Multivariate Applications of Bayesian Model Averaging

Robert B. Noble Jr.

Dissertation proposal submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Statistics

Eric P. Smith, Co-chair
Keying Ye, Co-Chair
Christine M. Anderson-Cook
Jeffrey B. Birch
Robert V. Foutz

November 29, 2000
Blacksburg, Virginia

Keywords: Bayesian Model Averaging, Canonical Variate Analysis,
Canonical Correlation Analysis, Principal Components Analysis

Multivariate Applications of Bayesian Model Averaging

Robert B. Noble Jr.

(ABSTRACT)

The standard methodology when building statistical models has been to use one of several algorithms to systematically search the model space for a good model. If the number of variables is small then all possible models or best subset procedures may be used, but for data sets with a large number of variables, a stepwise procedure is usually implemented. The stepwise procedure of model selection was designed for its computational efficiency and is not guaranteed to find the best model with respect to any optimality criteria. While the model selected may not be the best possible of those in the model space, commonly it is almost as good as the best model. Many times there will be several models that exist that may be competitors of the best model in terms of the selection criterion, but classical model building dictates that a single model be chosen to the exclusion of all others. An alternative to this is Bayesian model averaging (BMA), which uses the information from all models based on how well each is supported by the data.

Using BMA allows a variance component due to the uncertainty of the model selection process to be estimated. The variance of any statistic of interest is conditional on the model selected so if there is model uncertainty then variance estimates should reflect this. BMA methodology can also be used for variable assessment since the probability that a given variable is active is readily obtained from the individual model posterior probabilities.

The multivariate methods considered in this research are principal components analysis (PCA), canonical variate analysis (CVA), and canonical correlation analysis (CCA). Each method is viewed as a particular multivariate extension of univariate multiple regression. The marginal likelihood of a univariate multiple regression model has been approximated using the Bayes information criteria (BIC), hence the marginal likelihood for these multivariate extensions also makes use of this approximation.

One of the main criticisms of multivariate techniques in general is that they are difficult to interpret. To aid interpretation, BMA methodology is used to assess the contribution of each variable to the methods investigated. A second issue that is addressed is displaying of results of an analysis graphically. The goal here is to effectively convey the germane elements of an analysis when BMA is used in order to obtain a clearer picture of what conclusions should be drawn.

Finally, the model uncertainty variance component can be estimated using BMA. The variance due to model uncertainty is ignored when the standard model building tenets are used giving overly optimistic variance estimates. Even though the model attained via standard techniques may be adequate, in general, it would be difficult to argue that the chosen model is in fact "*the correct*" model. It seems more appropriate to incorporate the information from all plausible models that are well supported by the data to make decisions and to use variance estimates that account for the uncertainty in the model estimation as well as model selection.

Dedication

This project is dedicated to Dianne, Bob, and Nels.

Acknowledgments

I am grateful to the many individuals who have contributed not only to this project, but to my education in general.

In particular I thank Eric Smith, and Keying Ye, who each provided many hours of guidance and assistance throughout this project.

Committee members Christine Anderson-Cook, Jeffrey Birch, and Robert Foutz offered valued suggestions that made it possible for this research to take on it's final form.

I would also like to thank Susan Norton for supplying the data used to illustrate the methods and the U.S. Environmental Protection Agency for the grant, #CR 827820-01-0, that funded the project.

Contents

List of Tables

List of Figures

Chapter 1

Background on Model Building

1.1 Introduction

The techniques most commonly used in practice for building a linear regression model will be presented in this chapter. These popular methods result in the selection of a single model that is used for the purposes of prediction and/or description of a process. This manner of model building ignores the uncertainty that is an inherent part of the selection process. Recent advances in this area have addressed the problem using Bayesian model averaging (BMA), which assigns weights to each candidate model based on how well it is supported by the data.

Several model building techniques used for multivariate data can be viewed as special cases of multivariate linear regression which is an extension of the univariate model. The special cases considered are: principal components analysis, canonical discriminant analysis, and canonical correlation analysis. Model building is an early step in the analysis of data of these types and some of the commonly used methodologies will be discussed. Model uncertainty has not been accounted for in these multivariate settings so Bayesian model averaging is proposed as a solution.

1.2 Regression

Linear regression and its generalizations are among the most widely used methods in the sciences for finding relationships between explanatory and dependent variables. With the advent of faster and more powerful computers, researchers are collecting larger data sets both in sample size and in the number of possible predictor variables. To find relationships and develop models, researchers often use automated model selection methods and data-

mining to determine the set of covariates to be used for prediction and description of the process of interest. The selection of subsets of predictor variables is a basic part of building a linear regression model. The problem of model building and selection when using regression analysis with k independent variables has been approached in several ways.

The first method is the brute force method of fitting all combinations of possible models. A model performance criterion is computed for each model, such as *adjusted r-square*, *Mallow's Cp*, *PRESS*, or *APRESS* [47]. The model judged to be "best" based on some criterion is chosen in conjunction with also constructing a parsimonious model. Generally, evaluation of each individual model is not practical since the number of possible models grows exponentially with respect to k. Even for small values of k, say $k \approx 20$, another model selection technique must be employed since the cost in computer time is large enough so as to make the procedure impractical.

To alleviate the burden of assessing 2^k models, a compromise is attained by using a best subsets regression algorithm. One popular method is the leaps and bounds algorithm [15] which uses a relatively small number of operations to compute lower bounds for the residual sum of squares for all possible regressions without actually looking at each model individually. This is accomplished by using the fundamental inequality

$$RSS(A) \leq RSS(B)$$

where A is any set of independent variables and B is a subset of A. The efficiency of the leaps and bounds algorithm comes from the fact that once a model is deemed unacceptable, all models that are subsets of the model investigated are also removed without explicitly examining each submodel. Subset methods identify and display the best subset of K models. Popular computer packages implementing this algorithm, or some variation, allow for model evaluation based on *r-square*, *adjusted r-square* and *Mallow's Cp*. These criterion are allowed for model evaluation because these statistics are monotone functions of the *RSS* whereas statistics such as *PRESS* and *APRESS* are not. The final model is then chosen by the practitioner from the K models presented generally using some other criteria, such as simplicity or some first principles argument. Even though the number of required operations is much less than the all possible models method, this technique also becomes less practical as k grows larger.

Forward, backward, and stepwise methods are also popular and used very often in practice. These sequential techniques were designed to efficiently identify a reasonable subset of regressors in cases where k can be quite large. These procedures were motivated by computational efficiency and most likely will result in models that are not the "best" with respect to the model evaluation criterion of interest [47]. Each of these methods adds and/or drops regressor variables from the model based on an F-test.

Forward selection starts with the null model and examines the list of possible explanatory variables by performing regressions on each variable one at a time. The largest partial F-

statistic associated with the variable of interest that is greater than some critical value, F_{in}, is then entered into the model. The process is repeated by looking at the remaining variables and including each variable one at a time, in the model that (1) has the largest partial F-statistic as compared to the other candidate variables, and (2) has a partial F-statistic that exceeds F_{in}. The process stops when the partial F-statistic associated with each variable trying to enter the model is less than F_{in}.

The backward elimination procedure begins with a regression on all possible explanatory variables. Each parameter is evaluated on the basis of a partial F-statistic. If the smallest of all the partial F-statistics is found to be less than some critical value, F_{out}, then that variable is removed. The process is repeated on the remaining regressors until the smallest partial F-statistic exceeds the user specified F_{out}.

Stepwise regression combines elements of both backward elimination and forward selection. It begins like forward selection by examining the list of all possible explanatory variables and chooses the variable associated with the largest partial F-statistic that is greater than F_{in}. The list of remaining covariates is examined and as before, the variable associated with the largest partial F-statistic that is greater than F_{in} is also admitted to the model. At this point, the stepwise procedure begins to act like the backward elimination scheme. After adding the subsequent variables (after the first) to the model, each partial F is computed and if the smallest of these is less than F_{out} then the variable is removed. Once none of the remaining out-of-equation variables test as significant and all the variables in the model are judged to be necessary, the stepwise procedure terminates. It is obvious that the user must set $F_{in} \geq F_{out}$ in order for the process to eventually stop [10, 43].

In almost all instances some degree of collinearity will exist between the independent variables in a study. The effectiveness of the sequential procedures is negatively impacted as collinearity increases [46]. This condition can be eliminated by using principal components regression or ridge regression. In the case of principal components when each component is retained, no collinearity exists and the sequential procedures are guaranteed to find the same model given values F_{in} and F_{out}, and this model will be the "best" with respect to minimizing RSS for all possible models containing the same number of variables selected by the sequential procedure. A problem with this method is that all of the original possible covariates need to be kept to form the linear combinations to form the principal component scores, and this is counter to the idea of attaining a parsimonious model when rotating back to the original data space and prevents elimination of any variables.

1.3 Bayesian Model Averaging

1.3.1 Introduction

Several commonly used model building techniques were identified in the previous section. While the discussion of all model building methods was not exhaustive, a common feature to each of the procedures, and their variations, is to select a single model at the conclusion of the algorithm. A consequence of this action is the assumption that the chosen model is the correct model. Standard statistical protocol at this point is to proceed as if the selected model had generated the data. This approach ignores the uncertainty in model selection leading to over-confident inferences and decisions that are more risky than appears on the surface [22]. Bayesian model averaging (BMA) provides a mechanism to account for model uncertainty.

Suppose we want to build a regression model and have k predictors $X_1, X_2, \ldots, X_k$, and T possible models of interest. For each model we can compute the posterior probability of the model M_i given the data D using Bayes rule [16]

$$P(M_i|D) = \frac{P(D|M_i)P(M_i)}{\sum_{j=1}^{T} P(D|M_j)P(M_j)} \tag{1.1}$$

For the sake of completeness we note that all probabilities are implicitly conditional on $\mathcal{M}$, the set of all models being considered. In equation 1.1 we must compute $P(D|M_i)$ which is interpreted as the probability of the data given the model, independent of the unknown parameters, and $P(M_i)$ which is the prior probability that model M_i is the correct model. In the absence of any prior information all models are generally assumed to be equally likely, so $P(M_i) = T^{-1}$ which reduces equation 1.1 to

$$P(M_i|D) = \frac{P(D|M_i)}{\sum_{j=1}^{T} P(D|M_j)} \tag{1.2}$$

The quantity $P(D|M_i)$ is the marginal likelihood of the data and is obtained by

$$P(D|M_i) = \int P(D|\theta_i, M_i)P(\theta_i|M_i)d\theta_i \tag{1.3}$$

where θ_i is the unknown parameter vector for model M_i, $P(D|\theta_i, M_i)$ is the likelihood and $P(\theta_i|M_i)$ is the prior probability density assumed for θ_i. In normal multiple linear regression, $\theta_i = (\beta_i, \sigma^2)$ and β_i the vector of slope parameters such that

$$\beta_{ij} = \begin{cases} \beta_j & \text{if } x_j \in M_i \\ 0 & \text{if } x_j \notin M_i \end{cases}$$

Hoeting [21] showed that the integral in equation 1.3 may be evaluated analytically in some special cases [8] such as multiple linear regression assuming *iid* normal errors using conjugate normal-gamma priors as follows

$$\begin{aligned} P(D|\mu_i, V, X_i, M_i) &= \frac{\Gamma\left(\frac{\nu+n}{2}\right)(\nu\lambda)^{0.5\nu}}{\mu^{0.5n}\Gamma\left(\frac{\nu}{2}\right)|I + X_i V_i X_i'|^{0.5}} \\ &\quad \left[\lambda\nu + (Y - X_i\mu_i)'(I + X_i V_i X_i')^{-1}(Y - X_i\mu_i)\right]^{-0.5(\nu+n)} \end{aligned} \tag{1.4}$$

where X_i is the data matrix, μ_i is the prior mean vector for β, and V_i is the variance matrix for β corresponding to model i. This is an n dimensional non-central Student's t distribution with ν degrees of freedom, mean $X\mu$, and variance $[\nu/(\nu-2)]\lambda(I + XVX')$. The prior distributions for the unknown regression parameters needed to obtain equation 1.4 are $\beta \sim \mathrm{N}(\mu, \sigma^2 V)$, and $\frac{\nu\lambda}{\sigma^2} \sim \chi^2_\nu$ where ν, λ, V, and μ are hyperparameters to be chosen or calibrated based on the data.

For more general situations, Raftery [55] proposed using the *Bayes Information Criterion* (BIC) as an accurate approximation to the integrated likelihood. For normal linear regression,

$$BIC_j = n\ln(1 - r_j^2) + k_j \ln n$$

for model j where r_j^2 is the usual r-square for the model, k_j is the number of regressors in model j, and n is the number of observations. The marginal likelihood of model j is then approximated by

$$P(D|M_j) \propto e^{-0.5BIC_j} \tag{1.5}$$

After the marginal likelihood is approximated, the posterior probability of each model can be approximated using equation 1.1. The individual models are then weighted by their posterior probabilities so that the various quantities of interest may be estimated. If Δ is the quantity of interest, such as a prediction in a regression model, then its posterior distribution given data D is

$$P(\Delta|D) = \sum_{i=1}^{T} P(\Delta|M_i, D)P(M_i|D) \tag{1.6}$$

Equation 1.6 is the weighted average of the posterior distributions under each of the models considered where the weights are the posterior model probability. The posterior mean and variance of Δ are

$$\begin{aligned} E[\Delta|D] &= \sum_{i=1}^{T} E[\Delta|M_i, D]P(M_i|D) \\ Var[\Delta|D] &= E_M(Var(\Delta|D, M)) + Var_M(E(\Delta|D, M)) \end{aligned}$$

where $Var_M(E(\Delta|D,M))$ is the model uncertainty variance component [9]. Averaging over all models in this way provides better out of sample predictive ability, as measured by a logarithmic scoring rule, than any single model M_i, conditional on $\mathcal{M}$ [22].

1.3.2 Example

Suppose we have 20 observations with the following correlation structure

	X_1	X_2	X_3
Y	0.85	0.85	0.25
X_1		0.97	0.10
X_2			0.20

Five commonly used methods were used to select a model given this correlation structure. The methods used are forward, backward and stepwise regression, best adjusted r-square, and the PRESS statistic.

The three sequential methods, forward, backward, and stepwise, make use of partial F-tests. The F-statistics can be formed from the *r-square* statistic as

$$F = \frac{r^2}{1-r^2}\frac{n-p-1}{p}$$

where n is the sample size, p is the number of variables in the model, and

$$r^2 = \frac{S_{yx}S_{xx}^{-1}S_{xy}}{S_{yy}}$$

where $S_{..}$ are the sums of squares and cross product matrices, or equivalently, the corresponding correlation matrices [60]. Since we have the correlation matrix then no data is required for the sequential tests. The default values used by SAS 8.00 for entry or removal of a variable were used. Any variable will stay in the model if the p-value for the partial F-test is less than 0.10 when the backward method is used. In the forward method, variables are added if their partial F-test results in a p-value less than 0.5. For the stepwise method, a variable is added if the p-value for the partial F-test is less than 0.15 whereas any variable already in the model with a p-value greater than 0.15 is removed.

The next model selection criterion considered is based on choosing the model with the highest *adjusted r-square*. The adjusted r-square is commonly used as a model selection criterion since it accounts not only for variance explained, but also for number of variables in the model and is defined as

$$r^2_{adj} = 1-(1-r^2)\frac{n-1}{n-p-1}$$

Table 1.1: Posterior model probabilities and conventional model selection choices.

Model	$E(y)$	$-0.5BIC$	$P(M_i\|D)$	*adj r-sq*	*PRESS*	Sequential
1	β_0	0.0000	0.0000	0.0000	21.0526	
2	$\beta_0 + \beta_1 X_1$	12.2404	0.1642	0.7071	**6.5380**	Backward
3	$\beta_0 + \beta_2 X_2$	12.2404	0.1642	0.7071	6.5600	Stepwise
4	$\beta_0 + \beta_1 X_1 + \beta_2 X_2$	12.0660	0.1379	0.7022	7.1195	
5	$\beta_0 + \beta_3 X_3$	0.0664	0.0000	0.0104	22.1375	
6	$\beta_0 + \beta_1 X_1 + \beta_3 X_3$	12.7051	0.2613	**0.7206**	6.6504	
7	$\beta_0 + \beta_2 X_2 + \beta_3 X_3$	11.9047	0.1174	0.6973	7.2279	
8	$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$	12.1837	0.1551	0.7048	7.5265	Forward

Note: The *PRESS* statistic for each model is data dependent so the listed values are averages based on 500 samples. All other columns are constant for given correlation matrix.

where r^2 is the model r-square, n is the sample size, and p is the number of variables in the model [10]. This method chooses the model with the highest adjusted r-square and as with the sequential procedures, only requires either the sums of squares or correlation matrices to compute.

The final model selection criterion considered is based on the choosing the model that minimizes the *PRESS* statistic defined as

$$\text{PRESS} \;=\; \sum_{i=1}^{n} \left(\frac{e_i}{1 - h_{ii}} \right)^2$$

where e_i is the residual for the i^{th} observation and h_{ii} is the i^{th} diagonal element of the hat matrix [47]. Since this statistic is based on residuals and location of the observed values of the independent variables, then data must be generated to compute it.

The BIC based posterior model probability approximations along with the results of the various selection criteria are shown in table 1.1. We see that model 2 in table 1.1 has the lowest *PRESS* statistic and was also chosen using the backward procedure. Model 6 achieved the highest *adjusted r-square*, but the stepwise procedure chose model 3 and the forward selection resulted in model 8. It is not at all clear which model should be the final model chosen to be the "best". The model space consists of only eight models but four are deemed "best" using the different selection criteria.

The competition between models is also apparent upon examination of the posterior model probabilities. The models chosen based on *PRESS*, *adjusted r-square*, stepwise, and backward elimination each have a high posterior probability. A major benefit at this point is that we are not forced to choose which is the correct model because we will use the weighted average

Table 1.2: Sampling distribution of $\hat{\beta}_3$

Model	$E(y)$	$P(M_i\|D)$	$\hat{\beta}_3$	Std.Err.$(\hat{\beta}_3)$	Distribution
1	β_0	0.0000	0.00000	0.00000	Point
2	$\beta_0 + \beta_1 X_1$	0.1642	0.00000	0.00000	Point
3	$\beta_0 + \beta_2 X_2$	0.1642	0.00000	0.00000	Point
4	$\beta_0 + \beta_1 X_1 + \beta_2 X_2$	0.1379	0.00000	0.00000	Point
5	$\beta_0 + \beta_3 X_3$	0.0000	0.25000	0.22822	t_{18}
6	$\beta_0 + \beta_1 X_1 + \beta_3 X_3$	0.2613	0.16667	0.12188	t_{17}
7	$\beta_0 + \beta_2 X_2 + \beta_3 X_3$	0.1174	0.08333	0.12882	t_{17}
8	$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$	0.1551	0.14875	0.13845	t_{16}

of each model to estimate the quantities of interest. To illustrate this idea, suppose we want to compare the sampling distribution of $\hat{\beta}_3$ for model 6 (highest *r-square* model) with that achieved using BMA. Table 1.2 shows the distribution of $\hat{\beta}_3$ for each of the models considered. Upon choosing model 6 we have that

$$\frac{\hat{\beta}_3 - 0.16667}{0.12188} \sim t_{18}$$

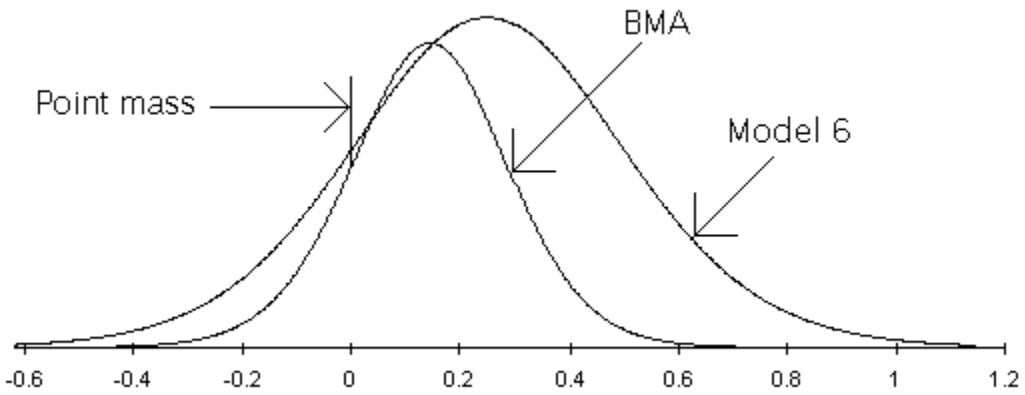
We see that the posterior probability of models 1 and 5 is zero (to four decimal places) so their weight is zero and hence do not contribute to sum. Models 2–4 are point masses at zero with total probability of 0.4662. The remainder of the density is the weighted sum comprised of the three t_ν distributions centered on their respective parameter estimate and rescaled by the standard error of the estimate. Both the BMA and the conventional sampling distributions are shown in figure 1.1.

If variable assessment is the objective, then at this point we could construct $(1-\alpha)100\%$ confidence intervals or develop interval probabilities for $\hat{\beta}_3$.

1.3.3 Implementation

Regardless of the method used to compute a posterior probability for a particular model, we see that all models must be evaluated individually. Just as in the brute force method of all possible models described in the first section, this becomes impractical for k much larger than say 20. When it becomes too difficult to enumerate and evaluate each possible model a suitable subset of the most likely models can be constructed to be used in equation 1.6. Madigan and Raftery [38] propose a method they call "Occam's window" which eliminates any model that is much less likely than the best model. Since science is an iterative process

Figure 1.1: $\hat{\beta}_3$ sampling distribution for Model 6 and BMA



in which models that predict far less well than their competitors are discarded, then they maintain that equation (1.6) should also not include models that are not supported by the data. The reduced set of models is then

$$\mathcal{A} = \left\{ M_k : \frac{\max\{P(M_l|D)\}}{P(M_k|D)} \leq c \right\}$$

for some constant c that depends on the context of the problem and may range from 10 up to perhaps 1000. The set is then further reduced by the principle of "Occam's razor". This principle would exclude any model that is less likely than any more simple submodel nested within it. Consequently the final collection of models included in equation 1.6 is

$$\mathcal{B} = \left\{ M_k : \nexists M_l \in \mathcal{A}, M_l \subset M_k, \frac{P(M_l|D)}{P(M_k|D)} > 1 \right\}$$

and equation 1.6 is replaced by

$$P(\Delta|D) = \sum_{M_i \in \mathcal{B}} P(\Delta|M_i, D) P(M_i|D) \tag{1.7}$$

Madigan and Raftery [38] outline an algorithm for implementing this procedure and a modification of the method by Raftery [55] incorporates the leaps and bounds method of model selection by Furnival and Wilson [15]. Since this method is analogous to the best subset regression discussed in Section 1.2, it also becomes impractical as k grows larger. Raftery and Volinsky (1996) published S-plus code (available in the StatLib index at http://lib.stat.cmu.edu/S/bicreg) which limits $k \leq 30$.

Madigan and York [39] proposed approximating 1.6 by using Markov chain Monte Carlo model composition (MC^3). This method generates a stochastic process which moves through model space. If $\mathcal{M}$ is the space of models under consideration, then a Markov chain $\{M(t)\}, t = 1, 2, \ldots$ can be constructed with state space $\mathcal{M}$ and stationary distribution $\Pr(M_i|D)$. Since the transition matrix is finite and can be constructed to be irreducible, then by applying the ergodic theorem, for any function $g(M_i)$ defined on $\mathcal{M}$, $E(g(M))$ can be estimated by drawing from the Markov chain for $t = 1, 2, \ldots, N$, and

$$\hat{G} \;=\; \frac{1}{N}\sum_{i=1}^{N} g(M(t))$$

which is a simulation-consistent estimate of $E(g(M))$ (i.e. $\hat{G} \to E(g(M))$ almost surely) [64, 39, 49]. Raftery, Madigan, and Hoeting [56] define a neighborhood $\text{nbd}(M)$ for each $M \in \mathcal{M}$ that consists of the model M itself and the set of models with either one variable more or one variable fewer than M. They define a transition matrix $\mathbf{q}$ by setting $\mathbf{q}(M \to M') = 0$ for all $M' \notin \text{nbd}(M)$ and $\mathbf{q}(M \to M')$ constant for all $M' \in \text{nbd}(M)$. If the chain is currently in state M, then proceed by randomly picking a model, M', from the neighborhood of M such that all models in the neighborhood are equally likely to be chosen. The model is then accepted with probability

$$\min\left\{1, \frac{P(M'|D)}{P(M|D)}\right\}$$

and the process moves to state M' otherwise the state stays in M [56, 39, 20]. Since this process is stochastic, a large number of models may be visited only a few times with the large majority of iterations taking place within only a few states. At this point we can appeal to "Occam's window" and possibly to "Occam's razor" discussed previously to eliminate those models that are not supported by the data as compared to the best model, and not supported by the data as well as one of it's nested submodels. This technique will greatly reduce the number of models in equation 1.6 so as to make it more manageable and conform to the principles and philosophies of model building.

1.4 Multivariate Models

1.4.1 Introduction

Model selection is also an important part of multivariate modeling. In this research, three multivariate techniques will be viewed as extensions of the univariate multiple regression model; they are principal components analysis (PCA), canonical variate analysis (CVA) (also known as canonical discriminant analysis), and canonical correlation analysis (CCA).

The three methods are similar in that each is a dimension reduction technique and can be viewed as a special case of multivariate regression [26]. Though there are similarities, the goals of each model are quite different. Of the three, CVA is most similar to multivariate regression because the model is identical to a one way MANOVA. The goal of CVA is to describe the group mean separation in a small number of meaningful dimensions. The variables used to form a CCA model are treated symmetrically, meaning that neither is considered dependent or independent. The purpose in this situation is to identify and describe the meaningful linear relationships that exist between the two sets of variables. PCA consists of one set of variables and is used to create a number latent variables, called principal components. The goal is to adequately describe a high dimensional variable space with a small number of latent constructs. In describing the components it becomes necessary to determine which natural variables are important in their construction.

1.4.2 Principal Components Analysis (PCA)

Principal components is a multivariate technique classified as a non-dependent variable method. The goal in principal components analysis is to create a set of orthogonal variables (components) from some given data by creating linear combinations of the original data that maximizes the variance of each new variable.

The goal is accomplished by rotating either the centered data or centered and scaled data, so that the axis along which the variance is maximum coincides with the axis of the first principal component. The next step is to rotate the data orthogonally to the first component's axis so as to maximize the remaining variance in the second principal component. This process is repeated until a zero eigenvalue is encountered or the number of components equals the number of variables in the original data.

The rotation is accomplished by using either the covariance matrix or the correlation matrix of the original data. The choice of which to use results in different solutions. If all the variables are not measured on the same scale it is common to remove the units by using the correlation matrix to perform the rotation.

Given an $(n \times k)$ data matrix $X = [\underline{x}_1 \quad \underline{x}_2 \quad \dots \quad \underline{x}_k]$ the i^{th} principal component is then equal to

$$\underline{y}_i = a_{i1}\underline{x}_1 + a_{i2}\underline{x}_2 + \cdots + a_{ik}\underline{x}_k$$

where $\underline{a}_i'\underline{a}_i = 1$ for all $i = 1, \dots, k$, and $\underline{a}_i'\underline{a}_j = 0$ for all $i \neq j$, and $var(\underline{y}_1) \geq var(\underline{y}_2) \geq \cdots \geq var(\underline{y}_k)$.

The standard procedure for maximizing a function of several variables subject to constraints is the method of *Lagrange multipliers* [13].

Using matrix notation the solution is outlined. Given an $n \times k$ matrix X, let $Z = S^{-0.5}(X - \underline{1}_n(\underline{1}_n'\underline{1}_n)^{-1}\underline{1}_n X)$ where $\underline{1}_n$ is $n \times 1$ column vector of ones and S is the sample covariance matrix of the data. Let $R = (n-1)^{-1}(Z'Z)$ be the sample correlation matrix. The i^{th} principal component is $y_i = Z\underline{e}_i$ and $var(y_i) = \lambda_i$ where $\underline{e}_i$ and λ_i are the i^{th} eigenvector and eigenvalue of R respectively.

In geometrical terms the first principal component defines the best fit line (in the least squares sense) to the k-dimensional observations in the sample, or equivalently, that it minimizes the total perpendicular sum of squares from the observations to the first component [13, 60]. The second and subsequent components are interpreted in the same with the add restriction that they are orthogonal to each of the previous components.

A common reason that PCA is performed is one of dimension reduction while maintaining a large amount of the original information. Following selection of the number of components, investigators assign some physical meaning to each of the retained components. Since each component is a linear combination of all of the original variables it becomes necessary to identify which variables are important to the construction of a component.

Many different methods have been proposed to accomplish the task of how many components to retain. Jolliffe [28, 29], Rencher [60], and Jackson [25] describe and investigate the properties of several of these methods. The techniques can be characterized by the methods that drive the particular procedure. Each method can be classified as heuristic, inferential, cluster analytic, or multiple correlation in nature. The heuristic methods include Scree graphs; discarding components with eigenvalues less than one (correlation matrix PCA); attaining a certain percent of the total variation by keeping the pc's with the largest eigenvalues; assigning one variable to each pc via the loading of the eigenvectors and then either keeping the variables that are connected with the largest eigenvalues or discarding the variables connected with the smallest eigenvalues. With respect to the inferential methods, sequential tests have been developed for testing equality of the smallest eigenvalues [24, 28], also the based on a probabilistic argument, the broken stick method has been used to identify components that should be discarded [14, 24]. The clustering methods differ from the heuristic and inferential methods because dimension reduction takes place within the variable space and not in the rotated component space. The cluster methods define a measure of association between the vectors called a link. Variables, or groups of variables that are strongly linked together form the clusters. The clustering process stops at the point where the link between all remaining clusters is below some threshold, and the number of clusters represents the number of components that are to be retained [29]. The multiple correlation methods use a linear regression approach to identify the number of components to retain.

Beale, Kendall, and Mann [3] developed interdependence analysis as an alternative to PCA and extended the use of regression to the interdependent variable situation by retaining the p variables that maximize the minimum r-square value when the p selected variables are regressed on the remaining $k - p$ variables. This method is the same as running all possible

regressions with each variable being treated in each model as either a dependent variable or an independent variable. Since this becomes impractical as k grows, Jollife [29] proposed a modification to the method by suggesting that the procedure be conducted in a stepwise regression fashion. McCabe [41] developed principal variables analysis (pva) to incorporate the ideas of model building and dimension reduction but this method does not involve a rotation to a new variable space and only removes variables that are explained adequately by the remaining ones.

After the number of components with significant structure is determined, practitioners often want to interpret the components by looking at eigenvector weights. The purpose of interpretation is to assign some meaning to these created components that matches up to some idea in reality. Variables that contribute heavily to a component are important to the construction of that component. Since the idea of component interpretation has not been addressed in the context of model building in the literature the method is developed and illustrated in chapter 2.

1.4.3 Canonical Variate Analysis (CVA)

The CVA model can be viewed from either of two equivalent points of view. The first approach takes a MANOVA perspective while the second is from a regression point of view.

Viewing the model as a one way MANOVA, suppose that a data set consists of n measurements of variables $Y_1, \cdots, Y_p$ on samples from g known populations. The MANOVA model is

$$\underline{y}_{ij} = \underline{\mu}_i + \underline{\epsilon}_{ij}, \quad i = 1, \cdots, g, \quad j = 1, \cdots, n_i$$

where

$$\begin{aligned} E[\underline{\epsilon}_{ij}] &= \underline{0} \\ Cov(\underline{\epsilon}_{ij}, \underline{\epsilon}_{i'j'}) &= \begin{cases} \Sigma & \text{if } i = i' \text{ and } j = j' \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

The "between" and "within" matrices, H and E, are then defined as

$$H = \sum_{i=1}^{g} n_i (\overline{\underline{y}}_{i\cdot} - \overline{\underline{y}}_{\cdot\cdot})(\overline{\underline{y}}_{i\cdot} - \overline{\underline{y}}_{\cdot\cdot})' \tag{1.8}$$

$$E = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\overline{\underline{y}}_{ij} - \overline{\underline{y}}_{i\cdot})(\overline{\underline{y}}_{ij} - \overline{\underline{y}}_{i\cdot})' \tag{1.9}$$

The four most commonly used statistics to test the hypothesis that the g mean vectors are equal are known as Wilks' Λ, Roy's greatest root, Pillai's trace test, and the Lawley-Hotelling

test. The test statistics for these tests can each be written in terms of the eigenvalues of $E^{-1}H$ where $\lambda_1 > \lambda_2 > \cdots > \lambda_k$, where $k = \min(p, g-1)$ and are shown below [60].

Wilk's lambda	$\Lambda = \prod_{i=1}^{k} \frac{1}{1+\lambda_i}$
Roy's root	$\theta = \frac{\lambda_1}{1+\lambda_1}$
Pillai's trace	$V = \sum_{i=1}^{k} \frac{\lambda_i}{1+\lambda_i}$
Lawley-Hotelling	$U = \sum_{i=1}^{k} \lambda_i$

Since the mean vectors are in a k-dimension space there are many possible mean configurations and none of the above tests is uniformly most powerful so all four are generally listed in the output of popular statistical software packages.

The squared canonical correlation for each canonical variate can be written in terms its associated eigenvalue as

$$r_i^2 = \frac{\lambda_i}{1+\lambda_i} \quad \text{for } i = 1, \cdots, k. \tag{1.10}$$

and is interpreted in the same way as the univariate *r-square* [27].

Alternatively, if we view that model from a multivariate regression perspective we suppose that a data set consists of n measurements of variables $Y_1, \cdots, Y_p$ on samples from g known populations. An $n \times (g-1)$ indicator matrix, X, can be constructed so that

$$X_{ij} = \begin{cases} 1 & \text{if } i^{th} \text{ observation is in group } j \\ 0 & \text{otherwise} \end{cases}$$

The linear model relating $Y = [\underline{y}_1 \ \cdots \ \underline{y}_p]$ to $X = [\underline{1} \ \underline{x}_1 \ \cdots \underline{x}_{g-1}]$ is obtained by

$$\underset{n \times p}{Y} = \underset{n \times g}{X} \ \underset{g \times p}{\beta} + \underset{n \times p}{\epsilon}$$

where

$$E(\epsilon) = \underset{n \times p}{0}$$

$$\text{cov}(\text{vec}(\epsilon)) = \underset{(p \times p)(n \times n)}{\Sigma \otimes \sigma^2 I}$$

Here Σ is a positive semi-definite matrix and $\sigma^2 I$ is the identity matrix multiplied by a scalar σ^2, and $\otimes$ stands for the Kronecker product [35]; that is the $np \times np$ matrix $\Sigma \otimes \sigma^2 I$ that can be partitioned into arrays of $p \times p$ matrices of the form

$$\begin{bmatrix} \sigma^2\Sigma & 0 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2\Sigma & 0 & 0 & \dots & 0 \\ 0 & 0 & \sigma^2\Sigma & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \sigma^2\Sigma \end{bmatrix}$$

This equation is the natural extension of the univariate linear model to the multivariate case where there are p dependent variates. The residuals of each p-variate observation are assumed to be independent, but since the different responses may be correlated, the residuals of the individual y values have unknown covariance matrix Σ. The standard test for $H_0 : \beta = 0$ is the likelihood ratio test where

$$\begin{aligned} \Lambda &= \frac{|Y'Y - \hat{B}'X'Y|}{|Y'Y - n^{-1}\overline{yy}'|} \\ &= \frac{|E|}{|E+H|} \\ &= \frac{1}{|I + E^{-1}H|} \\ &= |(I + E^{-1}H)^{-1}| \end{aligned}$$

now, if the eigenvalues of $E^{-1}H$ are $\lambda_1 > \cdots > \lambda_k$ then the eigenvalues of Λ are $\frac{1}{1+\lambda_1} < \cdots < \frac{1}{1+\lambda_k}$ hence we see that the likelihood ratio test is equal to the Wilks' lambda test. From the equation 1.10 we then have that the eigenvalues of Λ are equal to $(1 - r_1^2) < \cdots < (1 - r_k^2)$.

