recognition that unreserved maximization of likelihood provides an unsatisfactory method of choice between models that are differing appreciably in their parametric dimension. Both AIC and cross-validation provide techniques to account for this difference in parameter dimension. As it will be shown below they are in fact asymptotically equivalent. Stone defined $A(\alpha)$ as the density evaluated at the observations. He examined the setting where S is previously observed data, and α are the parameters for the model:

$$A(\alpha) = \sum_i \log f(y_i|x_i, \alpha, S).$$

This shows us that the log likelihood is dependent on the data S. Therefore our parameters are estimated based on the data available. For this setting it is more reasonable to use cross-validation methods where $f^{(i)}(y) = f(y|x_i, \alpha, S_{-i})$ where S_{-i} is data S with observation i removed. This gives us

$$A(\alpha) = \sum_i \log f(y_i|x_i, \alpha, S_{-i}),$$

Asymptotic equivalence can be shown by starting with fixing α, thus setting it to an arbitrary model. We then take the first and second order derivatives

of our log likelihoods, set our first derivative to zero and solve for our MLE.

$$l^{'} = (\frac{\delta l}{\delta \theta_1}, \ldots, \frac{\delta l}{\delta \theta_l}) = 0,$$

$$l^{''} = (\frac{\delta^2 l}{\delta \theta_i, \delta \theta_j}).$$

$$n^{-1} l''(\hat{\theta} + b_i(\hat{\theta}_{-i} - \hat{\theta})) \xrightarrow{p} E[l''(y|x, \theta_0)] = l_2$$

$$n^{-1} \sum_i l'(y_i|x_i, \hat{\theta} + a_i(\hat{\theta}_{-i} - \hat{\theta})) l'(y_i|x_i, \hat{\theta}_{-i})^T \xrightarrow{p} E[l'(y|x, \theta_0) l'(t|x, \theta_0)] = l_1$$

Stone argues that if we suppose that S is a random sample from the joint distribution of (x, y). Then, with additional conditions, A is asymptotically $= l(\hat{\theta}) - trace(l_2^{-1} l_1)$ and if the model α is a version of the true model, then it can be shown that $trace(l_2^{-1} l_1) = trace(I) = p$. Therefore, $A = l(\hat{\theta}) - p$ which is identical to the AIC formulation above without α. We saw in Section 1.7 that PLD is defined as $\log p(z, \hat{\theta}_y)$, reformulating in the context of CV we can describe it as $\sum_{i=1,\ldots,n} \log p(y_i, \hat{\theta}_{y(-i)}) = A(\alpha)$. This means that $A(\alpha)$ and AIC both approximate $EPLD$.

Chapter 2

Block Correlation

2.1 Methodology

Three methods were tested to identify the highest predictive accuracy given the correlation structure assigned. Our research has defined the three methods as corrected, naive, and misspecified. The corrected and naive method use preconditioning transformations of the data to account for our dependence in the the linear model before fitting the model. The two methods implement the transformations at different stages of the sampling for cross-validation. We use simulations to compare the two methods in different situations to observe the efficacy of our corrected method versus the naive method. We've included a misspecified method where no preconditioning occurs and the dependency is ignored. The three methods are explained below in further detail.

2.1.1 Misspecified (Independence) Method for Cross-Validation

The misspecified approach completely ignores our correlation structure and assumes the data to be independent. LOO-CV is conducted without any transformation of the data.

2.1.2 Naive Method for Cross-Validation

The naive approach begins with the LOO-CV where a test sample of size $n = 1$ is removed from the original data. The remaining training set is then transformed using our square root inverse transformation. Then two models

are fit on the transformed train set.

$$\begin{aligned} y &= \mathbf{X}^T\beta + \epsilon \\ \mathbf{V}_{-i} &= \mathbf{V} \text{ with ith row and column removed} \\ y_{-i} &= y \text{ with ith row removed} \\ \tilde{y}_{-i} &= \mathbf{V}_{-i}^{-1/2} y_{-i}, \\ \mathbf{X}_{-i} &= \mathbf{X}_{-\mathbf{i}} \text{ with ith row removed} \\ \tilde{\mathbf{X}}_{-i} &= \mathbf{V}_{-i}^{-1/2}\mathbf{X}_{-i}. \end{aligned}$$

A prediction is generated using the training models an compared against our test set and the mean square error is recorded. What we use to select between the models is EPLD and in this simulation scenario our $\mathbf{V}$ is fixed and σ^2 is known and equal to 1. Therefore, we use sum of square error as a proxy for EPLD for ease of calculation for the reasons described in Section 1.6. We repeat the splitting process n times and the sum of square error for both models is used for comparison.

$$\text{CV Error} : \sum_{i=1}^{n}(y_i - \hat{y}_{-i})^2$$
$$\hat{y}_{-i} = x_i^T\hat{\beta}_{-i},$$

where $\hat{\beta}_{-i}$ is the estimate of β from a model fit with $\tilde{y}_{-i}$ and $\tilde{\mathbf{X}}_i$ and observation i removed.

2.1.3 Corrected Method for Cross-Validation

The corrected approach takes our original data and transforms it by multiplying with the square root inverse of our correlation matrix. This transformation occurs prior to the LOO-CV. Then a test sample of size $n = 1$ is removed from the transformed data and the models are fit on the remaining

train set.

$$\begin{aligned} \tilde{y} &= \mathbf{V^{-1/2}} y \\ \tilde{x} &= \mathbf{V^{-1/2}} x \\ \text{Fit}: \ \tilde{y}_j &= \tilde{x}_j \beta + \epsilon_j. \end{aligned}$$

$$\begin{aligned} \text{CVerror} &: \sum_{i=1}^{n} (\tilde{y}_i - \hat{\tilde{y}}_{-i})^2 \\ \hat{\tilde{y}}_{-i} &= [\tilde{x}_i]^T \hat{\beta}^*_{-i}. \end{aligned}$$

2.2 Simulation Procedure

We examine different blocked correlation structures with a variety of models to test which of our four methods selects the true generating model from two or many candidate models. The block model structure described in Section 1.2, with a $\sigma^2 = 1$. The four methods we compared are our corrected method in Section 2.1.3, naive method in Section 2.1.2, misspecified model in Section 2.1.1, and AIC in Section 1.8.

This section outlines the procedure conducted in the simulation functions developed for researching our model selection behavior. As stated previously in equation (1.3):

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \qquad \text{where} \ \ \mathbf{e} \sim N(0, \sigma^2 \mathbf{V}).$$

We begin our simulation by specifying the correlation structure to examine. We need to simulate $\mathbf{e}$ and $\boldsymbol{X}_i$ in the general linear model with covariance matrix $\mathbf{V}$. This involves setting number of blocks, size of blocks, values of ρ to examine, the generating model, the false model for testing, the correlation

structure of the response variables ($\mathbf{V}$), and the correlation between our covariates ($\mathbf{V_x}$). The first step takes our correlation structure and generates a data set based on that structure. We recognized that if $\boldsymbol{Z} \sim N(0, \boldsymbol{I})$ then $\boldsymbol{W} = \boldsymbol{V}^{1/2}\boldsymbol{Z} \sim N(0, \boldsymbol{V})$. In R we can generate $\boldsymbol{Z}$ with the rnorm() function and then apply the transformation $\boldsymbol{V^{1/2}Z}$ to get a $N(0, \boldsymbol{V})$ vector. To generate the data, we take our $\mathbf{V}$ matrix and build a square root inverse of the matrix ($\mathbf{V^{-1/2}}$). Proofs in Section 2.3.2 show that a square root and an inverse of the square root of $\mathbf{V}$ exist if $\rho < 1$. We multiply the $\mathbf{V^{-1/2}}$ with errors generated from a $\sim N(0, \sigma^2)$ distribution. This gives us our correlated errors. We then take our fixed β coefficient multiply them by our randomly generated covariates. This will give us a deterministic response variable, the errors calculated above are added to our response variable to add stochasticity. This data is then used to conduct cross validation with our four methods. We also examine the scenario where the ρ variable in our blocks within the $\mathbf{V}$ matrix are fixed and known vs. estimated. The estimation was done by fitting a mixed effect model and then estimating ρ from the model. Proof of the relationship between a mixed effect model and our ρ variable is in Section 1.9.

Four types of generated data sets were examined with varying parameters. The types include data from a constant correlation $\mathbf{V}$ matrix, a 44 observations in a 4 block $\mathbf{V}$ structure, and a 44 observations in a 2 block $\mathbf{V}$ structure, and finally a 88 observations in a 4 block $\mathbf{V}$ structure. Our selection criteria involve comparing the sum of the square error from all the procedures above for the true model and the false model. We select the smaller sum of square error between our true and false model. We repeat each procedure 2,000 times. We take the proportion of times our true model was selected over our false model given the simulation specifications. This proportion was selected as the measure for comparing method performance. These simulations are then repeated again across multiple values of ρ in the block structure, and different

values of β in our generating model. These repeated simulations are then plotted to see how our selection mechanism works as ρ goes from 0 to 0.99 for all our methods. In the results section below, I present different block structures and number of observations.

A variety of correlation structures were examined through the simulation of data. Our main focus was data generated with blocked correlation, but other correlation structures were also examined. Section 2.3 below overviews the case where the correlation exists as a single uniform block of the form:

$$\mathbf{V} = \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \rho & \ddots & \vdots \\ \rho & \dots & \rho & 1 \end{bmatrix}, \tag{2.1}$$

Blocked correlation has the following general form:

$$\mathbf{V} = \begin{bmatrix} V_1 & 0 & \dots & 0 \\ 0 & V_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & V_k \end{bmatrix}.$$

$$\mathbf{V_i} = \begin{bmatrix} 1 & \rho_i & \dots & \rho_i \\ \rho_i & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho_i \\ \rho_i & \dots & \rho_i & 1 \end{bmatrix} \qquad where\ i = 1, \dots, k,$$

The first model is the true model, the second model is a some false nested or unnested model. Then a prediction using the training models is used against

our test set and the sum square error is recorded. As previously stated SSE is used as a proxy to EPLD. This is repeated n times and the sum of square error for both models is used for comparison. Simulations are used to repeat the four methods above on m-simulated data sets. Data is generated from the same true model. At each iteration the CV sum of the square error is recorded, and the fitted model with the smaller SSE is selected. The proportion of times the true model, is recorded and compared across our four methods. The simulations also iterate across different values of ρ in the correlation structure. The trend of proportion of times the true model is selected, across different values of ρ, is examined in our results below.

2.3 Single Block Structure Results

The first correlation structure we examine, is a single block with a constant ρ variable on the off diagonals. For this simulation data is generated from $y = 1 + 0.5X_1 + \beta_2^* X_2 + \epsilon$ and we set $y = 1 + 0.5X_1 + \beta_2 X_2 + \epsilon$ as the true fitted model, and $y = 1 + 0.5X_1 + \epsilon$, as the false model for comparison. During the generation phase we alternate the value of β_2^* to note the effect of a larger coefficient.

2.3.1 Simulation results in constant correlation scenario

In this scenario, where our correlation matrix has the structure below:

$$\mathbf{V} = \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \rho & \ddots & \vdots \\ \rho & \dots & \rho & 1 \end{bmatrix}, \tag{2.2}$$

Our simulations, given in Figure 2.1, showed that all methods selected the true model equivalently across all potential values or ρ and alternating β^*s. This

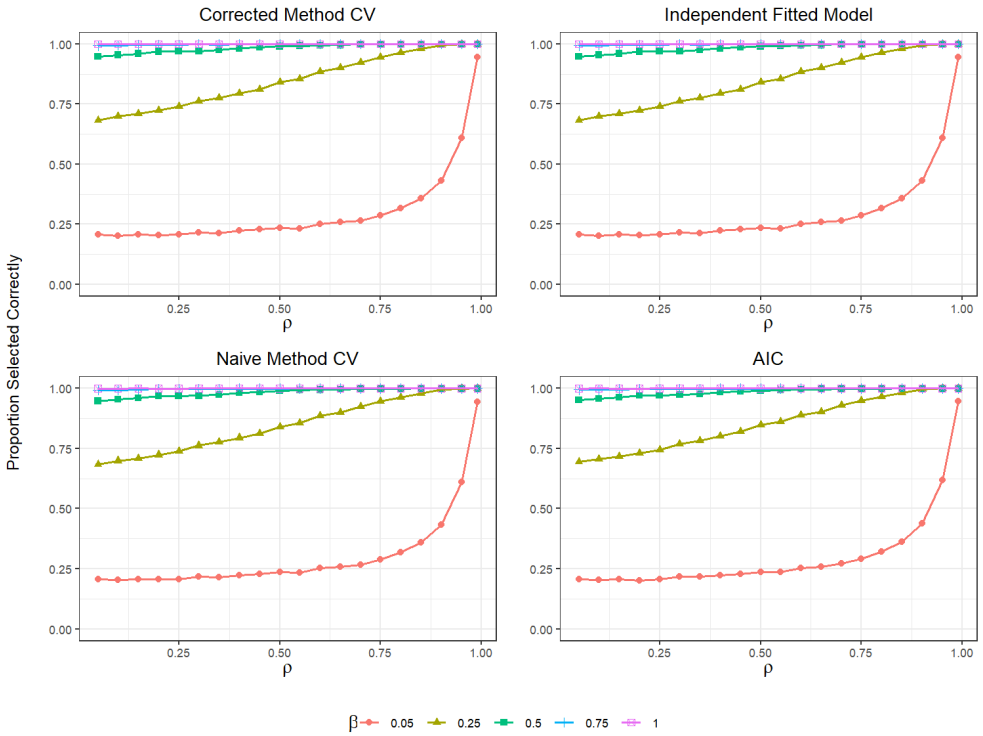

Figure 2.1: Simulation results showing proportion of times the true complex model is selected two candidate models, simple and complex. The correlation structure is a single block of size 44, the number of observations is 44, and $\mathbf{V_x} = \mathbf{I}$. A single line for each β_2^* value, across a variety of ρ values.

result was a surprise, but after examination we recognized this as a mathematical property of this correlation structure. This is due to the fact that with constant correlation your parameter estimates will remain the same but your intercept will be shifted. This behaviour is not seen when different correlation structures exist. With this single block correlation structure, we note that for the task of model selection, although the intercept is misspecified, all our methods perform equally well at selecting the true model across all values of ρ and β^*. Its also evident that for low values of ρ and β^* all our methods are biased towards selecting the simpler model. This is explained by for very low values of β_2^* not affecting our response variable during the generation phase,

but this effect of biasing towards the simpler model disappears as our values of ρ approach 1. The true model is always selected in the case where β_2^* is large enough $\sim$0.9.

2.3.2 Comparison of Corrected, Naive, and Misspecified Methods.

In this section below, I prove that the results hold generally for a constant correlation model no matter the parameters. I prove that when data contains a single block structure, the $\hat{\beta}_j$ parameter estimates are equivalent while the intercept $\hat{\beta}_0$ varies between the corrected, naive, and misspecified methods.

