Understanding Multicollinearity in Bayesian Model Averaging with BIC Approximation

by

Ran Wang

B.Sc., University of British Colombia, 2016

Project Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

in the

Department of Statistics and Actuarial Science

Faculty of Science

© Ran Wang 2018

SIMON FRASER UNIVERSITY

Spring 2018

Copyright in this work rests with the author. Please ensure that any reproduction or re-use is done in accordance with the relevant national copyright legislation.

Approval

Name:	**Ran Wang**
Degree:	**Master of Science (Statistics)**
Title:	**Understanding Multicollinearity in Bayesian Model Averaging with BIC Approximation**
Examining Committee:	**Chair:** Jinko Graham Professor
	Thomas M. Loughin Senior Supervisor Professor
	Lawrence McCandless Supervisor Associate Professor
	Richard Lockhart Examiner Professor
Date Defended:	**April 23, 2018**

Abstract

Bayesian model averaging (BMA) is a widely used method for model and variable selection. In particular, BMA with Bayesian Information Criterion (BIC) approximation is a frequentist view of model averaging which saves a massive amount of computation compared to the fully Bayesian approach. However, BMA with BIC approximation may give misleading results in linear regression models when multicollinearity is present. In this article, we explore the relationship between performance of BMA with BIC approximation and the true regression parameters and correlations among explanatory variables. Specifically, we derive approximate formulae in the context of a known regression model to predict the BMA behaviours from 3 aspects — model selection, variable importance and coefficient estimation. We use simulations to verify the accuracy of the approximations. Through mathematical analysis, we demonstrate that BMA may not identify the correct model as the highest probability model if the coefficient and correlation parameters combine to minimize the residual sum of squares of a wrong model. We find that if the regression parameters of important variables are relatively large, BMA is generally successful in model and variable selection. On the other hand, if the regression parameters of important variables are relatively small, BMA can be dangerous in predicting the best model or important variables, especially when the full model correlation matrix is close to singular.

The simulation studies suggest that our formulae are over-optimistic in predicting posterior probabilities of the true models and important variables. However, these formulae still provide us insights about the effect of collinearity on BMA.

Keywords: All subsets regression, Simulation, Model selection, Variable importance, Expected residual sum of squares.

Acknowledgements

I would like to thank my senior supervisor Prof. Tom Loughin for his guidance and patience. Thanks for all his support, motivation, encouragement and humour for the past two years. I cannot achieve what I have done in my research without him.

Besides my advisor, I would also like to thank the rest of my thesis committee. Thank you for all the your insightful comments and advice.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

In statistical analysis, regression is almost always performed in settings where the "true model" is unknown. By true model we mean whatever probability distribution generates the observed data (Hocking, 1996). The problem is often approached by assuming a convenient model structure that approximates the true model, and for which parameter estimation and inference are relatively easily performed (Davison and Demetrio, 2002). It is typical that the model we choose is not the true model, but we hope it is a "good model". A good model should capture the important effects of explanatory variables on the response, while filtering out noise from the particular dataset (Burnham and Anderson, 2002). However, in reality, we often do not know which model to choose that gives the best approximation to the truth. Parameter estimation and subsequent inferences based on the chosen model are typically performed without acknowledging this uncertainty.

An alternative approach is to propose several candidate models as approximations to the truth and let data decide which one is the best. In particular, in multiple linear regression, a primary question is often to identify which variables are important to the model. Many methods have been developed to perform model selection and variable selection in regression settings. For example, classical approaches include stepwise regression, all subsets regression, selection based on information criteria, and others (Hocking, 1996).

These methods often lead to the selection of a single model as the "best". The parameter estimates and conclusions reached after such a process depend on treating the winning model as if it had been the only one considered. As a result, "the consequent uncertainty is not usually incorporated into the inference" (Davison and Demetrio, 2002). This may "lead to underestimation of uncertainty about quantities of interest and hence to overoptimistic and biased inferences" (Davison and Demetrio, 2002).

With concern regarding the cost of ignoring model uncertainty, Bayesian model averaging (BMA) was proposed by Raftery (1995). BMA provides a way to combine models and

give further inference. It incorporates model uncertainty into the parameter estimates and inferences by estimating a posterior probability that each considered model is the correct one, given the data. Then inferences can be based on posterior means (weighted averages) of quantities across models (Raftery, 1995). Conducting fully Bayesian analysis is challenging, but BMA can be used in a frequentist way using an easily computed quantity, the Bayesian Information Criterion (BIC), to provide an approximation to the Bayes factor for each model. Such approximation makes BMA computationally simple and fast. Moreover, if we apply BMA to all subsets regression, it performs model selection (selecting the "best" model) and variable selection (selecting the most important variables) simultaneously. In other words, we can easily obtain the posterior probability of each model, and from this each variable, and hence have a measure of variable importance (Raftery, 1995).

The large sample behaviour of BMA is trustworthy. Wasserman (2000) claimed that BMA is consistent for model selection, which means the posterior probability of the correct model or the model closest to the true model converges to one as the sample size increases. An interesting question is how accurate BMA is in small samples. In particular, it is important to understand how BMA reacts when the covariates are highly correlated. Multicollinearity often causes difficulty for estimating regression parameters (Hocking, 1996), and hence may also disturb BMA in multiple regression.

The goal of this paper is to gain a better understanding of the effect of multicollinearity on BMA. Specifically, we focus on a 3 variable problem with a known true model where all subsets regression is used for model and variable selection. We use a combination of mathematical analysis and simulations to analyze the behaviour of the BIC approximation to BMA. We identify situations where BMA provides misleading results in model selection, variable importance and coefficient estimation. In particular, our results corroborate those from Ghosh and Ghattas (2015), who studied a similar problem from a different perspective. We show BMA may fail to select the correct model when great multicollinearity exists. Also, the posterior probabilities for unimportant variables may be inflated and coefficient estimates may give signs and/or values that do not make sense. From this study, we "pave the way" for developing preliminary diagnostics of BMA applications.

The outline of the paper is as follows. Chapter 2 reviews the BMA structure and BIC approximation. Chapter 3 provides details on the mathematical analysis to assess the effect of collinearity on BMA. The results of mathematical analysis are given in Chapter 4. Chapter 5 describes the simulation study used to corroborate the results from Chapter 4. We draw conclusions and discuss possible further research in Chapter 6.

Chapter 2

Background and Literature Review

2.1 A Review of BMA Structure from a Bayesian Perspective

Bayesian model averaging can be studied from either Bayesian or frequentist perspectives. Bayesian estimation expresses model and variable uncertainty and views all unknown parameters as random variables (Raftery, 1995). The comprehensive Bayesian approach to model selection and estimation starts by considering that data are generated from one of K candidate models $\mathcal{M}_1$, $\mathcal{M}_2$, ..., $\mathcal{M}_K$. For now, we allow these models to be arbitrarily different. Later we will focus on linear regression models that differ only in the variables they contain. We denote the set of all possible models the model space by $\mathcal{M}$. Each $\mathcal{M}_k$ $(k = 1, ..., K)$ has its corresponding model parameter vector θ_k. We assign a prior probability $pr(\mathcal{M}_k)$ to each model as the probability that $\mathcal{M}_k$ is the correct model, and a prior probability distribution $pr(\theta_k|\mathcal{M}_k)$ to the parameters of each model. Then $pr(D|\theta_k, \mathcal{M}_k)$ is the likelihood of the data under $\mathcal{M}_k$ (Hoeting et al., 1999).