As with the one way MANOVA model, the three other multivariate test statistics previously described can also be used to test $H_0 : \beta = 0$.

The intention of CVA is descriptive in nature [60] and may be characterized by any or all of the following:

- examine separation of the groups in a two dimensional plot,
- find a subset of the original variables that separates the groups almost as well as the entire original set,
- rank the variables in terms of their relative contribution to group separation,
- interpret the new dimensions represented by the discriminant functions.

Within the general model building framework, approximate partial F-statistics can be constructed from the Wilks' Λ statistic. The partial F-values are not associated with a single dimension of group separation, but constitutes an overall contribution of a particular variable to discriminate between the means [60]. Forward selection, backward elimination, and particularly stepwise discriminant analysis are the most commonly used tools for ranking variables and significance testing to screen which variables should potentially be included in the final model [46, 19, 60].

Significance testing of variables can also be done using multivariate regression. Constructing the dependent variable matrix so that each response vector corresponds to one of the $g-1$ degrees of freedom can be done as follows:

$$y_{ij} = \begin{cases} 1 & \text{if } j^{th} \text{ observation is in group } i \\ 0 & \text{otherwise} \end{cases} \quad i = 1, \ldots, g-1, \qquad j = 1, \ldots, n$$

Assuming normal errors, the usual test of $\beta = \mathbf{0}$ for multivariate regression is the likelihood ratio test which is exactly Wilks' Λ [60]. Consequently, model building for discriminant analysis is exactly that of model building for multivariate regression, using dummy variables for the dependent variables.

The model building procedures and inferences that can be drawn are analogous to those discussed for univariate [12]. The differences between the two models are in how the resulting model is interpreted. Using the definitions of H and E from equations 1.8 and 1.9, we obtain the discriminant functions as the eigenvectors of $E^{-1}H$. The relative importance of each discriminant function as it pertains to group mean separation is measured by the size of the associated eigenvalue. Everitt and Dunn [13] show that the significance of each canonical discriminant function can be tested by constructing a sequence of Wilks' lambda tests based on the relative decreasing size of the eigenvalues. It is generally desired that the group separation can adequately be described in as few as two dimensions.

If a variable does not significantly contribute to the separation of the group means then it should be discarded from the model. In the special case where $g = 2$ then Mardia [40] describes a test using the *Mahalanobis distance* between the groups. In the more general case of $g > 2$ all pairwise comparisons between the groups inflates the Type I error rate so a more appropriate test should be used [54].

By extending the principles of Bayesian model averaging to discriminant analysis we will account for the uncertainty in the model selection process just as was done in the univariate regression case previously discussed in section 1.3. The posterior probability for any given model will be estimated using the BIC as in the univariate case.

1.4.4 Canonical Correlation Analysis (CCA)

Canonical correlation analysis is concerned with the amount of linear relationship between two sets of variables. If $p+q$ variables are measured on n objects, then the data may be split into two matrices X and Y with dimensions $n \times p$ and $n \times q$ respectively. In this instance, X and Y have a symmetric relationship in that neither represents the dependent variable, so without loss of generality, let $p \leq q$. If R is the overall correlation matrix then it can be partitioned into four submatrices as

$$R = \begin{bmatrix} R_{xx} & R_{xy} \\ R_{yx} & R_{yy} \end{bmatrix} \tag{1.11}$$

where R_{xx} and R_{yy} represent the correlations or interdependencies within X and Y respectively, and $R_{xy} = R'_{yx}$ represents the correlations between X and Y. The problem is to assess the between group correlation structure assuming that R_{xy} contains at least some nonzero entries [37].

The i^{th} canonical variates are linear combinations $\underline{a}_i$ and $\underline{b}_i$ of the variables in X and Y such that if $U_i = \underline{a}'_i X$ and $V_i = \underline{b}'_i Y$, then $corr(U_i, V_i) = r_i$ is maximized in absolute magnitude.

If R_{xx} and R_{yy} are of full rank then let

$$\begin{aligned} Z_p &= R_{xx}^{-0.5} R_{xy} R_{yy}^{-1} R_{yx} R_{xx}^{-0.5} \\ Z_q &= R_{yy}^{-0.5} R_{yx} R_{xx}^{-1} R_{xy} R_{yy}^{-0.5} \end{aligned}$$

Let $\underline{a}'_i = \underline{e}'_i R_{xx}^{-0.5}$ and $\underline{b}'_i = \underline{f}'_i R_{yy}^{-0.5}$ where $\underline{e}_i$ and $\underline{f}_i$ are the i^{th} eigenvectors of Z_p and Z_q respectively, and the canonical correlations are the corresponding eigenvalues of Z_p, so $r_1^2 \geq r_2^2 \geq \cdots \geq r_p^2$ [26].

Canonical variates have the following properties:

$$\begin{aligned} var(U_i) = var(V_i) &= 1 \\ corr(U_i, U_j) &= 0 \quad i \neq j \\ corr(V_i, V_j) &= 0 \quad i \neq j \\ corr(U_i, V_j) &= 0 \quad i \neq j \end{aligned}$$

There is a direct link between the canonical variate coefficients and multivariate regression coefficients. The matrix of the regression coefficients of the y's regressed on the x's (corrected for their means) can be written as $\hat{\beta} = S_{xx}^{-1} S_{xy}$. This matrix can be used to relate $\underline{a}_i$ and $\underline{b}_i$: $\underline{b}_i = \hat{\beta} \underline{a}_i$. By regression the x's on the y's we can obtain a similar relationship between $\underline{a}_i$ and $\underline{b}_i$: $\underline{a}_i = S S_{yy}^{-1} S_{yx} \underline{b}_i$ [60].

One of the primary goals after running a CCA is interpretation of the variates. Three common tools to aid in interpretation of the canonical variates are: (1) standardized coefficients,

(2) correlation between the original variable and the canonical variate, and (3) rotation of the canonical variate coefficients [60].

Johnson and Wichern [26] and Rencher [60] warn against using the correlation between the original variable and the canonical variate since they provide univariate information only and do not indicate how the variables contribute jointly to the analysis. Rencher [59] shows that rotation introduces correlations between the canonical variates so the gain in interpretability is offset by the increase in complexity caused by the interrelationships between the canonical variates.

The assumption that each variable in X and Y contributes in some way and should be included in the analysis may not be justified. If its inclusion makes interpretation more difficult since its contribution may be on the magnitude of sample error then the variable should be eliminated. BMA will provide a mechanism to identify sets of matrices $(X_1, Y_1), \ldots, (X_T, Y_T)$, where $T = 2^{p+q}$, that are supported by the data and also identify those combinations of the variables that are not supported by the data. Using this information will account for the uncertainty of whether a variable is useful in connecting X and Y. Also, changes in the canonical variates can indicate how confident we can be that the interpretations made are reasonable.

Chapter 2

Principal Components Analysis (PCA)

2.1 Introduction

2.1.1 Background

In principal components analysis (PCA), the original variables in a multivariate data set are rotated and new variables called *principal components* are created. The method of rotation used in PCA has many optimal properties [41] but is generally performed because of two desirable properties. Firstly, the components are orthogonal, implying independence of these new variables under the assumption of multivariate normality. Secondly, the new axes represent directions of maximum variability. This second property is interpreted as providing a more parsimonious description of the data because any component associated with a small amount of the total variability may be discarded without a substantial loss of information.

It is important to note that the parsimony previously mentioned is in terms of the principal components only since each sample component is made up of a linear combination of *all* of the original variables. It is often desirable to ascribe some interpretation to each of the retained components via inspection of the eigenvector weights associated with each of the original variables. Large magnitude loadings are interpreted as the variable in question being important with respect to construction of the given component. When principal components is performed on the covariance matrix, Anderson [1] and Girshik [17] derived the large sample distribution theory for the sample eigenvalues and associated eigenvectors hence formal tests may be constructed to test for zero contribution to a given component. The large sample results do not extend to the case where principal components is performed

using the correlation matrix [25]. The cause of this complication is twofold; the correlations are functions of the elements of the covariance matrix, and the sum of the eigenvalues is equal to the number of variables. To circumvent the need to rely on large sample results and distributional assumptions, Lambert *et al.* [36] proposed using a bootstrap solution to the problem. Empirical distributions were formed from the bootstrap sample for each eigenvalue and eigenvector element. The number of "meaningful" components was determined based on the Guttman-Kaiser criteria, and the eigenvectors elements whose bootstrapped empirical interval did not include zero identified.

2.1.2 Outline

In subsequent sections of this chapter, limitations of the existing methods of model building with respect to PCA are discussed, and improvements proposed to overcome these limitations. The notion of model building within the context of principal components analysis is described. Bayesian model averaging (BMA) and Markov chain Monte Carlo model composition (MC^3) is introduced and applied to principal components analysis. The properties of the PCA model obtained using BMA are investigated via a power study performed on several patterned correlation matrices. Finally, the BMAPCA method is used to analyze a set of environmental data [51].

2.2 Limitations

Commonly in practice, when eigenvectors from a PCA are interpreted, only the magnitude of the point estimate of the individual elements is considered and there is no measure of variability to formalize any decision. Lambert *et al.* [36] proposed using a bootstrap solution to the problem of relying only on point estimates. The individual observations are sampled with replacement for each iteration of the bootstrap simulation. A PCA is run on each bootstrap sample and the sample eigenvectors and eigenvalues are recorded. Empirical $(1-\alpha)100\%$ confidence intervals are then constructed for each eigenvector element and eigenvalue by trimming $0.5\alpha\%$ of the sample values from each tail of the simulation distribution. With each bootstrap sample there is potential for the eigenvectors to change signs and for individual components to swap positions with respect to the original data set. While Lambert addressed the issue of eigenvectors potentially reversing signs, there was no mention of the problem of component position swapping. There then must be an algorithm in place to insure that the eigenvectors generated from each bootstrap sample are assigned to the "most correct" component based on some criterion. Given original data eigenvectors $\underline{e}_1, \cdots, \underline{e}_p$ with eigenvalues $l_1 > \cdots > l_p$ and bootstrap eigenvectors $\underline{b}_1, \cdots, \underline{b}_p$, we propose matching the bootstrap eigenvectors to the eigenvectors obtained from the full sample so

that

$$\sum_{i=1}^{p} l_i |\underline{e}_i' \underline{b}_i|$$

is maximized which has a range from 0 to p. It weights each bootstrap eigenvector assignment by the full sample eigenvalue magnitude thus putting more emphasis on agreement in the larger components. The occurrence of component swapping depends on the degree of overlap in the distributions of the eigenvalues connected to the component. If there is a large amount of overlap then component swapping is a much larger concern.

While the bootstrap does allow for empirical confidence intervals to be created, it is implemented in an intrinsically univariate manner by considering each empirical confidence interval individually. Since each element within an eigenvector is correlated with the other elements [26] it is possible that one variable may contribute significantly only if another is excluded, and vice versa. Also, since no variables are ever actually being excluded from the model when using the bootstrap there really is no model or variable selection.

2.3 Model selection

Beale *et al.* [3], Jolliffe [29], and McCabe [41] discuss variable selection for nondependent variable data sets. Beale *et al.* developed Interdependence Analysis (IA) and Jolliffe investigated the various model selection strategies pertaining to it. McCabe introduced the method of Principal Variables Analysis (PVA) and outlined variable selection strategies. Both IA and PVA do not use a rotation as PCA does, but instead uses the original variables and regression methods to remove redundant variables thereby creating a parsimonious data set. The notion of model building and variable selection in the context of PCA has not been addressed as such in the literature, but when eigenvector weights are interpreted, variable selection has in fact occurred for each component interpreted.

If a subset of the original variables contribute significantly to a given component then the eigenvector weights and hence PC scores may be better estimated if the noncontributing variables were not present during estimation.

A measure of the impact that a variable or group of variables has on a particular principal component is how much the proportion of variance explained by that component changes with various variable configurations (i.e. models). Let R be the sample correlation matrix obtained from the measured variables $X_1, \cdots, X_p$. Now, R can be rewritten in terms of its spectral decomposition as

$$\begin{aligned} R &= ELE' \\ &= \sum_{i=1}^{p} \underline{e}_i l_i \underline{e}_i' \end{aligned}$$

where E is the matrix composed of eigenvectors $\underline{e}_1, \cdots, \underline{e}_p$, and L is the diagonal matrix of eigenvalues with diagonal elements $l_1 \geq l_2 \geq \cdots \geq l_p \geq 0$ [60]. Each component, after the first, is conditioned on all previous components since the variance is maximized subject to the constraint that the component in question be orthogonal to previously determined components [28]. With this in mind, it is important to note that if a variable does not contribute significantly to the second component, for example, it may be a major contributor to the first component. Hence any variables impact on previous components must be preserved while investigating subsequent components. To accomplish this we make use of the spectral decomposition of R in a sequential manner. Suppose the jth component is under investigation, so to preserve the structure in the previous $j-1$ components we construct R_j by

$$\begin{aligned} R_j &= R - \sum_{i=1}^{j-1} \underline{e}_i l_i \underline{e}_i' \\ &= \sum_{i=j}^{p} \underline{e}_i l_i \underline{e}_i' \end{aligned}$$

and the eigenvalues of R_j are $l_j \geq l_{j+1} \geq \cdots \geq l_p \geq 0$ with associated eigenvectors $\underline{e}_j, \cdots, \underline{e}_p$.

A model is specified by the variables contained in it. The goal is to identify models that capture the significant portion of the eigenvector structure while eliminating those variables that play only a spurious role in the construction. The model space $\mathcal{M}$ is made up of the 2^p possible variable configurations for each component. Model $M_i \in \mathcal{M}$ can be represented by $\underline{\delta}_i$, a p dimensional column vector such that

$$\delta_{ik} = \begin{cases} 1 & \text{if } X_k \in M_i \\ 0 & \text{if } X_k \notin M_i \end{cases} \qquad \text{for } k = 1, \cdots, p$$

so the model index number "i" is

$$i = \sum_{j=1}^{p} 2^{j-1} \delta_{ij}$$

For example, if $p = 6$ then the model with variables X_2, X_4 and X_5 has $\underline{\delta}' = [0 \;\; 1 \;\; 0 \;\; 1 \;\; 1 \;\; 0]$ and $i = 2^{2-1} + 2^{4-1} + 2^{5-1} = 26$ hence M_{26} is the designation for this configuration.

To impose any model, M_i, onto R_j we construct R_j^i as

$$R_j^i = diag(\underline{\delta}_i) R_j diag(\underline{\delta}_i)$$

where

$$diag(\underline{\delta}_i) = \begin{bmatrix} \delta_{1i} & 0 & \cdots & 0 \\ 0 & \delta_{2i} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \delta_{pi} \end{bmatrix}$$

This construction essentially zeroes out the rows and columns of R_j corresponding to each $X_k \notin M_i$. This is important since the effect of the variables that are not in M_i are eliminated from R_j, but the dimension of R_j^i is still $p \times p$.

The proportion of variance explained by the jth component given model M_i and the previous components is then equal to

$$r_i^2 = \frac{l^*}{\sum_{k=j}^{p} l_k} \tag{2.1}$$

where l^* is the eigenvalue of R_j^i whose associated eigenvector has the largest magnitude correlation with the first eigenvector of R_j and $l_j, \cdots l_p$ are the eigenvalues of R_j. The rational for using l^* comes from the fact that if the variables in M_i are actually important to the jth component, then the components of R_j^i are likely to swap places so it is necessary to use the eigenvalue of R_j^i whose associated eigenvector is most in line with the first eigenvector of R_j (i.e. the jth eigenvector or R). The denominator of equation 2.1 is the total variance minus that accounted for in the previous $j-1$ components.

For example, suppose

$$R = \begin{bmatrix} 1 & 0.8 & 0.5 & 0.4 \\ 0.8 & 1 & 0.4 & 0.3 \\ 0.5 & 0.4 & 1 & 0.0 \\ 0.4 & 0.3 & 0.0 & 1 \end{bmatrix}$$

and we wish to determine which variables contribute significantly to the construction of the second component. The eigenvector associated with the second component is

$$\underline{e}_2' = [0.010 \quad -0.012 \quad -0.612 \quad 0.790]$$

It appears variables that X_3 and X_4 are most important in the construction of this component. Given there are four variables, the table below shows the results for the sixteen possible models. The results shown in table 2.1 show that the models can be grouped into basically three categories; those that contain X_3 and X_4 (i.e. $M_{12}, \cdots, M_{15}$), those that contain either X_3 or X_4 (i.e. $M_4, \cdots, M_{11}$), and those that contain neither X_3 or X_4 (i.e. $M_0, \cdots, M_3$). Using the r-square measure shown in equation 2.1, we see that the models that contain both X_3 and X_4 explain a substantially larger proportion of the variance than models in the other two groups. Also, in the models where both X_3 and X_4 are excluded, none of the variance is accounted for (up to four decimals of precision).

2.4 Bayesian Model Averaging (BMA)

When a model selection method such as an all possible regressions or stepwise procedure is used, the single model obtained is assumed to be the *correct* model. All future inferences

Table 2.1: Example Model *r-square* Measures

Model	Active Variables				r_i^2
M_{15}	X_1	X_2	X_3	X_4	0.5888
M_{14}		X_2	X_3	X_4	0.5888
M_{13}	X_1		X_3	X_4	0.5887
M_{12}			X_3	X_4	0.5887
M_{11}	X_1	X_2		X_4	0.4674
M_{10}		X_2		X_4	0.4650
M_9	X_1			X_4	0.4497
M_8				X_4	0.4462
M_7	X_1	X_2	X_3		0.3938
M_6		X_2	X_3		0.3873
M_5	X_1		X_3		0.3637
M_4			X_3		0.3536
M_3	X_1	X_2			0.0000
M_2		X_2			0.0000
M_1	X_1				0.0000
M_0		Null			0.0000

and predictions that are made with the model do not account for the uncertainty involved in the selection process. Alternatively, the model obtained using BMA does incorporate the variance component associated with the uncertainty of model building.

There is a standard Bayesian solution to the problem of accounting for model uncertainty. If the model space is $\mathcal{M} = \{M_1, \cdots, M_T\}$ then the posterior probability of M_i given the data $\mathbf{X}$ is given by

$$P(M_i|\mathbf{X}) = \frac{P(\mathbf{X}|M_i)P(M_i)}{\sum_{M_j \in \mathcal{M}} P(\mathbf{X}|M_j)P(M_j)} \tag{2.2}$$

where $P(M_i)$ denotes the prior probability of each model and $P(\mathbf{X}|M_i)$ is the marginal likelihood of the data. Generally, each model has been assumed to be equally likely *a priori*, so equation 2.2 simplifies to

$$P(M_i|\mathbf{X}) = \frac{P(\mathbf{X}|M_i)}{\sum_{M_j \in \mathcal{M}} P(\mathbf{X}|M_j)}$$

Now, the marginal likelihood of the data is

$$P(\mathbf{X}|M_i) = \int P(\mathbf{X}|M_i, \underline{\theta}_i)\pi(\underline{\theta}_i)d\underline{\theta}_i \tag{2.3}$$

where $\underline{\theta}_i$ is the unknown model parameters with joint prior density $\pi(\underline{\theta}_i)$. Hoeting [21] shows for univariate multiple regression that the marginal likelihood follows an n dimensional non-central Student's t-distribution when proper conjugate priors are used. Using the multivariate non-central t-distribution can be used if hyperparameters are chosen so that the prior density it calibrated to the data. Raftery [55] approximates equation 2.3 using the *Bayes Information Criterion* (BIC) which is a penalized likelihood measure. For the case of linear regression Raftery shows that

$$\begin{aligned} P(\mathbf{X}|M_i) &\propto \exp\left(-0.5BIC_i\right) \\ &= \exp\left(-0.5(n\ln(1-r_i^2)+p_i\ln n)\right) \end{aligned} \tag{2.4}$$

where r_i^2 is the model *r-square*, p_i is the number of independent variables in model M_i, and n is the number of observations. Using Raftery's approximation requires no calibration to the data and is composed of readily available regression information. By normalizing we get the posterior probability of model M_i given the data is

$$P(M_i|\mathbf{X}) \approx \frac{\exp\left(-0.5BIC_i\right)}{\sum_{M_j\in\mathcal{M}}\exp\left(-0.5BIC_j\right)} \tag{2.5}$$

In the context of linear regression, any model deemed to be less likely than the intercept only model represents a subset of the variables that is less desireable than doing nothing at all. The models that do have value are those that are better than the null model. If a model, M_i has a posterior probability greater than the null model then

$$\begin{aligned} P(M_i|X) &> P(M_0|X) \\ \exp(-0.5BIC_i) &> \exp(-0.5BIC_0) \\ -0.5BIC_i &> -0.5BIC_0 \\ BIC_i &< BIC_0 \\ n\ln(1-r_i^2)+p_i\ln n &< n\ln(1-r_0^2)+p_0\ln n \end{aligned}$$

Now, $r_0^2=0$ and $p_0=0$, so, solving for r_i^2 we get

$$\begin{aligned} n\ln(1-r_i^2)+p_i\ln n &< 0 \\ n\ln(1-r_i^2) &< -p_i\ln n \\ \ln(1-r_i^2) &< \frac{-p_i}{n}\ln n \\ 1-r_i^2 &< \exp\left\{\frac{-p_i}{n}\ln n\right\} \\ r_i^2 &> 1-\exp\left\{\frac{-p_i}{n}\ln n\right\} \\ r_i^2 &> 1-n^{-p_i/n} \end{aligned}$$

So any model with an r-square greater than $1-n^{-p_i/n}$ will be more likely than the null model whereas if the r-square is smaller the model would most likely be discarded due to its poor performance. For a PCA application, the proportion of variance explained, from equation 2.1, is analogous to the r-square of a regression model. We want to allow for the possibility that all p variables may significantly contribute to a given component so if we model the first component, for example, replacing model r-square with proportion of variance explained using the eigenvalues, we have

$$\begin{aligned} \frac{\lambda_1}{\sum_{i=1}^{p}\lambda_i} &> 1-n^{-p/n} \\ p^{-1}\lambda_1 &> 1-n^{-p/n} \\ \lambda_1 &> p-pn^{-p/n} \end{aligned}$$

Suppose for example that $p=20$ and $n=250$ then if $l_1 < 7.141$ then the full model would be less likely than the null model even though it is quite plausible that each variable may significantly contribution to a component associated with such a large eigenvalue.

Since the BIC may penalize for parsimony too heavily for this application, we propose a penalty term that will let the posterior probability of the full model be *as likely as* the null model. If r_f^2 is the proportion of the variance explained by the full model, then let the *truncated information criterion* or TIC for model M_i be defined as

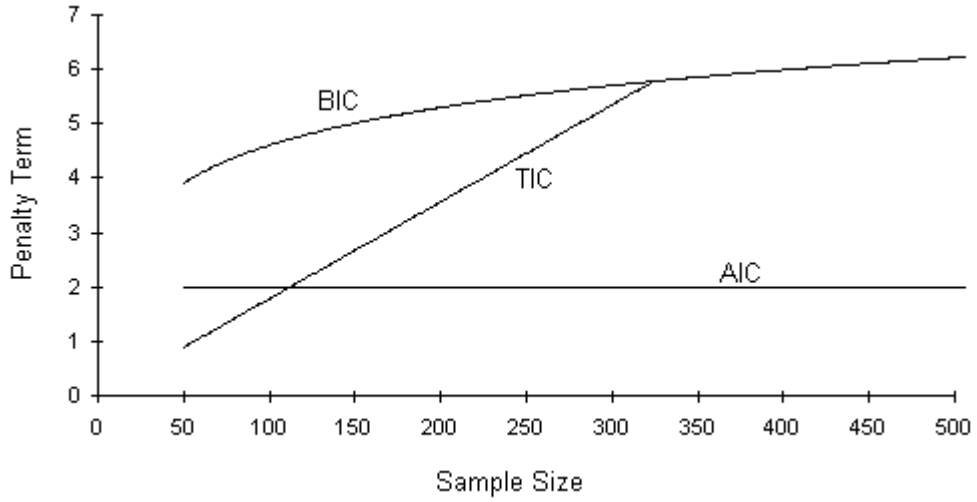
$$TIC_i = n\ln(1-r_i^2)+p_i a_n \tag{2.6}$$

where $a_n = \min\left\{-np^{-1}\ln(1-r_f^2), \ln n\right\}$. This penalty term lets the full model be at least as likely as the null model since

$$\begin{aligned} P(M_{full}|\text{data}) &\geq P(M_0|\text{data}) \\ \exp(-0.5TIC_{full}) &\geq \exp(-0.5TIC_0) \\ TIC_{full} &\leq TIC_0 \\ n\ln(1-r_f^2)+pa_n &\leq 0 \\ a_n &\leq -\frac{n}{p}\ln(1-r_f^2) \end{aligned}$$

There exists some finite n_0 such that $\ln n_0 < -n_0 p^{-1}\ln(1-r_f^2)$, and for all $n \geq n_0$ the sample size is large enough so that the standard form of the BIC allows the full model to be more likely than the null model hence the form of the penalty term chosen for the TIC. To illustrate the size of the penalty term, suppose that $p=20$ and $\lambda_1 = 6$, then figure 2.1 shows the penalty terms associated with the BIC, TIC and AIC (Akaike's information criterion). The penalty using TIC is less then that of the AIC for all $n < 113$. For $113 < n < 325$ the penalty term for the TIC is higher (more conservative) than that of the AIC, but not as much as BIC, but for all $n \geq 325$, $TIC = BIC$. Since $\lim_{n\to\infty} a_n = \infty$ and

Figure 2.1: Bayes, truncated, and Akaike information criteria penalty term for $p = 20$, $\lambda_1 = 6$



$\lim_{n\to\infty} n^{-1}a_n = 0$ then TIC falls into the class of *generalized information criteria* (GIC) (of which BIC is also a special case) [63, 48] and therefore shares the same asymptotic properties as BIC. Furthermore, there exists finite n_0 such that TIC = BIC for all $n > n_0$. Using the same form as shown in equation 2.5 the posterior probability of model M_i for component j will be estimated by

$$P(M_i|R_j^i) \approx \frac{\exp(-0.5TIC_i)}{\sum_{M_k\in\mathcal{M}} \exp(-0.5TIC_k)} \tag{2.7}$$

Posterior model probabilities obtained from a marginal likelihood approximated using a generalized information criteria (GIC) assuming a uniform prior on the model space is equivalent to posterior model probabilities obtained using the Bayes information criteria (BIC) where the prior assumed for the model space is defined by

$$P(M_i) = \frac{\exp(-0.5p_i(a - \ln n))}{\sum_{j=0}^{p} \frac{p!}{j!(p-j)!} \exp(-0.5j(a - \ln n))}$$

where $a = a_n$ (proof shown in A.1).

2.5 Stochastic search of model space

Stepwise methods are often used when it is not practical to evaluate a large number of possible models. While this algorithm is convienient, it is deterministic for a given data set. Slight perturbations in the data can result in a very different "best" model. By adopting a stochastic search of the model space we are able to identify the models best supported by the data almost surely [64, 49, 20]. Madigan and York [39] implemented the Markov chain Monte Carlo model composition (MC^3) procedure on graphical models. Hoeting [21] and Hoeting, Madigan, and Raftery [22] applied the method to univariate multiple regression. One advantage in using the MC^3 approach is that the model selection process is stochastic, and each model will be visited during the simulation in proportion to how well it is supported by the data. Models will be visited during the simulation in proportion to how well they are supported by the data, so we get a summary of all the best models and not just a single snapshot that is obtained using a stepwise procedure.

The stochastic search of the model space is made necessary by the potentially enormous number of models in the denominator of equation 2.7. The MC^3 method is used to reduce the number of terms in this sum by focusing on the most probable models and eliminating those models that are not supported by the data.

The states of the Markov chain to be sampled from are the individual models in $\mathcal{M}$ hence the chain is discrete and finite. In order to insure the proper stationary distribution we must specify how to move from one model to another. This task is accomplished by forming neighborhoods around each model [39]. The neighborhood, centered at an arbitrary model M_i, denoted by $nbd(M_i)$, consists of model M_i and every other model that can be obtained by either addition or removal of a single variable to M_i.

The transition from one neighborhood to another is accomplished using Hasting's [20] method. Given that the current state of the Markov chain is $nbd(M_i)$, the models in $nbd(M_i)$ are sampled with equal probability. Suppose $M_k \in nbd(M_i)$ is proposed, then the move to $nbd(M_k)$ is accepted with probability

$$P_{acc} \;=\; \min\left\{1, \frac{P(M_k|R_j)}{P(M_i|R_j)}\right\} \tag{2.8}$$

Since the transition matrix is finite and irreducible, then by applying the ergodic theorem for Markov chains, any function $g(M_i)$ defined on $\mathcal{M}$, $E(g(M))$ can be estimated by drawing from the Markov chain for $t = 1, 2, \ldots, N$, and

$$\hat{G} \;=\; \frac{1}{N}\sum_{i=1}^{N} g(M(t))$$

which is a simulation-consistent estimate of $E(g(M))$ (i.e. $\hat{G} \to E(g(M))$ almost surely) [64, 39, 49]. In other words, the posterior probability of model M_i given R_j is approximated

by the proportion of iterations that the Markov chain spends in $nbd(M_i)$ and as the number of iterations goes to infinity then the estimate goes to $P(M_i|R_j)$ almost surely. The primary goal of using the MC^3 process in this particular application is not convergence to the stationary distribution, but rather to identify a subset of models that are most supported by the data. Let $\mathcal{M}^* \subset \mathcal{M}$ denote the models that are actually visited during the simulation. Any model $M_i \in \mathcal{M}$ and $M_i \notin \mathcal{M}^*$ has an estimated posterior probability of zero and is therefore eliminated from the sum in denominator of equation 2.7.

The posterior probability of any $M_i \in \mathcal{M}^*$ can be estimated by the number of times the Markov chain was in state $nbd(M_j)$ divided by the total number of draws from the chain which is only appropriate when convergence it attained. Alternatively, the posterior probability for the models can also be estimated by replacing $\mathcal{M}$ with $\mathcal{M}^*$ in equation 2.7 since the TIC must be computed for each model visited during the simulation of the chain. Since the goal is model identification, the convergence of the Markov chain is not necessary because as long as each good model is visted at least of once during the simulation then it's posterior probability will be estimated using the model's TIC.

To further reduce the number of models in the denominator of equation 2.7 we use the principle of Occam's razor. which holds that models which perform much less well than their competitors should be discarded [38]. The MC^3 algorithm eliminates most of the poor models by not visiting them, but there may be models in $\mathcal{M}^*$ that still are much less likely than the most probable model visited and are effectively discredited and should be eliminated. The reduced class of models is then defined by

$$\mathcal{M}^{**} \;=\; \left\{ M_k : M_i, M_k \in \mathcal{M}^*, \frac{\max_i P(M_i|R_j)}{P(M_k|R_j)} < C \right\}$$

Madigan and Raftery [38] adopted $C = 20$, but values from 10 to 1000 have been suggested with respect to the particular application. As a result, equation 2.7 can essentially be replaced by

$$P(M_i|R_j^i) \;\approx\; \frac{\exp\left(-0.5TIC_i\right)}{\sum_{M_k \in \mathcal{M}^{**}} \exp\left(-0.5TIC_k\right)} \tag{2.9}$$

2.6 Implementation

Suppose we have n measurements on each of the variables $X_1, \cdots, X_p$ and wish to identify the important contributors to the jth principal component. In order to construct a set of the most likely models $\mathcal{M}^{**} \subset \mathcal{M}$ we use the following algorithm to analyze the contribution of each variable to the jth eigenvector.

1. Compute the correlation matrix R

2. Construct $R_j = R - \sum_{i=1}^{j-1} \underline{e}_i l_i \underline{e}_i'$

3. Randomly choose $M_i \in \mathcal{M}$ as a starting point. Let all $M_i \in \mathcal{M}$ be equally likely and $\mathcal{M}^* \stackrel{set}{=} \emptyset$

4. Let $u \sim U(0,1)$. If $u < s$ then choose some $M_i \in \mathcal{M}$ at random where all M_i are equally likely

5. Record current neighborhood index $\mathcal{M}^* \stackrel{set}{=} \mathcal{M}^* \cup M_i$

6. Let $R_j^i = diag(\underline{\delta}_i) R_j diag(\underline{\delta}_i)$

7. Compute TIC_i

8. Randomly choose $M_k \in nbd(M_i)$ where each model in $nbd(M_i)$ is equally likely.

9. Compute R_j^k and TIC_k for model M_k.

10. Move to $nbd(M_k)$ with probability $P_{acc} = \min\left(1, \exp\left\{-0.5(TIC_i - TIC_k\right\}\right)$ or stay in $nbd(M_i)$ with probability $1 - P_{acc}$

11. Iterate steps 4–10 N times

12. Construct $\mathcal{M}^{**} = \{M_i : TIC_i < \min_{M_k \in \mathcal{M}^*} TIC_k + 2 \ln C\}$

13. Compute $P(M_i|R_j) = \frac{\exp(-0.5 TIC_i)}{\sum_{M_k \in \mathcal{M}^{**}} \exp(-0.5 TIC_k)}$ for all $M_i \in \mathcal{M}^{**}$

14. Compute $E[\underline{\delta}] = \sum_{M_i \in \mathcal{M}^{**}} \underline{\delta}_i P(M_i|R_j)$

2.6.1 Algorithm details

Steps **1** and **2** summarize the data with the sufficient statistic R and then focus on the jth component with R_j. In step **3** the starting point is determined at random by selecting a model $M_i \in \mathcal{M}$ where all models are equally likely. Since no model has been visited prior to the first sample from the chain, the class of models visited, $\mathcal{M}^*$, is set equal to the null set.

We define the distance between any two models, M_i and M_j to be

$$d_{ij} \quad = \quad (\underline{\delta}_i - \underline{\delta}_j)'(\underline{\delta}_i - \underline{\delta}_j)$$

For any model $M_i \in \mathcal{M}$, there are C_k^p models that are a distance $k \leq p$ units away from M_i. As sampling from the chain continues, by the nature of the process there is an emphasis on spending more iterations in the neighborhoods of the best models. If there are groups of neighborhoods containing good models that are far apart from one another it may take many iterations to achieve convergence . One method of assessing convergence is the use

of multiple sequences using overdispersed starting points [16]. By choosing random starting points for multiple sequences, the expected distance between any two start points is $0.5p$. We propose starting new sequences at random with probability s (we use $s = 0.01$) with the starting point of the new sequence being some model chosen at random. Therefore when a new sequence is triggered, the initial model in the new chain is some model in $\mathcal{M}$ as shown in step **4** and the MC^3 starts anew. Recall that our goal in sampling from the chain is model identification and not convergence. While the process cycles within a group of good models we randomly start the process over in a randomly determined spot in the model space in the hopes of finding other groups of likely models if they exist.

In step **5** the current model M_i that has been selected is unioned with $\mathcal{M}^*$ in order to record the history of the chain. In step **6** the model is imposed onto R_j and in step **7** the model TIC is computed. In step **8** a model within the neighborhood of model M_i is chosen at random where all models in the neighborhood are equally likely. The proposed model, M_k, is imposed onto R_j to get R_j^k and the chain moves to the neighborhood of M_k with probability

$$\begin{aligned} P_{acc} &= \min\left\{1, \frac{P(M_k|R_j)}{P(M_i|R_j)}\right\} \\ &= \min\left\{1, \frac{\exp(-.5TIC_k)}{\exp(-.5TIC_i)}\right\} \\ &= \min\left\{1, \exp(-.5(TIC_k - TIC_i))\right\} \end{aligned} \tag{2.10}$$

or stays in the neighborhood of M_i with probability $1 - P_{acc}$ which is shown if step **10**. Whether the chain moves to the neighborhood of M_k or stays in M_i, the set of models visited, $\mathcal{M}^*$ is updated. The transition probability in equation 2.10 has been used in other applications so that the correct stationary distribution is attained [20, 39, 21] but is used here because the best models are more likely to be visited. The random draws from the chain are repeated N times so that the models most supported by the data can be visited during the stochastic search process.