Proof:

1. We show the $\mathbf{V^{-1}}$ is of form:

$$[\mathbf{V^{-1}}]_{ij} = \begin{cases} a & i = j \\ b & i \neq j \end{cases}, \tag{2.3}$$

In the constant correlation scenario, the $\mathbf{V}$ matrix as:

$$[(1/\sigma^2)\mathbf{V}]_{ij} = \begin{cases} 1 & i = j \\ \rho & i \neq j \end{cases},$$

It suffices to show that for $\boldsymbol{V}^{-1}$ of the form (2.3), and $\boldsymbol{V}$ of the form (2.2) $\mathbf{V^{-1}V} = \boldsymbol{I}$.

$$[\mathbf{V^{-1}V}]_{ij} = \sum_k V_{ik}^{-1} V_{kj} = \begin{bmatrix} V_{i1}^{-1} & \dots & V_{in}^{-1} \end{bmatrix} \begin{bmatrix} V_{1j} \\ \vdots \\ V_{nj} \end{bmatrix},$$

This gives us that the result holds when $i \neq j$:

$$[\mathbf{V^{-1}V}]_{ij} = a\rho + b + \sum_{k \notin (i,j)} \mathbf{V^{-1}}_{ik} V_{kj} = a\rho + b + b\rho(n-2) = 0,$$

and when $i = j$:

$$[\mathbf{V^{-1}V}]_{ii} = a + \sum_{k \notin (i,j)} \mathbf{V^{-1}}_{ik} V_{kj} = a + b\rho(n-2) = 1,$$

Therefore the result holds if:

$$[\mathbf{V^{-1}V}]_{ij} = \begin{cases} b + \rho(a + b(n-2)) = 1 & i = j \\ a + b\rho(n-2) = 0 & i \neq j \end{cases},$$

If we multiply the bottom equation by ρ then subtract the top and bottom equations.

$$\rho a + b\rho^2(n-2) - b - \rho(a + b(n-2)) = -1$$

Rearranging things gives us:

$$b = -1/[\rho^2(n-2) - 1 - \rho(n-2)]$$

$$a = \rho(n-2)/[\rho^2(n-2) - 1 - \rho(n-2)]$$

This gives us that the $\mathbf{V^{-1}}$ matrix in the form:

$$[\mathbf{V^{-1}}]_{ij} = \begin{cases} a = \rho(n-2)/[\rho^2(n-2) - 1 - \rho(n-2)] & i = j \\ b = -1/[\rho^2(n-2) - 1 - \rho(n-2)] & i \neq j \end{cases},$$

2. We show the $\mathbf{V^{-1/2}}$ is of form:

$$[\mathbf{V^{-1/2}}]_{ij} = \begin{cases} c & i = j \\ d & i \neq j \end{cases},$$

If this is true then it must be the case that:

$$[\mathbf{V^{-1/2}V^{-1/2}}] = \mathbf{V^{-1}} = \begin{cases} a & i = j \\ b & i \neq j \end{cases} = \begin{bmatrix} V_{i1}^{-1/2} & \dots & V_{in}^{-1/2} \end{bmatrix} \begin{bmatrix} V_{1j}^{-1/2} \\ \vdots \\ V_{nj}^{-1/2} \end{bmatrix},$$

$$= \begin{cases} \mathbf{V^{-1/2}}_{ii}\mathbf{V^{-1/2}}_{ij} + \mathbf{V^{-1/2}}_{ij}\mathbf{V^{-1/2}}_{jj} + \sum_{k \notin (i,j)} \mathbf{V^{-1/2}}_{ik}\mathbf{V^{-1/2}}_{kj} & i \neq j \\ \mathbf{V^{-1/2}}_{ii}\mathbf{V^{-1/2}}_{jj} + \sum_{k \notin (i,j)} \mathbf{V^{-1/2}}_{ik}\mathbf{V^{-1/2}}_{kj} & i = j \end{cases}.$$

This gives that the result holds if and when $i \neq j$:

$$[\mathbf{V^{-1/2}V^{-1/2}}]_{ij} = cd + cd + \sum_{k \notin (i,j)} d^2 = 2cd + d^2(n-2) = b,$$

and when $i = j$:

$$[\mathbf{V^{-1/2}V^{-1/2}}]_{ii} = c^2 + \sum_{k \notin (i,j)} d^2 = c^2 + d^2(n-1) = a,$$

Therefore the result holds if:

$$[\mathbf{V^{-1/2}V^{-1/2}}]_{ij} = \begin{cases} c^2 + d^2(n-1) = a & i = j \\ 2cd + d^2(n-2) = b & i \neq j \end{cases}.$$

We subtract the top equation from the bottom, we get:

$$a - b = c^2 + d^2(n-1) - 2cd - d^2(n-2) = c^2 - 2cd + d^2 = (c-d)^2.$$

If we rearrange the equation above we get that:

$$c = \sqrt{a-b} + d.$$

Note that we only have a real solution only if $a > b$, this is a certainty when

$-1 < \rho < 1$, because a is the diagonal entry and b is the off-diagonal entry of the positive definite matrix $\mathbf{V}^{-1/2}$. For our research we constrict our ρ to being positive, so we examine the case where $0 < \rho < 1$. Plugging the equation for c back into the top equation gives us:

$$(\sqrt{a-b}+d)^2 - d^2(n-1) = a,$$

If we use the quadratic form we can rearrange the equation to get:

$$d = \frac{-2\sqrt{a-b} \pm \sqrt{4(a-b)+4b(n-1)}}{2(n-1)}$$

We require that $\mathbf{V^{-1/2}}$ be positive definite, therefore $c > d$. This gives us square-root inverse matrix $\mathbf{V^{-1/2}}$ of the form:

$$[\mathbf{V^{-1/2}}]_{ij} = \begin{cases} c = \sqrt{a-b} + \frac{-2\sqrt{a-b} \pm \sqrt{4(a-b)+4b(n-1)}}{2(n-1)} & i = j \\ d = \frac{-2\sqrt{a-b} \pm \sqrt{4(a-b)+4b(n-1)}}{2(n-1)} & i \neq j \end{cases}.$$

3. We show that the above implies $\tilde{y}_i = \alpha y_i - \gamma \bar{y}$ for some α, γ: To transform our y_i we multiply it by our $\mathbf{V^{-1/2}}$ matrix for $i = 1, \ldots, n$:

$$\tilde{y}_i = [\mathbf{V}^{-1/2} y]_i = cy_i + d\sum_{k \neq i} y_k = (c-d)y_i + d[\sum_{k \neq i} y_k - y_i],$$

$$(c-d)y_i + dn\bar{y} = \alpha y_i - \gamma \bar{y},$$

We apply the same logic for our x_{ij} to get the result:

$$\tilde{x}_{ij} = [\mathbf{V^{-1/2}} x]_{ij} = (c-d)x_{ij} + dn\bar{x}_{.j} = \alpha x_{ij} - \gamma \bar{x}_{.j}.$$

4. Show that the estimated transformed correlation coefficients $\tilde{\hat{\beta}}_j$ and $\tilde{\hat{\beta}}_0$ are minimizers: The least squares estimates $\hat{\beta}_j$ and $\hat{\beta}_0$ are minimizers

and so they satisfy that:

$$\sum_i (y_i - \sum_j x_{ij}\hat{\beta}_j - \hat{\beta}_0)^2 \leq \sum_i (y_i - \sum_j x_{ij}\beta_j^* - \beta_0^*)^2 \text{ , for arbitrary } \beta_j^*, \beta_0^*,$$

We replace our original observations with the transformed variables using the result from part 3.

$$\tilde{y_i} = \alpha y_i + \gamma \bar{y},$$

$$y_i = 1/\alpha \tilde{y_i} - \gamma/\alpha \bar{y},$$

$$\tilde{x}_{ij} = \alpha x_{ij} + \gamma \bar{x}_{.j},$$

$$x_{ij} = 1/\alpha \tilde{x}_{ij} - \gamma/\alpha \bar{x}_{.j}.$$

Given our corrected model is $y_i = \beta_0 + \sum_j x_{ij}\beta_j + \epsilon_i$ where $\epsilon \sim N(0, \sigma^2 \mathbf{V})$. We fit the transformed model:

$$\tilde{y_i} = \tilde{X_i}\tilde{\beta} + \tilde{\epsilon} = \tilde{x}_{i0}\beta_0^t + \tilde{x}_{i1}\tilde{\beta}_1 + \ldots + \tilde{x}_{ip}\tilde{\beta}_p + \tilde{\epsilon}_i = \tilde{x}_{i0}\beta_0^t + \sum_j {x_{ij}}^t \beta_j^t + \tilde{\epsilon}_i$$

$$\tilde{\epsilon}_i \sim N(0, \sigma^2 I),$$

$$\tilde{x}_{i0} = \sum_k [\mathbf{V^{-1/2}}]_{ik} 1 = [\mathbf{V^{-1/2}}]_{ii} + \sum_{k \neq i} [\mathbf{V^{-1/2}}]_{ik} = c + (n-1)d. \tag{2.4}$$

This shows us that x_{i0} is independent of i and can be treated as a constant. Replacing our y and x with there respective transformed variants yields inequality (2.5):

$$\sum_i ((\frac{1}{\alpha}\tilde{y_i} - \frac{\gamma}{\alpha}\bar{y}) - \sum_j (\frac{1}{\alpha}\tilde{x}_{ij} - \frac{\gamma}{\alpha}\bar{x}_{.j})\hat{\beta}_j - \hat{\beta}_0)^2 \leq \sum_i (y_i - \sum_j x_{ij}\beta_j^* - \beta_0^*)^2 \tag{2.5}$$

Rearranging the left hand side of (2.5) yields:

$$\sum_i((\frac{1}{\alpha}\tilde{y}_i - \frac{\gamma}{\alpha}\bar{y}) - \sum_j(\frac{1}{\alpha}\tilde{x}_{ij} - \frac{\gamma}{\alpha}\bar{x}_{.j})\hat{\beta}_j - \hat{\beta}_0)^2 = \frac{1}{\alpha^2}\sum_i(\tilde{y}_i - \sum_j \tilde{x}_{ij}\tilde{\hat{\beta}}_j - \tilde{x}_{i0}\tilde{\hat{\beta}}_0)^2,$$

Where $\tilde{\hat{\beta}}_j = \hat{\beta}_j$ and $\tilde{x}_{i0}\tilde{\hat{\beta}}_0 = -\gamma\bar{y} - \hat{\beta}_j\gamma\bar{x}_{.j} + \alpha\hat{\beta}_0$. Using the same logic above on the right hand side of (2.5) yield:

$$\frac{1}{\alpha^2}\sum_i(\tilde{y}_i - \sum_j \tilde{x}_{ij}\tilde{\beta}^*_j - \tilde{x}_{i0}\tilde{\beta}^*_0)^2,$$

Where $\tilde{\beta}^*_j = \beta^*_j$ and $\tilde{x}_{i0}\tilde{\beta}^*_0 = -\gamma\bar{y} - \beta^*_j\gamma\bar{x}_{.j} + \alpha\beta^*_0$. After cancellation inequality (2.5) becomes:

$$\sum_i(\tilde{y}_i - \sum_j \tilde{x}_{ij}\tilde{\hat{\beta}}_j - \tilde{x}_{i0}\tilde{\hat{\beta}}_0)^2 \leq \sum_i(\tilde{y}_i - \sum_j \tilde{x}_{ij}\tilde{\beta}^*_j - \tilde{x}_{i0}\tilde{\beta}^*_0)^2.$$

This holds for all $\tilde{\beta}^*_j$ and $\tilde{\beta}^*_0$, therefore $\tilde{\hat{\beta}}_j$ and $\tilde{\hat{\beta}}_0$ are minimizers. We know that for any arbitrary value of β^*_j and β^*_0 there exists a 1-1 transformation with $\tilde{\beta}^*_j$ and $\tilde{\beta}^*_0$.

$$\tilde{\beta}^*_j = g(\beta^*_j) = \beta^*_j$$

$$\tilde{x}_{i0}\tilde{\beta}^*_0 = g(\beta^*_0) = -\gamma\bar{y} - \beta^*_j\gamma\bar{x}_{.j} - \alpha\beta^*_0.$$

5. Leave-one-out comparison between our Corrected and Naive methods: a) Corrected: Leave one out $\tilde{y}_i, \tilde{x}_{ij}$ For our corrected method we perform the transformation before the leave-one-out stage. This gives us models of the form:

$$\tilde{\hat{y}}_i = \sum_j \tilde{x}_{ij}\tilde{\hat{\beta}}_{ji} + \tilde{x}_{i0}\tilde{\hat{\beta}}_{0j},$$

We showed in part 2. that

$$[\mathbf{V^{-1/2}}]_{ij} = \begin{cases} c & i = j \\ d & i \neq j \end{cases},$$

and that

$$[\mathbf{V^{-1/2}}y]_i = \tilde{y_i} = (c-d)y_i + dn\bar{y} = \alpha_1 y_i + \gamma_1 \bar{y},$$

and that

$$[\mathbf{V^{-1/2}}x]_{ij} = \tilde{x}_{ij} = (c-d)x_{ij} + dn\bar{x}_{.j} = \alpha_1 x_{ij} + \gamma_1 \bar{x}_{.j},$$

for some constants c, d, α_1. Therefore, $\tilde{\hat{y}}_{-i}$ is the observation after transformation we leave out one before the fitting:

$$\tilde{\hat{y}}_{-i} = \sum_j \tilde{x}_{-ij}\tilde{\hat{\beta}}_{j-i} + \tilde{x}_{i0}\tilde{\hat{\beta}}_{0-i},$$

Where $\tilde{\hat{\beta}}_{ji} = \hat{\beta}_{ji}$ and $\tilde{x}_{-i0}\tilde{\hat{\beta}}_{0-i} = -\gamma_1\bar{y} - \hat{\beta}_{ji}\gamma_1\bar{x}_{.ji} + \alpha_1\hat{\beta}_{0i}$. Now I can calculate the predictive error from $[\tilde{y}_i - \tilde{\hat{y}}_{-i}]$ by replacing the variables with their model parameters and functions of the original observation.

$$[(\alpha_1 y_i - \gamma_1\bar{y}) - (\tilde{x}_{-i0}\tilde{\hat{\beta}}_{0-i} + \sum_j \tilde{x}_{ij}\hat{\beta}_{ji})],$$

$$= [(\alpha_1 y_i - \gamma_1\bar{y}) + \gamma_1\bar{y} + \hat{\beta}_{ji}\gamma_1\bar{x}_{.j} + \alpha_1\hat{\beta}_{0i} - \sum_j \tilde{x}_{ij}\hat{\beta}_{ji})],$$

$$= [\alpha_1 y_i - \alpha_1\hat{\beta}_{0i} - \sum_j \alpha_1 x_{ij}\hat{\beta}_{ji}],$$

$$= \alpha_1[y_i - \hat{y}_i].$$

We've shown above that the predictive error between our corrected and independent methods are equivalent up to a scaling factor $[\tilde{y}_i - \hat{y}_i^t] = \alpha_1[y_i - \hat{y}_i]$.

b) Naive: Leave one out y_i, x_{ij} In the naive method we conduct the leave-one-out step first before the transformation. This gives us a model fit of:

$$\tilde{\hat{y}}_{-i} = \sum_j \tilde{x}_{-ij}\tilde{\hat{\beta}}_{j-i} + \tilde{x}_{i0}\tilde{\hat{\beta}}_{0-i}.$$

For the naive method we split the data prior to transformation, thus after the split we need to recompute $\mathbf{V^{-1/2}}$ based on all observations except i. If we take two new constants c_{n-1} and d_{n-1}, where they are the new constant calculated from data with an observation removed. It remains the case that:

$$[\mathbf{V^{-1/2}}]_{-ij} = \begin{cases} c_{n-1} & i = j \\ d_{n-1} & i \neq j \end{cases},$$

for some constant c, d that differ from those of $\mathbf{V^{-1/2}}$

$$[\mathbf{V^{-1/2}}y]_{-i} = \tilde{y}_i = (c_{n-1} - d_{n-1})y_i + d_{n-1}(n-1)\bar{y} = \alpha_2 y_i + \gamma_2 \bar{y},$$

$$[\mathbf{V^{-1/2}}]_{-ij}x_{ij} = \tilde{x}_{ij} = (c_{n-1} - d_{n-1})x_{ij} + d_{n-1}(n-1)\bar{x}_{.j} = \alpha_2 x_{ij} + \gamma_2 \bar{x}_{.j},$$

$$\tilde{\hat{y}}_{-i} = \sum_j \tilde{x}_{-ij}\tilde{\hat{\beta}}_{j-i} + \tilde{\hat{\beta}}_{0-i},$$

Where $\tilde{\hat{\beta}}_j = \hat{\beta}_j$ and $\tilde{\hat{\beta}}_0 = -\gamma_2\bar{y} - \hat{\beta}_j\gamma_2\bar{x}_{.j} + \alpha_2\hat{\beta}_0$, for some constant α_2 and γ_2. Now we can calculate the predictive error from $[\tilde{y}_i - \tilde{\hat{y}}_{-i}]$ by replacing the variables with their model parameters and functions of the original observation.

$$[(\alpha_2 y_i - \gamma_2\bar{y}) - (\tilde{x}_{-i0}\tilde{\hat{\beta}}_{0-i} + \sum_j \tilde{x}_{ij}\hat{\beta}_{ji})],$$

$$= [(\alpha_2 y_i - \gamma_2\bar{y}) + \gamma_2\bar{y} + \hat{\beta}_{ji}\gamma_2\bar{x}_{.j} + \alpha_2\hat{\beta}_{0(i)} - \sum_j \tilde{x}_{ij}\hat{\beta}_{ji})],$$

$$= [\alpha_2 y_i - \alpha_2\hat{\beta}_{0(i)} - \sum_j \alpha_2 x_{ij(i)}\hat{\beta}_{ji}],$$

$$= \alpha_2[y_i - \hat{y}_i].$$

We've shown above that the predictive error between our naive and independent methods are equivalent up to a scaling factor $[\tilde{y}_i - \tilde{\hat{y}}_i] = \alpha_2[y_i - \hat{y}_i]$. The proofs above have shown us that our predictive error for all three methods (corrected,naive,independent) are all equivalent up to a scaling factor. Thus, our selection criterion in the constant correlation scenario will always be the same regardless of method. This is confirmed by numeric results seen in the simulations and results for the constant correlation scenario.