The posterior probability for a model $\mathcal{M}_k$ is

$$pr(\mathcal{M}_k|D) = \frac{pr(D|\mathcal{M}_k)pr(\mathcal{M}_k)}{\sum_{l=1}^{K} pr(D|\mathcal{M}_l)pr(\mathcal{M}_l)} \tag{2.1}$$

using Bayes' theorem, where

$$pr(D|\mathcal{M}_k) = \int pr(D|\theta_k, \mathcal{M}_k)pr(\theta_k|\mathcal{M}_k)d\theta_k \tag{2.2}$$

is the integrated likelihood of model $\mathcal{M}_k$.

The posterior probabilities in (2.1) can be used as a straightforward model selection criterion by selecting the most likely model (Fragoso and Neto, 2014). Such a model is often referred as the highest probability model (HPM) (Ghosh and Ghattas, 2015).

Model averaging refers to the process of estimating some quantity under each model and then averaging the estimates according to the posterior probability of each model. For example, if Δ is some quantity of interest, then the posterior distribution of Δ given data D is

$$pr(\Delta|D) = \sum_{k=1}^{K} pr(\Delta|\mathcal{M}_k, D)pr(\mathcal{M}_k|D). \tag{2.3}$$

The posterior mean and variance of Δ are:

$$E(\Delta|D) = \sum_{k=1}^{K} E(\Delta|D, \mathcal{M}_k)pr(\mathcal{M}_k|D)$$

and

$$Var(\Delta|D) = \sum_{k=1}^{K} [Var(\Delta|D, \mathcal{M}_k) + E^2(\Delta|D, \mathcal{M}_k)]pr(\mathcal{M}_k|D) - E^2(\Delta|D)$$

where $E(\Delta|D, \mathcal{M}_k)$ and $Var(\Delta|D, \mathcal{M}_k)$ are the posterior expectation and variance of Δ under model $\mathcal{M}_k$. See Raftery (1993) and Draper (1995).

The posterior mean of Δ is the average of the expectations of Δ under each model under consideration, weighted by the posterior model probabilities.

2.2 A Frequentist View of Bayesian Model Averaging

Even though constructing a BMA analysis seems intuitively easy, there are some difficulties in its implementation. First, the size of the space of interesting models can grow very large. For example, in linear regression with p explanatory variables, all subsets regression results in 2^p models. Thus, the exhaustive summation in (2.3) becomes impractical with large p. One approach to tackle this problem is to average over a subset of models supported by the data and discard models with small posterior probabilities. To achieve this, the Occam's window method by Madigan and Raftery (1994) averages over a subset of models selected by the ratio of their posterior probabilities to the highest posterior probability, which greatly reduces the number of models.

Second, the process of carrying out the fully Bayesian computation is complicated. Tierney and Kadane (1986), found a method that uses a Laplace transformation to approximate integral (2.2) to simplify integration for generalized linear models and some other model

classes. However, the specification of priors is still left as a challenge. While most researchers seem to agree that the uniform prior on $pr(\mathcal{M}_k)$ is reasonable, there remains some discussion on the choice of $pr(\theta_k|\mathcal{M}_k)$.

This lack of consensus regarding priors on the model parameters allowed Raftery (1986, 1995) to develop a different perspective of BMA using a different set of priors. Specifically, Raftery (1986, 1995) introduced the BIC approximation to the Bayes factor, which allows posterior probabilities to be calculated with minimal effort.

BIC is defined as

$$BIC = -2\log(\hat{L}) + q\log(n)$$

where $\hat{L}$ is the maximized likelihood and q is the number of parameters.

In the context of linear regression with independent and identically distributed (iid) normal errors that we are interested in, suppose that $\mathcal{M}_k$, $k = 1, ..., K$, are linear regression models based on different subsets of p explanatory variables $X_1, \ldots, X_p$. Then the parameters θ_k consist of $q = p + 2$ parameters including regression parameters $\{\beta_i : X_i \in \mathcal{M}_k\}$, the intercept and the model variance. Then BIC for each model simplifies to

$$BIC_k = n\log(\frac{RSS_k}{n}) + q\log(n) + C$$

where RSS_k is the residual sum of squares under $\mathcal{M}_k$ and C is a constant that does not depend on q, or k.

It is beyond the scope of this paper to discuss the details of the derivation of this BIC approximation (see Section 4 in Raftery (1995) for details). Roughly, if we assume a uniform prior on each model (ie, $pr(\mathcal{M}_k) = \frac{1}{K}$, $k = 1, .., K$) and Jeffreys' priors (Kass and Wasserman, 1996) with carefully chosen constants for θ_k, (ie, $pr(\theta_k|\mathcal{M}_k)$ is a multivariate normal prior with mean at the maximum likelihood estimate and variance equal to the expected information matrix for one observation)(Raftery, 1995), the posterior probability can be approximated by:

$$\hat{pr}(\mathcal{M}_k|D) \approx \hat{\pi}_k = \frac{\exp(-\frac{1}{2}\Delta BIC_k)}{\sum_{l=1}^{K}\exp(-\frac{1}{2}\Delta BIC_l)} \tag{2.4}$$

where $\Delta BIC_k = BIC_k - BIC_{min}$, with $BIC_{min} = \min_k BIC_k$. We can also replace ΔBIC_k with BIC_k in the above formula (Raftery, 1995). We use the hat symbol to emphasize quantities that are obtained from data.

When $\hat{\pi}_k$ replaces $pr(\mathcal{M}_k|D)$ in (2.3), we can easily obtain a weighted average of the estimates of the interesting quantity Δ from different models. For example, in all subsets regression with p covariates, if Δ_k is an indicator for the presence of variable X_i in a model, one can use (2.3) to calculate the posterior probability that each variable belongs in model, and hence rank variable importance. Mathematically, we have

$$\hat{\gamma}_i = \hat{P}(\beta_i \neq 0) = \sum_{k=1}^{K} \hat{\pi}_k \mathbb{1}(X_i \in \mathcal{M}_k) \tag{2.5}$$

$(i = 1, ..., p)$, where $\mathbb{1}(X_i \in \mathcal{M}_k)$ is the covariate indicator. Barbieri and Berger (2004) refer γ_i as the posterior inclusion probability for variable X_i. Higher γ_i indicates higher importance of variable X_i.

On the other hand, when Δ_k represents a coefficient β_i that is common to all models, one can use (2.3) to obtain the posterior mean for coefficient estimation. The estimated posterior mean for β_i is

$$\hat{\beta}_i = \sum_{k=1}^{K} \hat{\pi}_k \hat{\beta}_{i,k}$$

where $\hat{\beta}_{i,k}$ is the estimate of β_i in model $\mathcal{M}_k$.