The purpose of step **11** is to insure that the best models are visited. Usually, in MCMC simulations, the number of iterations is chosen to achieve convergence to the proper stationary distribution and suggested values of N are on the order of 30000 [21]. In this particular application we are only interested in the neighborhoods that were actually visited throughout the simulation which make up the set $\mathcal{M}^*$ hence model identification is of greater importance than convergence so N may be as small as 5000 to attain the desired result. The justification for this is that the posterior probability of a model will not be estimated by the proportion of time the Markov chain spent in the neighborhood of the model, but instead will be approximated using the observed *TIC* for each model that is visited during the simulation. The assumption inherent in this approach is that all models that are most likely in $\mathcal{M}$ will be visited at least one time in 5000 iterations with the aid of the random restarts of the sequence (from step **4**).

Occam's razor is performed in step **12** which states that models in $\mathcal{M}^*$ that are C or more times less likely than the most likely model in the set have been essentially discredited and should be eliminated. Madigan and Raftery [38] adopted the value of $C = 20$ to eliminate models that were far less likely than the best model. We then have

$$\begin{aligned}
\mathcal{M}^{**} &= \left\{ M_i : \frac{\max_{M_k \in \mathcal{M}^*} P(M_k|R_j)}{P(M_i|R_j)} < C \right\} \\
&= \left\{ M_i : \frac{\max_{M_k \in \mathcal{M}^*} \exp(-0.5 TIC_k)}{\exp(-0.5 TIC_i)} < C \right\} \\
&= \left\{ M_i : \max_{M_k \in \mathcal{M}^*} \exp(-0.5(TIC_k - TIC_i)) < C \right\} \\
&= \left\{ M_i : \max_{M_k \in \mathcal{M}^*} -0.5(TIC_k - TIC_i) < \ln C \right\} \\
&= \left\{ M_i : \min_{M_k \in \mathcal{M}^*} TIC_k - TIC_i > -2 \ln C \right\} \\
&= \left\{ M_i : TIC_i < 2 \ln C + \min_{M_k \in \mathcal{M}^*} TIC_k \right\}
\end{aligned}$$

This is the step in the algorithm where models that were identified during the simulation but deemed unlikely in comparison to the best model, are removed.

In step **13** the potentially greatly reduced set $\mathcal{M}^{**} \subseteq \mathcal{M}^* \subseteq \mathcal{M}$ is then used to estimate the posterior probabilities of the most likely models and all models not in $\mathcal{M}^{**}$ have an estimated posterior probability of zero and are therefore eliminated from the denominator of equation 2.7.

Any variable that is in a given model has its corresponding position in the vector $\underline{\delta}$ set to one or it is set to zero if the variable is not present. The probability that a variable is a significant contributor to the component being analyzed can then be estimated by

$$\hat{E}[\underline{\delta}] = \sum_{M_i \in \mathcal{M}^{**}} \underline{\delta}_i P(M_i|R_j)$$

so if the estimated probability that any given variable should be in the model is greater than 0.5, it is more likely than not that the variable in question is a significant contributor to the eigenvector being investigated.

The elements of $\hat{E}[\underline{\delta}]$ are interpreted as the probability that the corresponding variable is active or contributes significantly to the component. If a group of variables has been identified *a priori* to be important in the construction of an index or latent variable then its overall importance to any given component can be evaluated as the weighted average of the activation probabilities of the variables that construct it. In the example that follows in section 2.9 nine of the variables are characterized as pertaining to water chemistry [51]. In the absence of any previously described weighting scheme, the level of water chemistry

contribution to any principal component can therefore be assessed as the average activation probability of the variables in the water chemistry group.

2.7 Power study

As with many multivariate procedures, there are too many scenarios possible to fully investigate how well a procedure performs in practice. In this section we consider the effect of sample size, number of contributing variables, and total number of variables on correct model identification at each of the following combination of levels

Factor	Levels
Total number of variables	20, 30
Sample Size	100, 400
Number of contributing variables	5, 10, 15

Within the above described framework, there are still too many correlation structures that could be investigated so the scope is narrowed further by looking at correlation matrices of the type

$$\Phi = \begin{bmatrix} B_a^r & 0_{a\times(p-a)} \\ 0_{(p-a)\times a} & I_{p-a} \end{bmatrix}$$

where p is the total number of variables in the data set, a is the number of contributing variables to the first eigenvector, and B_a^r is an $a \times a$ equicorrelation matrix with off diagonal elements equal to r. The eigenvalues and first eigenvector characteristics of Φ are

$$\begin{aligned} \lambda_1 &= r(a-1)+1 \\ \lambda_2 = \cdots = \lambda_{p-a+1} &= 1 \\ \lambda_{p-a+2} = \cdots = \lambda_p &= 1-r \end{aligned}$$

and

$$\underline{e}_1' = [\overbrace{a^{-0.5} \quad \cdots \quad a^{-0.5}}^{a} \quad \overbrace{0 \quad \cdots \quad 0}^{p-a}]$$

The number of variables, a, that can significantly contribute to the jth component is between $1 + \lfloor l_j \rfloor$ and p inclusive. The lower limit stems from the fact that each variable can add at most one unit to a given eigenvalue. Throughout the simulation, a is specified for each scenario. The eigenvalue under consideration can take on any positive value less than or equal to a based on the correlation between the a variables. We denote the *efficency of*

the contribution, $e \in [0,1]$, to be a measure of how well the contributing variables form the eigenvalue as

$$e = \frac{\lambda_j}{a}$$

Solving for λ_j we have that $\lambda_j = ae$. Recall that we are modeling the first component from an equicorrelation matrix so equating these results and solving for r we have

$$\begin{aligned} ae &= r(a-1)+1 \\ r &= \frac{ae-1}{a-1} \end{aligned}$$

Specification of an efficency rating is therefore equivalent to specifying the strength of the correlation structure in B_a^r but puts scenarios with differing numbers of contributing variables on equal footing in the sense that each is constructed using the same proportion of the total possible information attainable. For example, suppose $a = 10$ and $\lambda_1 = 6.4$ so the efficency is $6.4/10 = 0.64$ and $r = 0.6$. To construct B_a^r so that we use the same amount of available information (64%) if $a = 5$, then $\lambda_1 = 5 \times 0.64 = 3.2$ and $r = 0.55$ so we see that a smaller correlation is necessary when a is changed from 10 to 5 in order to obtain an eigenvalue that uses the same proportion of the total information available. The efficency values used in the simulation are 0.35, 0.50, and 0.65 and can be thought of as weak, moderate, and strong usage of total available information.

For each simulation scenario, identified by the 4-tuple (p, n, a, e), a random matrix $Z_{n\times p}$ is generated such that each element is an *iid* standard normal random variable. The data to be analyzed, $X_{n\times p}$, are obtained from the transformation $X = Z\Phi^{0.5}$. The true model for the first principal component is

$$\underline{\delta}' = [\overbrace{1 \ \cdots \ 1}^{a} \ \overbrace{0 \ \cdots \ 0}^{p-a}]$$

Errors in the model decided upon can come from either/both of two sources; (1) a variable that should not be in the model is included (type I), and (2) a variable that should be in the model is excluded (type II). The error rate for any particular data set is

$$ErrRate = p^{-1}(\underline{\delta} - \langle \hat{E}(\underline{\delta}) \rangle)'(\underline{\delta} - \langle \hat{E}(\underline{\delta}) \rangle)$$

where $\langle \cdot \rangle$ denotes the rounding operation. The estimated error rate is essentially just the proportion of variables misspecified.

Each data set was also analyzed using the BIC for error rate comparison. Every scenario was repeated 50 times. The average error rate and standard error estimate is recorded in the table 2.2.

Table 2.2: Error Rate Simulation Results

			n=100		n=400	
p	a	e	TIC ErrRate	BIC ErrRate	TIC ErrRate	BIC ErrRate
20	5	0.35	0.1225 (0.0168)	0.2500 (0.0000) †	0.0025 (0.0025)	0.2500 (0.0000) †
	10		0.0500 (0.0131)	0.5000 (0.0000) †	0.0000 (0.0000)	0.1675 (0.0286)
	15		0.1700 (0.0160)	0.7500 (0.0000) †	0.0225 (0.0077)	0.0575 (0.0159)
	5	0.50	0.0050 (0.0034)	0.2500 (0.0000) †	0.0000 (0.0000)	0.0000 (0.0000)
	10		0.0000 (0.0000)	0.5000 (0.0000) †	0.0000 (0.0000)	0.0000 (0.0000)
	15		0.0400 (0.0010)	0.7500 (0.0000) †	0.0000 (0.0000)	0.0000 (0.0000)
	5	0.65	0.0000 (0.0000)	0.2500 (0.0000) †	0.0000 (0.0000)	0.0000 (0.0000)
	10		0.0000 (0.0000)	0.5000 (0.0000) †	0.0000 (0.0000)	0.0000 (0.0000)
	15		0.0000 (0.0000)	0.7500 (0.0000) †	0.0000 (0.0000)	0.0000 (0.0000)
30	5	0.35	0.3533 (0.0247)	0.1667 (0.0000) †	0.0033 (0.0023)	0.1667 (0.0000) †
	10		0.0500 (0.0104)	0.3333 (0.0000) †	0.0017 (0.0017)	0.3333 (0.0000) †
	15		0.0317 (0.0062)	0.5000 (0.0000) †	0.0033 (0.0023)	0.5000 (0.0000) †
	5	0.50	0.1283 (0.0133)	0.1667 (0.0000) †	0.0000 (0.0000)	0.0000 (0.0000)
	10		0.0033 (0.0023)	0.3333 (0.0000) †	0.0000 (0.0000)	0.0000 (0.0000)
	15		0.0000 (0.0000)	0.5000 (0.0000) †	0.0000 (0.0000)	0.0000 (0.0000)
	5	0.65	0.0417 (0.0087)	0.1667 (0.0000) †	0.0000 (0.0000)	0.0000 (0.0000)
	10		0.0100 (0.0035)	0.3333 (0.0000) †	0.0000 (0.0000)	0.0000 (0.0000)
	15		0.0000 (0.0000)	0.5000 (0.0000) †	0.0000 (0.0000)	0.0000 (0.0000)

Note: Standard error estimates are shown in parenthesis.
†– Maximum type II error rate (i.e. no variables selected)

2.7.1 Discussion of results

For every (p, n, a, e) 4-tuple where $n = 100$, the BIC based decision resulted in no variables being selected in spite of the fact that there was structure to detect. These results illustrate how the BIC tends to be too conservative by using a large penalty term.

The TIC based decisions for $n = 100$ performed much better in every scenario investigated except $(30, 100, 5, 0.35)$. Upon further inspection we see that for this simulation configuration the eigenvalue attached to model

$$\underline{\delta}' = [\overbrace{1 \cdots 1}^{5} \overbrace{0 \cdots 0}^{25}]$$

is the largest about 80% of the time based on generation of 1000 data sets. In the 80% of cases where the desired model does occupy the first component, a 90% empirical confidence interval for the first eigenvalue is $[2.123, 2.714]$ which accounts for between 7 to 9% of the total variance. Many would argue this structure is too weak for principal components to be worthwhile. In the remaining 20% of cases where the model in question does not occupy the first component position, we would expect a large error rate since we are modeling the noise portion of the data rather than the structured portion.

As n increases, the effect of the penalty term is decreased so weaker structures can be detected. In the scenarios where $n = 400$, the penalty term for TIC was equal to BIC in all cases except those where $e = 0.35$. The probability of finding the correct model was quite high (at least 97.75%) using the TIC based posterior probabilities for each scenario. Once again, in the instances where the BIC based estimates did not perform well there is the situation such that the likelihood portion is too small to overcome the penalty term hence the high type II error rates.

As would be expected, when sample size and efficiency increase the probability of correct model identification increases. Also, as the number of variables increase, correct model identification becomes more difficult relative to sample size.

2.8 Graphical summarization of results

The random variable associated with the model space, $\mathcal{M}$, is discrete, finite, and univariate, yet conventional summary methods and graphical techniques are not useful in characterizing its properties. The difficulty in displaying summary or distributional graphics pertaining to this random variable stems from the fact that there is no unique ordering of the elements (models) of the space, and there are 2^p elements to evaluate. The potentially enormous number of elements aspect of the problem has been addressed by using MC^3 methodology in conjunction with the principle of Occam's razor in order to reduce the number of models

Table 2.3: Reduced Model Space and Posterior Probabilities

Model	Active Variables				$Pr(M_i\|X)$	$\ln Pr(M_i\|X)$
M_{12}			X_3	X_4	0.8324	-0.18344
M_{14}		X_2	X_3	X_4	0.0843	-2.47337
M_{13}	X_1		X_3	X_4	0.0833	-2.48531

to some manageable quantity by eliminating models that are far less likely than the best models.

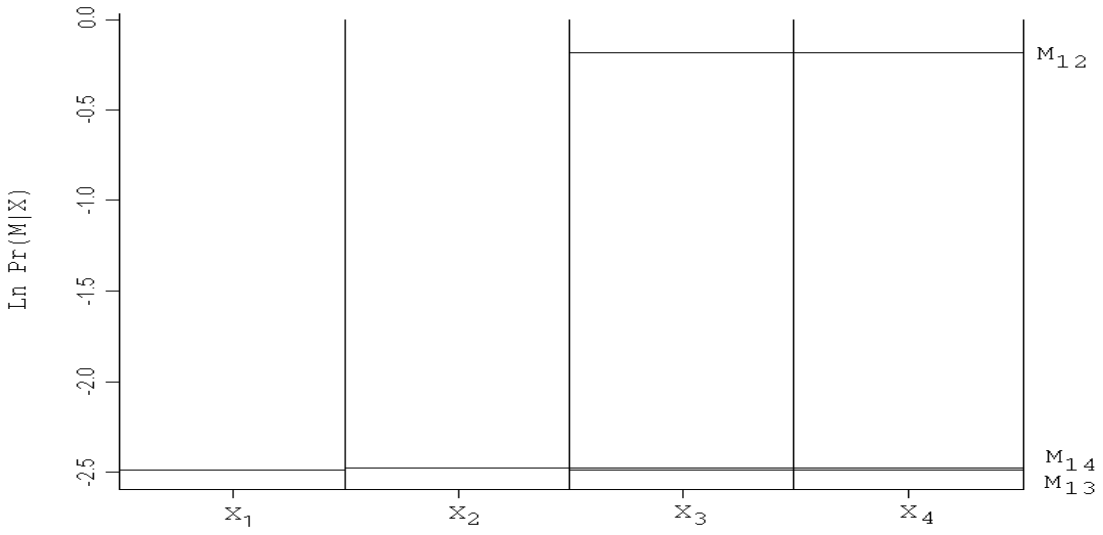
2.8.1 Summarizing the posterior model space

One graphical representation of the reduced model space that has been used is demonstrated by Clyde [6]. The models are sorted by either log posterior probability or posterior probability along the vertical axis and the variables that make up each model are aligned along the horizontal axis. For each model M_i in the reduced space, horizonal line segments across the areas corresponding to each $X_j \in M_i$ are drawn at the appropriate vertical level. For example, using the information from table 2.1, suppose $n = 100$ and the Occam's razor parameter used is 20, we get the following posterior model probabilities shown in table 2.3. The posterior probability information is also displayed graphically in figure 2.2. In this example, there are only three models in $\mathcal{M}^{**}$ so the graph is not very helpful other than to show that M_{12} is much more likely than either M_{13} or M_{14}, but as the cardinality of $\mathcal{M}^{**}$ grows, this graphical technique can be useful in identifying patterns between the variables.

2.8.2 Plotting individual scores

Using PCA to reduce the dimensionality of the data allows more convienient visualization of how observations may be related. Hopefully most of the available information is contained in the first few components so a plot of the first two scores, which is commonly done practice, will aid the practitioner in individual observation assessment. By bootstrapping the observations, empirical confidence regions can be constructed. Tukey proposed forming convex hulls around the observed values by using a multivariate analog to the univariate idea of trimming called "peeling" [40]. The peeling procedure consists of removing extreme points from the convex hull formed around the bootstrapped samples until a fixed percentage of points has been removed. The plot of the first two scores for a particular observation obtained from bootstrap samples is typically crescent shaped (i.e non-convex) so the convex hull method will be applied to the polar coordinates of the scores rather than the Cartesian coordinate score values since the in the transformed space, the distribution of points is typically convex

Figure 2.2: Log Posterior Probabilities versus Model Configuration



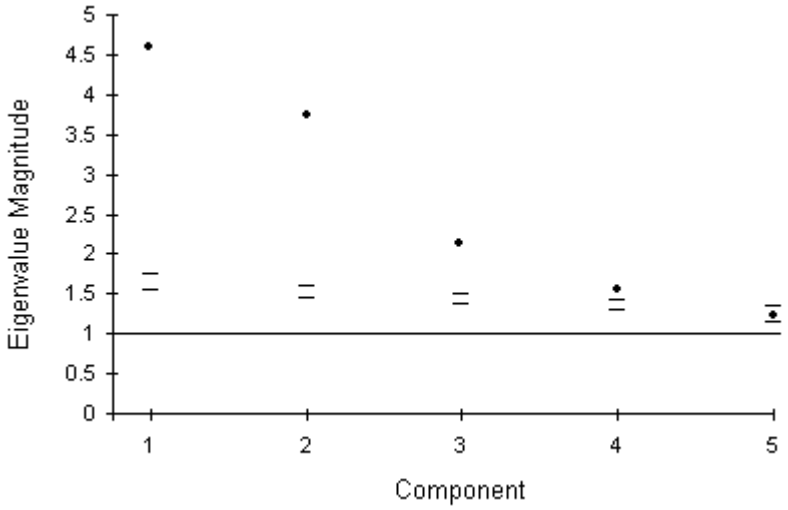
shaped.

To incorporate the variance component associated with model uncertainty into the formation of the empirical regions, the posterior model space is bootstrapped in addition to the individual observations. For each bootstrap sample, an individual model is randomly selected for each component based on their respective posterior probabilities. Using the bootstrap in this way not only accounts for variability due to sampling, but also the variance component associated with model uncertainty.

2.9 Application

Biological data were gathered from 1988 to 1994 by the Ohio Environmental protection Agency (EPA) over the Eastern Corn Belt Plains ecoregion of Ohio [50]. The data analyzed for this application consists of 20 variables identified as biological "stressor" variables at 178 sites in the region. Nine of the variables are characterized as pertaining to water chemistry and the remaining eleven variables are classified as habitat variables. In the original analysis of this set of data, various transformations were applied to individual variables to attain approximate univariate normality of the data [50]. Sections C.2 and C.3 show the names,

Figure 2.3: Scree Plot with 90% ECI for data with no structure



brief descriptions, and transformations used for the variables analyzed in this illustration.

To determine how many components should be retained a modified Guttman-Kaiser criteria was used. Using the standard Guttman-Kaiser criteria all components with sample eigenvalues greater than unity would be retained. If data were generated with no structure (i.e. *iid*) then since the components are ordered, several components would be retained using this criteria although no structure exists. We propose building empirical confidence limits for the sample eigenvalues based on a Monte Carlo simulation of *iid* data where the simulated data matrix has the same dimension as the true data matrix. If a component contains significant structure to be modeled then its eigenvalue should lie above the empirical interval of the eigenvalue based on no structure. Ten thousand *iid* standard normal data matrices of 178 observations and 20 variables were generated and 90% empirical intervals were constructed for the sample eigenvalues. The eigenvalues for the first five principal components along with 90% ECI are shown in figure 2.3. From the plot we see that the fifth eigenvalue does not exceed the 95-th percentile so we conclude that there is significant structure to be modeled in the first four principal components which accounts for 60.12% of the total variance.

The first four eigenvectors are shown in table 2.4. One commonly used method to identify the most important variables used to construct a given component is to choose those variables whose elements have magnitudes larger than the average magnitude within the component being interpreted [28]. The important variables for the contruction of the first four eigenvectors are also shown in table 2.4. The first component is made up of the habitat variables (CHANNEL, COVER, RIPSS, RIPARIAN, POOL, EMBSS, RIFEMSS, RIFFLE,

Figure 2.4: Activation Probabilities For Ohio Habitat and Water Chemistry Variables for the first four Principal Components

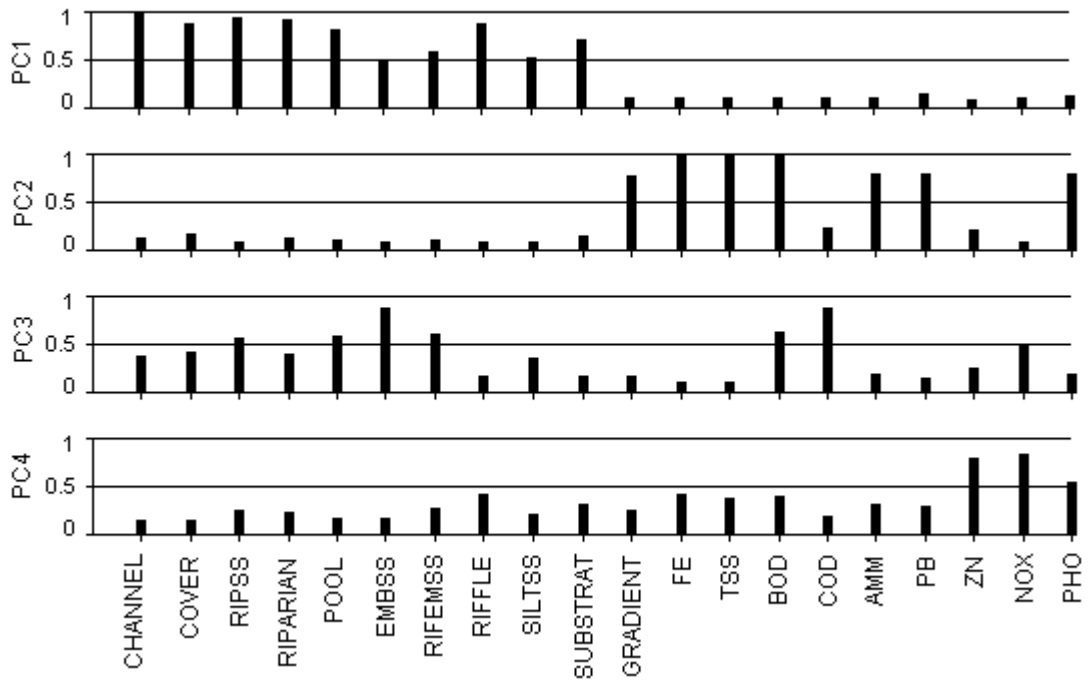


Table 2.4: First Four Eigenvectors in Standard PCA

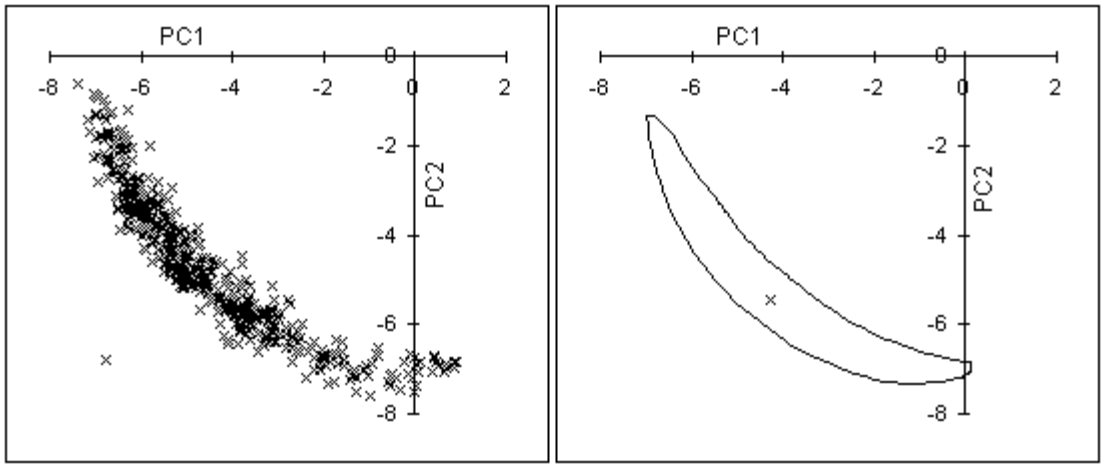
Variable	PC1	PC2	PC3	PC4	PC1	PC2	PC3	PC4
CHANNEL	0.3527	0.1232	0.2254	-0.0407	X		X	
COVER	0.2828	0.1583	0.2539	0.0495	X		X	
RIPSS	0.3242	0.0651	0.3079	-0.2170	X		X	X
RIPARIAN	0.3148	0.1087	0.2369	-0.1919	X		X	X
POOL	0.2627	0.1163	0.2899	0.1378	X		X	
EMBSS	0.2886	-0.1032	-0.4197	0.0457	X		X	
RIFEMSS	0.3056	-0.1153	-0.3003	0.1509	X		X	
RIFFLE	0.3591	-0.0238	-0.1119	0.2390	X			X
SILTSS	0.2657	0.0234	-0.2450	0.0863	X		X	
SUBSTRAT	0.2908	0.1000	-0.1341	0.1486	X			
GRADIENT	-0.0063	0.2840	-0.1069	-0.1073		X		
FE	0.0449	-0.4118	0.0246	0.2210		X		X
TSS	0.0787	-0.3962	-0.0027	0.2013		X		X
BOD	-0.0198	-0.4004	0.2391	0.2212		X	X	X
COD	-0.0940	-0.1114	0.3663	0.1350			X	
AMM	-0.0292	0.2974	-0.1340	0.2145		X		X
PB	-0.1421	0.3126	0.0540	0.1998		X		X
ZN	0.0371	-0.1781	-0.0995	-0.4823				X
NOX	0.0889	-0.0861	-0.1921	-0.4753				X
PHO	-0.1178	0.2999	-0.1476	0.2628		X		X

SILTSS, and SUBSTRAT) and the second component is primarily the water chemistry variables (GRADIENT, FE, TSS, BOD, AMM, PB, and PHO). The third component appears to be mainly a habititat component with 10 important variables (8 habitat, 2 water chemistry). Finally, the fourth PC is mainly a water chemistry variable with 11 (3 habitat and 8 chemistry) variables identified as important.

Using this technique to identify important variables is not satisfying for several reasons. Firstly, if the true structure of a given component is such that all variables are important to it's construction it will not be identified since there can be no eigenvector constructed where all elements are above average. Also, there is no theoretical justification for using the average magnitude so it is used solely because it is easy and has some heuristic appeal. Finally, interpretation of the components has value to the researcher and making this decision based on point estimates only is risky.

Figure 2.4 shows the estimated expected value of $\underline{\delta}$ for the first four principal components using BMA methodolgy developed in this chapter. Recall that the interpretation of the

Figure 2.5: 90% ECI of first two Principal Component scores for 1990 Swamp Creek location at 40.28N Latitude 84.28W longitude



estimated expected value of each element within the δ vector represents the probability that the corresponding variable contributes significantly to the given component. For example, we conclude that the variables CHANNEL, COVER, RIPSS, RIPARIAN, POOL, RIFEMSS, RIFFLE, SILTSS, and SUBSTRAT are the most likely contributors to the first component since each of their corresponding values are greater than 0.5. Norton [50] interprets this set of variables as stream structure corridor and siltation.

Figure 2.5 shows the sampling variability in the first two component scores for a particular site in the sample based on 500 bootstrap samples illustrating Tukey's method of confidence region construction. The left side of the figure shows the individual bootstrap samples and the right side illustrates the resulting region formed using Tukey's procedure [40] applied to the polar coordinate transformation of the bootstrap scores. Figure 2.6 shows the region formed without uncertainty compared to the region accounting for model of uncertainty about the scores for the same site. For this particlar set of data, the variance component due to sampling variability is much larger than the variance associated with model uncertainty but this does not hold true in general.

Plots of the first few principal component scores are commonly made in practice to compare observations to one another and to describe individual observations with respect to their relative location on each principal axis. Generally these comparisons and interpretations are made using the point estimate alone without consideration of the sampling variability or the variance due to model uncertainty. We see from figure 2.6 that the sampling variability is quite large for this observation and so any interpretation or description should reflect the level of uncertainty.

Figure 2.6: 90% ECI of first two Principal Component scores for 1990 Swamp Creek location at 40.28N Latitude 84.28W longitude. Left region formed with no uncertainty component, right region constructed accounting for model uncertainty

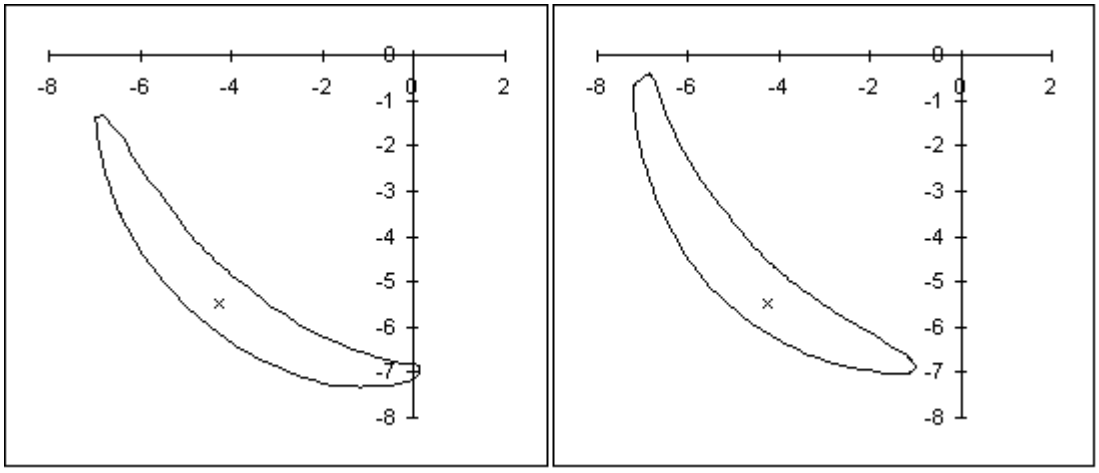


Figure 2.7: Posterior probabilities of first principal component model configurations

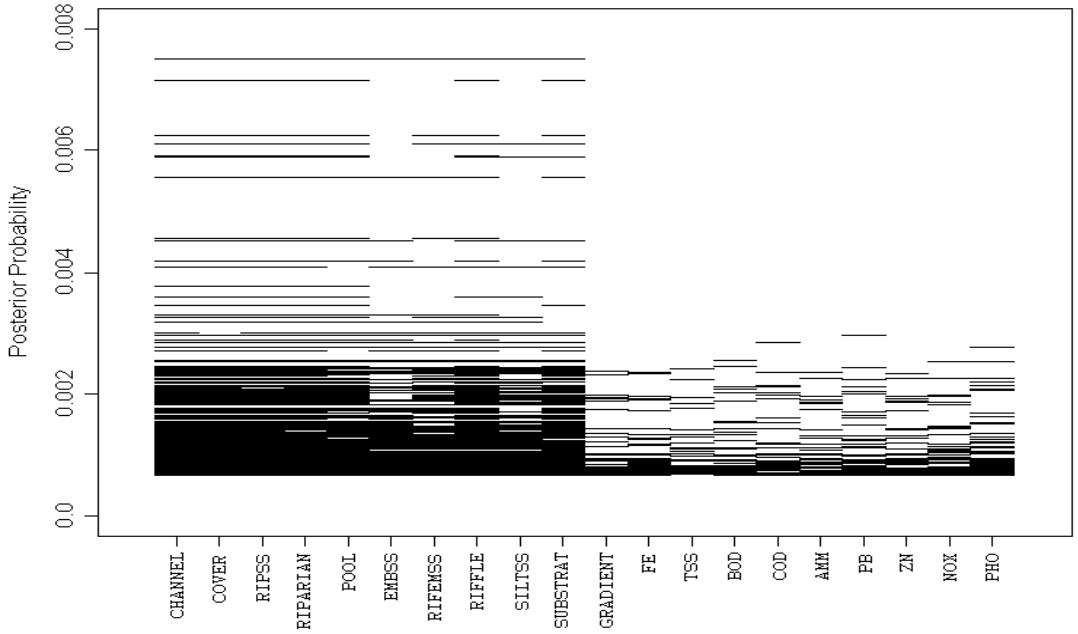
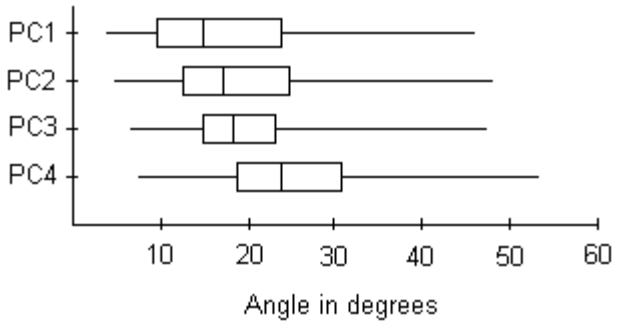


Figure 2.8: Distributions of the angle between the full sample eigenvectors and bootstrap eigenvectors for the first four principal components



Component swapping occured infrequently in the first four components of this data. The first and second component were interchanged with each other in 3.5% of the bootstrap samples, and the third and fourth swapped places in 1.6% of the samples. The fourth component also changed places with the fifth component 2.2% of the time and with the sixth component in 0.4% of the samples. The conclusion we can draw from this is that the eigenvalues are distinct since there is little overlap in their respective sampling distributions.

Figure 2.8 shows the distribution of angles between the first four eigenvectors from the full sample and the associated eigenvectors from the bootstrap samples. The lack of precision in the estimation of the eigenvectors is reflected in the large angular deviations. Since the component scores are constructed using the eigenvector coefficients, we why there is a large amount of variability in the pc score depicted in figure 2.6.

Using the activation probabilities from figure 2.4 and the variable classifications from tables C.2 and C.3, we compute the contribution of the water chemistry latent variable and habitat latent variable to each of the modeled principal components. The naive estimate for each was constructed by averaging the activation probabilities over the two classifications with the results shown in table 2.5. We see that habitat has an activation probability of 0.7056 whereas the water chemistry is relatively inconsequential with a value of 0.1088. The second component is mainly a water chemistry construction with an average activation probability of 0.6497 and the habitat is now inactive with a value of 0.1681. The third component is a mix of four active habitat and three water chemistry variables which is reflected in the two latent variables overall activation probabilities. Finally, the fourth component is inactive with respect to habitat (0.2324) and partially active with respect to water chemistry (0.4608).

Figure 2.7 shows the posterior probability of the models in $\mathcal{M}^{**}$ for the first principal component. Since this component was previously described as highly active with respect to habitat and inactive with respect to water chemistry, we see that the most highly probable models

Table 2.5: Contribution of Water Chemistry and Habitat to Principal Component construction

Component	Water Chemistry	Habitat
PC 1	0.7056	0.1088
PC 2	0.1681	0.6497
PC 3	0.4330	0.3359
PC 4	0.2324	0.4608

contain most of the habitat variables and the water chemistry variables are only active in lower probability variable configurations.

2.10 Comments

In the context of PCA, a variable is important to the construction of any selected component if it's deletion results in a substantial decrease in the associated eigenvalue. The importance of a variable on the retained components is only indirectly addressed by looking at the magnitude of point estimates of associated eigenvectors when standard PCA modeling methods are used. Finally, the effect of a variable on a component is viewed conditional on every other variable being present in the model since no variable is ever actually removed from the model.