2.4 Multi Block Structure

2.4.1 Four Blocks and the Complex Model as the Generating Model

In this simulation set we examine a blocked correlation structure with 4 blocks with block sizes equal to 11 in the correlation matrix. The ρ value is fixed in this simulation. This yields $n = 44$ as our number of observations. In the first iteration we looke at $\boldsymbol{y} = 1 + 0.5\boldsymbol{X_1} + \beta_2^*\boldsymbol{X_2} + \mathbf{e}$ as the true generating model, and select from two fitted models. $\boldsymbol{y} = \beta_0 + \beta_1\boldsymbol{X_1} + \beta_2\boldsymbol{X_2} + \mathbf{e}$, the complex model and $\boldsymbol{y} = \beta_0 + \beta_1\boldsymbol{X_1} + \mathbf{e}$, the simple model. During the generation phase we alternate the value of β_2^* to note the effect of a larger coefficient. The first simulation was run with our correlation between the covariates set to zero and we assigned $\mathbf{V_x}$, the covariance matrix between our covariates, as the identity matrix $\boldsymbol{I}$. We assumed our ρ value as fixed and known in this simulation. As we see in Figure 2.2., for large values of β_2^* all our methods select the true generating model almost exclusively. The behavior at lower values of β_2^* show a trend as ρ increases, all our methods increasingly select the true model. The rate of selecting is higher for the corrected and AIC methods, and they converge to 1 as ρ converges to 1. The bias towards selecting the simpler model for low values of β_2^* is still evident but disappears as ρ values are large. We note that increased performance in all methods

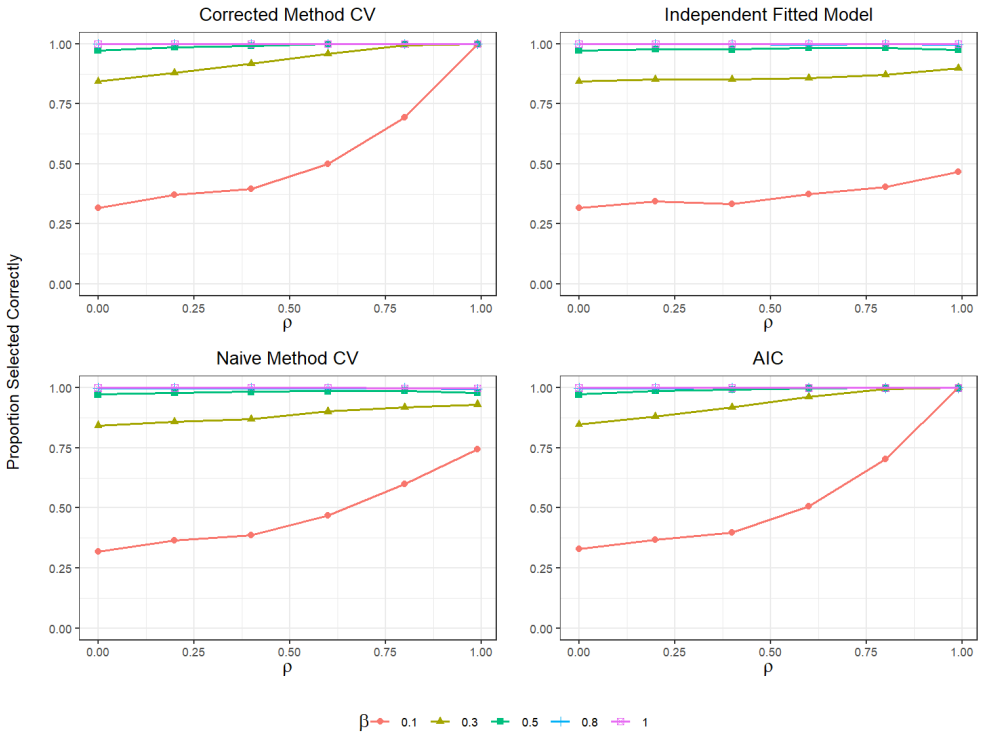

Figure 2.2: Simulation results showing proportion of times the true complex generating model is selected from two candidate models, simple and complex. The correlation structure consists of 4 equal sized blocks of size 11, the number of observations is 44, the true fixed ρ was used, and $\mathbf{V_x} = \mathbf{I}$. A single line for each β_2^* value, across a variety of ρ values.

as the correlation within blocks ρ increases. This is due to the relationship between the block model and the random effects model described in Section 1.5. To describe this more visually, we selected two random data sets from our simulations. One with a low value of $\rho = 0.2$ and the other $\rho = 0.99$. Plotting the x and y values for those two scenarios a clear pattern appears. We used the data from the simulations involving a single covariate to allow us to plot the data in 2-dimensions. Figure 2.3 shows us the plot of x and y for the case where $\beta_1^* = 0.2$ and $\rho = 0.2$ with 4 blocks of size 11 for a total number of observations of $n = 44$. As can be seen in Figure 2.3 all the data points are quite randomly scattered regardless of which block they belong to. This leads to the lower performance in selecting the correct model when the ρ value is low, as there is no clear distinction between data from different blocks. We note the slopes of the least squares line are inconsistent and averaging across them leads to a slope of zero. That estimated slope of zero is unrepresentative of the true slope which leads to the bad model selection outcomes. Figure 2.4 shows us the plot of x and y for the case where $\beta_1^* = 0.2$ and $\rho = 0.99$ with 4 blocks of size 11 for a total number of observations of $n = 44$. As can be seen in Figure 2.4 clusters of observations begin to appear for data from each block. As we did above, we use a least squares fit line for visual assistance. We use this least squares approach to approximate the random effects because the fitting procedure is difficult without knowing the true a_j from equation (1.4). Once the lines are fit, see Figure 2.4, we note that for all blocks a similar slope line appears with varying y-intercepts. This fits the the frame work of a random effects model defined in Section 1.4. This shows us that as the ρ value within blocks increases, clear slopes for each block appears. Averaging these slopes gives a better approximation to the true slope compared to the slopes from observations taken with a low value of ρ. This leads to the increase in proportion of times the true model is selected.

We now examine scenarios where our correlation in the covariates is not

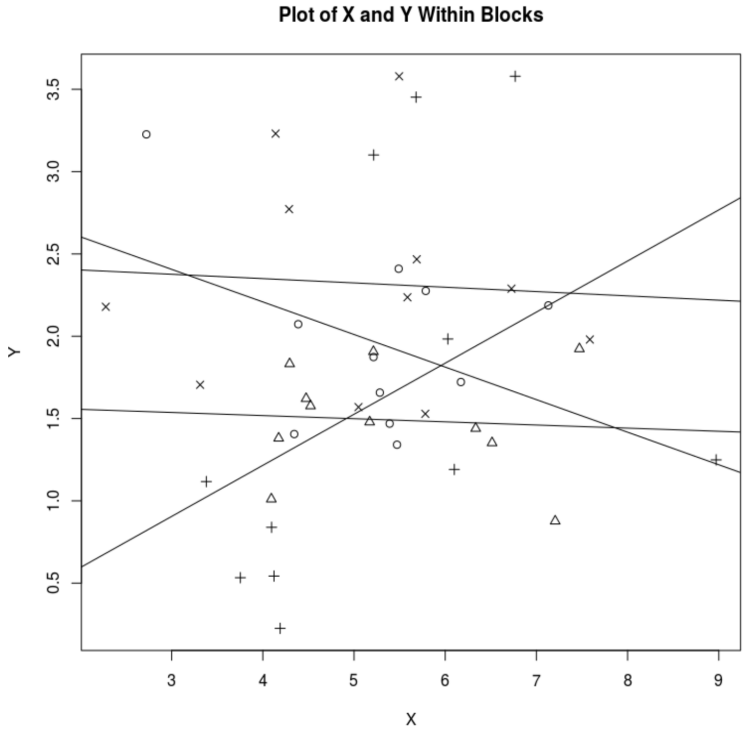

Figure 2.3: Plot of X and Y from a simulation with 4 equal sized blocks of size 11, the number of observations is 44, β_1^* value of 0.2, and ρ value of 0.2. With least squares fit lines added for each block.

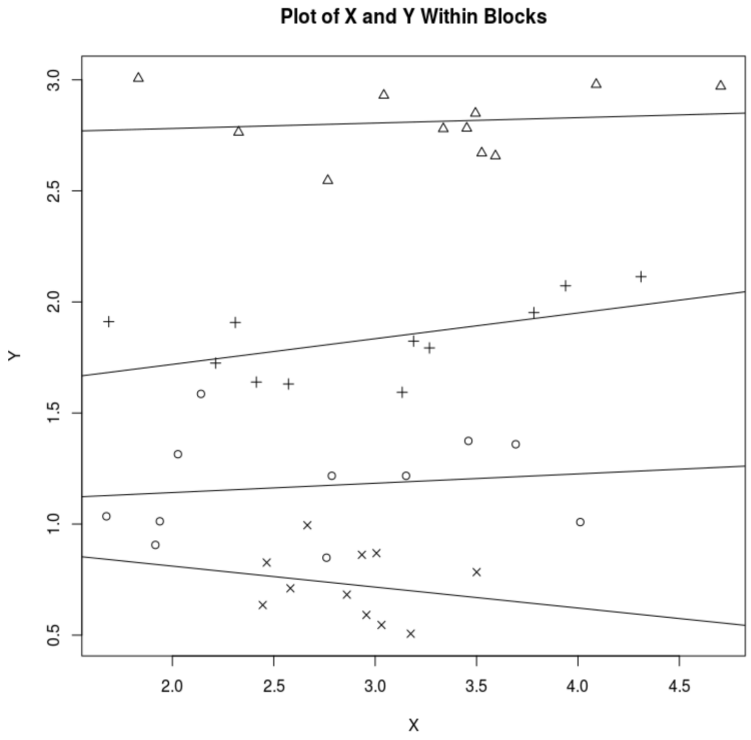

Figure 2.4: Plot of X and Y from a simulation with 4 equal sized blocks of size 11, the number of observations is 44, β_1^* value of 0.2, and ρ value of 0.99. With least squares fit lines added for each block.

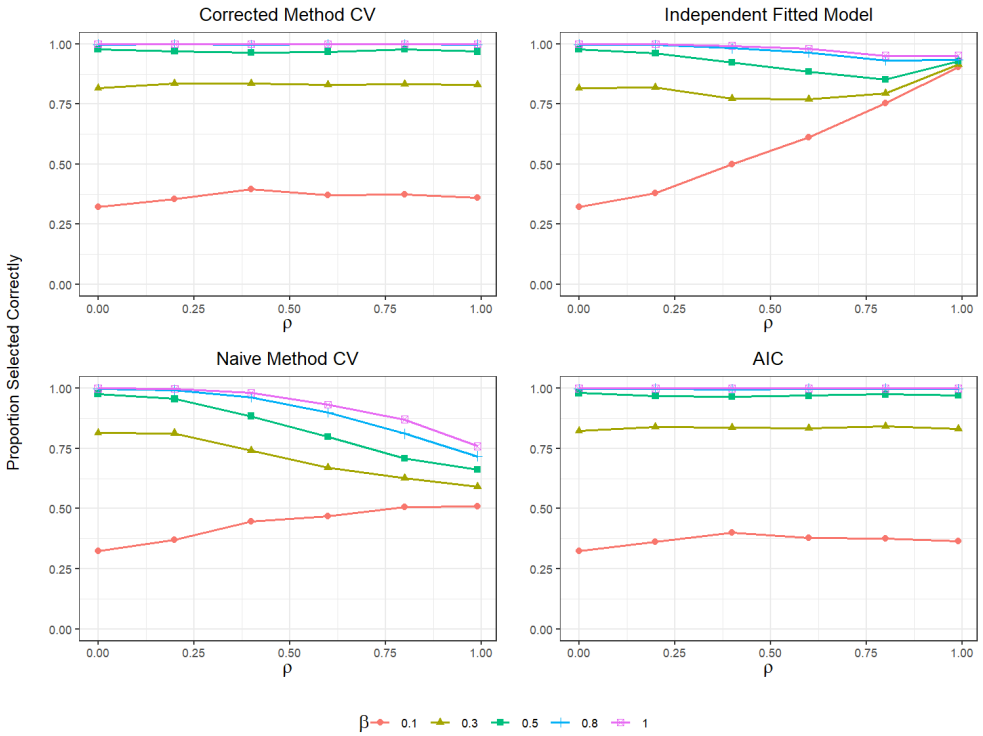

Figure 2.5: Simulation results showing proportion of times the true complex generating model is selected from two candidate models, simple and complex. The correlation structure consists of 4 equal sized blocks of size 11, the number of observations is 44, the true fixed ρ was used, and $\mathbf{V_x} = \mathbf{V}$. A single line for each β_2^* value, across a variety of ρ values.

zero. The first such simulation examined what occurs when our covariance matrix for our covariates is equivalent to the covariance matrix of our response. We are setting $\mathbf{V_x}$ as $\mathbf{V}$ and ρ as fixed and known. In this simulation we looked at $\boldsymbol{y} = 1 + 0.5\boldsymbol{X_1} + \beta_2^*\boldsymbol{X_2} + \mathbf{e}$ as the true generating model. We fitted $\boldsymbol{y} = \beta_0 + \beta_1\boldsymbol{X_1} + \beta_2\boldsymbol{X_2} + \mathbf{e}$, complex model, as the true fitted and $\boldsymbol{y} = \beta_0 + \beta_1\boldsymbol{X_1} + \mathbf{e}$, as the false fitted model for comparison. The simulations yielded very different results compared to Figure 2.2. For our corrected and AIC methods we note a constant selection rate regardless of the value of ρ, this is generally true when $\boldsymbol{V_x} = \boldsymbol{V}$. As our β_2^* increases it selects the true

model more often. The misspecified method seems to perform best under these conditions as the rate of selecting the true model increases with an increased value of ρ. This is a very curious result and occurs when $\mathbf{V_x} = \mathbf{V}$. For the naive method a strange behaviour occurs where for low values of β_2^* and high values of ρ it selects the true model more often than our corrected method. Alternatively, for high values of β_2 and high values of ρ, it selects the true model less often.

2.4.2 Four Blocks and the Simple Model as the Generating Model

In this simulation set we examine a blocked correlation structure with 4 blocks with block sizes equal to 11 in the correlation matrix. The ρ value is fixed in this simulation. We switch our true generating model as $\boldsymbol{y} = 1 + \beta_1^* \boldsymbol{X_1} + \mathbf{e}$, simple model. The fitted models for comparison are $\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{X_1} + \mathbf{e}$, simple model, as the correct fitted model and $\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{X_1} + \beta_2 \boldsymbol{X_2} + \mathbf{e}$, as the false fitted model for comparison. We note that the true generating simple model is selected more frequently than the complex model. The trends for our corrected, misspecified, and AIC methods are selecting the true model with a probability of 0.8 over all values of ρ and β. The naive method has a surprising behavior. The mean number of times the true model is selected seems to decrease as ρ increases.

The tables below show the differences in performance when our generating model is either the complex $\boldsymbol{y} = 1 + 0.5\boldsymbol{X_1} + \beta_2^* \boldsymbol{X_2} + \mathbf{e}$ as the true generating model or $\boldsymbol{y} = 1 + \beta_1^* \boldsymbol{X_1} + \mathbf{e}$, simple model, as the true generating model. Table 2.1 shows the differences between our methods when the generating model is complex or simple, and when $\mathbf{V_x} = \mathbf{I}$. It is clear to see that across all methods, the complex model as the generating model has better performance as β^* moves away from 0. A similar result is seen when $\mathbf{V_x} = \mathbf{V}$, Table 2.2 below shows the results. We recognize that in both cases, the proportion of times the true model is selected is agnostic to the value of correlation within

Proportion of Times the True Model is Selected					
Generating Model	ρ	Corrected	Independent	Naive	AIC
Complex	0.00	0.90	0.90	0.90	0.91
Simple	0.00	0.83	0.83	0.83	0.83
Complex	0.20	0.92	0.90	0.91	0.93
Simple	0.20	0.83	0.82	0.81	0.83
Complex	0.40	0.95	0.91	0.93	0.96
Simple	0.40	0.83	0.82	0.81	0.84
Complex	0.60	0.98	0.91	0.94	0.98
Simple	0.60	0.84	0.82	0.78	0.83
Complex	0.80	1.00	0.92	0.95	1.00
Simple	0.80	0.83	0.82	0.75	0.82
Complex	0.99	1.00	0.93	0.96	1.00
Simple	0.99	0.84	0.82	0.58	0.84

Table 2.1: The table outlines the proportion of times the true model is selected when the correlation structure consists of 4 equal sized block of size 11, and total number of observations is equal to 44. It examines both when the simple or complex models are the generating model, $\mathbf{V_x} = \mathbf{I}$ and $\beta_2 = 0.5$.

the blocks. I would also note the bias towards selecting the complex model in the independence method when the simple model is the generating model.