2.3 Properties of the BIC Approximation to BMA

When people use the BIC approximation instead of the fully Bayesian approach, they are often unaware that there are some fundamental differences between these two. In particular, the performance of BMA greatly depends on the choice of priors (both the priors on regression hyperparameters and model probability priors) (Fernandez et al., 1998). One of the desired properties of any model- or variable-selection tool is consistency. In BMA, assuming that the correct model is in our model space, we would like to have the posterior probability of the correct model to converge to 1 as the sample size increases (Fernandez et al.1998). Fernandez et al. (1998) theoretically determined that some prior combinations would lead to consistency, but some other cases failed. When applying the BIC approximation to BMA, Wasserman (2000) claimed that under weak conditions, the approximation achieves the desired consistency. In addition, under regularity conditions, the posterior probability of the model that contains the closest approximation to the true distribution tends to 1 if the true model is not in the model space considered (Wasserman, 2000). Wasserman (2000) also

pointed out that when the true model is in two or more nested models, BMA favours the parsimonious model and convergence happens faster if candidate models are not nested.

Even though BIC approximation to BMA is consistent, various critiques arose to question the validity of BIC approximation under finite samples. In particular, Weakliem (1999) argued that the Jeffreys' prior is too spread out, which causes the BIC approximation to be overly conservative. In response to Weakliem's doubt, Raftery (1999) admitted that "BIC is likely to be conservative relative to Bayes factors based on informative priors". It follows that, in most cases, "if BIC finds evidence for an effect, we should agree that data support the effect and not necessarily conversely". Wasserman (2000) also claimed that BIC approximation seems to work well in well-behaved problems with moderate to large sample sizes, but can break down in irregular cases.

Although it is known that the BIC approximation has some defects, Fragoso and Neto's (2015) review on BMA found that most articles use the BIC approximation since the BIC values can be easily obtained from most software. The use of BIC approximation eliminates many computational difficulties with the Markov chain Monte Carlo process that are encountered in the fully Bayesian approaches. It makes BMA analysis straightforward and immediately available from maximum likelihood estimates (Fragoso and Neto, 2015). Such significant computational savings often outweigh the defects of BIC approximation.

2.4 BMA and All Subsets Regression under Collinearity

All subsets regression is a common technique for model and variable selection in linear regression and related techniques. It estimates models for all possible combinations of explanatory variables and determines which set performs the best according to some predetermined criteria such as Akaike's information criterion (AIC) or BIC (Elliot et al., 2016). Similar to all other regression problems, collinearity is among the top concerns by researchers working with this method. When collineaity exists, the common criteria in all subsets regression can have very low success rate in identifying the correct model (Becker et al., 2014). Even when the correct model is chosen, collinearity can inflate the variance of ordinary least squares parameter estimates and yield incorrect sign and magnitude of the estimates (Hocking, 1996).

As we mentioned earlier, BMA is more reliable for model and variable selection than selecting a single model in all subsets regression since it incorporates selection uncertainty. Unfortunately, it is known that BMA with regression models can inherit problems with collinearity. Ghattas and Ghosh (2015) demonstrated via real data analysis and simulation studies that some priors may be more adversely affected under strong collinearity than others. They studied 4 different priors for regression coefficients and found that some priors led

to markedly better predictive performance than others. The HPM can also be different if different priors are chosen. Second, the prediction performance of the HPM and the median probability model (MPM) (Barbieri and Berger, 2004) can be greatly affected by collinearity. The MPM is defined to be the model consisting of those variables whose γ_i values are at least 0.5. Barbieri and Berger (2004) claimed that "the MPM considerably outperforms the HPM in terms of predictive performance". In contrast, Ghattas and Ghosh (2015) showed that under collinearity the HPM could provide better prediction than MPM.

We are unaware of any research done explicitly on the effect of collinearity when BMA is applied with BIC approximation. We aim to address this gap.

Chapter 3

Mathematical Analysis of BMA on All Subsets Linear Regression Model

This chapter describes the mathematical analysis we did to study the collinearity effect on BMA applications. The definition of collinearity effect largely depends on the goal of the study. In this paper, we focus on its effect from three aspects: model selection, variable importance and coefficient estimation. These aspects correspond to the most popular usages of BMA methods (Fragoso and Neto, 2015). In particular, we looked into the 3 variable all subsets regression model with various coefficients and correlation structures. We first conducted a mathematical approximation of the entire BMA process and identified cases when BMA gives wrong or misleading results in expectation. All cases were then tested via simulation studies described in Chapter 5. We hope the mathematical analysis can explain the simulation results, and hence offer insight for when BMA can be safely applied on regression models.

3.1 Model Setups

To fix ideas, suppose there are three covariates and the true linear model that generates data is

$$\boldsymbol{Y} = \beta_1 \boldsymbol{X_1} + \beta_2 \boldsymbol{X_2} + \boldsymbol{\epsilon} \tag{3.1}$$

where $\boldsymbol{Y} = (y_1, ..., y_n)'$ is the vector of responses. Also, $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ are n-vectors of values from covariates X_1 and X_2. A third variable, X_3 is measured but is not related to the response. As usual, the random errors $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I)$ are assumed to be independent. The

variance of ϵ_i $(i = 1, ..., n)$ is σ^2 and is fixed at 1. The variance term is not our primary interest, and our analysis can be easily generalized to cases where ϵ_i has variance other than 1. We also fixed sample size $n = 100$ throughout the analysis.

Let $\boldsymbol{X} = (\boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3})'$. Without loss of generality, assume all covariates are standardized. This significantly simplifies the subsequent calculations because $\boldsymbol{X'X}$ can be expressed in terms of correlation matrix.

$$\boldsymbol{X'X} = nR = n \begin{bmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{13} & r_{23} & 1 \end{bmatrix}$$

where r_{ij} is the sample correlation between X_i and X_j $(i, j = 1, 2, 3)$. We will show that we can approximate the entire BMA/BIC process using functions of only the parameters, β_1, β_2, r_{12}, r_{13} and r_{23}. We will investigate the BMA behaviour under different parameter combination settings.

It is important to know the constraints on the correlation matrix of 3 variables. For each particular pair of r_{ij} and r_{jl}, Cholesky-decomposition of correlation matrix shows r_{il} is bounded by

$$r_{ij}r_{jl} \pm \sqrt{(1 - r_{ij}^2)(1 - r_{jl}^2)}$$

$(i, j, l = 1, 2, 3)$. The correlation matrix is singular when r_{il} is at the boundaries.

All subsets regression takes $2^3 = 8$ possible models $\mathcal{M}_k$ $(k = 1, ..., 8)$ on the 3 variables. The 8 models are:

- $\mathcal{M}_1$: $\boldsymbol{Y} = \boldsymbol{\epsilon}$
- $\mathcal{M}_2$: $\boldsymbol{Y} = \beta_3\boldsymbol{X_3} + \boldsymbol{\epsilon}$
- $\mathcal{M}_3$: $\boldsymbol{Y} = \beta_2\boldsymbol{X_2} + \boldsymbol{\epsilon}$
- $\mathcal{M}_4$: $\boldsymbol{Y} = \beta_1\boldsymbol{X_1} + \boldsymbol{\epsilon}$
- $\mathcal{M}_5$: $\boldsymbol{Y} = \beta_2\boldsymbol{X_2} + \beta_3\boldsymbol{X_3} + \boldsymbol{\epsilon}$
- $\mathcal{M}_6$: $\boldsymbol{Y} = \beta_1\boldsymbol{X_1} + \beta_3\boldsymbol{X_3} + \boldsymbol{\epsilon}$
- $\mathcal{M}_7$: $\boldsymbol{Y} = \beta_1\boldsymbol{X_1} + \beta_2\boldsymbol{X_2} + \boldsymbol{\epsilon}$
- $\mathcal{M}_8$: $\boldsymbol{Y} = \beta_1\boldsymbol{X_1} + \beta_2\boldsymbol{X_2} + \beta_3\boldsymbol{X_3} + \boldsymbol{\epsilon}$

$\mathcal{M}_7$ is the true model. These 8 models form the model space that we will later use in BMA applications.