Using BMA the impact of a variable on a given component is directly measured in conjunction with how each variable contributes in the presence of the others. The models with the highest posterior probabilities are those variable configurations that jointly represent, as parsimoniously as possible, the meaningful portion of each component. By averaging over the model space we can estimate the probability that a variable is important to the construction of a given component.

If the variables can be naturally partitioned into groups (such as chemical and habitat variables in the application) then the overall contribution of these partitions may be used to aid in the interpretation. The overall importance of a group to the construction of a component can be assessed by computing the average probability that a variable within the group is active. Using this technique enables the investigator to use the natural variable groupings as an aid in the interpretation of a component.

Chapter 3

Canonical Variate Analysis (CVA)

3.1 Introduction

3.1.1 Background

Historically, two basic forms of discriminant analysis have emerged with each form having a different goal. Predictive discriminant analysis (PDA) can be traced to work by Karl Pearson, and others, on group distances [23]. In the 1930's R. A. Fisher translated multivariate group distance into a linear combinations of variables to aid in group discrimination. In PDA the usefulness of a variable is determined by how well it helps to predict which group an observation belongs. The other form of discriminant analysis, also known as canonical variate analysis (CVA), did not appear until the 1960's [23]. The purpose of CVA is descriptive rather than predictive and the goal is to describe group mean separation in a dimensionally reduced space.

Canonical variate analysis (sometimes called canonical discriminant analysis) creates linear combinations of the original variables called canonical variates. The first canonical variate is constructed so that the mean separation between the groups is maximized with respect to the likelihood ratio. The remaining canonical variates are formed so that the mean separation between the groups is maximized with the added constraint that they be orthogonal to previously defined canonical variates [60].

The format for data where CVA may be of use consists of measured variables $Y_1, \cdots, Y_p$ on g populations. While the addition of any measured variable to the data can not reduce the separation already accomplished by the other variables, there may be the measured variables that do not significantly contribute to group mean separation and therefore should not be included in the model [23]. McKay and Campbell [42] categorized and compared various model building and variable selection techniques. The methods were classified into

one of three groups; techniques associated with examination of magnitudes of the coefficients used to construct the canonical variates, techniques involving sequential F tests, and techniques involving all possible subsets. In practice, methods using sequential F tests are most commonly employed. These methods include forward selection, backward elimination and stepwise selection. Sequential multivariate model building techniques suffer the same criticisms as their univariate analogs [46]. A more recent criticism of classical model building in general is that it ignores the uncertainty involved in model selection [57, 56]. Typically a single model is chosen in some manner and is assumed to be the "correct" model to the exclusion of all other competitors. A consequence of the assumption that the correct model has been chosen is that all future inferences are valid only if the assumption is true.

3.1.2 Outline

In the following sections of this chapter, the computations involved with CVA are described and the popular stepwise method is outlined. Bayesian model averaging (BMA) for CVA is developed and a stochastic model search algorithm outlined. Some popular methods used to interpret results are discussed and BMA results are used as suggested improvements to aid in interpretation. The method is then illustrated in the analysis of a set of environmental data [51].

3.2 The Model

The CVA model can be viewed from either of two equivalent points of view. The first approach takes a MANOVA perspective while the second is from a regression point of view.

3.2.1 MANOVA

Suppose that a data set consists of n measurements of variables $Y_1, \cdots, Y_p$ on samples from g known populations. The MANOVA model is

$$\underline{y}_{ij} = \underline{\mu}_i + \underline{\epsilon}_{ij}, \quad i = 1, \cdots, g, \quad j = 1, \cdots, n_i$$

where

$$\begin{aligned} E[\underline{\epsilon}_{ij}] &= \underline{0} \\ Cov(\underline{\epsilon}_{ij}, \underline{\epsilon}_{i'j'}) &= \begin{cases} \Sigma & \text{if } i = i' \text{ and } j = j' \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

The "between" and "within" matrices, H and E, are then defined as

$$\begin{aligned} H &= \sum_{i=1}^{g} n_i(\underline{\overline{y}}_{i\cdot} - \underline{\overline{y}}_{\cdot\cdot})(\underline{\overline{y}}_{i\cdot} - \underline{\overline{y}}_{\cdot\cdot})' \\ E &= \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\underline{\overline{y}}_{ij} - \underline{\overline{y}}_{i\cdot})(\underline{\overline{y}}_{ij} - \underline{\overline{y}}_{i\cdot})' \end{aligned}$$

The four most commonly used statistics to test the hypothesis that the g mean vectors are equal are, Wilks' Λ, Roy's greatest root, Pillai's test, and the Lawley-Hotelling test. The test statistics for these tests can each be written in terms of the eigenvalues of $E^{-1}H$ where $\lambda_1 > \lambda_2 > \cdots > \lambda_k$, where $k = \min(p, g-1)$ and are shown below [60].

Wilk's lambda	$\Lambda = \prod_{i=1}^{k} \frac{1}{1+\lambda_i}$
Roy's	$\theta = \lambda_1$
Pillai's	$V = \sum_{i=1}^{k} \frac{\lambda_i}{1+\lambda_i}$
Lawley-Hotelling	$U = \sum_{i=1}^{k} \lambda_i$

Since the mean vectors are in a k-dimension space there are many possible mean configurations and a uniformly most powerful test does not exist so all four statistics are generally listed in the output of popular statistical software packages. Besides testing, it is also common to summarize group separation through correlation. Canonical correlations are the multivariate analog of Pearson's correlation coefficient. Using Pearson's correlation coefficient may be appropriate when we wish to measure the linear association between one "X" variable and one "Y" variable. When we have more than one "X" variables and more than one "Y" variables, the canonical correlations are appropriate. The squared canonical correlation for each canonical variate can be written in terms of its associated eigenvalue as

$$r_i^2 = \frac{\lambda_i}{1+\lambda_i} \quad \text{for } i = 1, \cdots, k. \tag{3.1}$$

and is interpreted in the same way as the univariate *r-square* in a multiple regression context.

3.2.2 Multivariate regression

Another view of CVA may be based on multivariate regression. The data set consists of n measurements of variables $Y_1, \cdots, Y_p$ on samples from g known populations. An $n \times (g-1)$ indicator matrix, X, can be constructed so that

$$X_{ij} = \begin{cases} 1 & \text{if } i^{th} \text{ observation is in group } j \\ 0 & \text{otherwise} \end{cases}$$

The linear model relating $Y = [\underline{y}_1 \ \cdots \ \underline{y}_p]$ to $X = [\underline{1} \ \underline{x}_1 \ \cdots \underline{x}_{g-1}]$ is obtained by

$$\underset{n\times p}{Y} = \underset{n\times g}{X}\ \underset{g\times p}{\beta} + \underset{n\times p}{\epsilon}$$

where

$$E(\epsilon) = \underset{n\times p}{0}$$

$$\mathrm{cov}(\mathrm{vec}(\epsilon)) = \underset{(p\times p)(n\times n)}{\Sigma\otimes\sigma^2 I}$$

Here Σ is a positive semi-definite matrix and $\sigma^2 I$ is the identity matrix multiplied by a scalar σ^2, and $\otimes$ stands for the Kronecker product [35]; that is the $np\times np$ matrix $\Sigma\otimes\sigma^2 I$ that can be partitioned into an array of $p\times p$ matrices of the form

$$\begin{bmatrix} \sigma^2\Sigma & 0 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2\Sigma & 0 & 0 & \dots & 0 \\ 0 & 0 & \sigma^2\Sigma & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \sigma^2\Sigma \end{bmatrix}$$

This equation is the natural extension of the univariate linear model to the multivariate case where there are p dependent variates. The error vectors, $\underline{\epsilon_i}$ of each p-variate observation are assumed to be independent, but since the different responses may be correlated, the residuals of the individual y values have unknown covariance matrix Σ. The standard test for $H_0 : \beta = 0$ is the likelihood ratio test where

$$\begin{aligned} \Lambda &= \frac{|Y'Y - \hat{B}'X'Y|}{|Y'Y - n^{-1}\underline{\overline{y}}\underline{\overline{y}}'|} \\ &= \frac{|E|}{|E+H|} \\ &= \frac{1}{|I+E^{-1}H|} \\ &= |(I+E^{-1}H)^{-1}| \end{aligned}$$

now, if the eigenvalues of $E^{-1}H$ are $\lambda_1 > \cdots > \lambda_k$ then the eigenvalues of Λ are $\frac{1}{1+\lambda_1} < \cdots < \frac{1}{1+\lambda_k}$ hence we see that the likelihood ratio test is equal to the Wilks' lambda test. From the equation 3.1 we then have that the eigenvalues of Λ are equal to $(1-r_1^2) < \cdots < (1-r_k^2)$.

As with the MANOVA, the three other multivariate test statistics previously described can also be used to test $H_0 : \beta = 0$.

3.2.3 Computations

The calculation of the eigenvalues of $E^{-1}H$ with respect to the MANOVA model and the multivariate regression model was shown to have value in the previous two subsections. In general, the matrix $E^{-1}H$ is not symmetric and many algorithms for computing eigenvalues accept only symmetric matrices. It can be shown [60] that

$$S_{yy}^{-1/2} S_{yx} S_{xx}^{-1} S_{xy} S_{yy}^{-1/2} \tag{3.2}$$

is symmetric and that the first $k = \min(p, g-1)$ eigenvalues are the squared canonical correlations where $SS_{..}$ are the sums of squares and cross product matrices. Note that if $p = 1$ then we have the case of univariate multiple regression and equation 3.2 is a scalar and is equal to the model *r-square*.

3.3 Model selection

Currently the stepwise method of model selection is quite popular in practice. When the number of variables is large, sequential procedures have generally been the only practical way to construct a reasonable model. The stepwise method is an extension of the forward procedure and makes use of Wilks' lambda statistic and partial F-tests.

In the first step, p univariate regressions are run and if the model with the largest F-statistic is significant, then that is the first variable to be retained. If the best regression fails to be significant then the procedure stops. In the second step, the variable yielding the smallest partial Λ for each y adjusted for the first variable in the model is considered for entry into the model. The partial Λ-statistic is based on the full and reduced model test for a subset of y's [60]:

$$\Lambda(y_i|y_j) \quad = \quad \frac{\Lambda(y_i, y_j)}{\Lambda(y_j)}$$

for each $y_i \neq y_j$, and the y_i that minimizes $\Lambda(y_i|y_j)$ is the candidate to enter the model next. If the partial F-test is significant then the second variable is added, if not, then the procedure stops. At the point where there are two variables in the model, a partial F-test is performed to see if any variables may be removed. The process continues alternating between testing to add the best possible of the remaining variables and testing whether the weakest already present in the model should be deleted. When no variables can be added or removed, the process stops.

Now, since we are adding or removing one variable at a time, there is an exact F test where

$$F \quad = \quad \frac{1-\Lambda}{\Lambda} \frac{N-1-p^*}{g-1} \tag{3.3}$$

follows an F distribution with $g-1$ and $N-1-p^*$ degrees of freedom and p^* is the number of variables currently in the model.

Though each test can be conducted at the α level of significance, the over all level of significance of the model is unknown since the number of tests to be performed before hand is unknown. As such, all p-values obtained from various tests of interest do not have the usual interpretation.

3.4 Bayesian Model Averaging (BMA)

When a model selection method such as an all possible subsets or stepwise procedure is used, the single model obtained is assumed to be the *correct* model. All future inferences and predictions that are made with the model do not account for the uncertainty involved in the selection process. Alternatively, the model obtained using BMA does incorporate the variance component associated with the uncertainty of model building.

There is a standard Bayesian solution to the problem of accounting for model uncertainty. If the model space is $\mathcal{M} = \{M_1, \cdots, M_T\}$ then the posterior probability of M_i given the data Y is given by

$$P(M_i|Y) = \frac{P(Y|M_i)P(M_i)}{\sum_{M_j \in \mathcal{M}} P(Y|M_j)P(M_j)} \tag{3.4}$$

where $P(M_i)$ denotes the prior probability of each model and $P(Y|M_i)$ is the marginal likelihood of the data. Generally, each model has been assumed to be equally likely *a priori*, so equation 4.2 simplifies to

$$P(M_i|Y) = \frac{P(Y|M_i)}{\sum_{M_j \in \mathcal{M}} P(Y|M_j)}$$

Now, the marginal likelihood of the data is

$$P(Y|M_i) = \int P(Y|M_i, \underline{\theta}_i)\pi(\underline{\theta}_i)d\underline{\theta}_i \tag{3.5}$$

where $\underline{\theta}_i$ is the unknown model parameters with joint prior density $\pi(\underline{\theta}_i)$. Hoeting [21] shows for univariate multiple regression that the marginal likelihood follows an n dimensional non-central Student's t-distribution when proper conjugate priors are used. This result can be used if hyperparameters are chosen so that the prior density it calibrated to the data. Raftery [55] approximates equation 4.3 using the *Bayes Information Criterion* (BIC) which is a penalized likelihood measure. For the case of linear regression Raftery shows that

$$\begin{aligned} P(Y|M_i) &\propto \exp\left(-0.5BIC_i\right) \\ &= \exp\left(-0.5(n\ln(1-r_i^2) + p_i \ln n)\right) \end{aligned} \tag{3.6}$$

where r_i^2 is the model *r-square*, p_i is the number of independent variables in model M_i, and n is the number of observations. Using Raftery's approximation requires no calibration to the data and is composed of readily available regression information. By normalizing we get the posterior probability of model M_i given the data is

$$P(M_i|Y) \approx \frac{\exp(-0.5BIC_i)}{\sum_{M_j \in \mathcal{M}} \exp(-0.5BIC_j)} \tag{3.7}$$

In order to use the BIC approximation we must supply an *r-square* like measure of association. Cramer and Nicewander [7] showed how the four multivariate measures discussed in subsection 3.2.1 may be converted into invariant measures of multivariate association.

Wilk's lambda	$\eta_\Lambda^2 = 1 - \Lambda = 1 - \prod_{i=1}^{k}(1 - r_i^2)$
Roy's	$\eta_\theta^2 = \frac{\theta}{1+\theta} = r_1^2$
Pillai's	$\eta_V^2 = k^{-1}V = k^{-1}\sum_{i=1}^{k} r_i^2$
Lawley-Hotelling	$\eta_U^2 = \frac{k^{-1}U}{1+k^{-1}U} = \frac{k^{-1}\sum_{i=1}^{k} r_i^2(1-r_i^2)^{-1}}{1+k^{-1}\sum_{i=1}^{k} r_i^2(1-r_i^2)^{-1}}$

Even though none of the four multivariate tests statistics is uniformly most powerful, as previously stated, the measure of associations listed above provide ordered values so that $\eta_\Lambda^2 \geq \eta_\theta^2 \geq \eta_U^2 \geq \eta_V^2$ (proof shown in A.3).

The BIC penalizes heavily for parsimony so it makes sense to use η_Λ^2 as the *r-square* in the likelihood portion since it is the least conservative measure of association of those considered. Using one of the weaker measures of association would only serve to amplify the penalization.

3.5 Stochastic search of model space

Stepwise methods are often used when it is not practical to evaluate a large number of possible models. While this algorithm is convienient, it is deterministic for a given data set. Slight perturbations in the data can result in a very different "best" model. By adopting a stochastic search of the model space we are able to identify the models best supported by the data almost surely [64, 49, 20]. Madigan and York [39] implemented the Markov chain Monte Carlo model composition (MC^3) procedure on graphical models. Hoeting [21], and Hoeting, Madigan, and Raftery [22] applied the method to univariate multiple regression. One advantage in using the MC^3 approach is that the model selection process is stochastic, and each model will be visited during the simulation in proportion to how well it is supported by the data. Therefore, all good models will be visited more often than those that are not supported by the data, so we get a summary of all the best models and not just a single snapshot that is obtained using a stepwise procedure.

The stochastic search of the model space is made necessary by the potentially enormous number of models in the denominator of equation 4.5. The MC^3 method is used to reduce the number of terms in this sum by focusing on the most probable models and eliminating those models that are not supported by the data.

The states of the Markov chain to be sampled from are the individual models in $\mathcal{M}$ hence the chain is discrete and finite. In order to insure the proper stationary distribution we must specify how to move from one model to another. This task is accomplished by forming neighborhoods around each model [39]. The neighborhood, centered at an arbitrary model M_i, denoted by $nbd(M_i)$, consists of model M_i and every other model that can be obtained by either adding a single variable to M_i or removing a single variable from M_j.

The transition from one neighborhood to another is accomplished using Hasting's [20] method. Given that the current state of the Markov chain is $nbd(M_i)$, the models in $nbd(M_i)$ are sampled with equal probability. Suppose $M_k \in nbd(M_i)$ is proposed, then the move to $nbd(M_k)$ is accepted with probability

$$P_{acc} \;=\; \min\left\{1, \frac{P(M_k|Y)}{P(M_i|Y)}\right\} \tag{3.8}$$

Since the transition matrix is finite and irreducible, then by applying the ergodic theorem for Markov chains, any function $g(M_i)$ defined on $\mathcal{M}$, $E(g(M))$ can be estimated by drawing from the Markov chain for $t = 1, 2, \ldots, N$, and

$$\hat{G} \;=\; \frac{1}{N}\sum_{i=1}^{N} g(M(t))$$

which is a simulation-consistent estimate of $E(g(M))$ (i.e. $\hat{G} \to E(g(M))$ almost surely) [64, 39, 49]. In other words, the posterior probability of model M_i given R_j is approximated by the proportion of iterations that the Markov chain spends in $nbd(M_i)$ and as the number of iterations goes to infinity then the estimate goes to $P(M_i|Y)$ almost surely. The primary goal of using the MC^3 process in this particular application is not convergence to the stationary distribution, but rather to identify a subset of models that are most supported by the data. Let $\mathcal{M}^* \subset \mathcal{M}$ denote the models that are actually visited during the simulation. Any model $M_i \in \mathcal{M}$ and $M_i \notin \mathcal{M}^*$ has an estimated posterior probability of zero and is therefore eliminated from the sum in denominator of equation 4.5.

The posterior probability of any $M_i \in \mathcal{M}^*$ can be estimated by the number of times the Markov chain was in state $nbd(M_j)$ divided by the total number of draws from the chain which is only appropriate when convergence it attained. Alternatively, the posterior probability for the models can also be estimated by replacing $\mathcal{M}$ with $\mathcal{M}^*$ in equation 4.5 since the BIC must be computed for each model visited during the simulation of the chain.

To further reduce the number of models in the denominator of equation 4.5 we use the principle of Occam's razor which is the principle that holds that models which perform much

less well than their competitors should be discarded [38]. The MC^3 algorithm eliminates most of the poor models by not visiting them, but there may be models in $\mathcal{M}^*$ that still are much less likely than the most probable model visited and are effectively discredited and should be eliminated. The reduced class of models is then defined by

$$\mathcal{M}^{**} = \left\{ M_k : M_i, M_k \in \mathcal{M}^*, \frac{\max_i P(M_i|Y)}{P(M_k|Y)} < C \right\}$$

Madigan and Raftery [38] adopted $C = 20$, but values from 10 to 1000 have been suggested with respect to the particular application. As a result, equation 4.5 can essentially be replaced by

$$P(M_i|Y) \approx \frac{\exp(-0.5BIC_i)}{\sum_{M_k \in \mathcal{M}^{**}} \exp(-0.5BIC_k)} \tag{3.9}$$

3.6 Implementation

Suppose we have n measurements on each of the variables $Y_1, \cdots, Y_p$ on g known populations and wish to identify the most important variables that contribute to mean separation. In order to construct a set of the most likely models $\mathcal{M}^{**} \subset \mathcal{M}$ we use the following algorithm.

1. Randomly choose $M_i \in \mathcal{M}$ as a starting point. Let all $M_i \in \mathcal{M}$ be equally likely and $\mathcal{M}^* \stackrel{set}{=} \emptyset$
2. Record current neighborhood index $\mathcal{M}^* \stackrel{set}{=} \mathcal{M}^* \cup M_i$
3. Let Y be the matrix such that the columns represent the variables present in M_i
4. Compute the $k = \min(p, g-1)$ non-zero eigenvalues of $S_{yy}^{-1/2} S_{yx} S_{xx}^{-1} S_{xy} S_{yy}^{-1/2}$ and represent them by $r_1^2 > \cdots > r_k^2$
5. Compute $BIC_i = n \ln(\Lambda_i) + p_i \ln(n)$
6. Randomly choose $M_k \in nbd(M_i)$ where each model in $nbd(M_i)$ is equally likely.
7. Compute BIC_k for model M_k.
8. Move to $nbd(M_k)$ with probability $P_{acc} = \min(1, \exp\{-0.5(BIC_i - BIC_k)\})$ or stay in $nbd(M_i)$ with probability $1 - P_{acc}$
9. Let $u \sim U(0,1)$. If $u < s$ then choose some $M_i \in \mathcal{M}$ at random where all M_i are equally likely
10. Iterate steps 2–10 N times

11. Construct $\mathcal{M}^{**} = \{M_i : BIC_i \leq \min_{M_k \in \mathcal{M}^*} BIC_k + 2\ln C\}$

12. Compute $P(M_i|Y) = \frac{\exp(-0.5BIC_i)}{\sum_{M_k \in \mathcal{M}^{**}} \exp(-0.5BIC_k)}$ for all $M_i \in \mathcal{M}^{**}$

13. Compute $E[\underline{\delta}] = \sum_{M_i \in \mathcal{M}^{**}} \underline{\delta}_i P(M_i|Y)$

3.6.1 Algorithm details

In step **1** a model is chosen at random from $\mathcal{M}$ where all models are equally likely. The initial starting point for the Markov chain is accomplished by generating $\delta_1, \cdots, \delta_p \sim Bern(0.5)$. If $\delta_i = 1$ then Y_i is in the model alternatively, if $\delta_i = 0$ then Y_i is excluded.

Step **2** records the iteration history (i.e. models visited) throughout the simulation in set $\mathcal{M}^*$.

The p_i variables that are present in the current model make up the $n \times p_i$ dimensional matrix Y in step **3**.

In step **4** the $k = \min(p_i, g-1)$ squared canonical correlations obtained from model M_i are computed.

In step **5** the BIC for model M_i is calculated by

$$\begin{aligned} BIC_i &= n\ln(1-\eta^2_{\Lambda_i}) + p_i\ln(n) \\ &= n\ln(\Lambda_i) + p_i\ln(n) \end{aligned}$$

A model is chosen at random from $nbd(M_i)$ where all models are equally likely in step **6**. The model M_j is selected by generating $U \sim U(0,1)$ and defining $V = \lfloor pU \rfloor + 1$ then

$$\delta_V \stackrel{set}{=} 1 - \delta_V$$

which causes Y_V to be added to the model if it was previously excluded or removed if $Y_V \in M_i$.

The BIC for the proposed model, M_j, is computed in step **7** and the process moves to $nbd(M_j)$ with probability

$$\begin{aligned} P_{acc} &= \min\left\{1, \frac{P(M_j|Y)}{P(M_i|Y)}\right\} \\ &= \min\left\{1, \frac{\exp(-.5BIC_j)}{\exp(-.5BIC_i)}\right\} \\ &= \min\left\{1, \exp(-.5(BIC_j - BIC_i))\right\} \end{aligned} \tag{3.10}$$

or stays in the neighborhood of M_i with probability $1 - P_{acc}$ which is shown if step **8**.

We define the distance between any two models, M_i and M_j to be

$$d_{ij} = (\underline{\delta}_i - \underline{\delta}_j)'(\underline{\delta}_i - \underline{\delta}_j)$$

Throughout the MC^3 simulation, each step through the model space amounts to a jump of one unit of distance when a proposed move is accepted. As sampling from the chain continues, by the nature of the process there is an emphasis on spending more iterations in the neighborhoods of the best models. If there are groups of neighborhoods containing good models that are far apart from one another it may take many iterations to achieve convergence. One method of assessing convergence is the use of multiple sequences using overdispersed starting points [16]. By choosing random starting points for multiple sequences, the expected distance between any two start points is $0.5p$ since the distance between any two models chosen at random from $\mathcal{M}$ is a binomial random variable with parameters p and 0.5. We propose starting new sequences at random with probability s (we use $s = 0.01$) with the starting point of the new sequence being some model chosen at random. Therefore when a new sequence is triggered, the initial model in the new chain is some model in $\mathcal{M}$ as shown in step **9** and the MC^3 starts anew. Recall that our goal in sampling from the chain is model identification and not convergence. While the process cycles within a group of good models we randomly start the process over in a randomly determined spot in the model space in the hopes of finding other groups of likely models if they exist.

The purpose of step **10** is to insure that the best models are visited. Usually, in MCMC simulations, the number of iterations is chosen to achieve convergence to the proper stationary distribution and suggested values of N are on the order of 30000 [21]. In this particular application we are only interested in the neighborhoods that were actually visited throughout the simulation which make up the set $\mathcal{M}^*$ hence model identification is of greater importance than convergence so N maybe be as small as 5000 to attain the desired result. The justification for this is that the posterior probability of a model will not be estimated by the proportion of time the Markov chain spent in the neighborhood of the model, but instead will be approximated using the observed BIC for each model that is visited during the simulation. The assumption inherent in this approach is that all models that are most likely in $\mathcal{M}$ will be visited at least one time in 5000 iterations with the aid of the random restarts of the sequence from the previous step.

Occam's razor is performed in step **11** which states that models in $\mathcal{M}^*$ that are C or more times less likely than the most likely model in the set have been essentially discredited and should be eliminated. Madigan and Raftery [38] adopted the value of $C = 20$ to eliminate models that were far less likely than the best model. We then have

$$\begin{aligned}
\mathcal{M}^{**} &= \left\{ M_i : \frac{\max_{M_k \in \mathcal{M}^*} P(M_k|Y)}{P(M_i|Y)} \leq C \right\} \\
&= \left\{ M_i : \frac{\max_{M_k \in \mathcal{M}^*} \exp(-0.5BIC_k)}{\exp(-0.5BIC_i)} \leq C \right\}
\end{aligned}$$

$$
\begin{aligned}
&= \left\{ M_i : \max_{M_k \in \mathcal{M}^*} \exp(-0.5(BIC_k - BIC_i)) \leq C \right\} \\
&= \left\{ M_i : \max_{M_k \in \mathcal{M}^*} -0.5(BIC_k - BIC_i) \leq \ln C \right\} \\
&= \left\{ M_i : \min_{M_k \in \mathcal{M}^*} BIC_k - BIC_i \geq -2 \ln C \right\} \\
&= \left\{ M_i : BIC_i \leq 2 \ln C + \min_{M_k \in \mathcal{M}^*} BIC_k \right\}
\end{aligned}
$$

This is the step in the algorithm where models that were identified during the simulation but deemed unlikely in comparison to the best model are removed.

In step **12** the potentially greatly reduced set $\mathcal{M}^{**} \subseteq \mathcal{M}^* \subseteq \mathcal{M}$ is then used to estimate the posterior probabilities of the most likely models. All models not in $\mathcal{M}^{**}$ have an estimated posterior probability of zero and are therefore eliminated from the denominator of equation 4.5.

Any variable that is in a given model has its corresponding position in the vector $\underline{\delta}$ set to one or it is set to zero of the variable is not present. The probability that a variable is a significant contributor to group separation can be assessed by estimating the probability that the variable in question should be present by

$$
\hat{E}[\underline{\delta}] = \sum_{M_i \in \mathcal{M}^{**}} \underline{\delta}_i P(M_i|Y)
$$

so if the estimated probability that any given variable should be in the model is greater than 0.5, it is more likely than not that the variable in question is a significant contributor to group mean separation.

The elements of $\hat{E}[\underline{\delta}]$ are interpreted as the probability that the corresponding variable is active or contributes significantly to the model. If a group of variables has been identified *a priori* to be important in the construction of an index or latent variable then the its overall importance to any given component can be evaluated as the weighted average of the activation probabilities of the variables that construct it. In the example that follows in section 3.8 twenty environmental variables are used to discriminate between locations based a partition of the Index of Biotic Integrity. To illustrate the concept if assessing the contribution of a latent construct, nine of the twenty variables are characterized as pertaining to water chemistry [51]. In the absence of any previously described weighting scheme, the level of water chemistry contribution to group separation can be assessed by examining the average activation probability of the variables in the water chemistry group.

3.7 Interpretation

Canonical variate analysis is a descriptive tool used to help understand the nature of the separation between group means and to identify which explanatory variables contribute to discrimination between the groups. A difficult problem is the interpretation of the axes of separation and variable contribution to that axis. Rencher [60, 61] outlines three methods used to interprete variable contribution to group separation.

One method involves examination of the standardized coefficients. The standardized coefficients are obtained by multiplying the eigenvectors of $E^{-1}H$ by $diag(s_i)$, where s_i is the within sample standard deviation of the i^{th} variable [61, 44]. The standardized coefficients reveal the joint contribution of the individual variables to the discriminant functions [59]. The variables with large magnitude standardized coefficients are considered important in the construction of the discriminant function in the presence of the other variables in the model. If some variables are deleted or added, coefficients will change signs and/or magnitudes, but these descriptive measures depend jointly on all other variables in the model so this would be expected but makes interpretation difficult.

Another method involves calculating partial F-statistics for each variable in the presence of the other variables in the model. The partial F-statistic shows each variables contribution to Wilks' Λ after adjusting for the other variables in the model. In the case of more than two groups, the partial F-values are not associated with a single discriminant function but rather indicate the over all contribution to group separation hence this is not useful if it is desired to interpret individual discriminant functions.

The other popular method of interpretation uses correlations between the original variables and the canonical structures formed. Three sources supplied by commercial software packages known as total, between, and pooled within canonical structures. If $Y_1, \cdots, Y_p$ represent the original data vectors and $V_1, \cdots, V_s$ denote the canonical variates. Let

$$\begin{aligned} Y_{ijk} &= k^{th} \text{ observation in the } j^{th} \text{ group for variable } i \\ Y_{ij\cdot} &= n_j^{-1} \sum_{k=1}^{n_j} Y_{ijk} \\ Y_{i\cdot\cdot} &= g^{-1} \sum_{j=1}^{g} Y_{ij\cdot} \end{aligned}$$

with the appropriate counterparts in the canonical variates. The three canonical structures are then defined as

$$\begin{aligned} r_T(Y_a, V_b) &= \frac{\sum_{j=1}^{g}\sum_{k=1}^{n_j}(Y_{ajk} - Y_{a\cdot\cdot})(V_{bjk} - V_{b\cdot\cdot})}{\sqrt{\sum_{j=1}^{g}\sum_{k=1}^{n_j}(Y_{ajk} - Y_{a\cdot\cdot})^2 \sum_{j=1}^{g}\sum_{k=1}^{n_j}(V_{bjk} - V_{b\cdot\cdot})^2}} \\ r_B(Y_a, V_b) &= \frac{\sum_{j=1}^{g}(Y_{aj\cdot} - Y_{a\cdot\cdot})(V_{bj\cdot} - V_{b\cdot\cdot})}{\sqrt{\sum_{j=1}^{g}(Y_{aj\cdot} - Y_{a\cdot\cdot})^2 \sum_{j=1}^{g}(V_{bj\cdot} - V_{b\cdot\cdot})^2}} \end{aligned}$$

$$r_W(Y_a, V_b) = g^{-1} \sum_{j=1}^{g} \frac{\sum_{k=1}^{n_j}(Y_{ajk} - Y_{aj\cdot})(V_{bjk} - V_{bj\cdot})}{\sqrt{\sum_{k=1}^{n_j}(Y_{ajk} - Y_{aj\cdot})^2 \sum_{k=1}^{n_j}(V_{bjk} - V_{bj\cdot})^2}}$$

Of these three structures, the between structure seems to be most appealing in the sense that the goal of CVA is to describe mean separation and the between structure deals directly with group means. The between canonical structure is composed of the correlations between the group means of each explanatory variable with each set of group means of the canonical variates. It therefore is a measure of how well the group means for the explanatory variables agree with the canonical group means. For any Y_i, $i = 1, \cdots, p$, we have that

$$r_B^2(Y_i, V_1) + \cdots + r_B^2(Y_i, V_s) = 1$$

so $r_B^2(Y_i, V_j)$ can be interpreted as the percentage agreement of the group means profile of variable i with the j^{th} canonical group means. Since the correlation is scale invariant then these structures indicate how much agreement there is between the explanatory group means with the canonical mean profiles, but do not show the extent to which a given variable contributes to the overall group separation.

BMA adds to the interpretability of group mean separation at looking at all possible variable configurations and weighting each configuration by how well it is supported by the data. The variability modeled by this method is due to model uncertainty. For any quantity of interest, Δ, the total variance is

$$\begin{aligned}
\text{Var}(\Delta) &= E[\Delta^2] - E[\Delta]^2 \\
&= \sum_{M \in \mathcal{M}} E[\Delta^2|M]P(M) - \left(\sum_{M \in \mathcal{M}} E[\Delta|M]P(M) \right)^2 \\
&= \sum_{M \in \mathcal{M}} (E[\Delta^2|M] - E[\Delta|M]^2 + E[\Delta|M]^2)P(M) - \left(\sum_{M \in \mathcal{M}} E[\Delta|M]P(M) \right)^2 \\
&= \sum_{M \in \mathcal{M}} (E[\Delta^2|M] - E[\Delta|M]^2)P(M) + \sum_{M \in \mathcal{M}} E[\Delta|M]^2 P(M) - \left(\sum_{M \in \mathcal{M}} E[\Delta|M]P(M) \right)^2 \\
&= \sum_{M \in \mathcal{M}} \text{Var}(\Delta|M)P(M) + E_{\mathcal{M}}\left[E[\Delta|M]^2\right] - E_{\mathcal{M}}[E[\Delta|M]]^2 \\
&= E_{\mathcal{M}}[\text{Var}(\Delta|M)] + \text{Var}_{\mathcal{M}}(E[\Delta|M]) \\
&= \text{Within} + \text{Between}
\end{aligned}$$

The "within" variance component is the sample variation. It can be estimated for any given model using exact distribution theory, asymptotic distribution theory, or via some simulation technique such as bootstrapping depending on the parameter of interest. The overall estimate of this component is then obtained as the weighted average of the individual model sample variance estimates.

The "between" component is the model uncertainty variance component and accounts for parameter differences between each of the models considered. The quantity can be estimated using plug in estimates of the expected values and then taking the variance of those values based on the posterior probabilities of the models.

The quantities we are interested in fall into one of two categories; variable assessment, and the nature of group separation. Variable contribution to group separation will be illustrated using the between canonical structure scores along with the pooled within-class standardized canonical coefficients. The nature of group separation will be evaluated by forming confidence regions about the group means of the canonical variates along with the canonical scores associated with individual observations.

3.8 Application

Biological data were gathered from 1988 to 1994 by the Ohio Environmental protection Agency (EPA) over the Eastern Corn Belt Plains ecoregion of Ohio [50]. The status of a fish community was assessed at 178 sites in the region using Ohio EPA's Index of Biological Integrity (IBI). The IBI is the sum of twelve metrics measuring the quantities of various species of fish collected via electroshocking methods [11]. Each of the metrics are evaluated on a five point Likert scale so the IBI can take on any integer value from 12 to 60. The index reflects total native species composition, indicator species composition, pollutant intolerant and tolerant species composition, and fish condition. The higher the IBI score, the healthier the aquatic ecosystem; conversely, the lower the index, the poorer the health of the aquatic ecosystem [11]. We partition the sites by IBI score into one of four classifications as follows:

Group	Lower Limit	Upper Limit	Group Size
I	50	60	37
II	40	49	64
III	30	39	51
IV	12	29	26

The explanatory variables for this application consists of 20 variables identified as biological "stressor" variables. Nine of the variables are characterized as pertaining to water chemistry and the remaining eleven variables are classified as habitat variables. In the original analysis of this set of data, various transformations were applied to individual variables to attain approximate univariate normality of the data [50]. Sections C.2 and C.3 show the names, brief descriptions, and transformations used for the variables analyzed in this illustration.