2.4.3 Estimated $\hat{\rho}$

Next we examine a simulation where all our factors are the same as above, but we set ρ as estimated rather than fixed and known. We estimate $\hat{\rho}$ by using the equivalence between the random effects model and CV as described in Section 1.4. Figure 2.6 and 2.7 shows that the scenarios between the fixed and estimated, ρ and $\hat{\rho}$, behave identically, regardless of the correlation structure of $\mathbf{V_x}$. Further inspection of a variety of simulations showed that the selection

Proportion of Times the True Model is Selected					
Generating Model	ρ	Corrected	Independent	Naive	AIC
Complex	0.00	0.88	0.88	0.88	0.89
Simple	0.00	0.83	0.83	0.83	0.82
Complex	0.20	0.90	0.87	0.87	0.90
Simple	0.20	0.83	0.79	0.80	0.83
Complex	0.40	0.89	0.85	0.83	0.90
Simple	0.40	0.82	0.68	0.74	0.82
Complex	0.60	0.89	0.81	0.75	0.90
Simple	0.60	0.83	0.50	0.65	0.82
Complex	0.80	0.89	0.80	0.70	0.90
Simple	0.80	0.83	0.34	0.57	0.82
Complex	0.99	0.89	0.90	0.65	0.90
Simple	0.99	0.84	0.12	0.46	0.84

Table 2.2: The table outlines the proportion of times the true model is selected when the correlation structure consists of 4 equal sized block of size 11, and total number of observations is equal to 44. It examines both when the simple or complex models are the generating model, $\mathbf{V_x} = \mathbf{V}$ and $\beta_2^* = 0.5$.

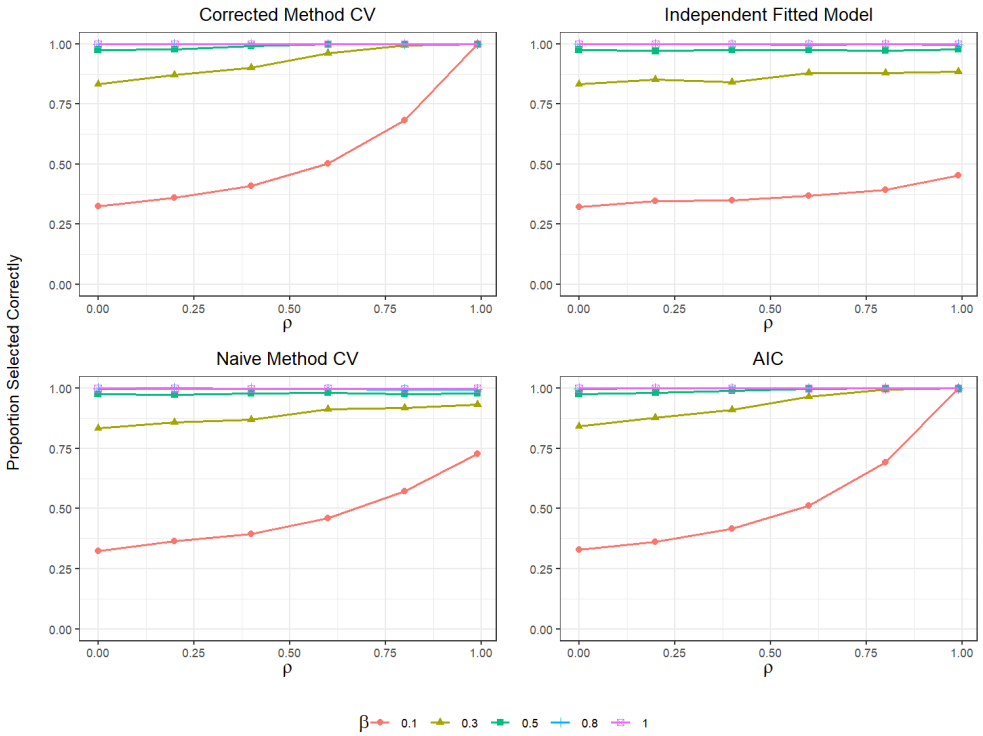

Figure 2.6: Simulation results showing proportion of times the true complex generating model is selected from two candidate models, simple and complex. The correlation structure consists of 4 equal sized blocks of size 11, the number of observations is 44, an estimated $\hat{\rho}$ was used, and $\mathbf{V_x} = \mathbf{I}$. A single line for each β_2^* value, across a variety of ρ values.

criteria is only marginally affected by the estimation of ρ at low values of ρ. Table 2.3 and 2.4 below summarizes all the results and methods, the values presented in the table represent proportion of times the true model is selected for each method with $\mathbf{V_x} = \mathbf{I}$ and $\mathbf{V_x} = \mathbf{V}$ accordingly.

2.4.4 Multiple Model Comparison

In this section we examine model selection using our techniques between a set of candidate models. This means that we select from a set that includes our generating model, a set of nested models, and unnested models. The

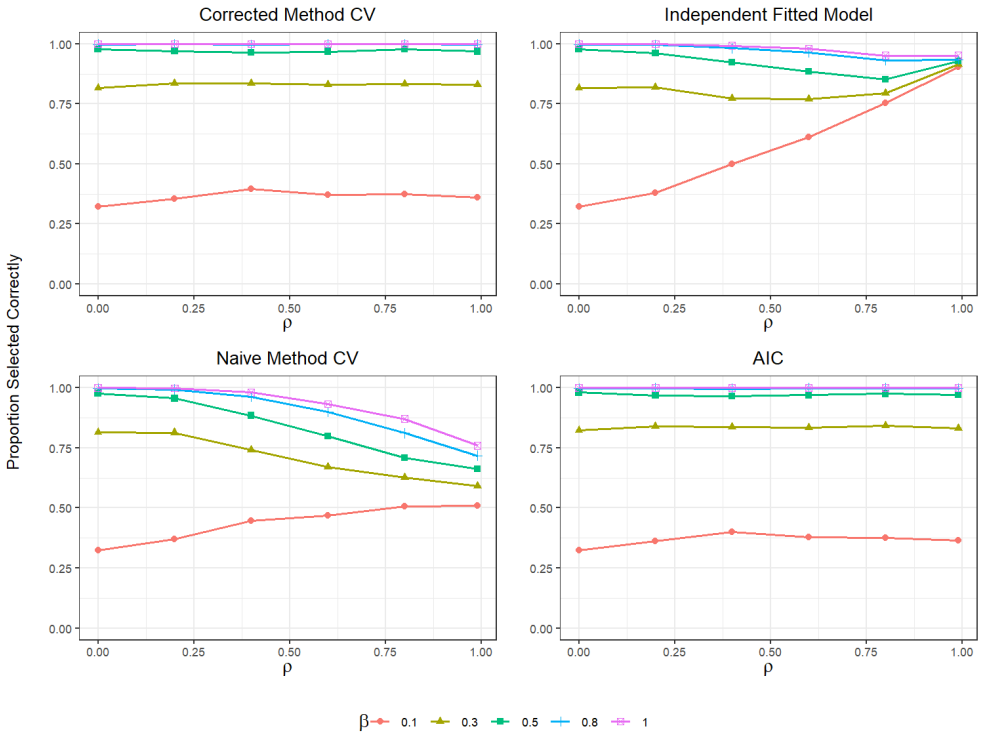

Figure 2.7: Simulation results showing proportion of times the true complex generating model is selected from two candidate models, simple and complex. The correlation structure consists of 4 equal sized blocks of size 11, the number of observations is 44, an estimated $\hat{\rho}$ was used, and $\mathbf{V_x} = \mathbf{V}$. A single line for each β_2^* value, across a variety of ρ values.

Proportion of Times the True Model is Selected					
ρ Type	ρ	Corrected	Independent	Naive	AIC
Fixed	0.00	0.90	0.90	0.90	0.91
Estimated	0.00	0.89	0.89	0.89	0.90
Fixed	0.20	0.92	0.90	0.91	0.93
Estimated	0.20	0.92	0.91	0.91	0.92
Fixed	0.40	0.95	0.91	0.93	0.96
Estimated	0.40	0.94	0.90	0.92	0.95
Fixed	0.60	0.98	0.91	0.94	0.98
Estimated	0.60	0.98	0.91	0.94	0.98
Fixed	0.80	1.00	0.92	0.95	1.00
Estimated	0.80	1.00	0.92	0.95	1.00
Fixed	0.99	1.00	0.93	0.96	1.00
Estimated	0.99	1.00	0.93	0.96	1.00

Table 2.3: The table outlines the proportion of times the true model is selected when the correlation structure consists of 4 equal sized block of size 11, and total number of observations is equal to 44. It examines both when the ρ value is fixed or estimated, $\mathbf{V_x} = \mathbf{I}$ and $\beta_2^* = 0.5$.

Proportion of Times the True Model is Selected					
ρ Type	ρ	Corrected	Independent	Naive	AIC
Fixed	0.00	0.88	0.88	0.88	0.89
Estimated	0.00	0.89	0.89	0.89	0.90
Fixed	0.20	0.90	0.87	0.87	0.90
Estimated	0.20	0.89	0.88	0.88	0.91
Fixed	0.40	0.89	0.85	0.83	0.90
Estimated	0.40	0.90	0.86	0.84	0.91
Fixed	0.60	0.89	0.81	0.75	0.90
Estimated	0.60	0.90	0.81	0.76	0.90
Fixed	0.80	0.89	0.80	0.70	0.90
Estimated	0.80	0.91	0.80	0.71	0.92
Fixed	0.99	0.89	0.90	0.65	0.90
Estimated	0.99	0.89	0.89	0.67	0.90

Table 2.4: The table outlines the proportion of times the true model is selected when the correlation structure consists of 4 equal sized block of size 11, and total number of observations is equal to 44. It examines both when the ρ value is fixed or estimated, $\mathbf{V_x} = \mathbf{I}$ and $\beta_2^* = 0.5$.

procedure is similar to above. I begin by comparing all my non generating models and taking the one with the smallest loss and comparing that against my generating model. The models examined are:

- Model 1: X_1
- Model 2: X_1, X_2 - Generating Model
- Model 3: X_1, X_2, X_3
- Model 4: X_1, X_3, X_4
- Model 5: X_4, X_5
- Model 6: X_2
- Model 7: X_2, X_3

For these multiple model comparisons I examine the case where $\mathbf{V_x} = \mathbf{I}$ and $\mathbf{V_x} = \mathbf{V}$, and the ρ is treated as fixed. Figure 2.8 and 2.9 show the results from these simulations: The behavior between the simulations conducted with $\mathbf{V_x} = \mathbf{I}$ and $\mathbf{V_x} = \mathbf{V}$ are distinctly different. Figure 2.8 shows the result from $\mathbf{V_x} = \mathbf{I}$. We can see that the proportion of times the true model is selected increases for our corrected and AIC methods as ρ increases. For low values of β_2^* the performances increases significantly as ρ converges to 1. The misspecified independent, method only exhibit slight increases as ρ converges to 1 but are generally performing far worse than our corrected and AIC methods at high values of ρ. The naive method has the most peculiar behaviour as for low values of β performance increases with ρ, but at higher values of β it exhibits much lower performance. Figure 2.9 shows us the results when $\mathbf{V_x} = \mathbf{V}$. As we can see the performance of selecting the true generating model in our corrected and AIC methods remain constant regardless of the value of ρ. The naive and misspecified both exhibit rapid decreases of

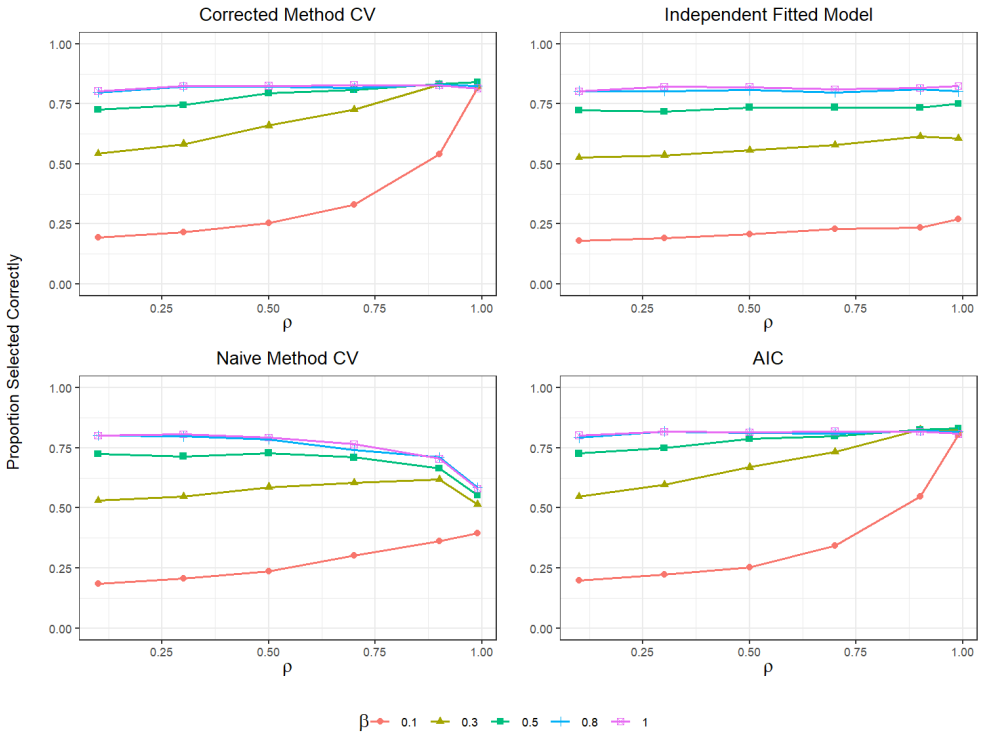

Figure 2.8: Simulation results showing proportion of times the true complex generating model is selected from many candidate models that are nested and unnested. The correlation structure consists of 4 equal sized blocks of size 11, the number of observations is 44, an estimated $\hat{\rho}$ was used, and $\mathbf{V_x} = \mathbf{I}$. A single line for each β_2^* value, across a variety of ρ values.

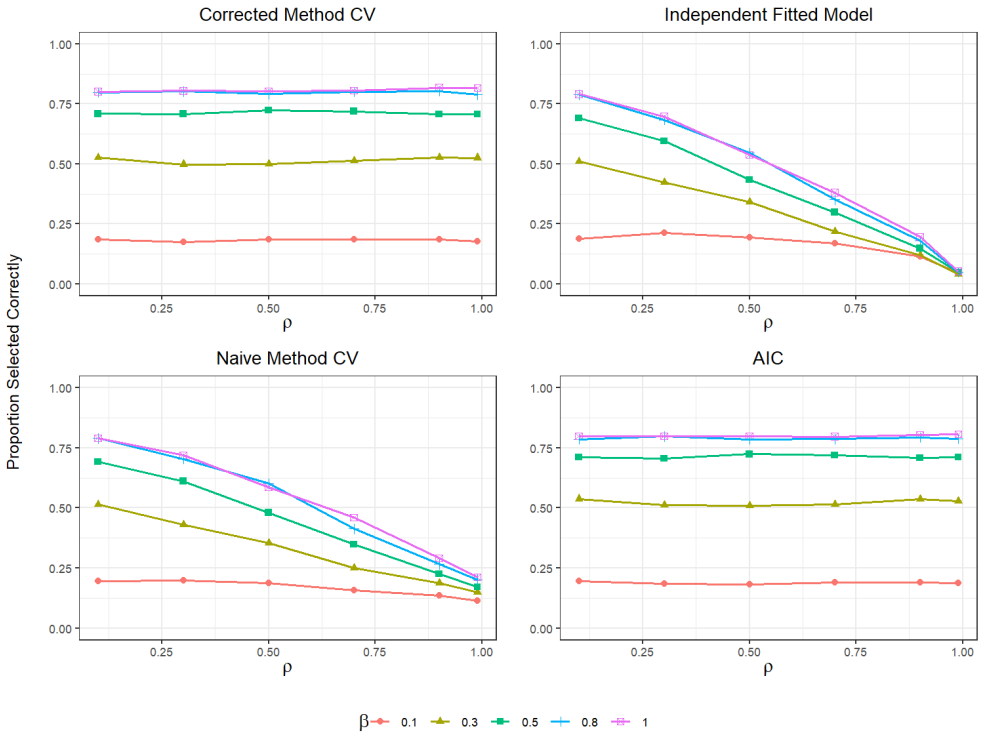

Figure 2.9: Simulation results showing proportion of times the true complex generating model is selected from many candidate models that are nested and unnested. The correlation structure consists of 4 equal sized blocks of size 11, the number of observations is 44, an estimated $\hat{\rho}$ was used, and $\mathbf{V_x} = \mathbf{V}$. A single line for each β_2^* value, across a variety of ρ values.

performance with an increase in the value of ρ. Table 2.5 below outlines the proportions for both cases:

2.4.5 Simulation with Reduced Number of Blocks

In this section, we repeat our simulations by fixing the number of observations but reduce the number of blocks. We assess the effect of number of blocks has on model selection performance. As above we examine a variety of correlation structures for our $\mathbf{V_x}$ and look at cases where ρ was fixed or estimated. Our number of observations remains fixed at 44. In this iteration we look at $y = 1 + 0.5X_1 + \beta_2^* X_2 + \mathbf{e}$ as the true generating model, and select from two fitted models. $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \mathbf{e}$, complex model as the correct fitted model and $y = \beta_0 + \beta_1 X_1 + \mathbf{e}$, simple model, as the false fitted model for comparison. During the generation phase we alternate the value of β_2^* to note the effect of a larger coefficient. The first simulation was run with our correlation between the covariates set to zero and we assigned $\mathbf{V_x} = \mathbf{I}$. We examine when the ρ value is fixed and estimated in Table 2.6. In this scenario we note that all four methods exhibit increased performance as ρ increases. Our corrected and AIC methods increase performance at a faster rate that the naive and misspecified methods. Table 2.7 looks at the same simulations but we set $\mathbf{V_x} = \mathbf{V}$. For the case where $\mathbf{V_x} = \mathbf{V}$ as seen in Table 2.7 for both the fixed and estimated ρ, the performance in selecting the true generating model remains constant for our corrected and AIC methods regardless of the value of ρ. The misspecified model appears to increase in performance as ρ converges to 1, this is a surprising result. The naive method on the other hand exhibits an increase in performance when ρ is in the low to medium range. As ρ gets large, the performance begins to decrease.