3.2 Expectation of Model Posterior Probability

All analysis for model selection, variable importance and coefficient estimation depends on the behaviours of model posterior probabilities (2.4), whereas model posterior probabilities are functions of BIC for each model and hence the corresponding residual sum of squares (RSS) in regression. We hope to understand the posterior probabilities by looking at their expectations.

For each model, the RSS is a quadratic form $\boldsymbol{Y'(I-H_k)Y}$, where $\boldsymbol{H_k = X_k(X_k'X_k)^{-1}X_k'}$ is the hat matrix of the fitted model $\mathcal{M}_k$ and $\boldsymbol{X_k}$ is the model covariate matrix. We derive the expectation of RSS:

$$E(RSS_k|\mathcal{M}_k) = \sigma^2(n-p) + \boldsymbol{(X_o\beta_o)'(I-H_k)(X_o\beta_o)}$$

where p is the number of nonzero β coefficients in the regression model. We denote $\boldsymbol{\beta_o} = (\beta_1, \beta_2)'$ and $\boldsymbol{X_o} = (\boldsymbol{X_1}, \boldsymbol{X_2})'$ as the regression coefficients and covariance of the true model.

Table 3.1 shows all $E(RSS)$ for all 8 models. The 0's and 1's in columns 2–4 are indicators for whether each covariate appears in the model.

Table 3.1: Expectation of RSS

Models	X_1	X_2	X_3	$E(RSS)$
$\mathcal{M}_1$	0	0	0	$n\beta_1^2 + 2nr_{12}\beta_1\beta_2 + n\beta_2^2 + \sigma^2 n$
$\mathcal{M}_2$	0	0	1	$n(1-r_{13}^2)\beta_1^2 + n(1-r_{23}^2)\beta_2^2 + 2n(r_{12}-r_{13}r_{23})\beta_1\beta_2 + \sigma^2(n-1)$
$\mathcal{M}_3$	0	1	0	$n\beta_1^2(1-r_{12}^2) + \sigma^2(n-1)$
$\mathcal{M}_4$	1	0	0	$n\beta_2^2(1-r_{12}^2) + \sigma^2(n-1)$
$\mathcal{M}_5$	0	1	1	$n\beta_1^2(1-\frac{r_{12}^2+r_{13}^2-2r_{12}r_{13}r_{23}}{1-r_{23}^2}) + \sigma^2(n-2)$
$\mathcal{M}_6$	1	0	1	$n\beta_2^2(1-\frac{r_{12}^2+r_{23}^2-2r_{12}r_{13}r_{23}}{1-r_{13}^2}) + \sigma^2(n-2)$
$\mathcal{M}_7$	1	1	0	$\sigma^2(n-2)$
$\mathcal{M}_8$	1	1	1	$\sigma^2(n-3)$

As shown in Table 3.1, the $E(RSS_k)$ for $\mathcal{M}_k$ is in the form of $f_k(\beta_i, r_{ij}, n) + \sigma^2(n-p)$, where $f_k(\beta_i, r_{ij}, n)$ is a function of coefficient parameters, variable correlations and sample size. The function $f_k(\beta_i, r_{ij}, n)$ is never negative and can be viewed as the bias in RSS. The bias is always 0 if the model we are fitting is or contains the true model. However, as we will see in Chapter 4, some combinations of β_i and r_{ij} values will result in very small bias in

$E(RSS_k)$ for some $k \neq 7$ and promote $\mathcal{M}_k$ to appear better than $\mathcal{M}_7$ even when it should not.

Similarly, the variance of the residual sum of squares for each model is

$$Var(RSS_k|\mathcal{M}_k) = 2\sigma^4(n-p) + 4\sigma^2(\boldsymbol{X_o\beta_o})'(\boldsymbol{I - H_k})(\boldsymbol{X_o\beta_o})$$

From there, we can further get the expectation of BIC by applying second order Taylor expansion (we drop the hat of $\hat{\pi}$ for easier notation).

$$\begin{aligned} E(BIC_k) &= E(n\log(\frac{RSS}{n}) + p\log(n)|\mathcal{M}_k) \\ &\approx n\log(\frac{E(RSS_k)}{n}) - \frac{n}{2}\frac{Var(RSS_k)}{E^2(RSS_k)} + p\log(n) \end{aligned}$$

Note that we express the penalty as $p\log(n)$ rather than $q\log(n)$, because $q = p + 2$ for all models. The additional $2\log(n)$ is common to all models and irrelevant to model comparisons.

The final step is to obtain the expectation of the model posterior probabilities, which is more complex. The exact expectation of posterior probability can be obtained by viewing (2.4) as a softmax mapping whose inputs are BIC_k. Assume $\boldsymbol{BIC}$ is a real valued vector indexed by k, and $-\frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{BIC}$ has mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Note that $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are greatly determined by the β_1, β_2, r_{12}, r_{13} and r_{23}. Daunizeau (2017) derived the approximation of the expected log-softmax based upon a second order Taylor expansion:

$$E(\log\pi_k(\boldsymbol{BIC_k})) \approx \log\pi_k(\boldsymbol{\mu}) + \frac{1}{2}tr[(\pi(\boldsymbol{\mu})\pi(\boldsymbol{\mu})^T - Diag(\pi(\boldsymbol{\mu})))\boldsymbol{\Sigma}] \tag{3.2}$$

where $\pi(\boldsymbol{x})$ is a vector whose entries are the softmax functions $\pi_k(x)$.

In the above approximation, the second term on the right hand side is independent of k. Therefore, under the same parameter and correlation setting,

$$E(\log\pi_k(\boldsymbol{BIC_k})) \approx \log\pi_k(\boldsymbol{\mu}) + C_1$$

for some constant C_1 across all 8 models.

Daunizeau (2017) claimed that the approximation does not ensure a proper normalization, i.e. $1 = \sum_k \exp E(\log\pi_k(\boldsymbol{BIC_k}))$ may not be satisfied.

Exponentiating both size of the previous equation, we have

$$\exp E(\log \pi_k(\boldsymbol{BIC_k})) \approx C_2 \pi_k(\boldsymbol{\mu})$$

For some constant C_2 independent of k. However, by Jensen's inequality,

$$\exp E(\log \pi_k(\boldsymbol{BIC_k})) \leq E(\pi_k(\boldsymbol{BIC_k}))$$

In order to fully obtain $E(\hat{\pi}_k)$, we need to work out the explicit expression of (3.2) which might be overly complicated for our purposes.