3.8.1 Standard analysis

Performing a classical analysis of the data, a model was selected using stepwise discriminant analysis. The chosen model has five habitat variables (POOL, RIPARIAN, SUBSTRAT, RIPSS, and COVER) and four water chemistry variables (AMM, NOX, ZN, and PHO). The model Wilks' Λ is 0.4805 (p-value $< .0001$) all the three canonical correlations (0.5971, 0.4010, 0.3322) were significant at the 0.01 level.

The canonical variate group means as functions of the aforementioned retained variables are shown in figure 3.1. The group mean profile for the first canonical variate is decreasing and approximately linear and accounts for 64% of the group separation. The group mean profile for the second canonical variate is quadratic shaped and accounts for 22% of the group separation.

Figure 3.2 shows the percent agreement between the group means of the explanatory variables with group mean structure of the canonical variates using the squared between canonical structure coefficients (BCSC). Since the canonical variates are orthogonal and have the same rank as the data, then each of the retained original variables contributes a percentage of it's mean separation to each variate. We see that the mean sepearation for the variable RIPARIAN, for example, is almost 100% explained by the first canonical variate, whereas NOX is practically evenly explained across each of the three canonical variates. For these data, we conclude that AMM, POOL, RIPARIAN, SUBSTRAT, ZN, RIPSS, and COVER are mainly in agreement with the mean profile of the first canonical variate. Variables NOX and PHO are essentially an even match with each of the canonical variate group mean profiles.

Figure 3.1: Canonical Group Means: CAN1 -vs- CAN2

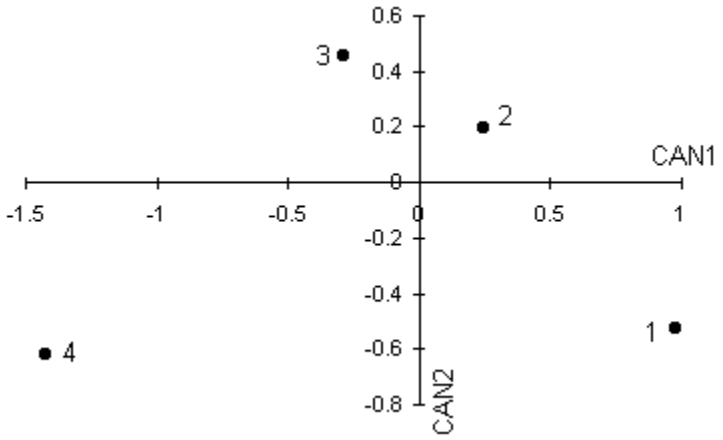
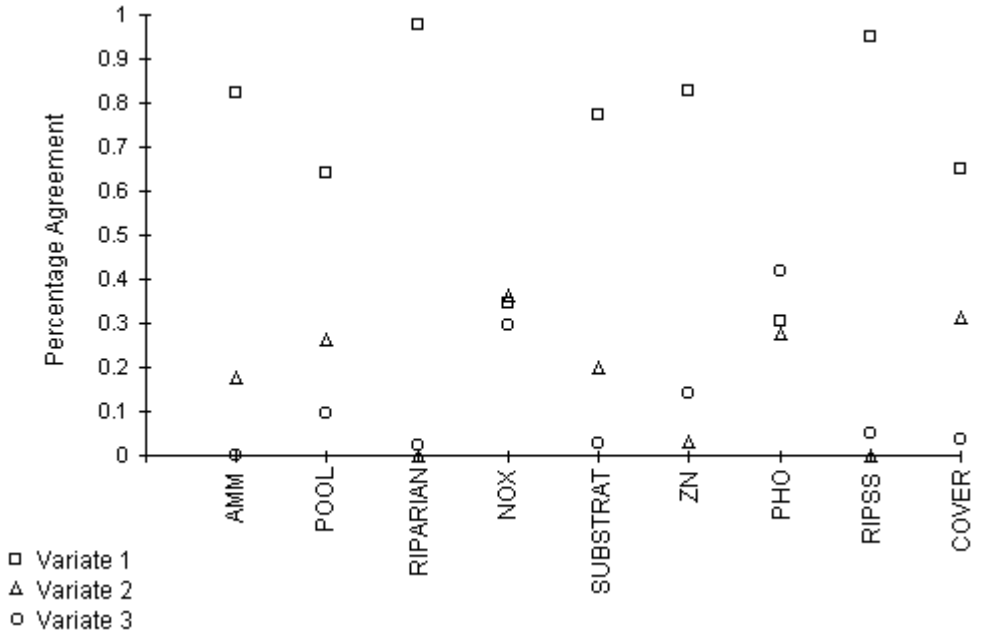


Figure 3.3 shows the magnitudes of the pooled within-class standardized canonical coeffi-

Figure 3.2: Squared Between Canonical Structure Coefficients



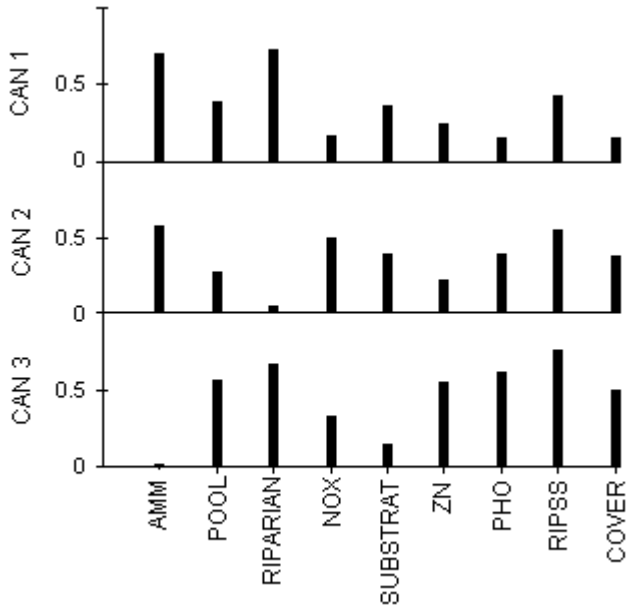
cients (PWCSCC). From this plot, one may conclude that AMM and RIPARIAN are important to the construction first canonical variate, AMM, NOX, and RIPSS contribute most heavily to the second canonical variate, and finally that POOL, RIPARIAN, ZN, PHO, and RIPSS are most important to the third canonical variate. This leaves SUBSTRAT and COVER as unassigned in terms of how each contributes to mean separation.

When comparing two possible conclusions drawn from either the BCSC or when using the PWCSCC, we see there is not much agreement for this data set. Also, since there is no measure of variablity what constitutes a large magnitude value is made based on a point estimate alone and is risky and the cutoff value for important versus unimportant somewhat arbitrary.

3.8.2 BMA analysis

The analysis will be approached in two different ways using BMA methodology. The goal of the first approach is variable assessment. In the variable assessment phase of the analysis, the posterior probability that a given variable is active in a model is estimated and these values are used to aid in the interpretation of the group separation. The second approach evaluates the magnitude of the model uncertainty variance component of the BCSC and the PWCSCC.

Figure 3.3: Pooled Within-Class Standardized Canonical Coefficient Magnitudes



Variable assessment

The contribution of each variable to group separation is made by the estimation of the probability that any given variable is present in the model. For each model $M_i \in \mathcal{M}^{**}$ we have the estimated posterior probability of the model, $P(M_i|Y)$, and the vector of indicator variables, $\underline{\delta}_i$, that represents which variables are present (1) and which are absent (0). The estimated posterior probability that a variable, Y_i, contributes significantly to group separation is given by

$$P(Y_i \text{ is active}) \quad = \quad P(\delta_{\cdot i} = 1) = \sum_{M_j \in \mathcal{M}^{**}} \delta_{ji} P(M_j|Y)$$

When $P(Y_i \text{ is active}) > 0.5$ it is more likely than not that variable Y_i contributes significantly to group separation.

One of the criticisms of using the between canonical structure coefficients is that they reflect only univariate information and not the joint contribution since they ignore the other variables in the model [59, 60]. We feel that another problem with this measure is that the amount of separation is not reflected by the measure. Rather it simply indicates how the information about separation for the variable is divided amongst the variates.

As previously defined, the loading associated with variable i with canonical variate j is denoted by $r_B(Y_i, V_j)$. As a correlation, each loading has a range between -1 and 1. This

value measures how correlated the shape of the mean profile of the variable is with the mean profile of the canonical variate in question, but does not the reflect the magnitude of group separation. Let

$$r_B^*(Y_i, V_j) \quad = \quad r_B(Y_i, V_j) P(Y_i \text{ is active})$$

be a scaled correlation. In the interpretation phase of an analysis, we look for large magnitude loadings hence in order for $|r_B^*(Y_i, V_j)|$ to be large, both the amount of agreement and the degree to which a variable discriminates must be large. Also, this criteria uses the posterior probability that a variable is active in a model and is made in the multivariate context since the probability takes into account the other variables in the model.

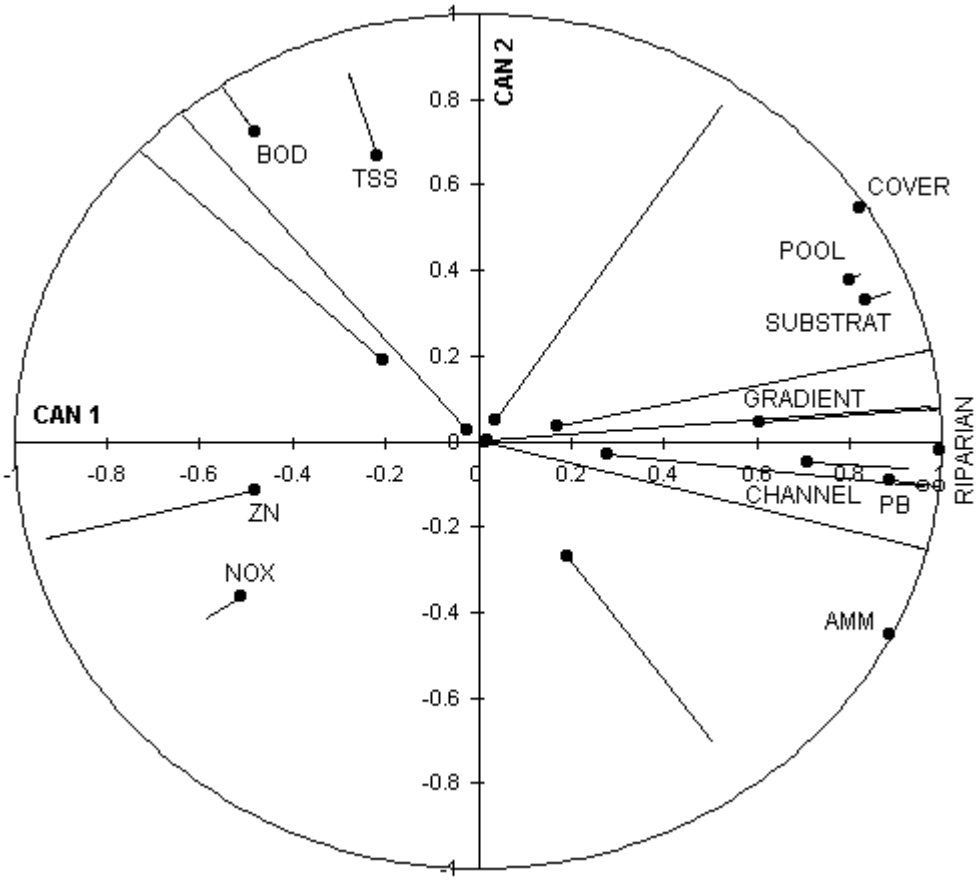
Table 3.1: Activation Probabilities

Variable	Prob	Std. Err.	Variable	Prob	Std. Err.
RIPARIAN	1.0000	0.0000	GRADIENT	0.6080	0.0022
AMM	1.0000	0.0000	ZN	0.5169	0.0036
POOL	0.9970	0.0006	PHO	0.3827	0.0025
COVER	0.9858	0.0009	RIPSS	0.2883	0.0027
SUBSTRAT	0.9439	0.0023	COD	0.2766	0.0025
PB	0.8947	0.0025	EMBSS	0.1741	0.0023
NOX	0.8729	0.0033	SILTSS	0.0631	0.0016
BOD	0.8705	0.0018	FE	0.0367	0.0017
TSS	0.7766	0.0035	RIFFLE	0.0190	0.0013
CHANNEL	0.7644	0.0026	RIFEMSS	0.0116	0.0007

Ten runs of 10000 iterations each were run using the modified MC^3 algorithm. The activation probabilities are shown in table 3.1 in decreasing order of importance. RIPSS and PHO were selected by the stepwise method but have low posterior probabilities of being active (i.e. less than 0.5). The variables PB, BOD, TSS, CHANNEL, and GRADIENT each had high posterior probabilities and were not selected by the stepwise model.

Figure 3.4 shows the BCSC for the first two canonical variates. Each variable is represented by a dot and a ray. The dot represents the scaled loading value along each of the first two canonical variate axis and is located at coordinate $(r_B^*(Y_i, V_1), r_B^*(Y_i, V_2))$. The ray extends from the scaled loading and terminates at the unscaled loading value at coordinate $(r_B(Y_i, V_1), r_B(Y_i, V_2))$. This graph emphasizes three major points of interest. Firstly, it shows how the individual variables are aligned along the first two canonical axis. For example, we could classify each variable as either primarily aligned with the first canonical variate (RIPARIAN, PB, CHANNEL, GRADIENT, and ZN), mainly aligned with the second canonical variate (TSS), or a mix of the two variates (AMM, POOL, COVER, SUBSTRAT, NOX,

Figure 3.4: Scaled Between Canonical Loadings for CAN 1 and CAN 2



and BOD). Next, the plot identifies variables that provide poor discrimination as the scaled loading coordinate will be closer to the origin when compared to a variable that provides stronger mean separation. For example, BOD is further from the origin than TSS so it provides a greater degree of mean separation. Finally, the coordinate value where the ray terminates indicates how much mean separation for a given variable is accounted for in the first two canonical variates. For example, the ray for ZN terminates closer to the outer circle (best possible separation) than the ray for NOX hence ZN is more in line with the first two canonical variates, but since the scaled loading for NOX is farther from the origin than the scaled loading for ZN then we conclude that NOX still provides stronger group separation.

Model uncertainty

Recall that the BMA estimated value of any quantity of interest, Δ, is the plug in estimate obtained from the estimated expected value

$$\begin{aligned}\hat{\Delta} &= \hat{E}(\Delta) \\ &= \sum_{M_i \in \mathcal{M}^{**}} \hat{E}(\Delta|M_i)P(M_i|Y)\end{aligned}$$

The BMA estimate of the BCSC along with the model uncertainty variance component estimates are shown for the three canonical variates in table 3.2. The average magnitude loading over the 20 variables and three canonical variates is 0.24698 and the underlined estimates in table 3.2 denote loadings with above average magnitudes. An interpretation that may be made from these values is that the variables with large magnitudes show which variables are important with respect to group separation and which canonical variate they are most closely identified with. For example, CHANNEL, RIPARIAN, GRADIENT, PB, and AMM are associated with CAN 1 only, COVER, POOL, SUBSTRAT, BOD, and NOX with CAN 1 and 2, and TSS with each of CAN 1–3.

Table 3.3 shows the BMA estimates of the PWCSCC. As with the BCSC, the overall average magnitude was computed and each score larger than 0.1532 is interpreted as important to the construction of the variate and is identified as such by being underlined in the table. We see that the most important variables based on this criterion are CHANNEL, COVER, RIPARIAN, POOL, SUBSTRAT, GRADIENT, TSS, AMM, PB, and NOX.

We see that there are only minor differences in the sets of variables selected as most important based on these two criteria. BOD was selected when the BCSC criteria was used but not when PWCSCC was used, but TSS was identified when using PWCSCC but not when the BCSC was used. Both BOD and TSS were identified as important when using the activation probabilities.

The means of the first two canonical variates are shown in figure 3.5. With the exception of the eigenvector for the second canonical variate switching signs, the overall mean pattern matches that shown in figure 3.1. This occurs because eigenvector identification is unique up to a scalar multiplier so the magnitude is generally scaled to unity, but in any given determination from a random sample, the signs may flip. The major difference between the two plots is that sample variability and the model uncertainty variance component have been added to the graph in figure 3.5. The sample variability for each group mean and canonical variate was estimated using 1000 bootstrap samples of the data set. The ellipses represent ± 2 standard deviations of the sampling variability. The plotted crosses show the model uncertainty in the means as 1000 independent draws were made from the posterior model space. From the estimated values of the variance components, shown in table 3.4, we see that sampling variability is between 1.2 and 2.5 orders of magnitude larger than the

Table 3.2: BMA Estimated Between Canonical Structure Coefficients

	CAN 1		CAN 2		CAN 3	
Variable	Est.	S.D.	Est.	S.D.	Est.	S.D.
CHANNEL	0.7430	0.0261	0.0254	0.0037	-0.0880	0.0177
COVER	0.7793	0.0061	0.5654	0.0104	0.1182	0.0080
RIPSS	0.2298	0.0278	0.0008	0.0012	-0.0205	0.0067
RIPARIAN	0.9851	0.0003	0.0251	0.0020	0.0919	0.0094
POOL	0.7955	0.0028	0.4681	0.0097	-0.1036	0.0225
EMBSS	0.1758	0.0248	0.0473	0.0067	0.0204	0.0045
RIFEMSS	0.0123	0.0073	-0.0020	0.0012	-0.0011	0.0007
RIFFLE	0.0173	0.0086	0.0035	0.0018	-0.0003	0.0006
SILTSS	0.0336	0.0081	0.0618	0.0149	-0.0017	0.0037
SUBSTRAT	0.8146	0.0147	0.3950	0.0109	-0.0345	0.0130
GRADIENT	0.5930	0.0326	0.0788	0.0047	0.0210	0.0020
FE	-0.0254	0.0087	0.0267	0.0091	0.0029	0.0018
TSS	-0.2792	0.0105	0.5566	0.0228	0.2637	0.0276
BOD	-0.5354	0.0136	0.6884	0.0174	0.0610	0.0047
COD	-0.2209	0.0232	0.1837	0.0196	0.0120	0.0031
AMM	0.9132	0.0008	-0.3840	0.0074	0.0123	0.0052
PB	0.8881	0.0206	-0.0395	0.0022	0.0504	0.0044
AN	-0.4568	0.0304	-0.1224	0.0090	-0.0111	0.0160
NOX	-0.5194	0.0129	-0.4651	0.0162	0.2445	0.0329
PHO	0.2024	0.0178	-0.2183	0.0219	0.0706	0.0211

variance due to model uncertainty for these data, but these size differences are data and model dependent and will not hold true in general.

Group 4 (IBI scores less than 30) appears to have the greatest degree of separation from the other groups whereas groups 2 and 3 have the most overlap implying they may be most similar with respect to the variables measured. Finally, as CAN 1 decreases IBI also tends to decrease. The interpretation of CAN 2 is less clear but since the mean profile has a quadratic shape it may be related to the square of CAN 1.

3.9 Conclusion

The classical approach to model building selects a single model which is equivalent to letting the posterior probability of that model be equal to one while all other models in the space

Table 3.3: BMA Estimated Pooled Within-Class Standardized Canonical Coefficients

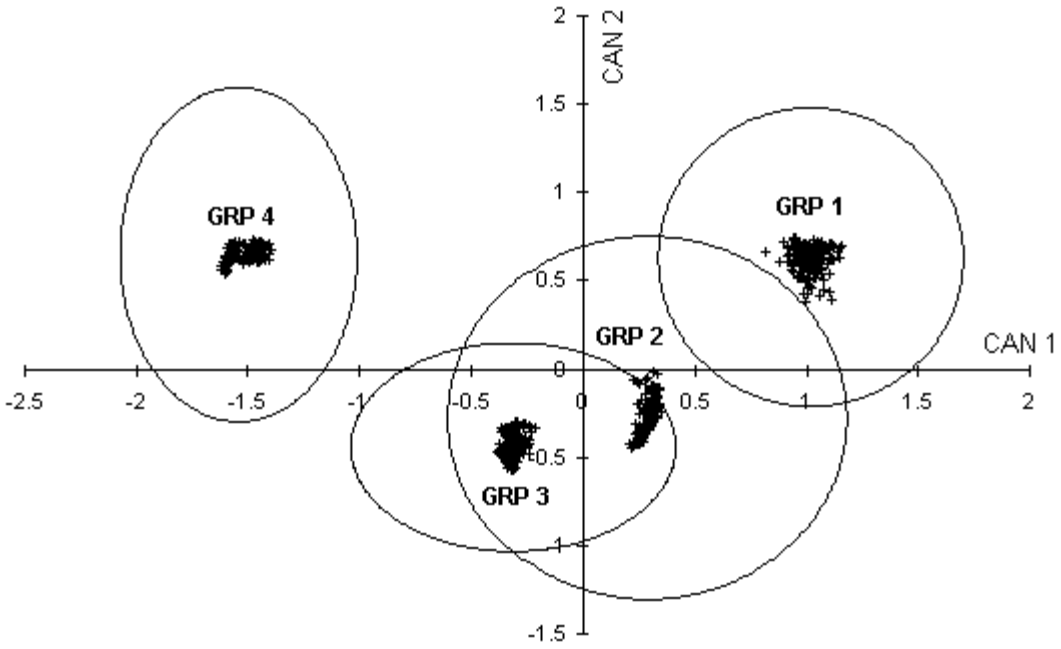
	CAN 1		CAN 2		CAN 3	
Variable	Est.	S.D.	Est.	S.D.	Est.	S.D.
CHANNEL	-0.0298	0.0037	-0.4106	0.0159	-0.2673	0.0356
COVER	-0.2107	0.0038	0.5485	0.0117	0.3630	0.0360
RIPSS	-0.0802	0.0098	-0.0811	0.0127	-0.0727	0.0204
RIPARIAN	0.5359	0.0067	-0.0614	0.0066	0.2699	0.0318
POOL	0.3839	0.0019	0.2441	0.0065	-0.2087	0.0328
EMBSS	0.0546	0.0081	0.0334	0.0053	0.0001	0.0024
RIFEMSS	-0.0012	0.0011	-0.0029	0.0020	-0.0012	0.0008
RIFFLE	0.0022	0.0011	-0.0024	0.0013	0.0006	0.0009
SILTSS	-0.0080	0.0023	0.0185	0.0045	-0.0024	0.0027
SUBSTRAT	0.3326	0.0065	0.3216	0.0099	-0.0129	0.0105
GRADIENT	0.0956	0.0058	0.2625	0.0148	0.0297	0.0085
FE	0.0101	0.0038	0.0018	0.0012	0.0010	0.0018
TSS	0.3444	0.0130	-0.0271	0.0058	0.1002	0.0121
BOD	-0.0311	0.0095	0.5966	0.0159	0.2524	0.0159
COD	-0.0756	0.0080	-0.0106	0.0041	-0.0060	0.0064
AMM	0.6526	0.0023	-0.3258	0.0095	0.0855	0.0133
PB	0.4044	0.0099	0.1082	0.0049	0.1439	0.0096
ZN	-0.0780	0.0063	-0.1127	0.0089	0.0140	0.0217
NOX	-0.1552	0.0044	-0.1780	0.0098	0.3002	0.0294
PHO	-0.0493	0.0047	-0.0875	0.0102	0.0865	0.0191

are excluded since they have probability zero. A direct result of selecting a single model is that the variance due to uncertainty is zero which does not accuratly reflect reality since the selected model was not chosen as such a priori. Bayesian model averaging provides a way to estimate and incorporate the previously ignored model uncertainty variance component or used for variable assessment.

When a model is developed empirically via some variable selection scheme, any variances that are estimated are overly optimistic since they are estimated based on the assumption that the model selected is correct. The model uncertainty variance component adjusts any variance estimate to reflect the fact that whatever the true model is, it is unknown and there may be several competitive models that adequately reflect what is happening with the data. The information from each good model is weighted based on its posterior probability, and estimates of desired quantities are formed with more appropriate variance estimates.

The joint contribution of each variable can be measured using BMA in the context of variable

Figure 3.5: Canonical Variate Means for CAN 1 and CAN 2



assessment. The posterior model space is summarized by computing the expected posterior probability that any given variable is active. The activation probabilities can then be used in conjuction with the standard tools that are used such as the structural loadings and standardized coefficients. The addition of the variable assessment information enhances and clarifies the interpretion thus adding value to the analysis.

Table 3.4: Estimated Sampling Variability and Model Uncertainty Variance Components for Canonical Group Means

Variate	Group	Sampling Variance	Model Uncertainty Variance	Order of Magnitude
1	1	0.11337	0.00352	1.51
1	2	0.19739	0.00114	2.24
1	3	0.13013	0.00086	2.18
1	4	0.06739	0.00327	1.31
2	1	0.17344	0.00535	1.51
2	2	0.25494	0.00927	1.44
2	3	0.08254	0.00458	1.26
2	4	0.22200	0.00166	2.12
3	1	0.33912	0.00301	2.05
3	2	0.29511	0.00091	2.51
3	3	0.43431	0.00669	1.81
3	4	0.16220	0.00472	1.54

Note: Order of magnitude $= \log_{10}\left(\frac{\text{Sampling Variance}}{\text{Model Uncertainty Variance}}\right)$

Chapter 4

Canonical Correlation Analysis (CCA)

4.1 Introduction

Situations exist in multivariate systems where measured variables are divided into two sets a priori and each set relates to a separate component of a system and some measure of linear association between the components is desired [34]. The data suitable for this type of analysis consists of n observations in $(p+q)$ numeric measurement variables and the division of the data into sets of p and q variables is made based on some external considerations. Canonical correlation analysis (CCA) is used to identify and quantify the linear association between the two sets of variables [26].

CCA creates $s = \min(p, q)$ new pairs of variables using linear combinations of the original variables from each set. The new variables, called canonical variates, are formed so that the first pair has the largest correlation of any linear combination of the original variables. Subsequent pairs also have maximized correlation subject to the constraint that they are uncorrelated with each previous pair [26]. Symbolically, given $X_1, \cdots, X_p$ and $Y_1, \cdots, Y_q$ then

$$\begin{aligned} U_i &= X\underline{a}_i \\ V_i &= Y\underline{b}_i \end{aligned}$$

for $i = 1, \cdots, s$ where

$$\begin{aligned} Corr(U_i, V_j) &= 0 \text{ if } i \neq j \\ Corr(U_i, U_j) &= 0 \text{ if } i \neq j \\ Corr(V_i, V_j) &= 0 \text{ if } i \neq j \\ Corr(U_i, V_i) &= \rho_i \end{aligned}$$

The vector of coefficients that make up a_i and b_i are the i^{th} eigenvectors of

$$\begin{aligned} M_p &= S_{xx}^{-1} S_{xy} S_{yy}^{-1} S_{yx} \qquad \text{and} \\ M_q &= S_{yy}^{-1} S_{yx} S_{xx}^{-1} S_{xy} \end{aligned}$$

respectively where S_{xx}, S_{yy}, S_{yx} and S_{xy} are the sums of squares and cross product matrices [34]. The canonical correlations are derived from the first s eigenvalues of either M_p or M_q by

$$\rho_i = \sqrt{\frac{\lambda_i}{1+\lambda_i}}$$

The main hypothesis tested in a CCA model is $H_0 : \rho_1 = \cdots = \rho_s = 0$. The likelihood ratio is used as the test statistic and thsi results in the same test in multivariate regression, $H_0 : \beta = 0$, and also that used in canonical variate analysis $H_0 : \underline{\mu}_1 = \cdots = \underline{\mu}_g$.

The four most commonly used statistics to test the hypothesis that the s canonical correlations are equal zero are Wilks' Λ, Roy's greatest root, Pillai's test, and the Lawley-Hotelling test. The test statistics can each be written in terms of the eigenvalues of M_p or M_q where $\lambda_1 > \lambda_2 > \cdots > \lambda_s$, and $s = \min(p, q)$ as shown below [60].

Wilks' lambda	$\Lambda = \prod_{i=1}^{s} \frac{1}{1+\lambda_i}$
Roy's	$\theta = \lambda_1$
Pillai's	$V = \sum_{i=s}^{k} \frac{\lambda_i}{1+\lambda_i}$
Lawley-Hotelling	$U = \sum_{i=1}^{s} \lambda_i$

While inferences made are similar to those of multivariate regression and canonical variate analysis, CCA is fundamentally different since the variable sets are treated symmetrically because neither set is treated as dependent on the other [60].

4.2 The Model

The goal in model building is to identify subsets of variables that best contribute to the linear association between the sets of variables. CCA is a multivariate dimension reduction method where it is hoped that the linear association between two sets of variables may be summarized by a few pairs of optimally constructed variables. The procedure is not as straightforward as computing simple correlations since the correlation between sets is determined by adjusting for the within set correlation structure [35]. After the canonical variates are formed, some interpretation of their meaning is often desired.

Each model, M_i, is represented by vector $\underline{\delta}_i$ where

$$\delta_{ij} = \begin{cases} 0 & \text{if the } j^{th} \text{ variable is not in } M_i \\ 1 & \text{if the } j^{th} \text{ variable is in } M_i \end{cases}$$

where $j = 1, \cdots, (p+q)$, and the j^{th} variable is defined as

$$V_j = \begin{cases} X_j & \text{if the } j \leq p \\ Y_{j-p} & \text{if the } j > p \end{cases}$$

The models are numbered by the index "i" and is obtained by

$$i = \sum_{j=1}^{p+q} \delta_{ij} 2^{j-1}$$

There are $2^{p+q} - 3$ models in the model space that can be analyzed since three models are not analyzable; the model with no variables, the model with no X variables, and the model with no Y variables. The likelihood ratio is not defined for these models so we set Wilks' lambda equal to one so that these three models may be included in order complete to the model space.

4.2.1 Computations

The calculation of the eigenvalues of M_p or M_q was shown to have value in the previous subsection. In general, the matrix M_p and M_q are not symmetric and many algorithms for computing eigenvalues accept only symmetric matrices. It can be shown [60] that

$$S_{yy}^{-1/2} S_{yx} S_{xx}^{-1} S_{xy} S_{yy}^{-1/2} \quad \text{and}$$
$$S_{xx}^{-1/2} S_{xy} S_{yy}^{-1} S_{yx} S_{xx}^{-1/2}$$

are symmetric and that the first $s = \min(p, q)$ eigenvalues of each are the squared canonical correlations that relate the linear combinations of X and Y variables.

4.3 Model selection

Currently there has been little research devoted to model selection and CCA [2]. A stepwise method of model selection, which is a quite popular method of model building regression and canonical variate models, is easily adapted to accommodate the CCA framework. When the number of variables is large, sequential procedures have generally been the only practical

way to construct a reasonable model. The stepwise method is an extension of the forward procedure and makes use of Wilks' lambda statistic and partial F-tests.

Recall that the data consists of n observations of variables $X_1, \cdots, X_p, Y_1, \cdots, Y_q$. The first step of the stepwise procedure is to choose X_i, Y_j from all possible $p \times q$ variable combinations so that the squared correlation is maximized. Wilks' lambda is equal to $1 - r_{ij}^2$ and the exact $F-$test can be performed to check for significance. If the best combination fails to be significant then the procedure stops.

In the second step, the variable from either X or Y yielding the smallest partial Λ for each variable adjusted for the first two variables in the model is considered for entry into the model. The partial Λ-statistic is based on the full and reduced model test for a subset [60]:

$$\Lambda(\bullet|x_i, y_j) \quad = \quad \frac{\Lambda(\bullet, x_i, y_j)}{\Lambda(x_i, y_j)}$$

for each $\bullet \neq x_i, y_j$, and the variable that minimizes $\Lambda(\bullet|x_i, y_j)$ is the candidate to enter the model next. If the partial F-test is significant then the third variable is added, if not, then the procedure stops. At this point, when there are more than two variables in the model, a partial F-test is performed to see if any variables may be removed. The process continues alternating between testing to add the best possible of the remaining variables and testing whether the weakest already present in the model should be deleted. When no variables can be added or removed, the process stops.

Since one variable is either being added or removed one at a time, there is an exact F test where

$$F \quad = \quad \frac{1 - \Lambda}{\Lambda} \frac{n - p^* - q^*}{w} \tag{4.1}$$

follows an F distribution with w and $n - p^* - q^*$ degrees of freedom where

$$w = \begin{cases} p^* & \text{if an } X \text{ variable is being tested and} \\ q^* & \text{if a } Y \text{ variable is being tested} \end{cases}$$

and p^* and q^* are the number of variables currently in the model from X and Y respectively. Note that if either p^* or q^* is zero then the configuration would result in an Wilks' lambda equal to one and an $F-$statistic of zero so this condition would not occur during the variable elimination portion of the procedure.

Though each test can be conducted at the α level of significance, the overall level of significance of the model is unknown since the number of tests to be performed before hand is unknown. As such, all p-values obtained from various tests of interest do not have the usual interpretation.

While a stepwise method has been outlined above, no currently available statistical package allows for any form of model building in a CCA analysis. In practice, all variables from

both sets X and Y are generally kept in the model and no formal testing is done to remove variables that may not be contributing to existing linear associations.

4.4 Bayesian Model Averaging (BMA)

When a model selection method such as an all possible subsets or stepwise procedure is used, the single model obtained is assumed to be the *correct* model. All future inferences and predictions that are made with the model do not account for the uncertainty involved in the selection process. Alternatively, the model obtained using BMA does incorporate the variance component associated with the uncertainty of model building.

There is a standard Bayesian solution to the problem of accounting for model uncertainty. If the model space is $\mathcal{M} = \{M_1, \cdots, M_T\}$ then the posterior probability of M_i given the data is

$$P(M_i|\text{data}) = \frac{P(\text{data}|M_i)P(M_i)}{\sum_{M_j \in \mathcal{M}} P(\text{data}|M_j)P(M_j)} \tag{4.2}$$

where $P(M_i)$ denotes the prior probability of each model and $P(\text{data}|M_i)$ is the marginal likelihood of the data. Generally, each model has been assumed to be equally likely *a priori*, so equation 4.2 simplifies to

$$P(M_i|\text{data}) = \frac{P(\text{data}|M_i)}{\sum_{M_j \in \mathcal{M}} P(\text{data}|M_j)}$$

Now, the marginal likelihood of the data is

$$P(\text{data}|M_i) = \int P(\text{data}|M_i, \underline{\theta}_i)\pi(\underline{\theta}_i)d\underline{\theta}_i \tag{4.3}$$

where $\underline{\theta}_i$ is the unknown model parameters with joint prior density $\pi(\underline{\theta}_i)$. Hoeting [21] shows for univariate multiple regression that the marginal likelihood follows an n-dimensional non-central Student's t-distribution when proper conjugate priors are used. This result can be used if hyperparameters are chosen so that the prior density it calibrated to the data. Raftery [55] approximates equation 4.3 using the *Bayes Information Criterion* (BIC) which is a penalized likelihood measure. For the case of linear regression Raftery shows that

$$\begin{aligned} P(\text{data}|M_i) &\propto \exp\left(-0.5BIC_i\right) \\ &= \exp\Big(-0.5(n\ln(1-r_i^2) + p_i \ln n)\Big) \end{aligned} \tag{4.4}$$

where r_i^2 is the model *r-square*, p_i is the number of independent variables in model M_i, and n is the number of observations. Using Raftery's approximation requires no calibration to

the data and is composed of readily available regression information. By normalizing we get the posterior probability of model M_i given the data is

$$P(M_i|\text{data}) \approx \frac{\exp(-0.5BIC_i)}{\sum_{M_j \in \mathcal{M}} \exp(-0.5BIC_j)} \tag{4.5}$$

4.4.1 Specification on Measure of Association

In order to use the BIC approximation we must supply an *r-square* like measure of association. Cramer and Nicewander [7] showed how the four multivariate measures discussed in section 4.1 may be converted into invariant measures of multivariate association.