The results above show similar results between the case when our correlation matrix is 2 blocks of size 22 and 4 blocks of size 11. This leads me to believe that the number of blocks has no significant effect on model selection

Proportion of Times the True Model is Selected					
Covariates	ρ	Corrected	Independent	Naive	AIC
I	0.10	0.73	0.72	0.73	0.73
V	0.10	0.71	0.69	0.69	0.71
I	0.30	0.75	0.72	0.72	0.75
V	0.30	0.71	0.60	0.61	0.71
I	0.50	0.80	0.74	0.73	0.79
V	0.50	0.72	0.44	0.48	0.72
I	0.70	0.81	0.74	0.71	0.80
V	0.70	0.72	0.30	0.35	0.72
I	0.90	0.83	0.74	0.67	0.83
V	0.90	0.71	0.15	0.23	0.71
I	0.99	0.84	0.75	0.55	0.83
V	0.99	0.71	0.05	0.17	0.71

Table 2.5: The table outlines the proportion of times the true model is selected when selecting from many models, nested and unnested, the correlation structure consists of 4 equal sized block of size 11, and total number of observations is equal to 44. It examines the scenarios where our covariates $\mathbf{V_x} = \mathbf{I}$ vs.$\mathbf{V_x} = \mathbf{V}$, for a fixed ρ and $\beta_2^* = 0.5$.

Proportion of Times the True Model is Selected					
ρ Type	ρ	Corrected	Independent	Naive	AIC
Fixed	0.00	0.94	0.94	0.94	0.94
Estimated	0.00	0.99	0.99	0.99	0.99
Fixed	0.20	0.94	0.93	0.93	0.95
Estimated	0.20	0.99	0.99	0.99	0.99
Fixed	0.40	0.98	0.96	0.96	0.98
Estimated	0.40	1.00	0.99	0.99	1.00
Fixed	0.60	1.00	0.98	0.98	1.00
Estimated	0.60	1.00	0.99	0.99	1.00
Fixed	0.80	1.00	0.98	0.99	1.00
Estimated	0.80	1.00	0.99	0.99	1.00
Fixed	0.99	1.00	0.97	0.98	1.00
Estimated	0.99	1.00	1.00	1.00	1.00

Table 2.6: The table outlines the proportion of times the true model is selected when the correlation structure consists of 2 equal sized block of size 22, and total number of observations is equal to 44. It examines both when the ρ value is fixed or estimated, $\mathbf{V_x} = \mathbf{I}$ and $\beta_2^* = 0.6$.

Proportion of Times the True Model is Selected					
ρ Type	ρ	Corrected	Independent	Naive	AIC
Fixed	0.00	0.93	0.93	0.93	0.94
Estimated	0.00	0.99	0.99	0.99	0.99
Fixed	0.20	0.95	0.93	0.92	0.95
Estimated	0.20	0.98	0.97	0.96	0.98
Fixed	0.40	0.95	0.90	0.87	0.95
Estimated	0.40	0.99	0.96	0.91	0.99
Fixed	0.60	0.94	0.85	0.77	0.93
Estimated	0.60	0.98	0.91	0.84	0.98
Fixed	0.80	0.94	0.83	0.69	0.93
Estimated	0.80	0.99	0.91	0.75	0.99
Fixed	0.99	0.95	0.87	0.61	0.92
Estimated	0.99	0.98	0.91	0.68	0.98

Table 2.7: The table outlines the proportion of times the true model is selected when the correlation structure consists of 2 equal sized block of size 22, and total number of observations is equal to 44. It examines both when the ρ value is fixed or estimated, $\mathbf{V_x} = \mathbf{V}$ and $\beta_2^* = 0.5$.

for cross validation in our 4 techniques.

2.4.6 Simulation with an Increased Number of Observations

In this section, we repeat our simulations by fixing the number of blocks but doubling our number of observations. We assess the effect increasing observations has on model selection performance. As above we examin a variety of correlation structures for our $\mathbf{V_x}$ and look at cases where ρ was fixed or estimated. Our number of blocks remains fixed with block sizes of 11.

In this iteration we look at $\mathbf{y} = 1+0.5\mathbf{X_1}+\beta_2^*\mathbf{X_2}+\mathbf{e}$ as the true generating model, and select from two fitted models. $\mathbf{y} = \beta_0 + \beta_1\mathbf{X_1} + \beta_2\mathbf{X_2} + \mathbf{e}$ as the correct fitted model and $\mathbf{y} = \beta_0 + \beta_1\mathbf{X_1} + \mathbf{e}$ as the false fitted model for comparison. During the generation phase we alternate the value of β_2^* to note the effect of a larger coefficient. The first simulation was run with our correlation between the covariates set to zero and we assigned $\mathbf{V_x} = \mathbf{I}$. From the simulation with a larger number of observations, it was evident that that proportion of times the true model is selected is approximately 1 at lower values of β_2^*. Table 2.8 below compares our methods at a lower threshold of $\beta_2^* = 0.3$. In the case where $\mathbf{V_x} = \mathbf{V}$ we observe the same results as in previous sections, where for our corrected and AIC methods the proportion of times the true model is selected remains constant regardless of the value of ρ. The behaviour of the naive and misspecified are also the same but the rates of convergence appear to be higher when we increase sample size. Table 2.9 compares our methods for fixed and estimated ρ when $\beta_2^* = 0.3$. We've shown that when $\mathbf{V_x} = \mathbf{I}$, performance is very good in model selection as in Table 2.8, compared to performance drop in our naive and independent techniques when $\mathbf{V_x} = \mathbf{V}$. The results from simulations with increased number of observations show that our selection criteria follow similar trends but with increased sensitivity towards the value of β_2^*. This means the convergence behaviour occurs at faster rates for lower values of β_2^*.

Proportion of Times the True Model is Selected					
ρ Type	ρ	Corrected	Independent	Naive	AIC
Fixed	0.00	0.97	0.97	0.97	0.97
Estimated	0.00	0.98	0.98	0.98	0.98
Fixed	0.20	0.99	0.98	0.98	0.99
Estimated	0.20	0.98	0.97	0.97	0.98
Fixed	0.40	0.99	0.98	0.98	0.99
Estimated	0.40	0.99	0.97	0.98	0.99
Fixed	0.60	1.00	0.98	0.99	1.00
Estimated	0.60	1.00	0.98	0.98	1.00
Fixed	0.80	1.00	0.98	0.99	1.00
Estimated	0.80	1.00	0.97	0.98	1.00
Fixed	0.99	1.00	0.98	0.98	1.00
Estimated	0.99	1.00	0.98	0.98	1.00

Table 2.8: The table outlines the proportion of times the true model is selected when the correlation structure consists of 4 equal sized block of size 22, and total number of observations is equal to 88. It examines both when the ρ value is fixed or estimated, $\mathbf{V_x} = \mathbf{I}$ and $\beta_2^* = 0.3$.

Proportion of Times the True Model is Selected					
ρ Type	ρ	Corrected	Independent	Naive	AIC
Fixed	0.00	0.98	0.98	0.98	0.98
Estimated	0.00	0.98	0.98	0.98	0.98
Fixed	0.20	0.97	0.95	0.94	0.97
Estimated	0.20	0.97	0.96	0.96	0.97
Fixed	0.40	0.97	0.93	0.89	0.98
Estimated	0.40	0.96	0.92	0.88	0.97
Fixed	0.60	0.97	0.88	0.79	0.97
Estimated	0.60	0.97	0.88	0.80	0.97
Fixed	0.80	0.98	0.88	0.71	0.98
Estimated	0.80	0.98	0.85	0.71	0.98
Fixed	0.99	0.97	0.92	0.66	0.97
Estimated	0.99	0.97	0.93	0.66	0.97

Table 2.9: The table outlines the proportion of times the true model is selected when the correlation structure consists of 4 equal sized block of size 22, and total number of observations is equal to 88. It examines both when the ρ value is fixed or estimated, $\mathbf{V_x} = \mathbf{V}$ and $\beta_2^* = 0.3$.

Chapter 3

Phylogenetic Correlation Structures

3.1 Phylogenetics Background

Phylogenetics is the field of science pertaining to the study of the evolution in and among species (12). In the study of phylogenetics we are often focused on a specific trait and its evolution through time and taxa. Our case study will focus on phylogenetics trait evolution (17). The correlation in traits of species is dependent on their evolutionary relationships. Those evolutionary relationships are represented in a phylogenetic tree. The phylogenetic tree is a graphical representation of the evolution of the aforementioned taxa (13). Phylogenetic trees take a variety of forms. To obtain results for a range of tree types, our research focused on Coalescent, Caterpillar, and Balanced Trees. The trees we examine are rooted and binary, which specifies that all species of interest to us share a single common ancestor, which is the root of the tree, and each species split is binary. The leaves of the tree are the species with observable traits. Each node is a unit of taxonomy (ie. a species). The splits from a node are binary a representation of two species splitting from their common ancestor. The edge lengths represented in a phylogenetic tree are an estimate of time that was needed for the species to evolve. In the sections below I outline the type of trees we examine in our analysis and describe the differences between them.

The generating process underlying the evolution of these traits is of particular interest to us. Statistical models are used to approximate how trait

evolution occurs through time. We examine two models specifically, Brownian-motion and Ornstein–Uhlenbeck processes. Sections below describe the two processes in more detail. Conditioned on a tree and a generating process we can generate data for simulation purposes. We use the data to select between two candidate models, one of which is the true generating process. We then adapt and use the methods we developed from Section 2.1, to identify which of them work best for selecting the true generating model.

3.1.1 Phylogenetic Trees

There are many forms of phylogenetic trees and a variety of parameters control the behaviour of the tree. Trees can be rooted, the common ancestor is known, or unrooted where there is no common ancestor (12). A tree can be binary (bifurcating) or multifurcating. Bifurcation states that when a species evolves and splits, the split occurs into two distinct species. While in a multifurcating tree, a ancestor can be split into many species at certain point in time. In our analysis we focus on rooted binary trees. Another important aspect of a tree is its shape. Tree shape describes the properties of a phylogeny. The shape of a tree consists of two parts, the topology and the edge lengths (14). These are known as clocklike trees. There exists a form of tree that is known as a non-clocklike tree, but we will not be examining them in the research. Topology refers to the branching pattern of a tree, while edge lengths represents the amount of time was required for the species to split.

Coalescent trees are one of the three tree structures we examine in our analysis. Coalescent trees are gene trees that treat traits as independent random variables generated from a coalescence process occurring along the lineages of the species tree. Since the multispecies coalescent model allows gene trees to vary across genes, coalescent methods have been popularly used to account for heterogeneous gene trees in phylogenomic data analysis (22). All

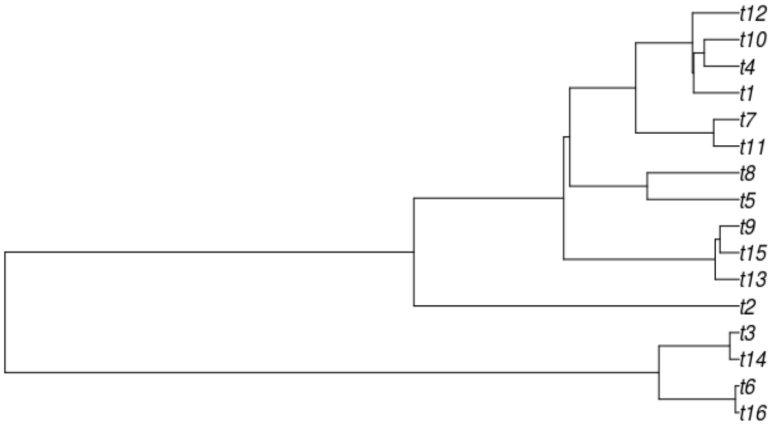

Figure 3.1: A Coalescent Tree with 16 nodes (species).

the trees we examine are a form of a rooted binary (bifurcating) tree, including the coalescent tree(15).This implies that within the species examined we can root back to a single common ancestor for all the species. While the bifurcation refers to the species splits. The edge lengths are re-scaled to the number of generations (15). This indicates that species evolution occurs in non standard time steps. Figure 3.1. gives an example of a coalescent tree. A caterpillar tree is another form of a rooted binary tree, but the behaviour of the tree is quite different than that of a coalescent tree. It is defined as a binary phylogenetic tree for which the induced subtree on the interior vertices forms a path graph (16).A path graph just states that the vertices along the tree can be listed in order. This means the tree is unbalanced, and only a single variant of a species from a split continuous to evolve and splits, while the other is restricted to only evolving. (14). Figure 3.2 gives us an example of a caterpillar tree with 16 nodes/species. The last tree we examine in our analysis is the balanced tree. Much like the previous two tree structures, it is a rooted binary tree (14). The key feature is the balanced nature of the tree.

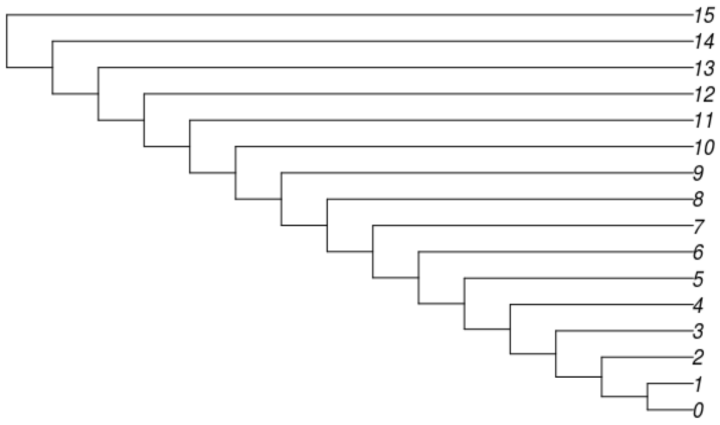

Figure 3.2: A Caterpillar Tree with 16 nodes (species).

This means that the edge lengths or time to species split is always equal. It takes a constant amount of time for a species to split into two distinct species (14). Similar to a coalescent tree, but the edge lengths are all constant. Figure 3.3. gives an example of a balanced tree.