There are several approximations to the expectation of posterior probability, each with its own drawbacks and imperfections. For analytical simplicity, we propose $\tilde{E}(\hat{\pi}_k)$ as an estimate of $E(\hat{\pi}_k)$.

$$\tilde{E}(\hat{\pi}_k) = \frac{\exp(-\frac{1}{2}E(BIC_k))}{\sum_{l=1}^{8} \exp(-\frac{1}{2}E(BIC_l))} \tag{3.3}$$

Even though $\tilde{E}(\hat{\pi}_k)$ is not the same as $E(\hat{\pi}_k)$, we will study its features using simulations and we hope it mimics $E(\hat{\pi}_k)$ reasonably well. If so, $\tilde{E}(\hat{\pi}_k)$ would allow us to explain the BMA process with less calculation.

3.3 A Study on (3.3)

Before we use $\tilde{E}(\hat{\pi}_k)$ to identify combinations of the parameters β_1, β_2 and r_{12}, r_{13}, r_{23}, we want to evaluate its accuracy. To do this, we evaluated $\tilde{E}(\hat{\pi}_k)$, $k = 1, ..., 8$, for 1,024 combinations of the parameters. We then simulated data from these same combinations and estimated each $\hat{\pi}_k$ empirically for comparison.

The parameters were chosen on a five dimensional grid to cover a wide range of parameters.

- $\beta_1 = -1, -0.5, 0.5, 1$
- $\beta_2 = -1, -0.5, 0.5, 1$
- $r_{12} = -0.9, -0.5, 0.5, 0.9$
- $r_{13} = -0.9, -0.5, 0.5, 0.9$
- For each pair of r_{12} and r_{13}. We divide the possible range of r_{23} into 16 equal length intervals and take r_{23} equal to the right point of either the 1st, 5th, 11th or 15th interval.

Then we additionally simulated data from each of these combinations to obtain an unbiased empirical estimate of model posterior probabilities. We fixed sample size at $n = 100$ and $\sigma^2 = 1$. We generated the explanatory variables $\boldsymbol{X_i}$ $(i = 1, 2, 3)$ from the multivariate normal density N(0,R), such that $cor(X_i, X_j) = r_{ij}$. The response variable $\boldsymbol{Y}$ was generated from the true model (3.1) and the procedure was repeated 100 times to generate 100 datasets. For each of the 100 runs, we fitted all 8 models. and obtained the estimates $\hat{\pi}_{k,c}$ from the model BICs as in (2.4) for each parameter combination c $(c = 1, \ldots, 1024)$.

We calculated $\tilde{E}(\hat{\pi}_{k,c})$ according to (3.3) for each k. We also calculated $\bar{\pi}_{k,c}$ as the mean of $\hat{\pi}_{k,c}$ over 100 runs. We then compared $\bar{\pi}_{k,c}$ with $\tilde{E}(\hat{\pi}_{k,c})$ for each k.

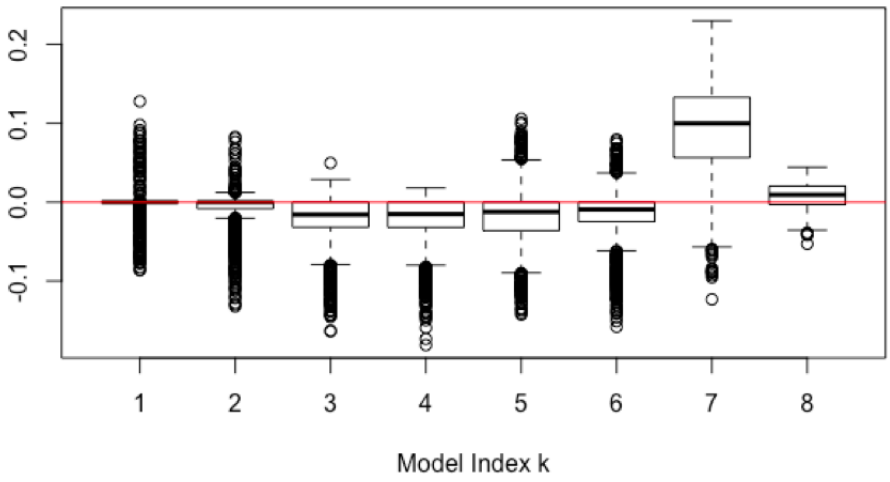


Figure 3.1: Box plots of the predicted model posterior probability minus the simulated mean of model posterior probability for all models.

The results are moderately satisfying. Figure 3.1 shows the box plots of $\tilde{E}(\hat{\pi}_{k,c}) - \bar{\pi}_{k,c}$, $k = 1, ..., 8$. For all the combinations we tested, $\bar{\pi}_{k,c}$ and $\tilde{E}(\hat{\pi}_{k,c})$ are very close for all models except the true model $\mathcal{M}_7$. We found $\tilde{E}(\hat{\pi}_k)$ tends to underestimate the posterior probabilities of $\mathcal{M}_3$ through $\mathcal{M}_6$, and overestimate the posterior probabilities of $\mathcal{M}_8$ slightly. The overestimation on $\mathcal{M}_7$ is more severe. $\mathcal{M}_3$ and $\mathcal{M}_4$ are the one-variable models which contain only one important variable. $\mathcal{M}_7$ and $\mathcal{M}_8$ are the models whose $E(RSS)$ have no bias term (see Table 3.1). For some parameter combinations, $\tilde{E}(\pi_7)$ and $\bar{\pi}_7$ can differ by at most 0.2, whereas they differ by at most 0.13 for other k. The two quantities are mostly consistent with each other in terms of model selection. For 85% of the times, $\tilde{E}(\hat{\pi}_{k,c})$ agrees with $\bar{\pi}_{k,c}$ in that they both assign the largest probability to the same model.

3.4 Computations for Variable Importance and Parameter Estimation

From the previous simulation, we see $\tilde{E}(\hat{\pi}_k)$ and $\bar{\pi}_k$ mostly agree on selecting the HPM. With this feature in mind, we will continue to use $\tilde{E}(\hat{\pi}_k)$ as a substitute for $E(\hat{\pi}_k)$ due to its simple structure.

In this section, we will consider variable importance and coefficient estimation as one group since both of them are applications of formula (2.3).

To assess variable importance, we get the expectation of $\hat{\gamma}_i$ in (2.5).

$$E(\hat{\gamma}_i) = E(\sum_{k=1}^{K} \hat{\pi}_k \mathbb{1}(X_i \in \mathcal{M}_k)) = \sum_{k=1}^{K} E(\hat{\pi}_k)\mathbb{1}(X_i \in \mathcal{M}_k)$$

Since $\tilde{E}(\hat{\pi}_k)$ is our estimate of $E(\hat{\pi}_k)$, we define

$$\tilde{E}(\hat{\gamma}_i) = \sum_{k=1}^{K} \tilde{E}(\hat{\pi}_k)\mathbb{1}(X_i \in \mathcal{M}_k) \tag{3.4}$$

as an estimate of $E(\hat{\gamma}_i)$.

A full analytic approach to understand the effect of BMA on coefficient estimation is again challenging. We first observe that, due to the structure of $\boldsymbol{X_k}$, the expectations of ordinary least squares (OLS) estimates $\hat{\boldsymbol{\beta}}_k$ under each specific model are easy to compute.