Wilk's lambda	$\eta^2_\Lambda = 1 - \Lambda = 1 - \prod_{i=1}^{k}(1 - r_i^2)$
Roy's	$\eta^2_\theta = \theta = r_1^2$
Pillai's	$\eta^2_V = k^{-1}V = k^{-1}\sum_{i=1}^{k} r_i^2$
Lawley-Hotelling	$\eta^2_U = \frac{k^{-1}U}{1+k^{-1}U} = \frac{k^{-1}\sum_{i=1}^{k} r_i^2(1-r_i^2)^{-1}}{1+k^{-1}\sum_{i=1}^{k} r_i^2(1-r_i^2)^{-1}}$

Even though none of the four multivariate tests statistics is uniformly most powerful, as previously stated, the measure of associations listed above provide ordered values so that $\eta^2_\Lambda \geq \eta^2_\theta \geq \eta^2_U \geq \eta^2_V$ (proof shown in A.3).

The BIC penalizes heavily for parsimony so it makes sense to use a measure of based on η^2_Λ as the *r-square* in the likelihood portion since it is the least conservative measure of association of those considered and using one of the weaker measures of association would only serve to amplify the penalization. However, there is a potential problem associated with inflation of this measure due to sample size and total number of variables since it is not uncommon for data sets where CCA is used to have large values of p and q relative to n. When no linear association exists the expected value [58] of the measure is

$$E[\eta^2_\Lambda] = 1 - \prod_{i=0}^{p-1}\left(1 - \frac{q}{n-i}\right)$$

To illustrate, suppose that $p = 30$, $q = 30$, and $n = 500$, then the expected value of $\eta^2_\Lambda = 0.8525$ assuming no association.

We propose a measure of association, η^2_*, where

$$\eta^2_* = 1 - \frac{\Lambda}{E[\Lambda]}$$

This measure of association is formed using the likelihood ratio but is adjusted for the p, q and n. From the above scenario posed where $p = q = 30$ and $n = 500$, three scenarios were

simulated 100 times using Monte Carlo methods. In each simulation, the data were normal *iid* with either no linear associations specified, the linear relationship $Corr^2(Y_1, X_1) = 0.5$, or the linear relationship $Corr^2(Y_1, X_1) = 0.8$ with the results shown below.

	No relationship	$Corr^2(Y_1, X_1) = 0.5$	$Corr^2(Y_1, X_1) = 0.8$
$\hat{E}(\eta_*^2)$	0.00175	0.49300	0.79736
$\hat{E}(\eta_\Lambda^2)$	0.85402	0.92586	0.97037

It is easily seen that the estimated expected values of η_*^2 are more reasonable for the scenarios posed than those values obtained from η_Λ^2.

4.4.2 Prior Specification on Model Space

The selection of what prior should be used on the model space is an open problem for BMA in general [6]. In many published applications of BMA, only a uniform prior on the model space has assumed (see Clyde [6], Hoeting [21, 22], Madigan [38] and Raftery [55, 56] for example). The *uniform* prior on the model space is equivalent to assuming that each of $p+q$ variables independently has a probability of 0.5 to be included in any randomly selected model since

$$\begin{aligned} P(M_i) &= \prod_{i=1}^{p+q} 0.5 \\ &= 0.5^{p+q} \\ &= 2^{-(p+q)} \end{aligned}$$

for all $M_i \in \mathcal{M}$.

A more informative prior can be obtained by assuming each variable independently has a probability of θ to be included in any model, so the prior takes on the form

$$P(M_i) = \theta^k (1-\theta)^{p+q-k}$$

where k is the number of variables in M_i for all $M_i \in \mathcal{M}$. Using this prior is equivalent to assuming a uniform prior on the model space and approximating the marginal likelihood with a generalized information criteria (GIC) [48]

$$P(\text{data}|M_i) \approx \frac{\exp(-0.5 GIC_i)}{\sum_{M_j \in \mathcal{M}} \exp(GIC_j)}$$

where

$$GIC_i = n \ln(1 - r_i^2) + k_i \left(\ln n - 2 \ln \frac{\theta}{1-\theta} \right)$$

where k_i is the number of variables in M_i (proof in section A.2). This prior will be referred to as the "*size*" prior since all models with the same number of variables are equally probable.

If it is assumed that each variable has a separate probability, θ_i, of being in a randomly selected model, then the prior probability for any model is given by

$$P(M_j) = \prod_{\delta_{ij}=1} \theta_i \prod_{\delta_{ij}=0} (1-\theta_i)$$

where

$$\delta_{ij} = \begin{cases} 0 & \text{if variable } i \text{ is not in } M_j \\ 1 & \text{if variable } i \text{ is in } M_j \end{cases}$$

A variable that is believed to be more important would be assigned a higher probability of inclusion than one that is considered to be less important. This prior will be referred to as the "*variable*" prior since the variables have separate probabilities of being active.

Another prior that could be considered would be the totally informed prior such that

$$P(M_i) = \pi_i \geq 0 \quad \forall M_i \in \mathcal{M}$$

hence each model is individually assigned a prior probability π_i. This prior will be referred to as the "*individual*" prior since each model is dealt with individually.

The distribution of models under the *uniform* and *size* priors follows a binomial distribution with parameters $p+q$ and θ (where $\theta = 0.5$ for *uniform* prior). As such, the *a priori* number of variables that are expected to be important in the model is $\theta(p+q)$. Under the *uniform* prior this translates to the belief that half the variables are important, so if there is some reason to believe that this is not reasonable or desirable, then the *size* prior may be more appropriate. If it is believed before hand that certain variables are more or less important than others, then the *variable* prior may be more appropriate to use. Since the model space will usually be too large to assign individual probabilities to models then the *individual* prior may not be practical. Also, without expert knowledge, the values assigned to models using this prior may be meaningless.

The four priors described require different knowledge and expectations of the model space. The order in which the priors were presented represents the need for an increasing amount of knowledge about the model space. In a situation where there may be only a few candidate models, the *individual* prior may be appropriate if expert opinion or empirical results drive the assignment of the prior probabilities for each model. In applications such as multiple regression or canonical variate analysis, the *variable* prior may be appropriate if expert opinion or information from pilot studies is available on the relative contributions of each variable. For situations where the probability of inclusion for individual variables does not have such a straight-forward interpretation, such as principal components analysis and canonical correlation analysis, then the *size* or *uniform* priors may be more appropriate.

4.5 Stochastic search of model space

Stepwise methods are often used when it is not practical to evaluate a large number of possible models. While this algorithm is convienient, it is deterministic for a given data set. Slight perturbations in the data can result in a very different "best" model. By adopting a stochastic search of the model space we are able to identify the models best supported by the data almost surely [64, 49, 20]. Madigan and York [39] implemented the Markov chain Monte Carlo model composition (MC^3) procedure on graphical models. Hoeting [21], and Hoeting, Madigan, and Raftery [22] applied the method to univariate multiple regression. One advantage in using the MC^3 approach is that the model selection process is stochastic, and each model will be visited during the simulation in proportion to how well it is supported by the data. Therefore, all good models will be visited more often than those that are not supported by the data, so we get a summary of all the best models and not just a single snapshot that is obtained using a stepwise procedure.

The stochastic search of the model space is made necessary by the potentially enormous number of models in the denominator of equation 4.5. The MC^3 method is used to reduce the number of terms in this sum by focusing on the most probable models and eliminating those models that are not supported by the data.

The states of the Markov chain to be sampled from are the individual models in $\mathcal{M}$ hence the chain is discrete and finite. In order to insure the proper stationary distribution we must specify how to move from one model to another. This task is accomplished by forming neighborhoods around each model [39]. The neighborhood, centered at an arbitrary model M_i, denoted by $nbd(M_i)$, consists of model M_i and every other model that can be obtained by either adding a single variable to M_i or removing a single variable from M_j.

The transition from one neighborhood to another is accomplished using Hasting's [20] method. Given that the current state of the Markov chain is $nbd(M_i)$, the models in $nbd(M_i)$ are sampled with equal probability. Suppose $M_k \in nbd(M_i)$ is proposed, then the move to $nbd(M_k)$ is accepted with probability

$$P_{acc} \;=\; \min\left\{1, \frac{P(M_k|\text{data})}{P(M_i|\text{data})}\right\} \tag{4.6}$$

Since the transition matrix is finite and irreducible, then by applying the ergodic theorem for Markov chains, any function $g(M_i)$ defined on $\mathcal{M}$, $E(g(M))$ can be estimated by drawing from the Markov chain for $t = 1, 2, \ldots, N$, and

$$\hat{G} \;=\; \frac{1}{N}\sum_{i=1}^{N} g(M(t))$$

which is a simulation-consistent estimate of $E(g(M))$ (i.e. $\hat{G} \to E(g(M))$ almost surely) [64, 39, 49]. In other words, the posterior probability of model M_i given R_j is approximated

by the proportion of iterations that the Markov chain spends in $nbd(M_i)$ and as the number of iterations goes to infinity then the estimate goes to $P(M_i|Y)$ almost surely. The primary goal of using the MC^3 process in this particular application is not convergence to the stationary distribution, but rather to identify a subset of models that are most supported by the data. Let $\mathcal{M}^* \subset \mathcal{M}$ denote the models that are actually visited during the simulation. Any model $M_i \in \mathcal{M}$ and $M_i \notin \mathcal{M}^*$ has an estimated posterior probability of zero and is therefore eliminated from the sum in the denominator of equation 4.5.

The posterior probability of any $M_i \in \mathcal{M}^*$ can be estimated by the number of times the Markov chain was in state $nbd(M_j)$ divided by the total number of draws from the chain which is only appropriate when convergence it attained. Alternatively, the posterior probability for the models can also be estimated by replacing $\mathcal{M}$ with $\mathcal{M}^*$ in equation 4.5 since the BIC must be computed for each model visited during the simulation of the chain.

To further reduce the number of models in the denominator of equation 4.5 we use the principle of Occam's razor that holds that models which perform much less well than their competitors should be discarded [38]. The MC^3 algorithm eliminates most of the poor models by not visiting them, but there may be models in $\mathcal{M}^*$ that still are much less likely than the most probable model visited and are effectively discredited and should be eliminated. The reduced class of models is then defined by

$$\mathcal{M}^{**} \;=\; \left\{ M_k : M_i, M_k \in \mathcal{M}^*, \frac{\max_i P(M_i|\text{data})}{P(M_k|\text{data})} < C \right\}$$

Madigan and Raftery [38] adopted $C = 20$, but values from 10 to 1000 have been suggested with respect to the particular application. As a result, equation 4.5 can essentially be replaced by

$$P(M_i|\text{data}) \;\approx\; \frac{\exp\left(-0.5BIC_i\right)}{\sum_{M_k \in \mathcal{M}^{**}} \exp\left(-0.5BIC_k\right)} \tag{4.7}$$

4.6 Implementation

Suppose we have n measurements on each of the variables $Y_1, \cdots, Y_p$ on g known populations and wish to identify the most important variables that contribute to mean separation. In order to construct a set of the most likely models $\mathcal{M}^{**} \subset \mathcal{M}$ we use the following algorithm.

1. Randomly choose $M_i \in \mathcal{M}$ as a starting point. Let all $M_i \in \mathcal{M}$ be equally likely and $\mathcal{M}^* \stackrel{set}{=} \emptyset$

2. Record current neighborhood index $\mathcal{M}^* \stackrel{set}{=} \mathcal{M}^* \cup M_i$

3. Let $V = [X \quad Y]$ be the matrix such that the columns represent the variables present in M_i
4. Compute the $s = \min(p, q)$ non-zero eigenvalues of $S_{yy}^{-1/2} S_{yx} S_{xx}^{-1} S_{xy} S_{yy}^{-1/2}$ and represent them by $r_1^2 > \cdots > r_s^2$
5. Compute BIC_i
6. Randomly choose $M_k \in nbd(M_i)$ where each model in $nbd(M_i)$ is equally likely.
7. Compute BIC_k for model M_k.
8. Move to $nbd(M_k)$ with probability $P_{acc} = \min\left(1, \exp\left\{-0.5(BIC_i - BIC_k)\right\}\right)$ or stay in $nbd(M_i)$ with probability $1 - P_{acc}$
9. Let $u \sim U(0, 1)$. If $u < p_{jump}$ then choose some $M_i \in \mathcal{M}$ at random where all M_i are equally likely
10. Iterate steps 2–10 N times
11. Construct $\mathcal{M}^{**} = \{M_i : BIC_i \leq \min_{M_k \in \mathcal{M}^*} BIC_k + 2 \ln C\}$
12. Compute $P(M_i|Y) = \frac{\exp(-0.5 BIC_i)}{\sum_{M_k \in \mathcal{M}^{**}} \exp(-0.5 BIC_k)}$ for all $M_i \in \mathcal{M}^{**}$
13. Compute $E[\underline{\delta}] = \sum_{M_i \in \mathcal{M}^{**}} \underline{\delta}_i P(M_i|\text{data})$

4.6.1 Algorithm details

In step **1** a model is chosen at random from $\mathcal{M}$ where all models are equally likely. The initial starting point for the Markov chain is accomplished by generating $\delta_1, \cdots, \delta_{p+q} \sim Bern(0.5)$. If $\delta_i = 1$ then V_i is in the model alternatively, if $\delta_i = 0$ then V_i is excluded where

$$V \quad = \quad [X_1 \quad \cdots \quad X_p \quad Y_1 \quad \cdots \quad Y_q]$$

Step **2** records the iteration history (i.e. models visited) throughout the simulation in set $\mathcal{M}^*$.

The $p_i + q_i$ variables that are present in the current model make up the $n \times (p_i + q_i)$ dimensional matrix V in step **3**.

In step **4** the $s = \min(p_i, q_i)$ squared canonical correlations obtained from model M_i are computed.

In step **5** the BIC for model M_i is calculated by

$$\begin{aligned} BIC_i &= n\ln(1-\eta^2_{*_i}) + (p_i+q_i)\ln(n) \\ &= n\ln\left(\frac{\Lambda_i}{E[\Lambda_i]}\right) + (p_i+q_i)\ln(n) \end{aligned}$$

A model is chosen at random from $nbd(M_i)$ where all models are equally likely in step **6**. The model M_j is selected by generating $U \sim U(0,1)$ and defining $W = \lfloor pU \rfloor + 1$ then

$$\delta_W \stackrel{set}{=} 1 - \delta_W$$

which causes V_W to be added to the model if it was previously excluded or removed if $V_W \in M_i$.

The BIC for the proposed model, M_j, is computed in step **7** and the process moves to $nbd(M_j)$ with probability

$$\begin{aligned} P_{acc} &= \min\left\{1, \frac{P(M_j|\text{data})}{P(M_i|\text{data})}\right\} \\ &= \min\left\{1, \frac{\exp(-.5BIC_j)}{\exp(-.5BIC_i)}\right\} \\ &= \min\left\{1, \exp(-.5(BIC_j - BIC_i))\right\} \end{aligned} \tag{4.8}$$

or stays in the neighborhood of M_i with probability $1 - P_{acc}$ which is shown if step **8**.

We define the distance between any two models, M_i and M_j to be the number of variables unique to either model

$$d_{ij} = (\underline{\delta}_i - \underline{\delta}_j)'(\underline{\delta}_i - \underline{\delta}_j)$$

Throughout the MC^3 simulation, each step through the model space amounts to a jump of one unit of distance when a proposed move is accepted. As sampling from the chain continues, by the nature of the process there is an emphasis on spending more iterations in the neighborhoods of the best models. If there are groups of neighborhoods containing good models that are far apart from one another it may take many iterations to achieve convergence. One method of assessing convergence is the use of multiple sequences using overdispersed starting points [16]. By choosing random starting points for multiple sequences, the expected distance between any two start points is $0.5(p+q)$ since the distance between any two models chosen at random from $\mathcal{M}$ is a binomial random variable with parameters $(p+q)$ and 0.5. We propose starting new sequences at random with probability p_{jump} (we use $p_{jump} = 0.01$) with the starting point of the new sequence being some model chosen at random. Therefore when a new sequence is triggered, the initial model in the new chain is some model in $\mathcal{M}$ as shown in step **9** and the MC^3 starts anew. Recall that our goal in sampling from the

chain is model identification and not convergence. While the process cycles within a group of good models we randomly start the process over in a randomly determined spot in the model space in the hopes of finding other groups of likely models if they exist.

The purpose of step **10** is to insure that the best models are visited. Usually, in MCMC simulations, the number of iterations is chosen to achieve convergence to the proper stationary distribution and suggested values of N are on the order of 30000 [21]. In this particular application we are only interested in the neighborhoods that were actually visited throughout the simulation which make up the set $\mathcal{M}^*$ hence model identification is of greater importance than convergence so N maybe be as small as 5000 to attain the desired result. The justification for this is that the posterior probability of a model will not be estimated by the proportion of time the Markov chain spent in the neighborhood of the model, but instead will be approximated using the observed BIC for each model that is visited during the simulation. The assumption inherent in this approach is that all models that are most likely in $\mathcal{M}$ will be visited at least one time in 5000 iterations with the aid of the random restarts of the sequence from the previous step.

Occam's razor is performed in step **11** which states that models in $\mathcal{M}^*$ that are C or more times less likely than the most likely model in the set have been essentially discredited and should be eliminated. Madigan and Raftery [38] adopted the value of $C = 20$ to eliminate models that were far less likely than the best model. We then have

$$\begin{aligned}
\mathcal{M}^{**} &= \left\{ M_i : \frac{\max_{M_k \in \mathcal{M}^*} P(M_k|\text{data})}{P(M_i|\text{data})} \leq C \right\} \\
&= \left\{ M_i : \frac{\max_{M_k \in \mathcal{M}^*} \exp(-0.5 BIC_k)}{\exp(-0.5 BIC_i)} \leq C \right\} \\
&= \left\{ M_i : \max_{M_k \in \mathcal{M}^*} \exp(-0.5(BIC_k - BIC_i)) \leq C \right\} \\
&= \left\{ M_i : \max_{M_k \in \mathcal{M}^*} -0.5(BIC_k - BIC_i) \leq \ln C \right\} \\
&= \left\{ M_i : \min_{M_k \in \mathcal{M}^*} BIC_k - BIC_i \geq -2 \ln C \right\} \\
&= \left\{ M_i : BIC_i \leq 2 \ln C + \min_{M_k \in \mathcal{M}^*} BIC_k \right\}
\end{aligned}$$

This is the step in the algorithm where models that were identified during the simulation but deemed unlikely in comparison to the best model are removed.

In step **12** the potentially greatly reduced set $\mathcal{M}^{**} \subseteq \mathcal{M}^* \subseteq \mathcal{M}$ is then used to estimate the posterior probabilities of the most likely models. All models not in $\mathcal{M}^{**}$ have an estimated posterior probability of zero and are therefore eliminated from the denominator of equation 4.5.

Any variable that is in a given model has its corresponding position in the vector $\underline{\delta}$ set to one

or it is set to zero of the variable is not present. The probability that a variable is a significant contributor to any linear relationships can be assessed by estimating the probability that the variable in question should be present by

$$\hat{E}[\underline{\delta}] = \sum_{M_i \in \mathcal{M}^{**}} \underline{\delta}_i P(M_i|\text{data})$$

so if the estimated probability that any given variable should be in the model is greater than 0.5, it is more likely than not that the variable in question is a significant contributor to intra-set correlation. The elements of $\hat{E}[\underline{\delta}]$ are interpreted as the probability that the corresponding variable is active or contributes significantly to the model.

4.7 Interpretation

Canonical correlation analysis is a descriptive tool used to help understand the nature of the linear association between two groups of variables. The most commonly used interpretation method involves examining the correlation between the original variables and canonical variates formed. Construction of the canonical variates was shown in section 4.1 and were denoted as

$$\begin{aligned} U_i &= X\underline{a}_i \\ V_i &= Y\underline{b}_i \end{aligned}$$

for $i = 1, \cdots, \min(p, q)$. For the purposes of interpretation, we are concerned with the correlations between pairs (X_j, U_i) for $j = 1, \cdots, p$, and (Y_k, V_i) for $k = 1, \cdots, q$. The rational for this method of interpretation is that variables that are highly correlated with a particular canonical variate can be considered important with respect to the construction of that variate. Rencher [59] criticizes this method since he claims it does not take into account the joint contribution of each variable but Al-Kandari and Jolliffe [2] dispute Rencher's conclusions and claim that the method does have value. The confusion and difficulty in deciding what measures constitute an appropriate interpretation tool in the context of multivariate methods is not unique to CCA and is more the rule rather than the exception.

The contribution that variable X_i makes to the canonical variate U_j can be estimated by the squared correlation, $corr^2(X_i, U_j)$ (and similarly for Y_i and V_j pairs). This quantity represents the proportion of the variable explained by a particular variate since

$$\begin{aligned} \sum_{j=1}^{\min(p,q)} corr^2(X_i, U_j) &= 1 \quad \text{for } i = 1, \cdots, p \\ \sum_{j=1}^{\min(p,q)} corr^2(Y_i, V_j) &= 1 \quad \text{for } i = 1, \cdots, q \end{aligned}$$

Levine [37] and others suggest using these correlations to interpret the variates. Since the variates are not directly observable, they are considered latent abstract constructs and can be interpreted by investigating what measurable variables are related to them. Rencher [59, 60] says that this method of interpretation does not take into account the joint contribution of each variable and therefore does not recommend it.

The contribution of each canonical variate pair (U_i, V_i) to the overall linear relationships that exist between X and Y can be measured by

$$PCT_j = \frac{\lambda_i}{\sum_{j=1}^{\min(p,q)} \lambda_j} \quad \text{for } i = 1, \cdots, \min(p, q)$$

where $\lambda_i = r_i^2(1 - r_i^2)^{-1}$ and r_i is the i^{th} canonical correlation. This quantity represents the percent of the total linear relationship that is accounted for by each variate pair.

Using BMA methodology, the most promising variable configurations are identified and models that do not adequately capture the linear associations between the two groups of variables are eliminated by assigning them posterior probabilities of zero. Two ways that BMA results may aid in the interpretation of CCA results are through Bayesian variable assessment and evaluation of model uncertainty.

Individual variables can be assessed by estimation of the posterior probability that they are in randomly selected model in the model space. This is accomplished using the model indicator vector $\underline{\delta}$, and the probability of each variable being active that is estimated by

$$\hat{E}[\underline{\delta}] = \sum_{M_i \in \mathcal{M}^{**}} \underline{\delta}_i P(M_i|\text{data})$$

Any variable with an estimated probability of 0.5 of being included in a randomly selected model may be eliminated since it is not significantly contributing to the linear associations between X and Y. Conversely, any variable with a high posterior probability (greater than 0.5) of being in a randomly selected model should be retained since it is significantly contributing to the linear associations between X and Y. The cutoff of 0.5 was selected because a probability greater than 0.5 can be interpreted as the variable is more likely to be included than excluded. Variable assessment takes into account the other variables in the model since each model has a posterior probability and if a variable is important jointly with the other variables then it should be present in models with high posterior probability.

The variance due to model uncertainty of any quantity of interest, Δ is

$$\begin{aligned}
\text{Var}(\Delta) &= E[\Delta^2] - E[\Delta]^2 \\
&= \sum_{M \in \mathcal{M}} E[\Delta^2|M]P(M) - \left(\sum_{M \in \mathcal{M}} E[\Delta|M]P(M) \right)^2 \\
&= \sum_{M \in \mathcal{M}} (E[\Delta^2|M] - E[\Delta|M]^2 + E[\Delta|M]^2)P(M) - \left(\sum_{M \in \mathcal{M}} E[\Delta|M]P(M) \right)^2
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{M\in\mathcal{M}} (E[\Delta^2|M] - E[\Delta|M]^2)P(M) + \sum_{M\in\mathcal{M}} E[\Delta|M]^2 P(M) - \left(\sum_{M\in\mathcal{M}} E[\Delta|M]P(M)\right)^2 \\
&= \sum_{M\in\mathcal{M}} \mathrm{Var}(\Delta|M)P(M) + E_{\mathcal{M}}\left[E[\Delta|M]^2\right] - E_{\mathcal{M}}[E[\Delta|M]]^2 \\
&= E_{\mathcal{M}}[\mathrm{Var}(\Delta|M)] + \mathrm{Var}_{\mathcal{M}}(E[\Delta|M])
\end{aligned}
$$

These two pieces may be described as the "within" and "between" variance components. The "within" component is the sample variance pooled over the model space. It can be estimated for any given model using exact distribution theory, asymptotic distribution theory, or via some simulation technique such as bootstrapping depending on the parameter of interest. The overall estimate of this component is then obtained as the weighted average of the individual model sample variance estimates. The "between" variance component represents the variance associated model uncertainty. If the various models are in general agreement with each other this component will be small but if the estimated quantity takes on very different values, this component may be large. These ideas are illustrated on the environmental data from the state of Ohio.

4.8 Application

Biological data were gathered from 1988 to 1994 by the Ohio Environmental protection Agency (EPA) over the Eastern Corn Belt Plains ecoregion of Ohio [50]. The X matrix for this application consists of 20 variables identified as biological "stressor" variables. Nine of the variables are characterized as pertaining to water chemistry and the remaining eleven variables are classified as habitat variables. The Y matrix is made up of 20 variables known as Benthic macroinvertebrate variables. In the original analysis of this set of data, various transformations were applied to individual variables to attain approximate univariate normality of the data [50]. Sections C.2, C.3, and C.4 show the names, brief descriptions, and transformations used for the variables analyzed in this illustration.

4.8.1 Standard analysis

The first four canonical variates were significant at the 0.05 level of significance. The test for a significant canonical correlation for the fifth variate and beyond gave a p-value of 0.0536. Arguably, there are either four or five significant linear relationships that can be interpreted, and since adding the fifth pair brings the total percent explained from 64% to 71% it will be included.

For this data, $\min(p, q) = \min(20, 20) = 20$, hence there are 20 canonical variate pairs formed. We are interested in determining which variables are most related to the first five

Table 4.1: Variables most strongly related to the significant canonical variates

Variable	R^x					Variable	R^y				
AMM	.	.	+	+	−	PERCAD	.	+	.	+	.
BOD	.	+	−	.	+	PERCR	.	.	.	.	.
COD	−	.	.	.	+	PERD	.	.	.	.	.
FE	+	.	−	.	+	PERDIPT	.	.	.	.	.
NOX	.	.	.	.	.	PERE	+	.	.	.	.
PB	.	−	.	+	.	PCFILTE	+	+	.	.	.
PHO	.	.	.	.	−	PCGATH	−	.	.	−	.
TSS	+	.	.	.	+	PERGL	.	+	.	.	.
ZN	.	.	.	.	.	INSNODI	+	.	.	+	.
CHANNEL	.	.	+	.	.	PERTOLN	.	.	.	.	.
COVER	.	.	.	.	.	PERMAY	.	.	.	.	.
EMBSS	+	.	.	−	.	NUMQUAL	.	.	+	+	−
GRADIENT	.	−	.	.	.	NUMTAXA	.	.	.	+	.
POOL	.	.	+	.	.	PEROL	.	.	.	.	.
RIFEMSS	+	.	.	.	.	PEROTS	.	.	.	.	.
RIFFLE	.	.	.	.	.	PEROTHDI	−	−	.	.	.
RIPARIAN	.	.	.	.	.	QUALEPT	+	+	+	+	−
RIPSS	.	+	.	.	.	PCSHRED	.	.	.	.	.
SILTSS	.	.	+	−	.	PERTANY	.	.	.	.	.
SUBSTRAT	.	.	+	.	.	PERTTS	−	.	.	.	.

variate pairs. In the standard analysis, there is no uniformly used method to determine which correlations are large and which are small. The method used here identifies all correlations larger in magnitude than the largest correlation in the non-signignifant canonical variates. To help summarize results, let

$$R^x_{ij} = \begin{cases} + & \text{if } corr^2(X_i, U_j) > \max_{k=6,\cdots,20} corr^2(X_i, U_k) \text{ and } corr(X_i, U_j) > 0 \\ - & \text{if } corr^2(X_i, U_j) > \max_{k=6,\cdots,20} corr^2(X_i, U_k) \text{ and } corr(X_i, U_j) < 0 \\ \cdot & \text{if } corr^2(X_i, U_j) < \max_{k=6,\cdots,20} corr^2(X_i, U_k) \end{cases}$$

$$R^y_{ij} = \begin{cases} + & \text{if } corr^2(Y_i, V_j) > \max_{k=6,\cdots,20} corr^2(Y_i, V_k) \text{ and } corr(Y_i, V_j) > 0 \\ - & \text{if } corr^2(Y_i, V_j) > \max_{k=6,\cdots,20} corr^2(Y_i, V_k) \text{ and } corr(Y_i, V_j) < 0 \\ \cdot & \text{if } corr^2(Y_i, V_j) < \max_{k=6,\cdots,20} corr^2(Y_i, V_k) \end{cases}$$

for $i = 1, \cdots, 20$, $j = 1, \cdots, 5$. This identifies any correlation in the first five variates that is larger in magnitude than the largest magnitude correlation of the remaining variates for each variable.

Interpretations can be made from the summarization shown in table 4.1. For example, the

first canonical variate can be characterized as the linear relationship between negative chemical oxygen demand (-COD), iron (FE), total suspended solids (TSS), the embeddedness subscore (EMBSS), and riffle embeddedness subscore (RIFEMSS) with percent that are erosional taxa (PERE), percent that are filter feeding insect taxa (PCFILTE), the negative of the percent that are gatering insect taxa (-PCGATH), number of insect taxa excluding dipterans (INSNODI), the negative of the percent that are dipterans and non-insects (-PEROTHDI), number of Ephemeroptera, Plecoptera, and Tricoptera taxa in qualitative dipnet sample (QUALEPT) and the negative of the percent that are toxic tolerant (-PERTTS). It is the hope that the relationship has meaning to the researcher in this area so that a more succinct interpretation can be reported.

4.8.2 BMA analysis

The analysis will be approached in two different ways using BMA methodology. The goal of the first approach is variable assessment. In the variable assessment phase of the analysis, the posterior probability that a given variable is active in a model is estimated and these values are used to aid in the interpretation of the intra-set correlation. The second approach evaluates the magnitude of the model uncertainty variance component associated with various qantities of interest.

Variable assessment

The contribution of each variable to the linear associations between the variables in X with those in Y is assessed by estimation of the probability that any given variable is present in a randomly chosen model. For each model $M_i \in \mathcal{M}^{**}$ we have the estimated posterior probability of the model, $P(M_i|\text{data})$, and the vector of indicator variables, $\underline{\delta}_i$, that represents which variables are present (1) and which are absent (0). The estimated posterior probability that a variable, Y_i for example, is active is given by

$$P(Y_i \text{ is active}) \;=\; P(\delta_{\cdot i} = 1) = \sum_{M_j \in \mathcal{M}^{**}} \delta_{ji} P(M_j|\text{data})$$

When $P(Y_i \text{ is active}) > 0.5$ it is more likely than not that variable Y_i contributes substantially to the linear relationship. We conclude that a variable is important if its activation probability is at least 0.5.

The "*size*" prior was assumed on the model space which has the form

$$P(M_i) \;=\; \theta^k (1-\theta)^{p+q-k}$$

where model M_i has k variables. Four values of θ were investigated: 0.2, 0.3, 0.4, and 0.5 (recall that $\theta = 0.5$ is the *uniform* prior). For each prior, 20 MC^3 simulations were run with

Table 4.2: Estimated activation probabilities and standard errors with $\theta = 0.2$

Variable	Prob	SE	Variable	Prob	SE
AMM	1.000	(0.000)	PERCAD	1.000	(0.000)
BOD	0.631	(0.074)	PERCR	0.448	(0.076)
COD	1.000	(0.000)	PERD	0.999	(0.001)
FE	0.878	(0.054)	PERDIPT	0.008	(0.004)
NOX	0.686	(0.050)	PERE	0.633	(0.050)
PB	0.385	(0.077)	PCFILTE	0.501	(0.076)
PHO	0.000	(0.000)	PCGATH	0.408	(0.077)
TSS	0.174	(0.057)	PERGL	1.000	(0.000)
ZN	0.578	(0.075)	INSNODI	0.259	(0.056)
CHANNEL	0.019	(0.015)	PERTOLN	0.290	(0.065)
COVER	0.762	(0.043)	PERMAY	0.647	(0.077)
EMBSS	0.327	(0.073)	NUMQUAL	0.958	(0.017)
GRADIENT	1.000	(0.000)	NUMTAXA	0.405	(0.051)
POOL	0.426	(0.076)	PEROL	0.028	(0.018)
RIFEMSS	0.713	(0.062)	PEROTS	0.969	(0.018)
RIFFLE	0.890	(0.043)	PEROTHDI	1.000	(0.000)
RIPARIAN	0.000	(0.000)	QUALEPT	1.000	(0.000)
RIPSS	0.965	(0.026)	PCSHRED	0.000	(0.000)
SILTSS	0.140	(0.048)	PERTANY	0.631	(0.075)
SUBSTRAT	0.081	(0.023)	PERTTS	0.002	(0.002)

5000 iterations in each. The estimated activation probabilites along with standard errors are shown in tables 4.2–4.5. The determination of importance for each variable over the range of priors investigated was generally the same. Differences in interpretation came about when $\theta = 0.2$. Embeddedness subscore (EMBSS), pool QHEI metrics (POOL), percent that are *Crictopus* (PERCR), percent that are gatering insect taxa (PCGATH), and total number of quantitative taxa (NUMTAX) would be interpreted as not important whereas for larger values of θ they would be retained.

The choice of θ does not appear to be very important for this particular data since the conclusion remains the same over a relatively wide range. The change in the activation probability over the range of $0.3 \leq \theta \leq 0.5$ ranged from 0.328 for PERTOLN to 0.000 for AMM, COD, FE, PHO, GRADIENT, PERCAD, PERD, PERGL, PEROTS, PEROTHDI, QUALEPT, and PCSHRED, and the median change on the activation probability for the 40 variables was 0.027.