3.1.2 Gaussian Models

Models for trait evolution assume conditionally independent evolution along distinct lineages given the trait value of their common ancestor. We start our description of the models used by describing models for the evolutionary process along the path from the root of the tree to a tip. Therefore is suffices to describe the evolutionary process through time for a single lineage. Trait evolution through time is a stochastic process which accounts for the trait at a specific time $X(t)$, $t \geq 0$. Such a stochastic process is called Gaussian if $X(t_1), \ldots, X(t_n)$ has a multivariate normal distribution for all $t_1, \ldots, t_n$ (19). We examine two Gaussian processes that can be used as models of trait evolution, Brownian-motion and Ornstein-Uhlenbeck. Brownian-motion (BM)

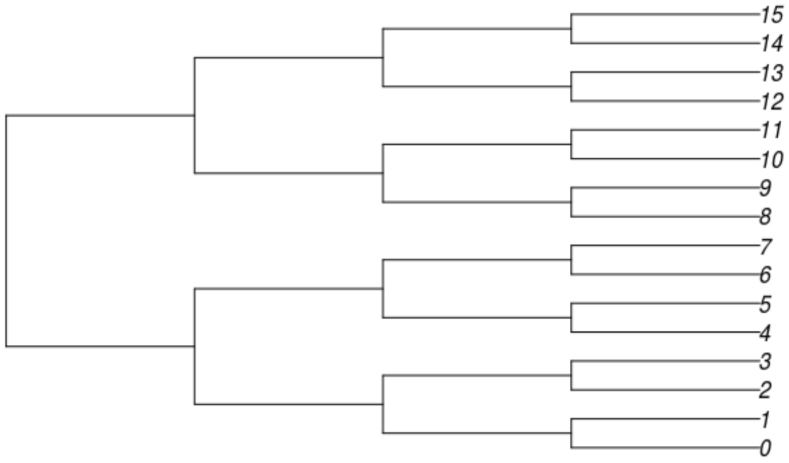

Figure 3.3: A Balanced Tree with 16 nodes (species).

is the simplest Gaussian process. BM is an extension of a random walk model (19). It does not account for natural selection or any other mitigating forces involved in trait evolution. The BM model states that traits will evolve along a tree following a Brownian-motion dynamic under which, after time t of evolution, the trait is normally distributed, centered at the ancestral value at time 0 and with variance proportional to t (20). BM deems the entire process as random in both directions and distance in a specified interval of time (17). Initially developed in physics to model the motion of particles in a fluid as a Brownian motion, these Brownian-motion dynamics can be extended to a variety of fields including finance, quantum mechanics, and in our case phylogenetic trait evolution. One appealing feature of BM models is that its easy to define and describe its statistical properties. As such, it is a widely used method of modelling phylogenetics. Brownian-motion modelling requires us to define two parameters. First, a trait value $X(t)$. At the root of our tree we have $X(0)$, the starting value of the trait at time zero with our root species.

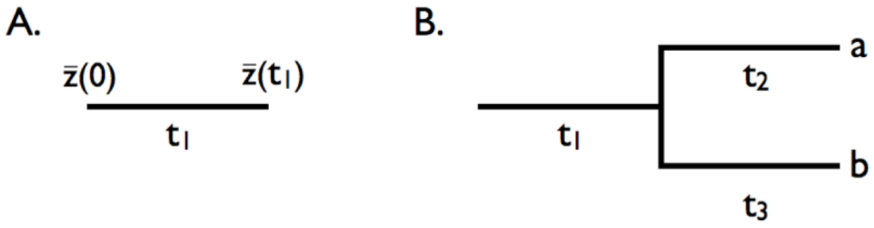

Figure 3.4: A Brownian-motion model for a simple tree in one time step.

Moving forward along any particular path from root to a tip, we consider a Brownian-motion process conditional on an $X(0)$. The second parameter is σ^2, the rate parameter at each step that determines how fast a trait shifts in a unit of time. The underlying statistical distribution associated with the BM model is a normal distribution with a mean of 0 and a variance of $\sigma^2 t$. If $X(t)$ is the value of a trait at time t, then we can derive three important properties to BM:

1. $E[X(t)] = X(0)$.
2. Independent Increments: $X(s) - X(t)$ and $X(v) - X(u)$ are independent for $u < v < t < s$.
3. $X(t) \sim N(X(0), \sigma^2 t)$.

These properties show us that the the variation along a tree from its node will increase if we increase our time interval t or if we increase our rate σ^2. All the traits will be centered around $X(0)$, the expected value of $E[X(t)] = X(0)$. A simple diagram at a single time step is provided in Figure 3.4 showing Brownian-motion. The key property to note about BM is the complexity of many small forces acting at once is reduced when we sum all those forces up. This sum follows the normal distribution. Alternatively, we can characterize BM through its multivariate normal distribution given root trait x_0, for any

fixed set of times $0 < t - 1 < \cdots t_N$.

$$[X(t_1), \cdots, X(t_N)]^T,$$

is multivariate normal with mean vector μ and covariance matrix V, where $\mu_j = x_0$ for all j and $V_{ij} = \sigma^2 min(t_i, t_j)'$. One of the main issues that arises when modelling with Brownian-motion is that the trait evolution is unbounded and does not account for natural selection pressures. This is unrealistic, and the Ornstein-Uhlenbeck (OU) process comes in to account for that. The BM process is actually a form of OU process that has the natural selection pressure set to 0 (17). If we assume that a trait has an optimal mean, the trait will converge toward that optimum at a specified rate. The parameters involved in the OU process are $X(t)$ the trait value at time t, σ^2, α the strength of the constraint towards the optimal value, and θ the optimal value of the trait (17). This process can be defined in terms of a stochastic differential equation as in (21):

$$dX(t) = -\alpha(X(t) - \mu)dt + \sigma dB_t$$

Where B_t is the BM Process. Similarly as for Brownian motion, the OU process can be characterized by its multivariate normal distribution, given root trait x_0, for any fixed set of times $0 < t_1 < \cdots < t_n$.

$$[X(t_1), \cdots, X(t_N)]^T,$$

is multivariate normal with mean μ and covariance $\mathbf{V}_{OU}$, where $\mu_j = x_0$ for all j and

$$V_{ij} = \frac{\sigma^2}{(2\alpha)} e^{-\alpha||t_j - t_i||} - e^{-\alpha(t_j + t_i)}$$

It's important to note that an OU process with an alpha parameter $\alpha = 0$, acts the same as a BM model due to the optimum having no effect on the model. We will use both OU and BM models in our analysis below and

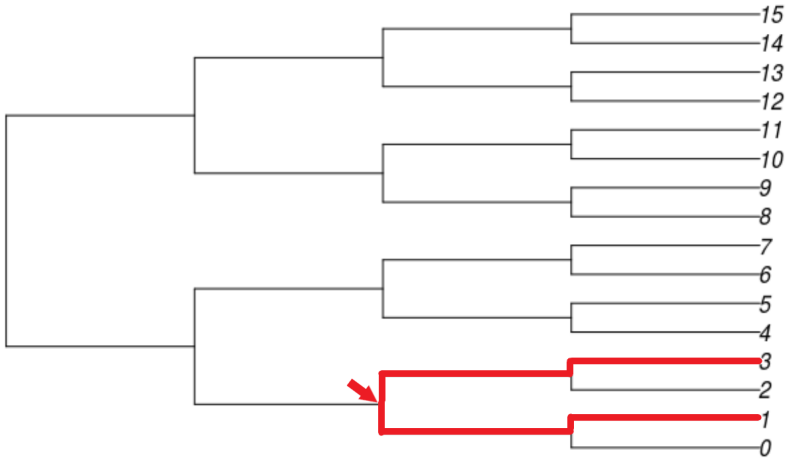

Figure 3.5: Path to shared ancestor between tip 1 and 3 along a balanced tree.

compare techniques for model selection with the two underlying models. We now consider how the Gaussian models along paths described above give rise to models for tip data, which is the only data we observe. The additional assumption is that given the trait for the last common ancestor of the pair of taxa i and j, evolution along the two separate paths from this common ancestor to taxa is independent and according to the Gaussian model. Figure 3.5 shows the location along the tree where the common ancestor occurs for tip 1 and 3. Let $y_1, \cdots, y_n$ denote the data at the tips. By the conditional independence assumption and the fact that for Gaussian processes, traits at any collection of times is multivariate normal, it follows that $[Y_1, \cdots, Y_n]$ are multivariate normal. All analysis is conditional upon the root value which we denote y_0. The means and variances of the Y_i require only knowing the model for the process from root to tip which we have described. This gives that $E[Y_j] = y_0 e^{-\alpha T} + \mu(1 - e^{-\alpha T})$ for the OU process where T is the common time

from root to tip. For the BM process, the $E[Y_j] = y_0$. The variances are each

$$\text{Var}(Y_j) = \frac{\sigma^2}{2\alpha}\{1 - e^{-2\alpha T}\}$$

To obtain the covariances we condition upon the character state Z at the last common ancestor of i and j as can be see in Figure 3.5. Let d_{ij} denote the distance between i and j. Then the time from the last common ancestor till i and j is $d_{ij}/2$. Since Y_j and Y_i are conditionally independent given Z,

$$\text{Cov}(Y_i, Y_j) = E[\text{Cov}(Y_i, Y_j|Z)] + \text{Cov}(E[Y_i|Z], E[Y_j|Z]) = \text{Cov}(E[Y_i|Z], E[Y_j|Z]) \tag{3.1}$$

For BM, $E[Y_i|Z] = Z$ so (3.1) gives that $\text{Cov}(Y_i, Y_j) = \text{Var}(Z) = \sigma^2 T_{ij}$ where T_{ij} is the time from the root to the last common ancestor of i and j. Let d_{ij} denote the distance between i and j. Then the time from the last common ancestor till i and j is $d_{ij}/2$. For the OU process, $E[Y_i|Z] = E[Y_j|Z] = Ze^{-\alpha d_{ij}/2} + \mu(1 - e^{-\alpha d_{ij}/2})$. Thus (3.1) gives that

$$\text{Cov}(Y_i, Y_j) = \text{Var}(Ze^{-\alpha d_{ij}/2} + \mu(1 - e^{-\alpha d_{ij}/2})) = e^{-\alpha d_{ij}}\text{Var}(Z) = \frac{\sigma^2}{2\alpha}\{1 - e^{-2\alpha T_{ij}}\}e^{-\alpha d_{ij}}$$

BM evolution of the trait variable along the tree results in normally distributed errors and in a covariance matrix governed by the tree structure, branch lengths and σ^2 . The covariance between two tips i and j is simply $\sigma^2 T_{ij}$, where T_{ij} is the shared time from the root of the tree to the tips (20). In summary, the model for the tip data is $\boldsymbol{Y} = \beta_0 \mathbf{1} + \boldsymbol{e}$ where $\beta_0 = e^{-\alpha T} + \mu(1 - e^{-\alpha T})$ and $\boldsymbol{e} \sim N(0, \sigma^2 \mathbf{V})$ where

$$V_{ij} = \sigma^2 T_{ij}$$

for the BM model and

$$V_{ij} = \frac{\sigma^2}{2\alpha} e^{-\alpha d_{ij}}(1 - e^{-2\alpha T_{ij}}),$$

for the OU model.

3.2 Simulation Procedure

This section outlines the simulation procedure conducted. We begin the simulation by assigning parameters to a generating model. Among the parameters are the underlying tree structure, the generating model, and the number of trait observations (tips). The trees are all normalized to the size of 1 for direct comparison. We begin by assigning the underlying tree structure and generating the covariance matrix associated with that tree. The following steps are repeated 1000 times. Using the specified tree and covariance structure, we generate data from either a BM or an OU model. There is no covariates in the data generation, only a mean value and an underlying model.

$$\mathbf{Y} \sim \mathbf{1}, \text{ BM or OU generating model where } \mathbf{e} \sim N(0, \sigma^2 \boldsymbol{V}),$$

With the generated data I fit two models, one using the true underlying model, the other with the false model. For each of these two fitted models I obtain the estimated covariance matrix $\hat{\mathbf{V}}$. The covariance for the fitted Brownian-motion model $\hat{\mathbf{V}}$ is equivalent to the covariance matrix from the tree structure. While the covariance for the fitted OU model $\hat{\mathbf{V}}_{OU}$ is equal to $1/2/\hat{\alpha} * (1 - e^{-2*\hat{\alpha}*\hat{V})} * e^{-\hat{\alpha}*D}$. We extract the AIC for both fitted models. For both the $\hat{\mathbf{V}}$ matrices I estimated above, I take the square root inverse of it $\hat{\mathbf{V}}^{-1/2}$. I use the $\hat{\mathbf{V}}^{-1/2}$ matrix to transform my observations either before splitting for cross validation or after split. This is analogous to our corrected and naive techniques described in Section 2.1.

3.2.1 Methods

Predictive Log Density Estimation for Model Selection

The performance of our techniques in selecting the true model in our phylogenetics example under performs when using the SSE as our selection criterion. As we noted previously, in the case of phylogenetic model selection, the equivalency between the Predictive Log Density and sum of square error for model selection breaks down. The predictive likelihood is defined as:

$$p(z, \hat{\theta}_y) = \frac{1}{(2\pi)\hat{\sigma}^n \left|\hat{\mathbf{V}}\right|^{1/2}} e^{(-(z-\hat{\mu})^T \hat{V}^{-1}(z-\hat{\mu})/2\hat{\sigma^2})},$$

The log predictive likelihood being:

$$log(p(z, \hat{\theta}_y)) = -\frac{n}{2}\log(2\pi) - n\log(\hat{\sigma}) - \frac{n}{2}\log\left|\hat{\mathbf{V}}\right| - (z-\hat{\mu})^T\hat{\mathbf{V}}(z-\hat{\mu})/2\hat{\sigma}^2.$$

When choosing between different generating models, BM or OU, the covariance matrix $\hat{\mathbf{V}}$ changes. This affects our equation below and the transformations no longer create an equivalency between SSE loss and EPLD. To adjust for this, we directly use EPLD to avoid bias introduced by using SSE loss. To assess the accuracy of the estimated log likelihood, we begin by defining a target Predictive Log Density for both the BM and OU model fits. This is done by generating 1000 training observations and 1000 testing observations from the generating model with parameters equal to that of the simulation. The training sets generated at this stage are also used as our data in the model selection simulation to allow for a one to one comparisons. For a given training set we fit a BM and OU model based on our parameters. That model is then used to calculate the Expected Predictive Log Density on each of the testing sets, and the mean Estimated EPLD for all the 1000 testing sets is stored. This process is repeated for each of the training sets, then averaged over test sets.

For each training set the mean calculated above is set as our target Predictive Log Density for that data set. The target predictive log density is a simulation approximation to the conditional predictive log density, $E[p(z, \hat{\theta}_y)|y]$. Cross-validation is more directly an approximation to the conditional predictive log density, which is on average the unconditional EPLD. The target is of interest in model selection performance because it provides a sort of best-case scenario for model selection using cross-validation. The contrast between its performance and the performance of cross-validation gives some sense of the effect of estimation of conditional EPLD and the goodness of conditional EPLD for model selection.

Two targets are generated one for each model fit, BM and OU. We then calculate the AIC for the fitted BM and OU models, as defined in 1.8. We extract the Expected Predictive Log Density for our corrected method by fitting models on the training set. The PLD is then calculated for the single observation. This is repeated n times and the sum for all observations becomes our Corrected method Predictive Log Density for both the BM and OU models. Comparisons are made between the four Expected Predictive Log Density estimates and the two target Expected Predictive Log Density's. The selection criterion was based on the magnitude of the EPLD. We select the model with the maximized EPLD.

Corrected Method

In the corrected method, we transform our observations y by multiplying with $\hat{\mathbf{V}}^{-1/2}$. After transformation, the data is split into two sets, training and testing. Training retains $n-1$ observations while the testing set contains 1 observations. I fit a BM and OU model using my training set. I predict on the testing observation and calculate a loss. This step is repeated n times and the loss at each iteration is summed. The loss between the two models, BM and OU, are compared and the model with the smaller loss is selected.

As described below, in the phylogenetic context, using a sum of square error loss is inappropriate. Rather than a loss we use Expected Predictive Log Density (EPLD) to select between models. The model with the largest EPLD is selected.

Naive Method

In the Naive method, the data is split into two sets, training and testing. Training retains $n-1$ observations while the testing set contains 1 observations. We then remove the tip from the tree associated with the observation from the testing set. We transform the observations y from our training set with $\hat{\mathbf{V}}^{-1/2}$ associated with each of the models, $\hat{\mathbf{V}}^{-1/2}$ and $\hat{\mathbf{V}}_{OU}^{-1/2}$. The models are fit and EPLD is calculated from the testing observation. This process is repeated n times and the EPLD at each iteration is summed. The EPLD between the two models, BM and OU, are compared and the model with the larger EPLD is selected. As in the corrected method, we use EPLD rather than SSE as the selection metric. The results of the two techniques are outlined below in Section 3.3. We compare between AIC and our two methods and show the proportion of times the true generating model being selected is given a specific tree structure, n, and α.