Table 3.2 shows the expectation of OLS estimates $E(\hat{\boldsymbol{\beta}}_k|\mathcal{M}_k) = E((\boldsymbol{X_k'X_k})^{-1}\boldsymbol{X_k'Y}|\mathcal{M}_k)$ under each model.

However,

$$E(\hat{\beta}_i) = E(\sum_{k=1}^{K} \hat{\pi}_k \hat{\beta}_{i,k})$$

$(i = 1, 2, 3)$. $\hat{\pi}_k$ and $\hat{\beta}_{i,k}$ may not be independent. Thus, the joint distribution of $\hat{\pi}_k$ and $\hat{\beta}_{i,k}$ is needed in order to compute the expectation of coefficient estimates exactly. We have no formal estimate of this joint distribution. Instead, to achieve easy calculation and structure, we use

$$\tilde{E}(\hat{\beta}_i) = \sum_{k=1}^{K} \tilde{E}(\hat{\pi}_k)E(\hat{\beta}_{i,k}) \tag{3.5}$$

to approximate the $E(\hat{\beta}_i)$.

Table 3.2: Expectation of Model Parameters

Models	X_1	X_2	X_3	$E(\hat{\beta}_1)$	$E(\hat{\beta}_2)$	$E(\hat{\beta}_3)$
$\mathcal{M}_1$	0	0	0	0	0	0
$\mathcal{M}_2$	0	0	1	0	0	$r_{13}\beta_1 + r_{23}\beta_2$
$\mathcal{M}_3$	0	1	0	0	$r_{12}\beta_1 + \beta_2$	0
$\mathcal{M}_4$	1	0	0	$\beta_1 + r_{12}\beta_2$	0	0
$\mathcal{M}_5$	0	1	1	0	$\beta_2 + \frac{r_{12}-r_{23}r_{13}}{1-r_{23}^2}\beta_1$	$\frac{r_{13}-r_{12}r_{23}}{1-r_{23}^2}\beta_1$
$\mathcal{M}_6$	1	0	1	$\beta_1 + \frac{r_{12}-r_{23}r_{13}}{1-r_{23}^2}\beta_2$	0	$\frac{r_{23}-r_{12}r_{13}}{1-r_{13}^2}\beta_2$
$\mathcal{M}_7$	1	1	0	β_1	β_2	0
$\mathcal{M}_8$	1	1	1	β_1	β_2	0

3.5 Methods for Main Study

Using the mathematical derivations developed in the previous two sections, we now want to broadly explore how BMA works across a wide range of parameter combinations. We analyze the BMA behaviour in model selection, variable importance and coefficient estimation, respectively. To do so, we test 96,800 combinations of β_1, β_2, r_{12}, r_{13}, and r_{23} on a five dimensional grid, maintaining $n = 100$ and σ^2=1. We chose the parameter combinations as follows:

1. β_1 and β_2: ±1, ±0.8, ±0.6, ±0.4, ±0.2.
2. r_{12} and r_{13}: ±0.9, ±0.7, ±0.5, ±0.3, ±0.1, 0.
3. For each pair of r_{12} and r_{13}, we chose 8 points for r_{23} by taking the midpoints of 8 equal intervals of its range.

The choice of parameter combinations is an extension of the parameters used in the simulation study in Section 4.3.

We calculate $\tilde{E}(\hat{\pi}_{k,c})$ $(k = 1, .., 8,\ c = 1, \ldots, 96800)$, $\tilde{E}(\hat{\gamma}_{i,c})$ and $\tilde{E}(\hat{\beta}_{i,c})$ $(i = 1, 2, 3)$ according to equations (3.3), (3.4) and (3.5). We wish to identify dangerous cases of BMA in model selection and variable importance. By dangerous cases, we mean the cases where the formulas predict that BMA may give wrong or misleading answers on average.

We say that BMA is successful for model selection if the true model is chosen as the HPM according to $\tilde{E}(\hat{\pi}_k)$. Otherwise it is dangerous to use for model selection. When it is dangerous, we use Table 3.1 and other mathematical results to explain why the correct

model is not chosen. To better understand BMA's performance on model selection, we define the Margin m as

$$m(\tilde{E}(\hat{\boldsymbol{\pi}})) = \tilde{E}(\hat{\pi}_7) - \max\{\tilde{E}(\hat{\pi}_1), \tilde{E}(\hat{\pi}_2), ..., \tilde{E}(\hat{\pi}_6), \tilde{E}(\hat{\pi}_8)\}$$

where $\tilde{E}(\hat{\boldsymbol{\pi}})$ is a vector whose entries are $\tilde{E}(\hat{\pi}_k)$.

The Margin measures the minimum amount of posterior probability by which the true model is correctly classified. Higher positive Margin means the posterior probability is more accumulated at the true model and the true model is more clearly distinguished from the rest. A negative Margin indicates a dangerous case for model selection.

Regarding variable importance, we follow Barbieri and Berger (2004) that a variable is identified as important if $\gamma_i > 0.5$. Then we categorize a successful case c of variable importance if $\tilde{E}(\hat{\gamma}_{1,c}) > 0.5$, $\tilde{E}(\hat{\gamma}_{2,c}) > 0.5$ and $\tilde{E}(\hat{\gamma}_{3,c}) < 0.5$ are satisfied simultaneously. Otherwise, the case is classified as dangerous for variable importance.

We will classify all the 96,800 cases into 4 groups according to their BMA performance in the math analysis: successful in model selection and variable importance (SS), successful in model selection and dangerous in variable importance (SD), dangerous in model selection and successful in variable importance (DS), dangerous in model selection and dangerous variable importance (DD). Within each group, we study coefficient estimation. If $\tilde{E}(\hat{\beta}_{3,c})$ is largely different from 0 or $\tilde{E}(\hat{\beta}_{i,c})$ $(i = 1, 2)$ is largely different from its corresponding parameter β_i, BMA is less satisfying and gives biased coefficient estimates. We define the Approximate Bias D for each parameter as:

$$D(\tilde{E}(\hat{\beta}_i)) = \tilde{E}(\hat{\beta}_i) - \beta_i$$

where β_3 is taken as 0. Approximate Bias measures how biased our coefficient estimates are.

All the above analyses are based on the mathematical derivations and approximations in Chapter 3. It can be viewed as our prediction of BMA performance for each parameter combination setting. We conduct simulation studies in Chapter 5 to verify how well these formulae predict BMA results.

Chapter 4

Results of Math Analysis

4.1 Observations of BMA in Model Selection

We summarize the model selection results of the 96,800 parameter combinations in Table 4.1.