Figures 4.1 and 4.2 show the scaled correlation values for the first two canonical variates

Table 4.3: Estimated activation probabilities and standard errors with $\theta = 0.3$

Variable	Prob	SE	Variable	Prob	SE
AMM	1.000	(0.000)	PERCAD	1.000	(0.000)
BOD	0.987	(0.006)	PERCR	0.972	(0.013)
COD	1.000	(0.000)	PERD	1.000	(0.000)
FE	1.000	(0.000)	PERDIPT	0.013	(0.010)
NOX	0.789	(0.053)	PERE	0.756	(0.044)
PB	0.032	(0.026)	PCFILTE	0.975	(0.020)
PHO	0.000	(0.000)	PCGATH	0.942	(0.037)
TSS	0.020	(0.010)	PERGL	1.000	(0.000)
ZN	0.993	(0.003)	INSNODI	0.578	(0.068)
CHANNEL	0.092	(0.049)	PERTOLN	0.130	(0.045)
COVER	0.889	(0.045)	PERMAY	0.996	(0.002)
EMBSS	0.865	(0.046)	NUMQUAL	0.947	(0.018)
GRADIENT	1.000	(0.000)	NUMTAXA	0.499	(0.069)
POOL	0.952	(0.017)	PEROL	0.012	(0.005)
RIFEMSS	0.739	(0.045)	PEROTS	1.000	(0.000)
RIFFLE	0.942	(0.019)	PEROTHDI	1.000	(0.000)
RIPARIAN	0.000	(0.000)	QUALEPT	1.000	(0.000)
RIPSS	0.908	(0.051)	PCSHRED	0.000	(0.000)
SILTSS	0.102	(0.032)	PERTANY	0.983	(0.009)
SUBSTRAT	0.037	(0.016)	PERTTS	0.000	(0.000)

Table 4.4: Estimated activation probabilities and standard errors with $\theta = 0.4$

Variable	Prob	SE	Variable	Prob	SE
AMM	1.000	(0.000)	PERCAD	1.000	(0.000)
BOD	1.000	(0.000)	PERCR	1.000	(0.000)
COD	1.000	(0.000)	PERD	1.000	(0.000)
FE	1.000	(0.000)	PERDIPT	0.021	(0.015)
NOX	0.963	(0.011)	PERE	0.897	(0.020)
PB	0.003	(0.003)	PCFILTE	1.000	(0.000)
PHO	0.000	(0.000)	PCGATH	0.997	(0.003)
TSS	0.058	(0.036)	PERGL	1.000	(0.000)
ZN	1.000	(0.000)	INSNODI	0.546	(0.059)
CHANNEL	0.017	(0.008)	PERTOLN	0.310	(0.045)
COVER	0.982	(0.007)	PERMAY	1.000	(0.000)
EMBSS	0.892	(0.039)	NUMQUAL	0.986	(0.006)
GRADIENT	1.000	(0.000)	NUMTAXA	0.557	(0.054)
POOL	0.997	(0.003)	PEROL	0.097	(0.053)
RIFEMSS	0.899	(0.030)	PEROTS	1.000	(0.000)
RIFFLE	0.994	(0.004)	PEROTHDI	1.000	(0.000)
RIPARIAN	0.000	(0.000)	QUALEPT	1.000	(0.000)
RIPSS	0.985	(0.008)	PCSHRED	0.000	(0.000)
SILTSS	0.146	(0.045)	PERTANY	0.996	(0.003)
SUBSTRAT	0.093	(0.017)	PERTTS	0.001	(0.001)

Table 4.5: Estimated activation probabilities and standard errors with $\theta = 0.5$

Variable	Prob	SE	Variable	Prob	SE
AMM	1.000	(0.000)	PERCAD	1.000	(0.000)
BOD	1.000	(0.000)	PERCR	1.000	(0.000)
COD	1.000	(0.000)	PERD	1.000	(0.000)
FE	1.000	(0.000)	PERDIPT	0.036	(0.013)
NOX	0.988	(0.004)	PERE	0.960	(0.012)
PB	0.072	(0.022)	PCFILTE	1.000	(0.000)
PHO	0.000	(0.000)	PCGATH	1.000	(0.000)
TSS	0.128	(0.026)	PERGL	1.000	(0.000)
ZN	1.000	(0.000)	INSNODI	0.525	(0.045)
CHANNEL	0.011	(0.004)	PERTOLN	0.458	(0.023)
COVER	0.999	(0.001)	PERMAY	1.000	(0.000)
EMBSS	0.838	(0.038)	NUMQUAL	0.983	(0.007)
GRADIENT	1.000	(0.000)	NUMTAXA	0.700	(0.041)
POOL	0.994	(0.003)	PEROL	0.116	(0.026)
RIFEMSS	0.932	(0.013)	PEROTS	1.000	(0.000)
RIFFLE	0.997	(0.002)	PEROTHDI	1.000	(0.000)
RIPARIAN	0.015	(0.005)	QUALEPT	1.000	(0.000)
RIPSS	0.990	(0.003)	PCSHRED	0.000	(0.000)
SILTSS	0.267	(0.054)	PERTANY	0.999	(0.001)
SUBSTRAT	0.090	(0.020)	PERTTS	0.000	(0.000)

Figure 4.1: Scaled Correlation plot for Habitat and Chemical Variables for first two variates

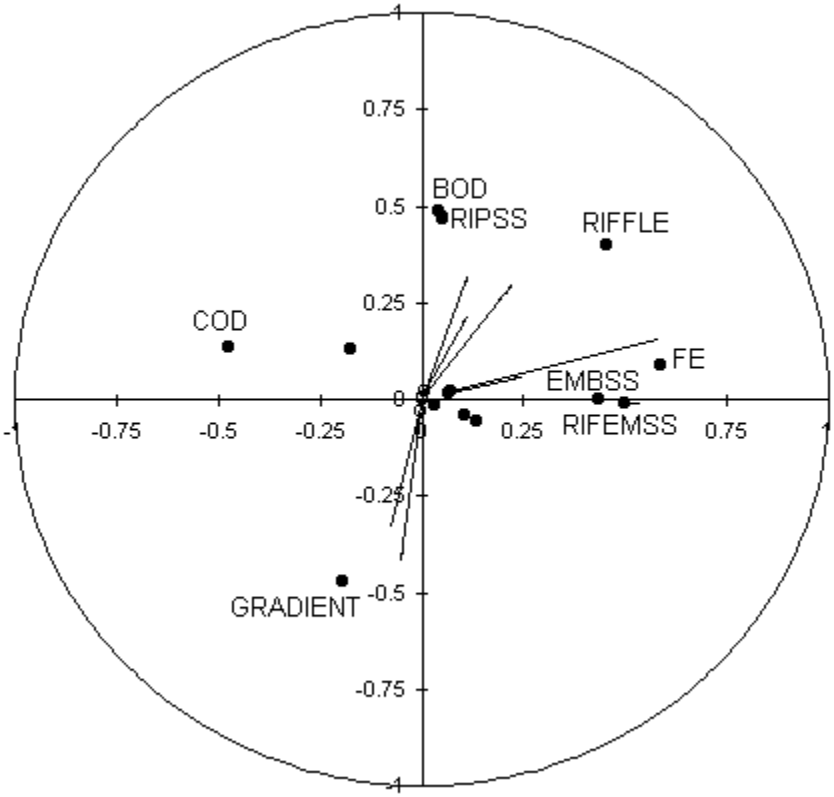
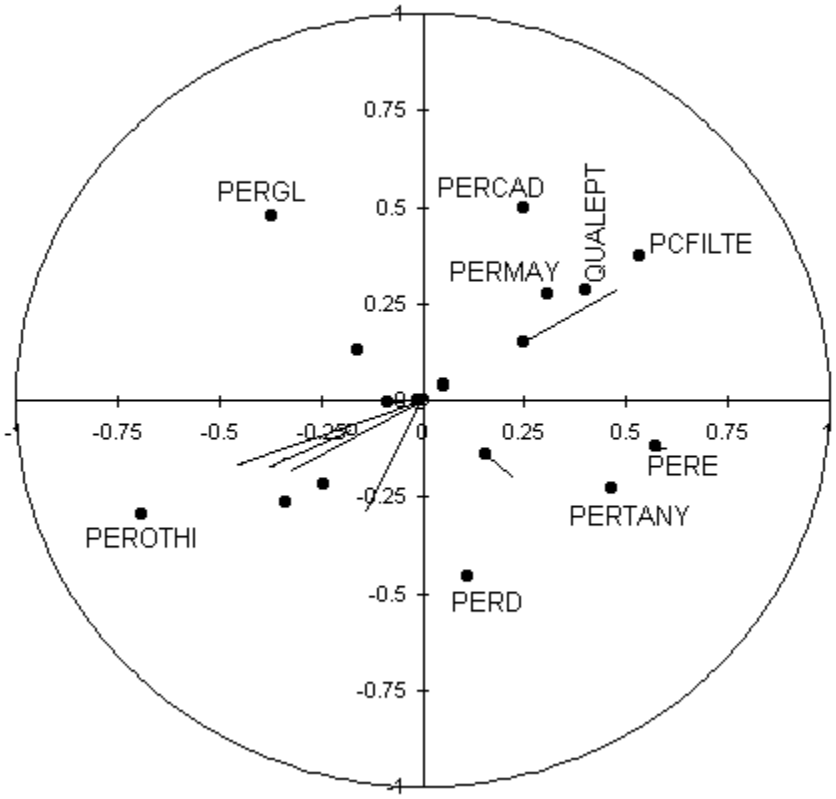


Figure 4.2: Scaled Correlation plot for Benthic Macroinvertebrate Variables for first two variates



when $\theta = 0.5$. Each variable is represented by a dot and a ray. The dot represents the scaled correlation value along each of the first two canonical variate axes. We define the scaled correlation to be the the full model correlation between a variable and the canonical variate multiplied by the probability that the variable is active, symbolically as

$$\begin{aligned} P(Y_i \text{ is active})Corr(Y_i, V_j) \qquad & \text{and} \\ P(X_i \text{ is active})Corr(X_i, U_j) \qquad & \text{and} \end{aligned}$$

The ray extends from the scaled correlation and terminates at the unscaled correlation. This graph emphasizes three major points of interest. Firstly, it shows how the individual variables are aligned along the first two canonical axis. For example, we could classify each variable as either primarily aligned with the first canonical variate (FE, EMBSS, RIFEMSS, COD, PERE, and PEROTHI), mainly aligned with the second canonical variate (BOD, RIPSS, GRADIENT, PERCAD, and PERD), or a mix of the two variates (RIFFLE, PERMAY, QUALEPT, PCFILTE, and PERGL). Next, the plot identifies variables that provide poor linear association as the scaled correlation coordinate will be closer to the origin when compared to a variable that provides stronger linear associations. This means that variables with long rays are being excluded since they have a low posterior probability of being included. Finally, the coordinate value where the ray terminates indicates how much linear association for a given variable is accounted for in the first two canonical variates. In this example, most activation probabilites are close to either one or zero so the rays for the active variables have no length or are very short. Since RIFFLE is closest of all habitat and chemical variables to the outer circle, then it contributes more to the first two variates than any of the other variables.

Model uncertainty

Recall that the BMA estimated value of any quantity of interest, Δ, is the plug-in estimate obtained from the estimated expected value

$$\begin{aligned} \hat{\Delta} &= \hat{E}(\Delta) \\ &= \sum_{M_i \in \mathcal{M}^{**}} \hat{E}(\Delta|M_i)P(M_i|Y) \end{aligned}$$

and that the variability is estimated by

$$\begin{aligned} \widehat{Var}(\Delta) &= \widehat{E_M}(\widehat{Var}(\Delta|M)) + \widehat{Var}_M(\hat{E}(\Delta|M)) \\ &= \text{Sampling Variation} + \text{Model Uncertainty Variation} \end{aligned}$$

These estimated variance components of the correlation coefficients associated with the first variate pair are shown in table 4.6. The sampling variability was estimated by obtaining the correlation coefficients from 10000 bootstrap samples of the observations using the full

model. The model uncertainty variance component was computed using the modified MC^3 algorithm and Occam's razor to obtain a subset of models, $\mathcal{M}^{**}$. The "Order" column listed in the table shows the order of maganitude difference between variation due to sampling and that due to model uncertainty by

$$\text{Order} \quad = \quad \log_{10} \frac{\text{Sampling Variance}}{\text{Model Uncertainty Variance}}$$

For the variables percent "shredding" insect taxa (PCSHRED), percent toxic tolerant (PERTTS), and total phosphorus (PHO), the estimated variance due to uncertainty is zero so no order can be esitmated. Of the remaining 37 variables, only number of insect taxa excluding dipterans (INSNODI) had more uncertainty variation than sampling variation. For the other variables, sampling variability was larger with the largest order estimate being 3.11 for percent dipterans and non-insects (PEROTHDI).

Table 4.6: Estimated variablity due to model uncertainty and sampling variation of the correlation coeffient of the first canonical variate

Variable Name	Uncert. StdDev	Samp. StdDev	Order	Variable Name	Uncert. StdDev	Samp. StdDev	Order
PERCAD	0.0268	0.2430	1.92	AMM	0.0134	0.2276	2.46
PERCR	0.0235	0.1959	1.84	BOD	0.0325	0.3057	1.95
PERD	0.0431	0.2770	1.62	COD	0.0187	0.2887	2.38
PERDIPT	0.0205	0.1849	1.91	FE	0.0119	0.3524	2.94
PERE	0.1373	0.3082	0.70	NOX	0.0170	0.1535	1.91
PCFILTE	0.0181	0.2455	2.26	PB	0.0383	0.2627	1.67
PCGATH	0.0153	0.1880	2.18	PHO	0.0000	0.2293	∞
PERGL	0.0433	0.3456	1.80	TSS	0.2288	0.3163	0.28
INSNODI	0.2680	0.2304	-0.13	ZN	0.0150	0.1856	2.19
PERTOLN	0.1988	0.2127	0.06	CHANNEL	0.0508	0.2060	1.22
PERMAY	0.0142	0.1760	2.19	COVER	0.0158	0.1904	2.16
NUMQUAL	0.0120	0.1977	2.43	EMBSS	0.1828	0.2614	0.31
NUMTAXA	0.0562	0.2328	1.23	GRADIENT	0.0334	0.2822	1.85
PEROL	0.0091	0.1420	2.38	POOL	0.0339	0.2090	1.58
PEROTS	0.0123	0.1769	2.32	RIFEMSS	0.1449	0.2749	0.56
PEROTHDI	0.0085	0.3042	3.11	RIFFLE	0.0645	0.2490	1.17
QUALEPT	0.0152	0.2269	2.35	RIPARIAN	0.0247	0.2049	1.84
PCSHRED	0.0000	0.1977	∞	RIPSS	0.0488	0.2638	1.47
PERTANY	0.0251	0.2997	2.15	SILTSS	0.1223	0.2062	0.45
PERTTS	0.0000	0.2290	∞	SUBSTRAT	0.0756	0.2098	0.89

4.9 Conclusion

The classical approach to model building selects a single model which is equivalent to letting the posterior probability of that model be equal to one while all other models in the space are excluded since they have probability zero. A direct result of selecting a single model is that the variance due to uncertainty is zero which does not accurately reflect reality since the selected model was not chosen as such a priori. Bayesian model averaging provides a way to estimate and incorporate the previously ignored model uncertainty variance component or used for variable assessment.

When a model is developed empirically via some variable selection scheme, any variances that are estimated are overly optimistic since they are estimated based on the assumption that the model selected is correct. The model uncertainty variance component adjusts any variance estimate to reflect the fact that whatever the true model is, it is unknown and there may be several competitive models that adequately reflect what is happening with the data. The information from each good model is weighted based on its posterior probability, and estimates of desired quantities are formed with more appropriate variance estimates.

The joint contribution of each variable can be measured using BMA in the context of variable assessment. The posterior model space is summarized by computing the expected posterior probability that any given variable is active. The activation probabilities can then be used in conjuction with the standard tools that are used such as the structural loadings and standardized coefficients. The addition of the variable assessment information enhances and clarifies the interpretation thus adding value to the analysis.

Chapter 5

Future Research

The primary goal of this research was to use BMA methodology to make interpretation of the output from the multivariate methods of principal components, canonical variate and canonical correlation analysis easier. By assigning posterior probabilities to the various models based on how well each is supported by the data, patterns may be found that would that would not be detected by examining a single model obtained by standard methods.

The BMA methodology allows for estimation of the variance due to the uncertainty of which model is correct. Standard model building practices have focused on the selection of a single model and ignores model uncertainty by then assuming the correct model has been found. In multivariate methods such as canonical correlation and principal components analysis, for example, no formal model selection is generally performed, but user interpretation of the output decides which variables are important based on the practitioners personal experience.

BMA can also be used as a variable assessment tool. The probability that a given variable is present in a randomly selected model is estimated by the sum of the posterior probabilities of the models where it is present. A high probability of a particular variable being present is interpreted to mean that the variable in question is important.

Kass and Wasserman [32] show that BIC is an especially accurate approximation to a Bayes factor where the prior on the unknown parameters is elliptically symmetric with density

$$\pi_{\psi}(\psi) \quad = \quad |\Sigma_{\psi}|^{-1/2} f\left((\psi - \psi_0)' \Sigma_{\psi}^{-1} (\psi - \psi_0)\right)$$

where $|\Sigma_{\psi}|^{-1}$ is a block diagonal Fisher information matrix. They point out that while these results do not strictly apply to linear models where the sampling is not *iid*, they are "confident that a rigorous extension is possible in such situations". In this current research, each multivariate method is a special case of multivariate regression so the accuracy and limitations of the BIC is of interest as is still an open area of research.

The BIC is the basis for approximating the posterior model probabilities and it is composed

of a likelihood term and a penalty term. Within the likelihood term, some measure of association must be specified. For univariate regression the choice is the usual r-square, but a single measure of association does not exist in the multivariate context. It was shown in section A.3 that the measure of association based on Wilks' lambda is uniformly larger than those proposed based on Lawley-Hotelling, Pillai, and Roy. Even though the measure based on Wilks' lambda has desirable properties and is heuristically appealing since it is the likelihood ratio, it cannot be used capriciously. As $\min(p, q)$ grows for a fixed n, the asymptotic expected value of Wilks' lambda decreases under the hypothesis of no structure or relationships. To correct for this artifical inflation of the measure of association, a measure based on Wilks' lambda adjusted for sample size and number of variables was introduced. Rencher [60, 61] points out that the various multivariate measures of association do not appear to be measuring the the same level of association. This statement is based on the fact that in practice the commonly used measures usually cover a wide range of values over the interval $[0, 1]$. In this research, some progress was made in understanding how the commonly used measures relate to one another, but more is required to fully investigate the properties and limitations of these statistics.

In the current BMA literature, a uniform prior on the model space has been assumed throughout. This research has defined four general forms, and all priors on the model space may be classified as either *uniform*, *size*, *variable*, or *individual*. The form chosen will most likely be made based on the level of prior knowledge the pratitioner is willing to assume. Future research in this area may take the *size* class of priors and assign a prior distribution on the activation parameter, θ, creating a hierarchical model. The effect of the hierarchical structure would be to enable a more flexible prior on the model space without necessarily having a precise value of θ in mind. This idea can then be easily extended to priors on the individual θ_i for the *variable* class of priors.

Convergence to the posterior distribution is the goal of MCMC methods. In this research the goal was changed to model identification rather than convergence. A modification of the procedure allowed for random hops to other parts of the model space in order to more easily identify all good models in a fewer number of iterations. Future research in this area may compare the convergence rate of the standard MCMC to that of the modified MCMC in order to quantify the improvement.

Appendix A

Proofs and Derivations

A.1 Information criteria and prior specification: Duality Theorem

Posterior model probabilities obtained from a marginal likelihood approximated using a generalized information criteria (GIC) assuming a uniform prior on the model space is equivalent to posterior model probabilities obtained using the Bayes information criteria (BIC) where the prior assumed for the model space is defined by

$$P(M_i) = \frac{\exp(-0.5p_i(a - \ln n))}{\sum_{j=0}^{p} \frac{p!}{j!(p-j)!} \exp(-0.5j(a - \ln n))}$$

where a is the penalty term multiplier in the GIC, p_i is the number of variables in model M_i, and n denotes the sample size.

Proof
From Bayes rule we have that posterior probability of M_i is

$$P(M_i|\text{data}) = \frac{P(\text{data}|M_i)P(M_i)}{\sum_{M_j \in \mathcal{M}} P(\text{data}|M_j)P(M_j)} \tag{A.1}$$

Raftery [55] shows that

$$P(\text{data}|M_i) \approx \frac{\exp(-0.5BIC_i)}{\sum_{M_j \in \mathcal{M}} \exp(-0.5BIC_j)} \tag{A.2}$$

and we assume the prior distribution on the model space to be

$$P(M_i) = \frac{\exp(-0.5p_i(a - \ln n))}{\sum_{j=0}^{p} \frac{p!}{j!(p-j)!} \exp(-0.5j(a - \ln n))} \tag{A.3}$$

Substituting equations A.2 and A.3 into A.1 we get

$$\begin{aligned} P(M_i|\text{data}) &\approx \frac{\frac{\exp(-0.5BIC_i)}{\sum_{M_j\in\mathcal{M}}\exp(-0.5BIC_j)}\frac{\exp(-0.5p_i(a-\ln n))}{\sum_{k=0}^{p}\frac{p!}{k!(p-k)!}\exp(-0.5k(a-\ln n))}}{\sum_{M_j\in\mathcal{M}}\left(\frac{\exp(-0.5BIC_j)}{\sum_{M_k\in\mathcal{M}}\exp(-0.5BIC_k)}\frac{\exp(-0.5p_i(a-\ln n))}{\sum_{k=0}^{p}\frac{p!}{k!(p-k)!}\exp(-0.5k(a-\ln n))}\right)} \\ &= \frac{\exp(-0.5BIC_i)\exp(-0.5p_i(a-\ln n))}{\sum_{M_j\in\mathcal{M}}\exp(-0.5BIC_j)\exp(-0.5p_j(a-\ln n))} \\ &= \frac{\exp(-0.5BIC_i-0.5p_i(a-\ln n))}{\sum_{M_j\in\mathcal{M}}\exp(-0.5BIC_j-0.5p_j(a-\ln n))} \end{aligned} \tag{A.4}$$

Since $BIC_i = n\ln(1-r_i^2)+p_i\ln n$, then by substitution into A.4 we get

$$\begin{aligned} P(M_i|\text{data}) &\approx \frac{\exp(-0.5(n\ln(1-r_i^2)+p_i\ln n)-0.5p_i(a-\ln n))}{\sum_{M_j\in\mathcal{M}}\exp(-0.5(n\ln(1-r_j^2)+p_j\ln n)-0.5p_j(a-\ln n))} \\ &= \frac{\exp(-0.5(n\ln(1-r_i^2)+p_i\ln n+ap_i-p_i\ln n))}{\sum_{M_j\in\mathcal{M}}\exp(-0.5(n\ln(1-r_i^2)+p_j\ln n+ap_j-p_j\ln n))} \\ &= \frac{\exp(-0.5(n\ln(1-r_i^2)+ap_i))}{\sum_{M_j\in\mathcal{M}}\exp(-0.5(n\ln(1-r_i^2)+ap_j))} \end{aligned} \tag{A.5}$$

Nishii [48] defines the GIC to be

$$GIC_i = n\ln(1-r_i^2)+ap_i$$

where $\lim_{n\to\infty} n^{-1}a = 0$, and substitution into A.5 gets

$$\begin{aligned} P(M_i|\text{data}) &\approx \frac{\exp(-0.5GIC_i)}{\sum_{M_j\in\mathcal{M}}\exp(-0.5GIC_j)} \\ &= \frac{P(\text{data}|M_i)}{\sum_{M_j\in\mathcal{M}}P(\text{data}|M_j)} \\ &= \frac{P(\text{data}|M_i)P(M)}{\sum_{M_j\in\mathcal{M}}P(\text{data}|M_j)P(M)} \end{aligned}$$

Hence $P(M_i) = P(M) = 2^{-p} \quad \forall M_i \in \mathcal{M}$.

A.2 *A priori* probability of model inclusion for individual variables and choice of information criterion

If the prior on model space $\mathcal{M}$ is defined so that variable X_j is in a randomly selected model with probability θ for $j = 1, \cdots, p$ then the equivalent posterior model space is obtained by approximating the marginal likelihood using a generalized information criterion (GIC) with penalty multiplier a and uniform prior on the model space if

$$a = \ln n - 2\ln\left(\frac{\theta}{1-\theta}\right)$$

Proof: Suppose $P(M_i|\text{data}) \propto \exp(-0.5BIC_i)$ and let π_1 and π_2 denote two priors on the model space $\mathcal{M}$ such that

$$\begin{aligned}\pi_1(M_i) &= \theta^k(1-\theta)^{p-k} \\ \pi_2(M_i) &= \frac{\exp(-0.5k(a-\ln n))}{\sum_{j=0}^{p}\frac{p!}{j!(p-j)!}\exp(-0.5j(a-\ln n))}\end{aligned}$$

where k represents the number of variables in model M_i.

Let the vector $\underline{\delta}_i$ be defined as

$$\delta_{ij} = \begin{cases} 0 & \text{if } X_j \in M_i \\ 1 & \text{if } X_j \notin M_i \end{cases}$$

hence $\underline{\delta}_i'\underline{\delta}_i$ is the number of variables in model M_i. There are C_k^p models with k variables so

$$\begin{aligned}\frac{P_1(\underline{\delta}'\underline{\delta} = k)}{P_1(\underline{\delta}'\underline{\delta} = k-1)} &= k\frac{\theta}{1-\theta} \\ \frac{P_2(\underline{\delta}'\underline{\delta} = k)}{P_2(\underline{\delta}'\underline{\delta} = k-1)} &= k\exp(-0.5(a-\ln n))\end{aligned}$$

and if

$$\frac{P_1(\underline{\delta}'\underline{\delta} = k)}{P_1(\underline{\delta}'\underline{\delta} = k-1)} \overset{set}{=} \frac{P_2(\underline{\delta}'\underline{\delta} = k)}{P_2(\underline{\delta}'\underline{\delta} = k-1)}$$

for $k = 1, \cdots, p$, then

$$a = \ln n - 2\ln\left(\frac{\theta}{1-\theta}\right)$$

or equivalently

$$\theta = 1 - \frac{1}{1 + \exp(-0.5(a - \ln n))}$$

Choose any $M_i \in \mathcal{M}$ and

$$\begin{aligned}
\pi_2(M_i) &= \frac{\exp(-0.5k(a - \ln n))}{\sum_{j=0}^{p} \frac{p!}{j!(p-j)!} \exp(-0.5j(a - \ln n))} \\
&= \frac{\exp\left(-0.5k\left(\ln n - 2\ln\left(\frac{\theta}{1-\theta}\right) - \ln n\right)\right)}{\sum_{j=0}^{p} \frac{p!}{j!(p-j)!} \exp\left(-0.5j\left(\ln n - 2\ln\left(\frac{\theta}{1-\theta}\right) - \ln n\right)\right)} \\
&= \frac{\exp\left(k\ln\left(\frac{\theta}{1-\theta}\right)\right)}{\sum_{j=0}^{p} \frac{p!}{j!(p-j)!} \exp\left(j\ln\left(\frac{\theta}{1-\theta}\right)\right)} \\
&= \frac{\theta^k(1-\theta)^{-k}}{\sum_{j=0}^{p} \frac{p!}{j!(p-j)!} \left(\frac{\theta}{1-\theta}\right)^j} \\
&= \frac{\theta^k(1-\theta)^{-k}}{\left(1 + \frac{\theta}{1-\theta}\right)^p} \\
&= \frac{\theta^k(1-\theta)^{-k}}{(1-\theta)^{-p}} \\
&= \theta^k(1-\theta)^{p-k} \\
&= \pi_1(M_i)
\end{aligned}$$

From the result proven in section A.1, if a GIC is used to approximate the marginal likelihood then using

$$\begin{aligned}
P(M_i|\text{data}) &\propto \exp(-0.5GIC_i) \\
&= \exp\left(-0.5\left(n\ln(1-r_i^2) + p_i\left(\ln n - 2\ln\left(\frac{\theta}{1-\theta}\right)\right)\right)\right)
\end{aligned}$$

with a uniform prior on $\mathcal{M}$ is equivalent to using the BIC to approximate the marginal likelihood and using a prior of the form

$$P(M_i) = \theta^{p_i}(1-\theta)^{p-p_i}$$

for all $M_i \in \mathcal{M}$.

A.3 Ordering of multivariate measures of association

The following multivariate measures of association

Wilk's lambda	$\eta^2_\Lambda = 1 - \Lambda = 1 - \prod_{i=1}^k (1 - r_i^2)$
Roy's	$\eta^2_\theta = \frac{\theta}{1+\theta} = r_1^2$
Pillai's	$\eta^2_V = k^{-1}V = k^{-1}\sum_{i=1}^k r_i^2$
Lawley-Hotelling	$\eta^2_U = \frac{k^{-1}U}{1+k^{-1}U} = \frac{k^{-1}\sum_{i=1}^k r_i^2(1-r_i^2)^{-1}}{1+k^{-1}\sum_{i=1}^k r_i^2(1-r_i^2)^{-1}}$

are ordered values so that $\eta^2_\Lambda \geq \eta^2_\theta \geq \eta^2_U \geq \eta^2_V$.

Proof: Let the squared canonical correlations be denoted by $r_1^2 \geq \cdots \geq r_k^2$.
Case 1: Wilks' versus Roy's

$$\begin{aligned}
& 0 \leq r_i^2 \leq 1 \quad \text{for } i = 1, \cdots, k \\
\Rightarrow \quad & 0 \leq 1 - r_i^2 \leq 1 \\
\Rightarrow \quad & 0 \leq \prod_{i=2}^k (1 - r_i^2) \leq 1 \\
\Rightarrow \quad & 0 \leq \prod_{i=1}^k (1 - r_i^2) \leq 1 - r_1^2 \\
\Rightarrow \quad & r_1^2 \leq 1 - \prod_{i=1}^k (1 - r_i^2) \\
\Rightarrow \quad & \eta^2_\theta \leq \eta^2_\Lambda
\end{aligned}$$

Case 2: Roy's versus Lawley-Hotelling's

$$\begin{aligned}
& r_1^2 \geq r_i^2 \quad \text{for } i = 2, \cdots, k \\
\Rightarrow \quad & \frac{r_1^2}{1 - r_1^2} \geq \frac{r_i^2}{1 - r_i^2} \\
\Rightarrow \quad & (k-1)\frac{r_1^2}{1 - r_1^2} \geq \sum_{i=2}^k \frac{r_i^2}{1 - r_i^2} \\
\Rightarrow \quad & k\frac{r_1^2}{1 - r_1^2} \geq \sum_{i=1}^k \frac{r_i^2}{1 - r_i^2} \\
\Rightarrow \quad & \frac{r_1^2}{1 - r_1^2} \geq k^{-1}\sum_{i=1}^k \frac{r_i^2}{1 - r_i^2} \\
\Rightarrow \quad & \frac{\frac{r_1^2}{1-r_1^2}}{1 + \frac{r_1^2}{1-r_1^2}} \geq \frac{k^{-1}\sum_{i=1}^k \frac{r_i^2}{1-r_i^2}}{1 + k^{-1}\sum_{i=1}^k \frac{r_i^2}{1-r_i^2}}
\end{aligned}$$

$$
\begin{aligned}
&\Rightarrow \quad r_1^2 = \frac{\frac{r_1^2}{1-r_1^2}}{1+\frac{r_1^2}{1-r_1^2}} \geq \frac{k^{-1}\sum_{i=1}^{k}\frac{r_i^2}{1-r_i^2}}{1+k^{-1}\sum_{i=1}^{k}\frac{r_i^2}{1-r_i^2}} \\
&\Rightarrow \quad \eta_\theta^2 \geq \eta_U^2
\end{aligned}
$$

Case 3: Lawley-Hotelling's versus Pillai's
The function $\varphi(r^2) = \frac{r^2}{1-r^2}$ is convex for $r^2 \in (0,1)$ since $\varphi''(r^2) > 0$. Using Jensen's inequality for convex functions [62] we have that

$$
\begin{aligned}
& k^{-1}\varphi(r_1^2) + \cdots + k^{-1}\varphi(r_k^2) \geq \varphi\left(k^{-1}\sum_{i=1}^{k} r_i^2\right) \\
\Rightarrow \quad & k^{-1}\sum_{i=1}^{k}\frac{r_i^2}{1-r_i^2} \geq \frac{k^{-1}\sum_{i=1}^{k} r_i^2}{1-k^{-1}\sum_{i=1}^{k} r_i^2} \\
\Rightarrow \quad & \frac{1}{k^{-1}\sum_{i=1}^{k}\frac{r_i^2}{1-r_i^2}} \leq \frac{1-k^{-1}\sum_{i=1}^{k} r_i^2}{k^{-1}\sum_{i=1}^{k} r_i^2} \\
\Rightarrow \quad & \frac{1}{k^{-1}\sum_{i=1}^{k}\frac{r_i^2}{1-r_i^2}} \leq \frac{1}{k^{-1}\sum_{i=1}^{k} r_i^2} - 1 \\
\Rightarrow \quad & 1 + \frac{1}{k^{-1}\sum_{i=1}^{k}\frac{r_i^2}{1-r_i^2}} \leq \frac{1}{k^{-1}\sum_{i=1}^{k} r_i^2} \\
\Rightarrow \quad & \frac{1+k^{-1}\sum_{i=1}^{k}\frac{r_i^2}{1-r_i^2}}{k^{-1}\sum_{i=1}^{k}\frac{r_i^2}{1-r_i^2}} \leq \frac{1}{k^{-1}\sum_{i=1}^{k} r_i^2} \\
\Rightarrow \quad & \frac{k^{-1}\sum_{i=1}^{k}\frac{r_i^2}{1-r_i^2}}{1+k^{-1}\sum_{i=1}^{k}\frac{r_i^2}{1-r_i^2}} \geq k^{-1}\sum_{i=1}^{k} r_i^2 \\
\Rightarrow \quad & \eta_U^2 \geq \eta_V^2
\end{aligned}
$$

Hence $\eta_\Lambda^2 \geq \eta_\theta^2 \geq \eta_U^2 \geq \eta_V^2$. Note that $\eta_\Lambda^2 = \eta_\theta^2$ iff $r_2 = \cdots = r_k = 0$, and $\eta_\theta^2 = \eta_U^2 = \eta_V^2$ iff $r_1 = \cdots = r_k$, so in practice strict inequality will always be observed.

A.4 Asymptotic properties of BIC^* using adjusted Wilks' Lambda

The information criterion defined as

$$BIC^* = n \ln \left(\frac{\Lambda}{E[\Lambda]} \right) + (p+q) \ln n$$

has same asymptotic properties as the standard BIC defined by

$$BIC = n \ln \Lambda + (p+q) \ln n$$

Proof
Nishii [48] shows that any information criteria defined as

$$GIC = n \ln \Lambda + (p+q) a_n$$

where $\lim_{n\to\infty} a_n = \infty$ and $\lim_{n\to\infty} n^{-1} a_n = 0$ share the same asymptotic properties. Swartz's criterion (BIC) formed using the likelihood ratio is

$$BIC = n \ln \Lambda + (p+q) \ln n$$

and has the same asymptotic properties as those in the class of generalized information criteria since

$$\begin{aligned} \lim_{n\to\infty} \ln n &= \infty \\ \lim_{n\to\infty} n^{-1} \ln n &= 0 \end{aligned}$$

Since

$$\begin{aligned} BIC^* &= n \ln \left(\frac{\Lambda}{E[\Lambda]} \right) + (p+q) \ln n \\ &= n \ln(\Lambda) - n \ln(E[\Lambda]) + (p+q) \ln n \\ &= n \ln(\Lambda) + (p+q) \left(\frac{-n}{p+q} \ln(E[\Lambda]) + \ln n \right) \\ &= n \ln(\Lambda) + (p+q) a_n \end{aligned}$$

then we must now show that

$$\begin{aligned} \lim_{n\to\infty} a_n &= \infty \\ \lim_{n\to\infty} n^{-1} a_n &= 0 \end{aligned}$$

where

$$a_n = \frac{-n}{p+q}\ln(E[\Lambda]) + \ln n$$

Now,

$$\ln E[\Lambda] = \sum_{i=0}^{p-1} \ln\left(1 - \frac{q}{n-i}\right)$$

hence

$$\begin{aligned}
\lim_{n\to\infty} a_n &= \lim_{n\to\infty}\left(\frac{-n}{p+q}\ln(E[\Lambda]) + \ln n\right) \\
&= \lim_{n\to\infty}\left(\frac{-n}{p+q}\sum_{i=0}^{p-1}\ln\left(1 - \frac{q}{n-i}\right) + \ln n\right) \\
&= \frac{1}{p+q}\sum_{i=0}^{p-1}\lim_{n\to\infty}\ln\left(1 - \frac{q}{n-i}\right)^{-n} + \lim_{n\to\infty}\ln n \\
&= \frac{pq}{p+q} + \lim_{n\to\infty}\ln n \\
&= \infty
\end{aligned}$$

Also,

$$\begin{aligned}
\lim_{n\to\infty} n^{-1}a_n &= \lim_{n\to\infty}\frac{1}{n}\left(\frac{-n}{p+q}\ln(E[\Lambda]) + \ln n\right) \\
&= \lim_{n\to\infty}\left(\frac{-1}{p+q}\ln(E[\Lambda]) + \frac{\ln n}{n}\right) \\
&= \frac{-1}{p+q}\lim_{n\to\infty}\ln(E[\Lambda]) + \lim_{n\to\infty}\frac{\ln n}{n} \\
&= \frac{-1}{p+q}\lim_{n\to\infty}\sum_{i=0}^{p-1}\ln\left(1 - \frac{q}{n-i}\right) + 0 \\
&= \frac{-1}{p+q}\sum_{i=0}^{p-1}\lim_{n\to\infty}\ln\left(1 - \frac{q}{n-i}\right) \\
&= 0
\end{aligned}$$

therefore BIC^* is in the class of generalized information criteria and has the same asymptotic properties as BIC.