3.3 Phylogenetics Results

3.3.1 Corrected Method Performance

For our simulations we examine a variety of combinations of parameters and test our methods at selecting the true generating model. We set a constant number of observations at $n = 32$. The parameters we examine include two generating models, BM and OU, and three tree structures, balanced, coalescent, and caterpillar, and finally two α values associated with the OU process. This yields a total of 9 unique combination, 3 with BM generating, 3 with OU generating and $\alpha = 1$, and 3 with OU generating and $\alpha = 10$. For each of

Phylogenetics Simulation Results Predictive LogLikelihood as Selection Criterion					
Model	Tree	Corrected	AIC	Naive	target
BM	Balanced	33.8	79.8	43.3	89.6
BM	Caterpillar	59.8	82.9	49.1	85
BM	Coel	57.2	60.6	55.2	95
OU	Balanced_alpha1	94.3	55.7	48	54.8
OU	Balanced_alpha10	100	100	26.5	99.9
OU	Caterpillar_alpha1	72.2	47.9	46.9	53.9
OU	Caterpillar_alpha10	99.1	97.5	32.4	99.2
OU	Coel_alpha1	60.8	57.2	37.9	21.1
OU	Coel_alpha10	99.1	99.4	25.8	67.1

Table 3.1: Table outlines the proportion of times the true model is selected for a variety of generating models and tree structures.

those combinations we assign them as the generating model parameters for the data, and test the methods we developed at being able to select for the true model from a set of candidate models. For each generating model, I assess the performance of our techniques. Table 3.1 outlines the results of our corrected, naive, and AIC methods when using the Predictive Log Density as our selection criterion. We also test the efficacy of using the target Predictive Log Density, our corrected and naive methods are estimates of our target. As can be seen in Table 3.1, the performance of our model selection is heavily dependent on the generating parameters. The sections below outline each generating model results. For the Brownian Motion model as the generating model, our corrected method heavily under performs for model selection, regardless of the tree involved. As expected the Naive method has the worst performance from all the methods examined. Using AIC performs well for a Balanced and Caterpillar generating tree but drops off heavily for the Coalescent tree. The target EPLD performs the best across all the trees and was an expected result as

all other methods are estimates of the target itself. The Ornstein–Uhlenbeck as the generating model yields interesting results that vary across trees, and α. Our corrected method does extremely well, especially in the situation of a large α parameter. We note a drop in performance as α becomes small. In the balanced tree scenario the drop in performance is minor but a huge drop in performance is seen for the Caterpillar and Coalescent trees. We will examine these situations of performance drop in further detail below. AIC performs in a similar manner to our corrected method when α is large, but as α approaches 1 for the generating process, AIC selection proportion drops heavily. The Naive method performs terribly in all scenarios and will be ignored for any further analysis as it is obviously terrible for model selection in a phylogenetic context. An interesting result is the selection proportion for the target EPLD as a selection criterion. With large α it performs decently well but falls off heavily with smaller values of α. This is highly concerning as the target is what the other methods are designed to be estimating. This behaviour requires additional examination and may pose a significant risk for application of any of our methods. Target behaviour is discussed more thoroughly in the section below. We examine how well each of our methods did at directly estimating the target EPLD for our simulations. Boxplots were generated to inform us of the distribution of estimates by method of the target EPLD. Figure 3.6 shows us the distributions when the BM model is the generating process. The plot shows us that for this scenario the AIC and Corrected methods fitted on a BM model do very well at estimating the target EPLD value for a fitted BM model. The Corrected method outperforms the AIC method when fitted on an OU model. AIC seems to underestimate target EPLD for a OU fitted model. The Corrected method for an OU fitted model does worse when the tree structure is coalescent. The behaviour of AIC and Corrected for a BM fitted model remains the same when the generating process is OU, see Figure 3.7. In this setting our Corrected method does far better at estimating the target EPLD

for a fitted OU model compared to the AIC method. This slightly seems to break down for the balanced tree.

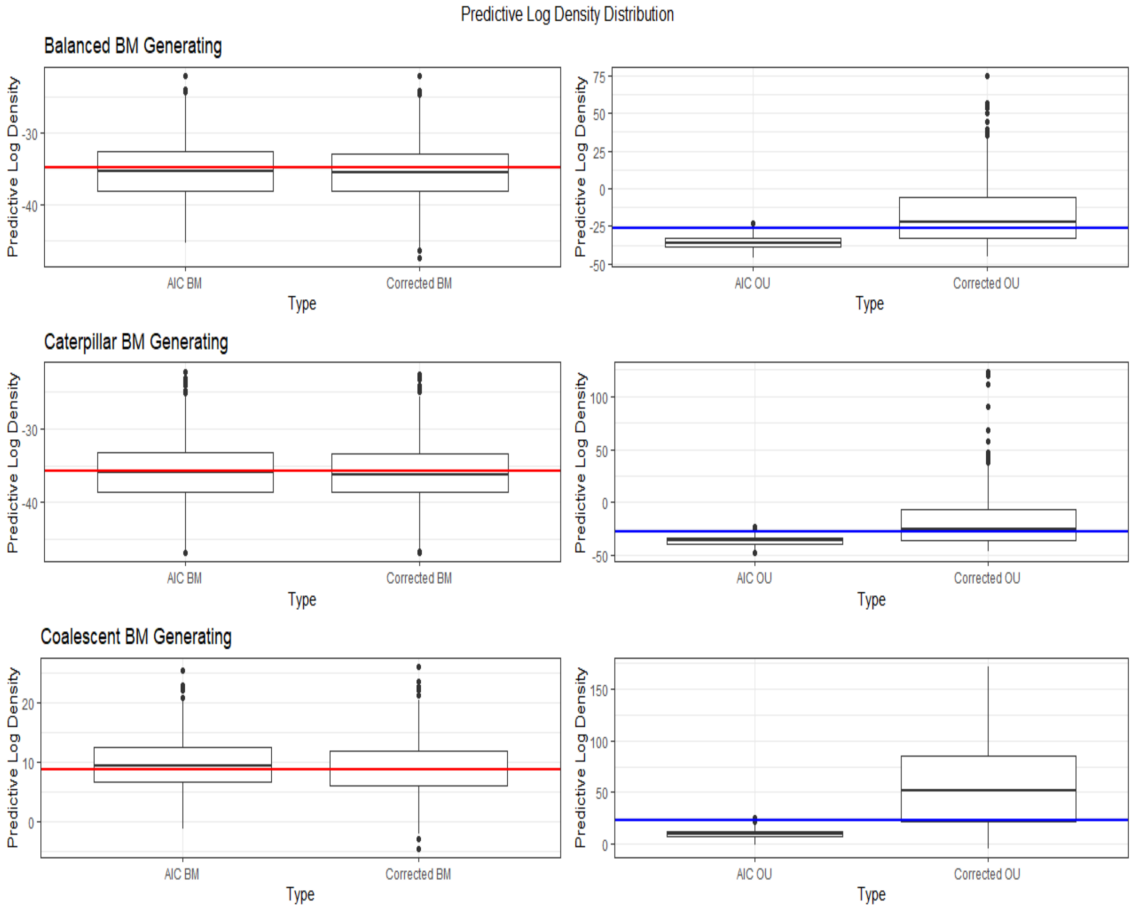

Figure 3.6: Boxplot outlining the distribution of EPLD by estimation method when the generating model is BM. Red line represents the target EPLD for the BM fitted model the methods are attempting to estimate. Blue line represents the target EPLD for the OU fitted model the methods are attempting to estimate.

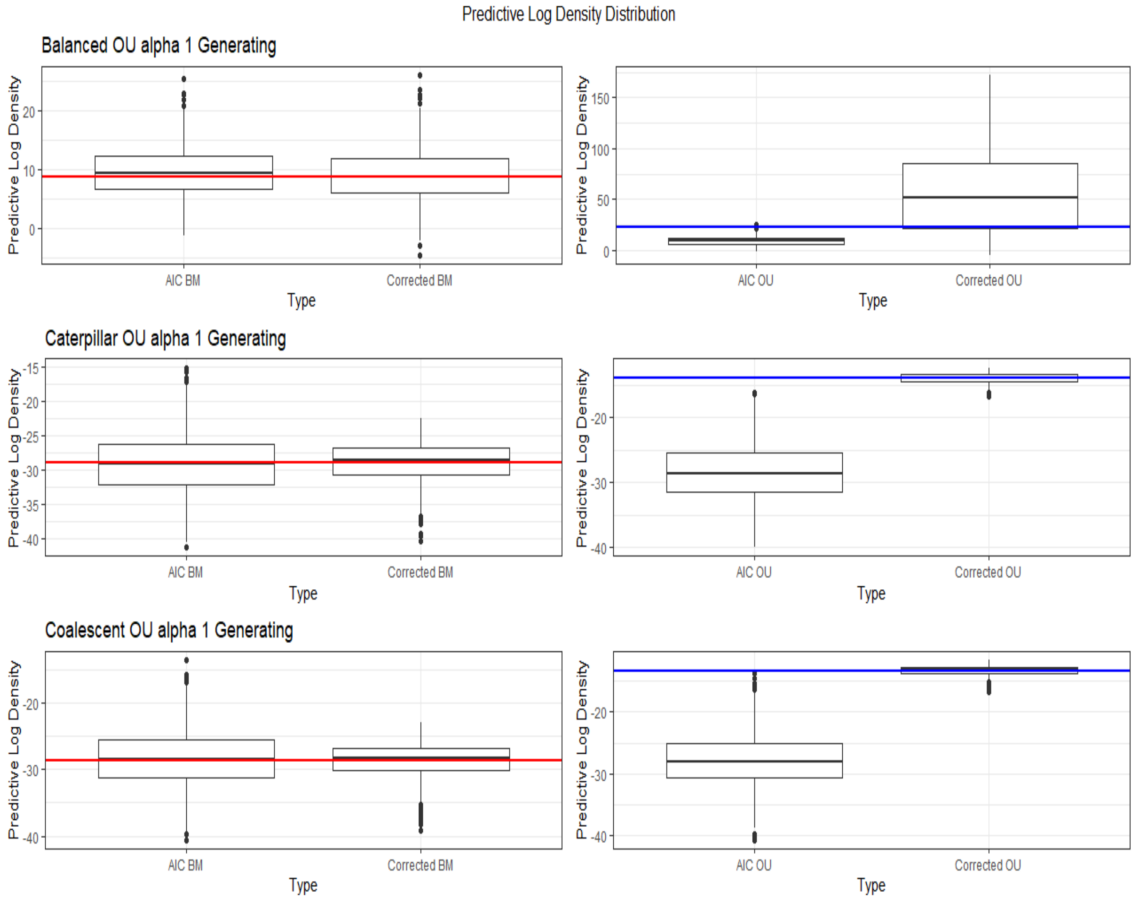

Figure 3.7: Boxplot outlining the distribution of EPLD by estimation method when the generating model is OU with $\alpha = 1$. Red line represents the target EPLD for the BM fitted model the methods are attempting to estimate. Blue line represents the target EPLD for the OU fitted model the methods are attempting to estimate.

3.3.2 Target Performance

As was shown in Table 3.1, we examine our target Predictive Log Density and its performance for selection. The target is defined as the conditional predictive log density, conditioned on an observed data set y.

$$E[p(z, \hat{\theta}_y|y)].$$

The target is generated by simulating 1000 training sets and 1000 testing sets from the same generating model. For each test set EPLD is calculated conditioned on parameter fits from a training set. The mean EPLD is retained. The process is repeated for all training sets and a mean of means is taken as our target value. In theory our target should perform best from all our techniques but the results show otherwise in certain circumstances. Its clearly a major issue as our corrected and naive methods are estimates of the target. A precise and unbiased estimate is useless if the target is inaccurate to begin with. Table 3.1 shows the target does a great job for selection when the true generating model is BM. It outperforms all methods for all tree structures. The reliability of the target drops off significantly for the generating model as OU and an α parameter that's small. Our analysis will focus on those cases.

As in the case of the corrected method, our initial assumption is that estimation of $\hat{\alpha}$ affects our estimation $\hat{\mathbf{V}}$ matrix negatively. We examine $\hat{\alpha}$ for the cases where the target did badly at selecting. The first case we examine is for the simulations where OU was the generating process, with an $\alpha = 1$ and a balanced tree. Referring to Table 3.1 in the previous section, the proportion of times our target EPLD selected the true model was 54.8%. This is in the line with the proportions for the Naive and AIC methods, but highly under performs the corrected method. This may seem an odd result, as the corrected method is an estimate of the target EPLD, its just that our corrected method overly prefers the OU model. Figure 3.8. shows the distribution of $\hat{\alpha}$ when

the target EPLD selects correctly and incorrectly. You clearly see when target does well at selecting the $\hat{\alpha}$ is small with a large portion of the density near 0 and the mean being approximately 1. This is not the case when it selects incorrectly. We note the distribution of $\hat{\alpha}$ is more right skewed, with most of the density larger than the true value $\alpha = 1$. This can be seen with the mean being approximately 4. Similar behaviour can be observed in Figure 3.9. for the $\hat{\alpha}$ with the caterpillar tree structure underlying the simulated data. For the cases where the wrong model is selected the $\hat{\alpha}$ estimates are even larger, and much of the density of those estimates overestimates $\hat{\alpha}$. This pushes the mean $\hat{\alpha}$ value to the right. This behaviour is even more prominent for the Coalescent tree scenario in Figure 3.10. The majority of the density of $\hat{\alpha}$ is larger than the true value of $\alpha = 1$. The Coalescent simulation scenario with OU generating and an $\alpha = 1$, gives us the worst performance in selecting based on target EPLD. Surprisingly the target EPLD under performs all of our methods. This is very contrary to expectations and a deeper dive into that simulation scenario is conducted below. Examining single datasets or subsets of the data for the coalescent tree simulation with an OU generating model and an $\alpha = 1$ can give us a better understanding of why the target EPLD is doing such a bad job at selecting the true model. As a first step, we extracted data sets and target EPLD values for data that had a very small $\hat{\alpha}$ value. The target EPLD for both the correct and incorrect model are essentially identical with a slight preference towards the true OU model. These diverge completely when $\hat{\alpha}$ is large and the simple BM, the incorrect model, is always selected and preferred. This tells us that an over estimation of $\hat{\alpha}$ has a serious effect on our variance-covariance matrix and thus effecting the efficacy of the target model. We now examine what effect does the large $\hat{\alpha}$ have on our $\mathbf{V}$ matrix. The first single data set we examine yielded $\hat{\alpha} = 12.21$. This value was not unusual when the generating model was OU with α=1, which is where the conditional EPLD did not perform well. The true variance of the data at the tip is $\mathbf{V} = 0.43$.

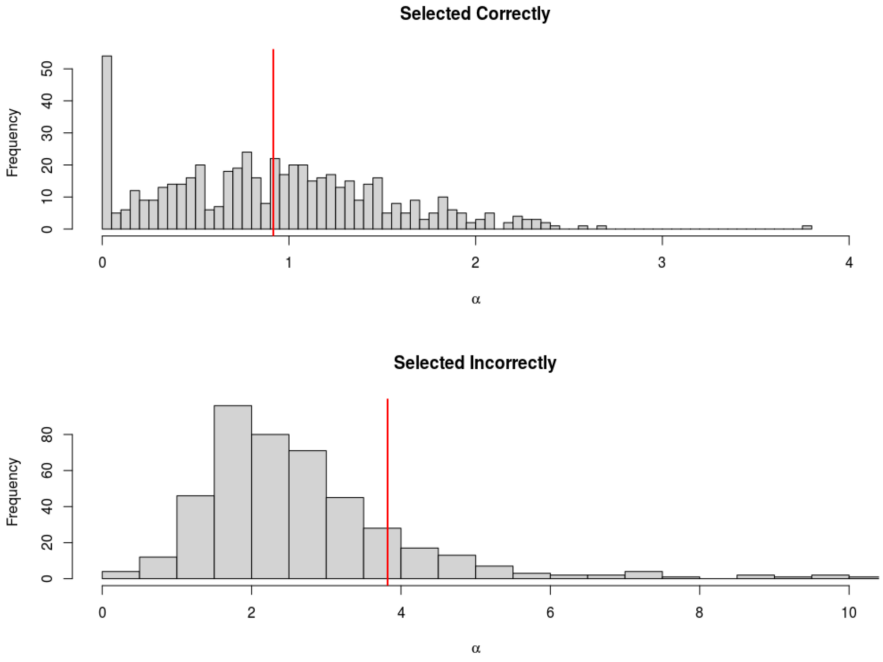

Figure 3.8: Histogram of $\hat{\alpha}$ estimates across simulations for a Balanced tree, $\alpha = 1$, and the OU model as the generating process.