Table 4.1: Frequency and Percentage of each Model being Selected as HPM

Models	X_1	X_2	X_3	**Frequency as HPM**	**Percentage as HPM**
$\mathcal{M}_1$	0	0	0	2,320	2.40%
$\mathcal{M}_2$	0	0	1	1,008	1.04%
$\mathcal{M}_3$	0	1	0	10,488	10.80%
$\mathcal{M}_4$	1	0	0	10,401	10.70%
$\mathcal{M}_5$	0	1	1	0	0%
$\mathcal{M}_6$	1	0	1	0	0%
$\mathcal{M}_7$	1	1	0	71,240	73.60%
$\mathcal{M}_8$	1	1	1	0	0%

We see that the true model $\mathcal{M}_7$ is correctly chosen as the expected HPM for only 73.60% of the parameter combinations. Also, $\mathcal{M}_3$ and $\mathcal{M}_4$ are tied as the HPM for 1,343 cases (1.39%). This happens when $\beta_1^2 = \beta_2^2$ and the two models have the same expected RSS. $\mathcal{M}_5$ and $\mathcal{M}_6$ never have the highest expected posterior probability, and never outperform $\mathcal{M}_7$. The same happens for $\mathcal{M}_8$. We will look into the details why $\mathcal{M}_1$ to $\mathcal{M}_4$ can have higher expected posterior probabilities.

It should be obvious from (2.4) that higher posterior probability corresponds to lower $E(BIC)$. And $E(BIC)$ can be approximately expressed in terms of $E(RSS)$ as the following:

$$E(BIC) \approx n\log(E(RSS)) - n\log(n) - \frac{2\sigma^2 n}{E(RSS)} + \frac{n\sigma^4(n-p)}{E^2(RSS)} + p\log(n)$$

That is, in our case, smaller $E(RSS)$ guarantees smaller $E(BIC)$ for a fixed p. For comparing models with different p, the smaller model is preferred as long as its $E(RSS)$ is not too much larger than that from the larger model. We can explain the observed cases from the last chapter by comparing the $E(RSS)$ formulae derived in Table 3.1.

Whenever there is very little signal from the explanatory variables in the data, BMA tends to select the most parsimonious model, $\mathcal{M}_1$. This happens when both β_1 and β_2 are close to 0, which makes the bias term in $E(RSS_1)$ close to 0.

The "most false" model is $\mathcal{M}_2$, since it contains one spare variable and neither of the important variables. This model beats models $\mathcal{M}_5$, $\mathcal{M}_6$, and $\mathcal{M}_8$ whenever β_2, β_1, or both, respectively, are very small, and/or when the magnitudes of the correlations between X_3 and X_1, X_2, or both, respectively, are very large. In these cases, the smaller penalty for $\mathcal{M}_2$ overcomes the tiny increase in bias in its $E(RSS)$. However, among one variables models, it is rather surprising to have $\mathcal{M}_2$ outperform $\mathcal{M}_3$ and $\mathcal{M}_4$.

Mathematically, we can see from Table 3.1, $E(RSS_2) < E(RSS_3)$ whenever

$$n(1-r_{13}^2)\beta_1^2 + n(1-r_{23}^2)\beta_2^2 + 2n(r_{12} - r_{13}r_{23})\beta_1\beta_2 < n\beta_1^2(1-r_{12}^2)$$

similarly for $\mathcal{M}_4$.

In other words, due to "unlucky" combinations of coefficient and correlation structure, $\mathcal{M}_2$ outperforms $\mathcal{M}_3$ when

$$|r_{12}\beta_1 + \beta_2| < |r_{13}\beta_1 + r_{23}\beta_2|$$

and $\mathcal{M}_2$ outperforms $\mathcal{M}_4$ when

$$|\beta_1 + r_{12}\beta_2| < |r_{13}\beta_1 + r_{23}\beta_2|$$

Thus, the strong correlation between the spare variable X_3 and an important variable, plus the weak correlation between the two important variables will deceive BMA to select $\mathcal{M}_2$ rather than $\mathcal{M}_3$ or $\mathcal{M}_4$.

Another unpleasant result that can be more misleading is when $\mathcal{M}_2$ outperforms $\mathcal{M}_7$. This happens when $E(BIC_7) > E(BIC_2)$ and hence a small bias in $E(RSS_2)$. If we look at the bias term $f_2(\beta_i, r_{ij}, n)$ from Table 3.1, in order for the bias to be small, we need low values of the coefficient parameters and/or large correlation between X_3 and each of X_1 and X_2, and as small a value of r_{12} as possible.

Thus, if $f_2(\beta_i, r_{ij}, n)$ is small enough such that $E(RSS_2) = \delta E(RSS_7)$ for some $\delta > 1$ and $\delta \approx 1$, we can show $E(BIC_7) > E(BIC_2)$ for sample size not too large.

We have

$$E(BIC_7) - E(BIC_2) = (\frac{1}{\delta} - 1)\frac{2n\sigma^2}{E(RSS_7)} + (1 - \frac{1}{\delta^2})\frac{n(n-1)\sigma^4}{E^2(RSS_7)} - \frac{n\sigma^4}{E^2(RSS_7)} + \log(n) - n\log(\delta) \tag{4.1}$$

Recall $E(RSS_7) = n - 2$ and $\sigma^2 = 1$. The Equation (4.1) shows, if $\delta \approx 1$ and n is not too large, the first two terms are close to 0 and $\frac{n}{E^2(RSS_7)} = O(\frac{1}{n})$. When $\log(n) > n\log(\delta) - \frac{n}{E^2(RSS_7)}$, (4.1) can be positive and hence allow $\mathcal{M}_2$ outperform $\mathcal{M}_7$ on average. However, as n increases, (4.1) will be guaranteed to become negative since $n\log(\delta)$ grows much faster than $\log(n)$. In other words, in expectation, as n increases $\mathcal{M}_2$ cannot outperform $\mathcal{M}_7$ anymore.

At last, we expect $\mathcal{M}_3$ (or $\mathcal{M}_4$) to outperform $\mathcal{M}_7$ in expectation when β_1 (or β_2) is small, or when $|r_{12}|$ is close to 1. In either of these two circumstances, BMA will favour the more parsimonious model compared to the true model.

$\mathcal{M}_5$ and $\mathcal{M}_6$ can never outperform $\mathcal{M}_7$ in expectation because the bias terms in the RSS, $f_5(\beta_i, r_{ij}, n)$ and $f_6(\beta_i, r_{ij}, n)$ are never negative. Hence, for models with $p = 2$, $E(RSS_7)$ is always smaller than $E(RSS_5)$ and $E(RSS_6)$, which guarantees a smaller $E(BIC_7)$ and a larger $\tilde{E}(\pi_7)$. Moreover, $\mathcal{M}_8$ can never outperform $\mathcal{M}_7$ because $\mathcal{M}_8$ contains the true model $\mathcal{M}_7$, and BMA will favour the parsimonious model.

4.2 Observations of BMA in Variable Importance and Coefficient Estimation

After getting all models' posterior probabilities for different parameter combinations, we now pay attention to the results for variable importance and coefficient estimation. It is desirable to have high $\tilde{E}(\hat{\gamma}_1)$ and $\tilde{E}(\hat{\gamma}_2)$, and low $\tilde{E}(\hat{\gamma}_3)$ for BMA to successfully classify the importance of all three variables. Out of the 96,800 parameter combinations, a rather modest 67.12% of them have both $\tilde{E}(\hat{\gamma}_1)$ and $\tilde{E}(\hat{\gamma}_2)$ bigger than 0.5 while $\tilde{E}(\hat{\gamma}_3) < 0.5$. For 96.83% of the combinations, at least one of X_1 or X_2 is recognized as important.