Appendix B

SAS Code

B.1 Principal Components Analysis

The PCABMA SAS macro analyzes data suitable to a standard principal components analysis. The data to be analyzed will be a SAS data set whose name is stored in parameter **datin**. The list of variables is stored in parameter **varlist**. The component number to be modeled is stored in parameter **comp**. The Occam's razor parameter, **occ**, which specifies the maximum ratio of the most probable model to the least probable model retained. The **its** parameter contains the desired number of iterations the MC^3 algorithm is to execute. The **datout** parameter is the name of the SAS data set where the retained models along with their posterior probability.

```
/*-----------------------------------------------------------+
 |  MACRO PCABMA                                             |
 |                                                           |
 |  Paramaters                                               |
 |    datin   = SAS dataset containing data to be analyzed   |
 |    varlist = List of variable names to be analyzed        |
 |    comp    = Component number to be investigated          |
 |    occ     = Occam's razor number                         |
 |    datout  = SAS dataset for models visited to be stored  |
 |    its     = number of iterations in MC^3                 |
 +-----------------------------------------------------------*/
%macro pcabma(datin=,varlist=,comp=,occ=,datout=,its=);

%let code = %str(
```

```
    /* construct Rj */
    rj = diag(delta)*r*diag(delta);

    /* compute eigen-info for model */
    l = eigval(rj);
    e = eigvec(rj);

    /* identify best component */
    corr = 0;
    do i = 1 to ncol(e);
      v = abs(e[,1]‘*te);
      if v>corr then do; corr=v; c=i; end;
    end;
    rsq = l[c,1]/d;

    /* calculate TIC */
    tic = n*log(1-rsq)+a*sum(delta);
);

proc iml;

  /* read in data */
  use &datin;
  read all var {&varlist} into x;

  /* p = # of variables, n = # of obs, u = n*1 vector of 1’s */
  p = ncol(x);
  n = nrow(x);
  u = j(n,1,1);

  /* center data */
  c = x-u*inv(u‘*u)*u‘*x;

  /* calculate correlation matrix */
  r = j(p,p,0);
  do i = 1 to p;
    do j = 1 to p;
      r[i,j] = (c[,i]‘*c[,j])/sqrt(c[,i]‘*c[,i]*c[,j]‘*c[,j]);
    end;
  end;
```

```
/* spectral decomposition of correlation matrix */
e = eigvec(r);
l = eigval(r);
r = j(p,p,0);
do i = &comp to p;
  r = r + e[,i]*l[i,1]*e[,i]‘;
end;

/* eigen-info for Rj */
e = eigvec(r);
l = eigval(r);
d = sum(l);

/* full model r-sq */
rsq = l[1,1]/d;

/* penalty term */
a = min(log(n),-(n/p)*log(1-rsq));

/* first eigenvector for full model */
te = e[,1];

/* pick random starting point in model space */
delta = j(1,p,0);
do i = 1 to p;
  if rannor(0)<.5 then delta[1,i]=1;
end;

/* store iteration history */
hist = j(1,p+1,.);

/* MC^3 */
do its = 1 to &its;

  /* random jump to a new spot in model space */
  if ranuni(0) < 0.01 then do;
    delta = j(1,p,0);
    do i = 1 to p;
      if rannor(0)<.5 then delta[1,i]=1; else delta[1,i]=0;
    end;
  end;
```

```
    /* TIC of neighborhood center */
    &code;
    tic0 = tic;

    /* proposal model in neighborhgood */
    modnum = int(ranuni(0)*p)+1;
    delta[1,modnum] = 1-delta[1,modnum];

    /* TIC for proposed model */
    &code;
    tic1 = tic;

    /* either accpet proposal or try again */
    pacc = min(1,exp(-.5*(tic1-tic0)));
    tic=tic1;
    if ranuni(0)>pacc then delta[1,modnum] = 1-delta[1,modnum];
                      else hist = hist//(delta||tic);
  end;

  /* output models visited */
  hist = hist[2:nrow(hist),];
  create temp var {&varlist tic};
  append from hist;

/* sort by TIC */
proc sort data=temp;
  by tic;
run;

proc iml;
  /* read in iteration history */
  use temp;
  read all var {&varlist tic} into m0;
  p = ncol(m0)-1;

  /* remove redundant model info */
  m1 = m0[1,];
  do i = 2 to nrow(m0);
    v = sum(abs(m0[i-1,1:p]-m0[i,1:p]));
    if v>.5 then m1 = m1//m0[i,];
```

```
  end;

  /* occam's razor */
  lim = min(m1[1,p+1])+2*log(&occ);
  m2 = m1[1,];
  do i = 2 to nrow(m1);
    if m1[i,p+1] < lim then m2 = m2//m1[i,];
  end;

  /* compute posterior probabilities */
  tic = m2[,p+1];
  prob = exp(-.5*(tic-max(tic)));
  prob = prob/sum(prob);

  m = m2[,1:p]||prob;
  create &datout var {&varlist postprob};
  append from m;

/* variable assessment using E(delta) */
proc means data=&datout mean;
  var &varlist;
  weight postprob;
run;

%mend;
```

B.2 Canonical Variate Analysis

The CVABMA SAS macro analyzes data suitable to a canonical variate analysis. The data to be analyzed will be a SAS data set whose name is stored in parameter **datin**. The list of possible discriminator variables is stored in parameter **x** and the set if $g-1$ group indicator variables are stored in parameter **y**. If the "*size*" prior is desired, then parameter **theta** can be set to some value between 0 and 1, or set to 0.5 if the "*uniform*" is desired. The Occam's razor parameter, **occ**, which specifies the maximum ratio of the most probable model to the least probable model retained. The **its** parameter contains the desired number of iterations the MC^3 algorithm is to execute. The **datout** parameter is the name of the SAS data set where the retained models along with their posterior probability.

```
/*-----------------------------------------------+
 |  Macro: CVABMA                                |
 |                                               |
 |  Parameters                                   |
 |    datain = SAS dataset to be analyzed        |
 |    datout = SAS dataset with results          |
 |    y      = group indicator variables (g-1)   |
 |    x      = list of discriminator variables   |
 |    theta  = prior prob of variable inclusion  |
 |    occ    = occam's razor parameter           |
 |    its    = number of iterations in MC^3      |
 |                                               |
 |  Output                                       |
 |    SAS dataset containing posterior model     |
 |      space with BIC values and posterior      |
 |      model probability                        |
 +-----------------------------------------------*/

%macro cvabma(datin=,datout=,y=,x=,theta=,occ=,its=);

%let code = %str(

    /*  build model from delta vector info */
    x = u;
    do i = 1 to k;
      if delta[1,i]=1 then x=x||d[,i];
    end;

    /* sums of squares and cross product matrices */
```

```
    sxx = x‘*x;
    syy = y‘*y;
    sxy = x‘*y;
    syx = sxy‘;

    /* number of variables in model matrix */
    p = ncol(x);
    q = ncol(y);

    /* squared canonical correlations */
    if p<q then do;
      r = inv(eigvec(sxx)*diag(sqrt(abs(eigval(sxx))))*eigvec(sxx)‘);
      cc = eigval(r*sxy*inv(syy)*syx*r);
    end; else do;
      r = inv(eigvec(syy)*diag(sqrt(abs(eigval(syy))))*eigvec(syy)‘);
      cc = eigval(r*syx*inv(sxx)*sxy*r);
    end;

    /* wilks’ lambda */
    wilks = exp(sum(log(1-cc)));

    /* expected value of wilks’ lambda */
    ewilks = j(p,1,.);
    do i = 0 to p-1;
      ewilks[i+1,1] = log(1-q/(n-i));
    end;
    ewilks = exp(sum(ewilks));

    /* log(Prob(data|M_i)Prob(M_i)) */
    lnp = -.5*(n*log(wilks/ewilks)+(p-1)*pen);
);

proc iml;
  /* read in data */
  use &datin;
  read all var {&y} into y;
  read all var {&x} into d;

  /* number of observations */
  n = nrow(y);
```

```
/* column vector of 1's */
u = j(n,1,1);

/* center data */
y = y - u*inv(u'*u)*u'*y;
d = d - u*inv(u'*u)*u'*d;

/* total number of explanatory variables */
k = ncol(d);

/* model indicator vector */
delta = j(1,k,0);

/* iteration history */
hist = delta||0;

/* select a random model */
do i = 1 to k;
  if ranuni(0)<.5 then delta[1,i]=1;
end;

/* penaly term for BIC with theta prior on variables */
pen = log(n) - 2*log( &theta / (1 - &theta));

/*--- MC^3 ---*/
do its = 1 to &its;

  /* compute -0.5BIC_0 */
  &code;
  lnp0 = lnp;

  /* propose jump to new neighborhood */
  m = int(ranuni(0)*k)+1;
  delta[1,m] = 1-delta[1,m];

  /* compute -0.5BIC_1 */
  &code;
  lnp1 = lnp;

  /* probability to accept move */
  p = lnp0//lnp1;
```

```
    p = exp(p-max(p));
    p = p/sum(p);
    pacc = min(1,p[2,1]/p[1,1]);

    /* accept proposed move? */
    lnp = lnp1;
    if ranuni(0)>pacc then do;
      delta[1,m] = 1-delta[1,m];
      lnp=lnp0;
    end;

    /* update iteration history */
    hist = hist//(delta||lnp);

    /* random hop to some other part of model space */
    if ranuni(0)<.01 then do;
      do i = 1 to k;
        if ranuni(0)<.5 then delta[1,i]=1; else delta[1,i]=0;
      end;
    end;
  end;

  /* output iteration history */
  hist = hist[2:nrow(hist),];
  create temp0 var {&x lnp};
  append from hist;

/* sort history so that best models are first in list */
proc sort data=temp0;
  by descending lnp;
run;

proc iml;
  /* read in sorted history */
  use temp0;
  read all var {&x lnp} into m0;

  /* keep most probable model */
  m1 = m0[1,];

  /* remove redundant model info */
```

```
  v = ncol(m0)-1;
  do i = 2 to nrow(m0);
    t = sum(abs(m0[i,1:v]-m0[i-1,1:v]));
    if t>0 then m1=m1//m0[i,];
  end;

  /*--- remove unlikly models using occam's razor ---*/
  v = ncol(m1);
  m2 = m1[1,];
  do i = 2 to nrow(m1);
    t = m1[1,v]-m1[i,v]-log(&occ);
    if t<0 then m2=m2//m1[i,];
  end;

  /* compute BIC for models in posterior space */
  bic = -2*m2[,v];
  do i = 1 to nrow(m2);
    bic[i,1] = bic[i,1]+2*sum(m2[i,1:(v-1)])*log(&theta/(1-&theta));
  end;

  /* compute posterior probabilities of remaining models */
  prob = exp(m2[,v]-max(m2[,v]));
  prob = prob/sum(prob);

  /* output posterior model space */
  m2 = m2[,1:(v-1)]||bic||prob;
  create &datout var {&x bic prob};
  append from m2;

/* variable assessment via E(delta) */
proc means data=&datout mean;
  var &x;
  weight prob;
run;

%mend;
```

B.3 Canonical Correlation Analysis

The CCABMA SAS macro analyzes data suitable to a canonical correlation analysis. The data to be analyzed will be a SAS data set whose name is stored in parameter **datin**. The list of one set of variable names is stored in parameter **x** and the other set of variable names is stored in parameter **y**. If the "*size*" prior is desired, then parameter **theta** can be set to some value between 0 and 1, or set to 0.5 if the "*uniform*" is desired. The Occam's razor parameter, **occ**, which specifies the maximum ratio of the most probable model to the least probable model retained. The **its** parameter contains the desired number of iterations the MC^3 algorithm is to execute. The **datout** parameter is the name of the SAS data set where the retained models along with their posterior probability.

```
/*-----------------------------------------------+
 |  Macro: CCABMA                                |
 |                                               |
 |  Parameters                                   |
 |    datain = SAS dataset to be analyzed        |
 |    datout = SAS dataset with results          |
 |    y      = list of variables                 |
 |    x      = list of variables                 |
 |    theta  = prior prob of variable inclusion  |
 |    occ    = occam's razor parameter           |
 |    its    = number of iterations in MC^3      |
 |                                               |
 |  Output                                       |
 |    SAS dataset containing posterior model     |
 |    space with BIC values and posterior        |
 |    model probability                          |
 +-----------------------------------------------*/

%macro ccabma(datin=,datout=,y=,x=,theta=,occ=,its=);
%let code = %str(
    /*  build X from delta vector info */
    x = u;
    do i = 1 to pt;
      if delta[1,i]=1 then x=x||xd[,i];
    end;

    /*  build Y from delta vector info */
    y = u;
    do i = 1 to qt;
```

```
      if delta[1,i+pt]=1 then y=y||yd[,i];
    end;

    /* sums of squares and cross product matrices */
    sxx = x‘*x;
    syy = y‘*y;
    sxy = x‘*y;
    syx = sxy‘;

    /* number of variables in model matrix */
    pi= ncol(x)-1;
    qi= ncol(y)-1;

    /* squared canonical correlations */
    if pi<qi then do;
      r = inv(eigvec(sxx)*diag(sqrt(abs(eigval(sxx))))*eigvec(sxx)‘);
      cc = eigval(r*sxy*inv(syy)*syx*r);
    end; else do;
      r = inv(eigvec(syy)*diag(sqrt(abs(eigval(syy))))*eigvec(syy)‘);
      cc = eigval(r*syx*inv(sxx)*sxy*r);
    end;

    /* wilks’ lambda */
    if nrow(cc)>1 then cc = cc[2:nrow(cc),1];
                  else cc=0;
    wilks = exp(sum(log(1-cc)));

    /* expected value of wilks’ lambda */
    ewilks = j(pi,1,.);
    do i = 0 to pi-1;
      ewilks[i+1,1] = log(1-qi/(n-i));
    end;
    ewilks = exp(sum(ewilks));

    /* log(Prob(data|M_i)Prob(M_i)) */
    lnp = -.5*(n*log(wilks/ewilks)+(pi+qi)*pen);
);

proc iml;
  /* read in data */
  use &datin;
```

```
read all var {&y} into yd;
read all var {&x} into xd;

/* number of observations  and column vector of 1's */
n = nrow(yd);
u = j(n,1,1);

/* center data */
yd = yd - u*inv(u‘*u)*u‘*yd;
xd = xd - u*inv(u‘*u)*u‘*xd;

/* total number of variables in X and Y */
pt= ncol(xd);
qt= ncol(yd);
pq= pt+qt;

/* model indicator vector and iteration history matrix */
delta = j(1,pq,0);
hist = delta||0;

/* select a random model */
do i = 1 to pq;
  if ranuni(0)<.5 then delta[1,i]=1;
end;

/* penaly term for BIC with theta prior on variables */
pen = log(n) - 2*log( &theta / (1 - &theta));

/*--- MC^3 ---*/
do its = 1 to &its;

  /* compute -0.5BIC_0 */
  &code;
  lnp0 = lnp;

  /* propose jump to new neighborhood */
  m = int(ranuni(0)*pq)+1;
  delta[1,m] = 1-delta[1,m];

  /* compute -0.5BIC_1 */
  &code;
```

```
    lnp1 = lnp;

    /* probability to accept move */
    p = lnp0//lnp1;
    p = exp(p-max(p));
    p = p/sum(p);
    pacc = min(1,p[2,1]/p[1,1]);

    /* accept proposed move? */
    lnp = lnp1;
    if ranuni(0)>pacc then do;
      delta[1,m] = 1-delta[1,m];
      lnp=lnp0;
    end;

    /* update iteration history */
    hist = hist//(delta||lnp);

    /* random hop to some other part of model space */
    if ranuni(0)<.01 then do;
      do i = 1 to pq;
        if ranuni(0)<.5 then delta[1,i]=1; else delta[1,i]=0;
      end;
    end;
  end;

  /* output iteration history */
  hist = hist[2:nrow(hist),];
  create temp0 var {&x &y lnp};
  append from hist;

/* sort history so that best models are first in list */
proc sort data=temp0;
  by descending lnp;
run;

proc iml;
  /* read in sorted history */
  use temp0;
  read all var {&x &y lnp} into m0;
```

```
  /* remove redundant model info */
  v = ncol(m0)-1;
  m1 = m0[1,];
  do i = 2 to nrow(m0);
    t = sum(abs(m0[i,1:v]-m0[i-1,1:v]));
    if t>0 then m1=m1//m0[i,];
  end;

  /*--- remove unlikly models using occam's razor ---*/
  v = ncol(m1);
  m2 = m1[1,];
  do i = 2 to nrow(m1);
    t = m1[1,v]-m1[i,v]-log(&occ);
    if t<0 then m2=m2//m1[i,];
  end;

  /*--- remove models less likely than null ---*/
  m3 = j(1,v,0);
  do i = 1 to nrow(m2);
    if m2[i,v]>0 then m3 = m3//m2[i,];
  end;
  if nrow(m3)>1 then m3=m3[2:nrow(m3),];

  /* compute BIC for models in posterior space */
  bic = -2*m3[,v];
  do i = 1 to nrow(m3);
    bic[i,1] = bic[i,1]+2*sum(m3[i,1:(v-1)])*log(&theta/(1-&theta));
  end;

  /* compute posterior probabilities of remaining models */
  prob = exp(m3[,v]-max(m3[,v]));
  prob = prob/sum(prob);

  /* output posterior model space */
  m3 = m3[,1:(v-1)]||bic||prob;
  create &datout var {&x &y bic prob};
  append from m3;

/* variable assessment via E(delta) */
proc means data=&datout mean;
  var &x &y;
```

```
  weight prob;
run;
%mend;
```

Appendix C

Variable Descriptions

C.1 Transformations

Any transformation of the variables used are those proposed by Norton [51] to achieve approximate univariate normality.

C.2 Chemical Variables

Variable	Description	Transformation
AMM	NH_3 and NH_4 conc.	$x^{-0.25}$
BOD	5-d biological oxygen demand	Log
COD	chemical oxygen demand	Log
FE	Iron conc.	Log
NOX	NO_2 and NO_3 conc.	Log
PB	Lead conc.	Inverse
PHO	Total phosphorus	$x^{-0.25}$
TSS	Total suspended solids (residue)	Log
ZN	Total zinc	Log

C.3 Habitat Variables

Variable	Description	Transformation
CHANNEL	QHEI score for channel metric	None
COVER	QHEI instream cover score	None
EMBSS	Embeddedness subscore	None
GRADIENT	Stream gradient	Log
POOL	Pool QHEI metrics	None
RIFEMSS	Riffle embeddedness subscore	None
RIFFLE	Riffle QHEI metric	None
RIPARIAN	Riparian QHEI metric	None
RIPSS	Riaparian width subscore	None
SILTSS	Silt subscore	None
SUBSTRAT	QHEI substrate metric	None

Note: Qualitative Habitat Evaluation Index (QHEI) measurements made using Ohio EPA's protocol [52]

C.4 Benthic Macroinvertebrate Variables

Variable	Description	Transformation
PERCAD	Percent that are caddisfly taxa	$x^{0.5}$
PERCR	Percent that are *Crictopus*	$x^{0.25}$
PERD	Percent that are depositional taxa	$x^{0.25}$
PERDIPT	Percent that are dipteran taxa	None
PERE	Percent that are erosional taxa	$\arcsin\sqrt{x}$
PCFILTE	Percent that are filter feeding insext taxa	None
PCGATH	Percent that are gatering insect taxa	None
PERGL	Percent that are *Glyptotendipes*	$x^{0.25}$
INSNODI	Number of insect taxa excluding dipterans	None
PERTOLN	Percent that are tolerant organisms	log
PERMAY	Percent that are mayfly taxa	$x^{0.5}$
NUMQUAL	Total number of taxa collected in the qualitative sample[1]	None
NUMTAXA	Total number of quantitative taxa	None
PEROL	Percent that are oligochaetes	$x^{0.25}$
PEROTS	Percent that are organic tolerant	$x^{0.25}$
PEROTHDI	Percent that are dipterans and non-insects	log
QUALEPT	Number of special taxa[2] in qualitative dipnet sample	None
PCSHRED	Percent that are "shredding" insect taxa	$x^{0.25}$
PERTANY	Percent that are tanytarsini midges	$x^{0.5}$
PERTTS	Percent that are toxic tolerant	$x^{0.25}$

Notes

1. NUMQUAL and QUALEPT are from data collected using a combination of dipnet and hand-picking. All other variables are from data collected in Hester-Dendry artificual substrate samples.

2. Special: Ephemeroptera, Plecoptera, and Tricoptera taxa.

Bibliography

[1] T. W. Anderson. (1963). Asymptotic theory for principal component analysis. *Annals of Mathematical Statistics*. Vol 34:122–148.

[2] N. M. Al-Kandari, I. T. Jolliffe. (1997). Variable selection and interpretation in canonical correlation analysis. *Communications in Statistics, Part B. Simulation and Computation*. Vol 26:873–900.

[3] E. M. L. Beale, M. G. Kendall, and D. W. Mann. (1967). The discarding of variables in multivariate analysis. *Biometrika*. Vol 54:357–366.

[4] P. J. Bickel and K. A. Doksum. (1977). *Mathematical Statistics*. Prentice Hall, Inc.

[5] B. P. Carlin and T. A. Louis. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*, Chapman and Hall.

[6] M. A. Clyde. (1999). Bayesian model averaging and model search strategies. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 6*. Oxford University Press. 157–185.

[7] E. M. Cramer and W. A. Nicewander. (1979). Some symmetric, invariant measures of multivariate association. *Pyschometrika*. Vol 44, No 1, 43–54.

[8] M. H. DeGroot. (1970). *Optimal Statistical Decisions*. McGraw-Hill Publishing Company.

[9] D. Draper. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society, Ser. B.*, Vol 57:45–97.

[10] N. R. Draper and H. Smith. (1981). *Applied Regression Analysis*. 2^{nd} Edition. John Wiley & Sons.

[11] S. D. Dyer, C. White-Hull, G. J. Carr, E. P. Smith, and X. Wang. (2000). Bottom-up and top-down approaches to assess multiple stressors over large geographic areas. *Environmental Toxicology and Chemistry*. Vol 19, No 4(2):1066–1075.

[12] R. A. Eisenbeis and R. B. Avery. (1972). *Discriminant Analysis and Classification Procedures*. Lexington Books.

[13] B. S. Everitt and G. Dunn. (1992). *Applied Multivariate Data Analysis*. Oxford University Press.

[14] S. Frontier. (1976). Etude de la decroissant valeurs propres dans une analyse en composantes principales: Comparison avec le modele du baton brise. *Journal of Experimental Marine Biology and Ecology*. Vol 25:67–74.

[15] G. M. Furnival and R. W. Wilson. (1974). Regression by leaps and bounds. *Technometrics*, Vol 16:499–511.

[16] A. Gelman, J. B. Carlin, H. S. Stern and D. B. Rubin. (1997). *Bayesian Data Analysis*, Chapman and Hall.

[17] M. A. Girshick. (1939). On the sampling theory of roots of determinantal equations. *Annals of Mathematical Statistics*. Vol 10:203–224.

[18] C. Goutis and G. Casella. (1999). Explaining the saddlepoint approximation. *The American Statistician*, Vol 53:216–224.

[19] D. J. Hand. (1981). *Discrimination and Classification*. John Wiley & Sons.

[20] W. K. Hastings. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*. Vol 57:97–109

[21] J. A. Hoeting. (1994). *Accounting for Model Uncertainty in Linear Regression*, PhD thesis, University of Washington.

[22] J. A. Hoeting, D. Madigan, A. Raftery, Chris T. Volinsky. (1998). Bayesian model averaging. Technical Report 335, Department of Statistics, University of Washington.

[23] C. J. Huberty. (1994). *Applied discriminant analysis*. Wiley & Sons.

[24] D. Jackson. (1993). Stopping rules in principal components analysis: A comparison of heuristic and statistical approaches. *Ecology*. 2204–2214.

[25] J. E. Jackson. (1991). *A User's Guide to Principal Components*. John Wiley & Sons.

[26] R. A. Johnson and D. W. Wichern. (1998). *Applied Multivariate Statistical Analysis*. 4^{th} edition. Prentice Hall.

[27] J. Jong and S. Kotz. (1999). On a relation between principal components and regression analysis. *The American Statistician*. Vol 53, No 4, 349–351.

[28] I. T. Jolliffe. (1986). *Principal Components Analysis*. Springer-Verlag.

[29] I. T. Jolliffe. (1972). Discarding variables in principal component analysis. I: Artificial data. *Applied Statistics*. Vol 21:160–173.

[30] R. E. Kass, L. Tierney, and J. B. Kadane. (1990). The validity of posterior expansions based on Laplace's method. In Seymor Geisser, James S. Hodges, S. James Press, and Arnold Zellner, editors, *Bayesian and Likelihood Methods in Statistics and Econometrics*, Vol 7:473–488, Elsevier Science Publishing Company.

[31] R. E. Kass and A. E. Raftery. (1995). Bayes factors. *Journal of the American Statistical Association*. Vol 90:773–795.

[32] R. E. Kass and L. Wasserman. (1995). A reference Bayesian test for nested hypothesis with large samples. *Journal of the American Statistical Association*. Vol 90:928–934.

[33] W. R. Klecka. (1980). Discriminant analysis. In John L. Sullivan and Richard G. Niemi, editors. *Quantitative Applications in the Social Sciences*. Sage Publications.

[34] W. J. Krzanowski. (1990). *Principles of Multivariate Analysis: A User's Perspective*. Oxford University Press.

[35] W. J. Krzanowski and F. H. C. Marriot. (1994). *Multivariate Analysis*. Part 1 of Kendall's Library of Statistics. Edward Arnold and Halsted Press.

[36] D. V. Lambert, A. R. Wildt, and R. M. Durand (1990). Assessing sampling variation relative to number-of-factors criteria. *Educational and Psychological Measurement*. Vol 50:33–49.

[37] M. S. Levine. (1977). Canonical analysis and factor comparison. In John L. Sullivan and Richard G. Niemi, editors. *Quantitative Applications in the Social Sciences*. Sage Publications.

[38] D. Madigan and A. E. Raftery. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, Vol 89:1535–1546.

[39] D. Madigan and J. York. (1993). Bayesian graphical models for discrete data. Technical Report 259, Department of Statistics, University of Washington.

[40] K. V. Mardia, J. T. Kent, and J. M. Bibby. (1979). *Multivariate Analysis*. Academic Press.

[41] G. P. McCabe. (1984). Principal variables. *Technometrics*. Vol 26, No 2: 137–144.

[42] R. J. McKay and N. A. Campbell. (1982). Variable selection techniques in discriminant analysis. *British Journal of Mathematical and Statistical Psychology*. Vol 35, 1–29.

[43] D. F. Morrison. (1983). *Applied Linear Statistical Models*. Prentice-Hall.

[44] R. O. Mueller and J. B. Cozad. (1993). Standardized discriminant coefficients: A rejoinder. *Journal of Educational Statistics*. Vol 18, No 1: 108–114.

[45] K. E. Muller and B. L. Peterson. (1984). Practical methods for computing power in testing the multivariate general linear hypothesis. *Computational Statistics & Data Analysis*. Vol 2, 134–158.

[46] G. D. Murray. (1977). A cautionary note on selection of variables on discriminant analysis. *Applied Statistics*. Vol 26, No 3:246–250.

[47] R. H. Myers. (1990). *Classical and Modern Regression with Applications*. 2^{nd} Edition. Duxbury Press.

[48] R. Nishii. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics*. Vol 12, No 2: 758–765.

[49] J. R. Norris. (1997). *Markov Chains*. Cambridge University Press.

[50] S. B. Norton, S. M. Cormier, M. Smith, R. C. Jones. (2000). Can biological assessments discriminate among types of stress? A case study from the Eastern Corn Belt Plains ecoregion. *Environmental Toxicology and Chemistry*. Vol 19, No 4(2):1113–1119.

[51] S. B. Norton. (1999). *Using biological monitoring data to distinguish among types of stress in streams of the eastern corn belt plains ecoregion*. Ph.D. thesis. George Mason University.

[52] Ohio Environmental Protection Agency. (1989). The qualitative habitat evaluation index (QHEI): Rational, methods, and application. Columbus, OH, USA.

[53] N. G. Polson. (1996). Convergence of Markov chain Monte Carlo algorithms. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 5*. Oxford University Press. 297–321.

[54] S. J. Press. (1972). *Applied Multivariate Analysis*. Holt, Rinehart, and Winston, Inc..

[55] A. E. Raftery. (1995). Bayesian model selection in social research. In Peter V. Marsden, editor, *Sociological Methodology*, Vol 25:111–195, Blackwell publishers.

[56] A. E. Raftery, D. Madigan, and J. A. Hoeting. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*. Vol 92:179–191.

[57] A. E. Raftery, D. Madigan, and C. T. Volinsky. (1996). Accounting for model uncertainty in survival analysis improves predictive performance. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 5*. Oxford University Press. 297–321.

[58] C. R. Rao. (1973). *Linear Statistical Inference and its Applications*. John Wiley & Sons.

[59] A. C. Rencher. (1992). Interpretation of canonical discriminant functions, canonical variates, and principal components. *American Statistician*. Vol 46:217–225.

[60] A. C. Rencher. (1995). *Methods of Multivariate Analysis*. John Wiley & Sons.

[61] A. C. Rencher. (1998). *Multivariate Statistical Inference and Applications*. John Wiley & Sons.

[62] A. N. Shiryayev. (1984). *Probability*. Springer-Verlag.

[63] G. Swartz. (1978). Estimating the dimension of a model. *The Annals of Statistics*. Vol 6:461–464.

[64] H. M. Taylor and S. Karlin. (1994). *An Introduction to Stochastic Modeling*. Revised Edition. Academic Press.

[65] R. A. Thisted. (1988). *Elements of Statistical Computing*. Chapman & Hall.

[66] L. Tierney and J. B. Kadane. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, Vol 81:82–86.

Robert B. Noble Jr

Education

- **Virginia Polytechnic Institute and State University**, Blacksburg, Virginia. Degree: Doctor of Philosophy in Applied Statistics. GPA 3.98/4.0. Expected graduation date: December 2000. Dissertation topic: Multivariate Applications of Bayesian Model Averaging.
- **University of Delaware**, Newark, Delaware. Degree: Master of Science in Applied Statistics. GPA 3.98/4.0, May 1997.
- **Castleton State College**, Castleton, Vermont. Degree: Bachelor of Arts in Mathematics, Summa Cum Laude. GPA 3.89/4.0, 4.0/4.0 in major, May 1994.
- **Community College of the Air Force**. Degree: Associate in Applied Science in Electronic Engineering Technology, July 1988.

Experience

- To start 1/01
  Project Biostatistican, DuPont Pharmaceutical, Wilmington, Delaware. Designs and analyzes projects in the Pre-clinical, Discovery, and Research & Development areas. Develops and evaluates statistical methods and protocol to aid in the identification of potentially beneficial medical compounds.
- 9/97-12/00
  Consultant, Statistics Laboratory, Virginia Tech. Designs and analyzes projects involving classical experimental design, linear and nonlinear regression, generalized linear models, response surfaces, empirical and hierarchical Bayesian models, sampling, multivariate techniques, and conjoint analysis.
- 5/99-7/99
  Instructor, Department of Statistics, Virginia Tech, Blacksburg, Virginia. Teaches introductory statistics course. Plans and presents lectures concerning elementary statistical concepts. Develops and grades exams, quizzes and homework assignments.

- 1/98-5/98
  Teaching Assistant, Department of Statistics, Virginia Tech, Blacksburg, Virginia. Conducts recitation sessions for introductory statistics. Grades homework and administers tests.

- 6/96-8/97
  Statistician, DuPont, Quality Management & Technology Center, Wilmington, Delaware. Designs and analyzes experimental and observational data for clients in automotive, printing and publishing, engineering polymers, and agricultural products strategic business units. Builds models from process data and theoretical principles. Determines optimal settings for multi-response processes. Analyzes toxicological data and contributes to reports to the Environmental Protection Agency (EPA). Designs and constructs Visual Basic programs used in Microsoft Excel to enable clients to perform on site analyzes, simulation, and/or process quality monitoring.

- 9/95-5/96
  Consultant, University of Delaware. Designs and analyzes projects involving experimental design, time series, spectral analysis, generalized linear models, linear and non-linear regression, sampling techniques, nonparametric techniques, and survival analysis.

- 7/95-8/95, 1/96-2/96
  Instructor, Department of Mathematical Sciences. University of Delaware, Newark, Delaware. Teaches introductory statistics course. Plans and presents lectures concerning elementary statistical concepts. Develops and grades exams, quizzes and homework assignments.

- 2/95-5/95
  Teaching Assistant, Department of Mathematical Sciences, University of Delaware, Newark, Delaware. Conducts lab sessions for introductory statistics course using Systat statistical software. Designs and grades lab reports, and administers tests.

- 4/92-1/95
  Electro-Mechanical Technician, Kalow Technologies, North Clarendon, Vermont. Troubleshoots and repairs a wide variety of microprocessor controlled machines. Modifies existing circuits and mechanical devices to comply with customer specifications. Researches and initiates acquisition of materials.

- 6/89-4/92
  Machinist / Toolmaker, General Electric Company, Aircraft Engines Business Group, Rutland , Vermont. Designs, modifies, and manufactures tooling and related gages used in the production of precision airfoils. Monitors part quality and initiates appropriate corrective actions to maintain close, interrelated tolerances. Completed extensive

apprenticeship program which incorporated engineering and design courses along with an understanding of numerous manufacturing processes. Increased the accuracy and efficiency of several processes at substantial savings to the company.

- 4/85-4/89
  Precision Measurement Equipment Specialist, United States Air Force. Isolates malfunctions, repairs and calibrates test, measurement, and diagnostic equipment. Certified to perform calibrations traceable to the National Institute of Standards and Technology. Conducts extremely precise measurements of resistance, inductance, capacitance, voltage and current. Repairs oscilloscopes, function and signal generators, frequency synthesizers, distortion analyzers, spectrum analyzers, thermal transfer standards, power indication devices, multimeters, torque wrenches, along with force, mass, pressure, and linear displacement standards.

Honors

- Klaus Hinkelmann Outstanding Graduate Assistant Award, 2000
- Publication referee for *Journal of the American Statistical Association*
- Publication referee for *Journal of Quality Technology.*
- President, Virginia Alpha Chapter of Mu Sigma Rho, 2000
- Vice President, Virginia Alpha Chapter of Mu Sigma Rho, 1999
- Awarded full tuition waiver and teaching assistantship, Virginia Tech, 1997–2000
- Awarded full tuition waiver and teaching assistantship, University of Delaware, 1995–1997
- Graduated summa cum laude, Castleton State College, 1994
- Dean's List: Fall 1992, Spring 1993, Fall 1993, Castleton State College
- President's List: Spring 1994, Castleton State College
- Academic Excellence Award in Mathematics 1992-1993, Castleton State College.