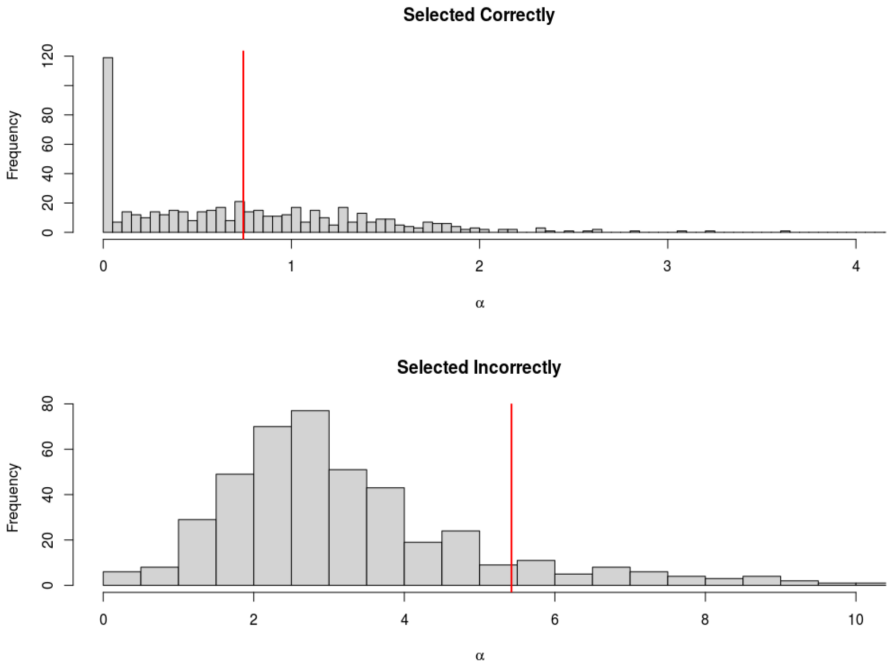

Figure 3.9: Histogram of $\hat{\alpha}$ estimates across simulations for a Caterpillar tree, $\alpha = 1$, and the OU model as the generating process.

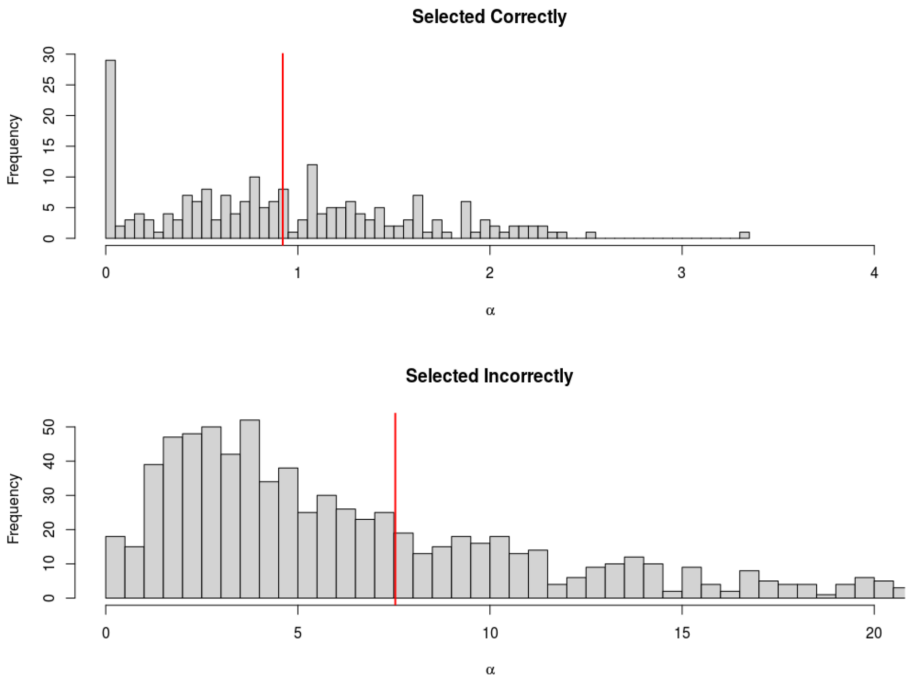

Figure 3.10: Histogram of $\hat{\alpha}$ estimates across simulations for a Coalescent tree, $\alpha = 1$, and the OU model as the generating process.

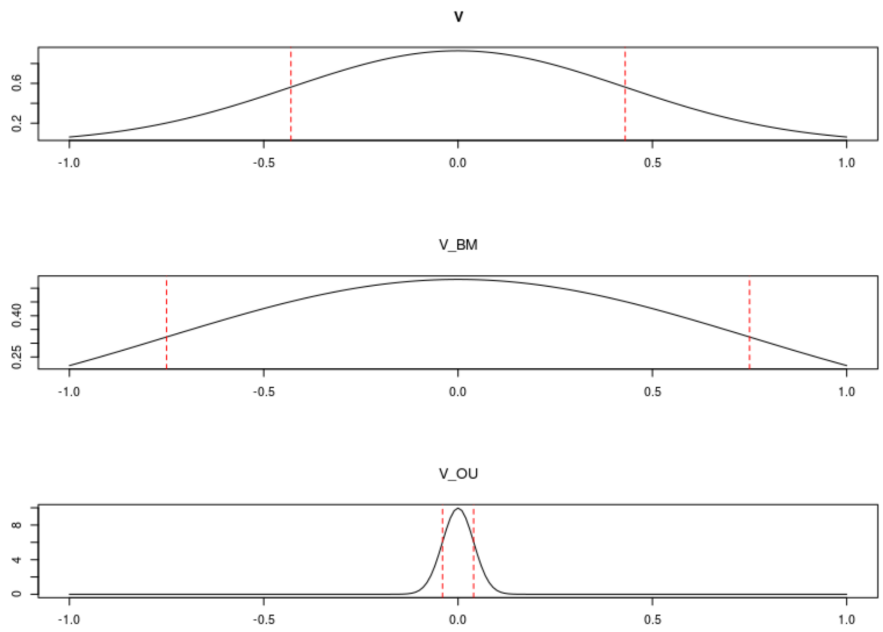

Figure 3.11: Density plot showing the differences in $\mathbf{V}, \hat{\mathbf{V}}, \hat{\mathbf{V}}_{OU}$ for overestimated $\hat{\alpha}$.

While the estimates are $\hat{\mathbf{V}} = 0.75$, $\hat{\mathbf{V}}_{OU} = 0.04$ for the BM and OU fits respectively. This shows us that with large $\hat{\alpha}$ estimates the estimated variance for the true OU fit highly undervalues the true variance at the tip. To describe this behaviour Figure 3.11. plots normal densities generated randomly with the same mean $\mu = 0$, and 3 different variances, $\mathbf{V}, \hat{\mathbf{V}}, \hat{\mathbf{V}}_{OU}$. Lets note that the effect of a bad estimate $\hat{\alpha}$ has a significant effect on $\hat{\mathbf{V}}_{OU}$, it reduces to near 0. Figure 3.11 plots lines at the value of σ from the mean. Thus, the data using the $\hat{\mathbf{V}}$ from the misspecified BM model contains a larger range but captures more data from the true range compared to the data generated from $\hat{\mathbf{V}}_{OU}$. This was examined for other data sets with large $\hat{\alpha}$ values and a similar result occurs. I repeat this analysis for $\hat{\alpha}$ values that are reasonable and within range of the true α value. The data set selected had an $\hat{\alpha} = 1.1$ and the estimates are $\hat{\mathbf{V}} = 0.61$, $\hat{\mathbf{V}}_{OU} = 0.33$ for the BM and OU fits respectively.

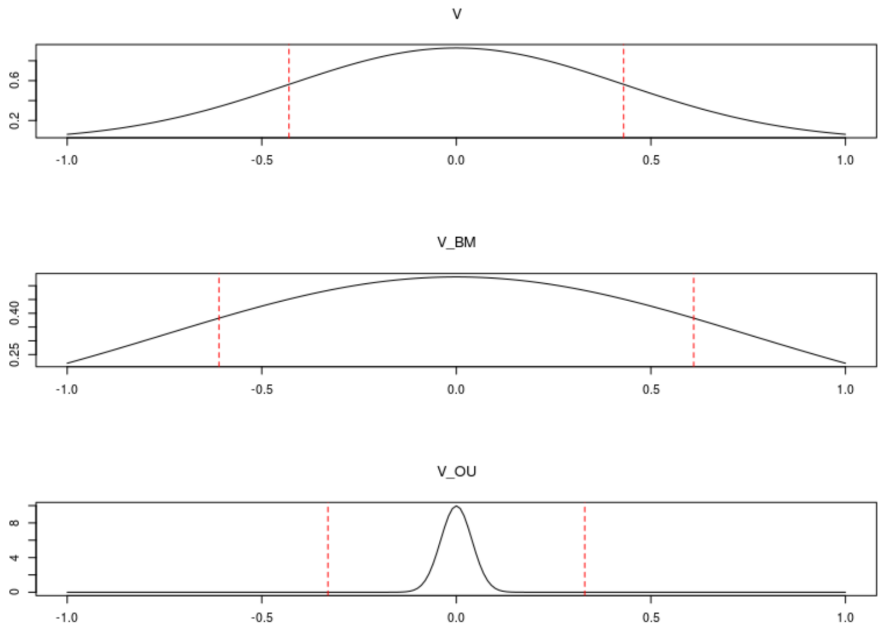

Figure 3.12: Density plot showing the differences in $\mathbf{V}, \hat{\mathbf{V}}, \hat{\mathbf{V}}_{OU}$ for appropriate $\hat{\alpha}$ estimate.

As we can see with an $\hat{\alpha}$ close to the true value, our OU fit does a much better job estimated the true Variance of the data at the tip.This was repeated with multiple datasets containing reasonable $\hat{\alpha}$ estimates and similar results were observed. Figure 3.12 shows the density plots for $\mathbf{V}, \hat{\mathbf{V}}, \hat{\mathbf{V}}_{OU}$. These plots show us that for an appropriate $\hat{\alpha}$ estimate, the estimated variance from the OU fit generates similar data to the true variance. Knowing that the incorrect estimation of $\hat{\alpha}$ leads to severely undervalued variance which in turn leads to a decrease in model selection performance, we ran the simulations with α as fixed. This allows us to gauge the magnitude of the effect of estimating the parameters in model selection. The results from these simulations confirmed our previous result and showed us that estimation of parameters can have a large effect on our target EPLD. Table 3.2 shows the proportion of times the correct model was selected across our methods for simulations with fixed parameters. The effect of fixing the parameters for our corrected and naive

Phylogenetics Simulation Results Predictive LogLikelihood as Selection Criterion and Fixed Parameters					
Model	Tree	Corrected	AIC	Naive	target
BM	Balanced	33.8	79.8	43.3	89.6
BM	Caterpillar	59.8	82.9	49.1	85
BM	Coel	57.2	60.6	55.2	95
OU	Balanced_alpha1	84.8	55.7	56.6	84.4
OU	Balanced_alpha10	98.7	100	29.5	99.8
OU	Caterpillar_alpha1	69	47.9	57.7	81.4
OU	Caterpillar_alpha10	96.5	97.5	32.3	99.3
OU	Coel_alpha1	74.3	57.2	50	64.9
OU	Coel_alpha10	96.6	99.4	31.5	97.8

Table 3.2: Table outlines the proportion of times the true model is selected for a variety of generating models and tree structures, with fixed parameters.

methods were slightly affected by fixing the parameters. The major changes occur when using the target EPLD for model selection. We note a significant increase in proportion of times the correct model is selected when using the target EPLD for selection with fixed parameters.

Chapter 4

Conclusion

This research aimed at comparing between different model selection techniques when data is correlated. We focus on correlated data with a block correlation structure. We propose an adjustment to classical Cross-Validation when data is correlated, naming it the corrected method. This method was compared against traditional CV which we named Naive and CV with a false independence assumption. We examine the efficacy of the technique for GLMs and Phylogenetic models. In the context of GLMs we found that when data was uncorrelated, all our methods produced equivalent model selection outcomes. When blocked correlation was introduced our corrected and AIC methods stood out as the most effective methods for model selection. In fact both methods produced identical results, and we have shown in Section 1.9 the relationships between our corrected CV and AIC. This was more evident as the magnitude of the correlation increased. Our initial assumptions had us convinced that this behavior would occur in an inverse manner but as we have shown with our research as correlation increased the data behaved far more like a random effects model. We saw this behavior regardless of varying the sample size and the number of blocks in the correlation matrix. In the Phylogenetics sections we compared our selection efficacy with a variety of tree structures and two generating model types, BM and OU. We identified when predictive log density is no longer equivalent to square error loss. We showed that estimating EPLD directly does a better job than using SSE. A procedure was built out to identify the target EPLD. What we found generally is our

corrected method did not perform well in many circumstances, but in some instances did better than AIC. We also found that the target EPLD was by far the best selection criterion to go by, but even the target failed at select the true model in the case of an OU generating process and a coalescent tree structure. These results are seen in Table 3.1. Our research examine why this happens, and found that when an OU process is the generating process, we estimate the $\hat{\alpha}$ parameter. The estimate $\hat{\alpha}$ had a significant effect in calculating the target EPLD, and the more it deviated from the true α value the worse the method did at selecting. The results showing the effect of incorrectly estimated $\hat{\alpha}$ are troubling and can have implications in the field of phylogenetics. The implications are vast and further research is required to understand and remedy the effect of inappropriately estimating $\hat{\alpha}$ on the variance.

Bibliography

[1] Cox D. R. (2006), Principles of Statistical Inference. Cambridge University Press, Chapter 2.

[2] Feller W. (1971), "Stochastic Independence" An Introduction to Probability Theory and Its Applications. J. Willey and Sons, Volume 2.

[3] Stoica P. & Selen Y. (2004), Model-Order Selection: A Review of Information Criterion Rules. IEEE Signal Processing Magazine, Volume 21, Issue 4, 36-47.

[4] Mardia K.V. , Kent J.T. & Bibby J.M. (1979), Multivariate Analysis. Biometrical Journal. Volume 24, Issue 5, 502-502.

[5] Cohen, J., Cohen, P., West, S.G., Aiken, L.S. (2002), Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences (3rd ed.). Routledge. 18-100.

[6] Arlot S. (2010), A survey of cross-validation procedures for model selection. Statistics Surveys. Vol. 4. 40–79.

[7] Akaike H. (1994), Implications of Informational Point of View on the Development of Statistical Science. In: Parzen E., Tanabe K., Kitagawa G. (eds) Selected Papers of Hirotugu Akaike. Springer Series in Statistics (Perspectives in Statistics). Springer, New York.

[8] Stone M. (1974), An Asymptotic Equivalence of Choice of Model by Cross-validation and Akaike's Criterion. Royal Statistical Society. Volume 36, Issue 2, 111-133.

[9] Alain, C. (2014), Optimal cross-validation in density estimation with the L^2-loss. The Annals of Statistics. 42(5) 1879-1910.

[10] Hastie T., Tibshirani R., & Friedman J. (2009),Elements of Statistical Learning: data mining, inference, and prediction. 2nd Edition. Springer Series in Statistics. 241-257.

[11] Molinaro A.M., Simon R., & Pfeiffer R. M. (2005), Prediction error estimation: a comparison of resampling methods. Bioinformatics. Volume 21, Issue 15, 3301–3307.

[12] Woese C.R., & Fox G.E. (1977). Phylogenetic structure of the procaryote domain: The primary kingdoms. Proceedings of the National Academy of Sciences. National Academy of Sciences Nov. Volume, Issue 11, 5088-5090.

[13] Hug L.A., Baker B.J., Anantharaman K., Brown C.T., Probst A.J., Castelle C.J., Butterfield C.N., Hernsdorf A.W., Amano Y., Ise K., Suzuki Y., Dudek N., Relman D.A., Finstad K.M., Amundson R., Thomas B.C., & Banfield J.F. (2016). A new view of the tree of life. Nature Microbiology. 1:16048.

[14] Hallinan N.M. (2011), Tree Shape: Phylogenies & Macroevolution. System Biology. Volume 60, Issue 6, 735-746.

[15] Felsentstein J. (2004), Inferring Phylogenies. Sinauer Associates Inc. Sunderland, Massachusetts 1-36.

[16] Dávila F.M., Domelevo E.J.B., Lemoine F., Truszkowski J., & Gascuel O. (2019), Distribution and asymptotic behavior of the phylogenetic transfer distance. Journal of Mathematical Biology. Volume 79, Issue 2, 485–508.

[17] Harmon L.J. (2018), Phylogenetic Comparative Methods learning from trees. https://doi.org/10.32942/osf.io/e3xnr

[18] Gardiner J.C., Luo Z., Roman L.A. (2009), Fixed effects, random effects and GEE: What are the differences? Statistical Medicine. Volume 28, Issue 2, 221-239.

[19] Ross S.M. (2014), Introduction to Probability Models 11th Edition. Elsevier Inc. 607-645.

[20] Lam S.T.H, & Ane C. (2013), Asymptotic Theory with Hierarchical Autocorrelation: Ornstein-Uhlenbeck Tree Models. Annals of Statistics, Volume 41, Issue 2, 957-981.

[21] Ikeda N. & Watanabe S. (1981), Stochastic Differential Equations and Diffusion Processes. North Holland Mathematical Library. Volume 24, 1-464.

[22] Liang L.,Shaoyuan W.,Lili Y. (2015), Coalescent Methods for Estimating Species Trees from Phylogenomic Data. Journal of Systematics and Evolution. Volume 53, Issue 5, 380-390.