For cases where the true model is selected, variable importance is generally also successful. For 89.33% of the cases when BMA selects the correct model, it is also successful in variable importance. Among the 25,560 cases when BMA is dangerous in model selection, only 1,336 of them correctly select important variables. Even worse, 496 of them select X_3 instead of X_1 or X_2. This scenario happens when BMA puts too much weight on the models which contain X_3, especially for the cases that $\mathcal{M}_2$ is selected as HPM.

We divided the cases into 4 groups based on their model selection and variable importance performance based on our formulas (Table 4.2). Obviously, there is strong association between the two criteria.

Table 4.2: Number of Cases Within Each Group

	Variable Importance			
Model Selection		Successful	Dangerous	Total
	Successful	63640	7600	71240
	Dangerous	1336	24224	25560
	Total	64976	31824	96800

We describe the features of the parameter combinations for each group in Table 4.3. We define the coefficient parameters as High if $|\beta_i|$ is mostly higher than 0.5 for the parameter combinations classified in each group, and Low if $|\beta_i|$ is mostly lower than 0.5. We define it as Lower if there are more values less than 0.5. We also calculate the determinant of $\boldsymbol{X'X}$ for each parameter combination within each group and report the median of the determinants for each group (See Appendix A). Smaller determinant of $\boldsymbol{X'X}$ implies the correlation matrix from a parameter combination is closer to its boundary.

Table 4.3: Coefficient Estimation of Each Group

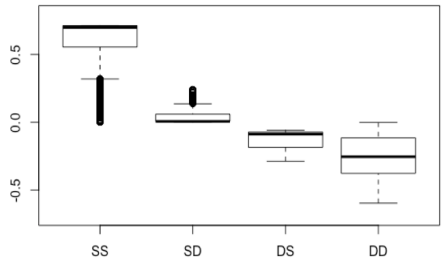
	β_1	β_2	Median of determinant
SS	High	High	0.33
SD	Lower	Lower	0.21
DS	Low	Low	0.14
DD	Lower	Lower	0.13

From Table 4.3, we summarize main findings for our predictions of BMA behaviours in model selection and variable importance. First, if both β_i $(i = 1, 2)$ are substantially different from 0 and the determinant is large, our prediction will suggest BMA to be more likely successful in both model selection and variable importance. Second, if both β_i are close to 0 and the determinant is relatively small, our analysis suggests those are probably the dangerous cases in at least one of model selection or variable importance.

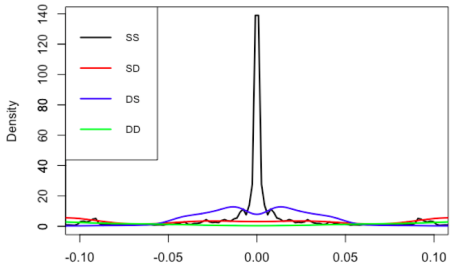
The predicted Margin for each group is also shown in Figure 4.1(a). The Margins are negative for group DS and DD because they do not select the correct model. The medians of Margin for group SS and DD are more clearly away from 0. In the SD and DS groups, when BMA is dangerous in just one of model selection or variable importance, the Margins are closer to 0. It means that in the SD and DS groups, the posterior probability of the true model is mildly different from the other models (small margin).

It is rather intuitive that if $\mathcal{M}_7$ is clearly better than others, then automatically, β_1 and β_2 get high probabilities. Thus, a high margin predisposes the case to be successful in variable

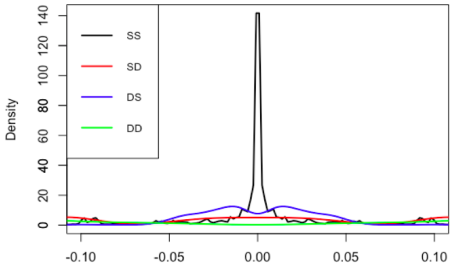
Figure 4.1: Mathematical Analysis Results of the 4 Groups



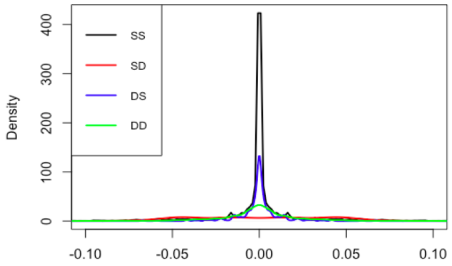
(a) Predicted Margin for each group



(b) Predicted Approximate Bias for approximate expectation of β_1 coefficient estimator: $D(\tilde{E}(\hat{\beta}_1))$



(c) Predicted Approximate Bias for approximate expectation of β_2 coefficient estimator: $D(\tilde{E}(\hat{\beta}_2))$



(d) Predicted Approximate Bias for approximate expectation of β_3 coefficient estimator: $D(\tilde{E}(\hat{\beta}_3))$

selection. If the Margin isn't high, then other models rather than the true model have high probabilities too, and they may promote β_3 or fail to promote both β_1 and β_2.

Figures 4.1(b), 4.1(c) and 4.1(d) show the estimated density curves for the Approximate Bias of each coefficient β_i. We are not surprised that when BMA is dangerous in model selection and/or variable importance, the densities of $\tilde{E}(\hat{\beta}_i)$ $(i = 1, 2, 3)$ are more spread out, whereas for group SS, the densities of $D(\tilde{E}(\hat{\beta}_i)$ have clear peaks at 0. Among the groups except for SS, DS has smaller Approximate Bias for all β_i compared to the groups that are dangerous in variable importance.

Chapter 5

Simulation Study

Our mathematical analysis predicts BMA performance over different parameter combinations. We now describe the simulation study used to verify the observations we noticed earlier.

We take the same parameter combinations (β_1, β_2, r_{12}, r_{13}, r_{23}) used in the mathematical analysis. For each parameter combination, $c = 1, \ldots, 96800$, we simulated the explanatory variables $\boldsymbol{X_i}$ $(i = 1, 2, 3)$ from the multivariate normal N(0,R), such that $cor(X_i, X_j) = r_{ij}$. The response variable $\boldsymbol{Y}$ was generated from the true model (3.1) and the process was repeated 100 times to generate 100 datasets. For each of the 100 runs, we fitted all 8 models and obtained the following statistics.

- $\hat{\pi}_{k,c}$ is the realized model posterior probability. It is calculated from the model BICs as in (2.4).
- Variable importance for X_i $(i = 1, 2, 3)$ is $\hat{\gamma}_{i,c} = \sum_{k=1}^{8} \hat{\pi}_{k,c} \mathbb{1}(X_i \in \mathcal{M}_k)$.
- $\hat{\beta}_{i,c}$ $(i = 1, 2, 3)$ is the estimated model coefficient for each variable. It is the sum of $\hat{\beta}_{i,k}$ weighted by model posterior probabilities $\hat{\pi}_{k,c}$. See (3.5).

We define the average of each of the 3 statistics of the 100 runs as $\bar{\pi}_{k,c}$, $\bar{\gamma}_{i,c}$ and $\bar{\beta}_{i,c}$, which are the counterparts of $\tilde{E}(\hat{\pi}_{k,c})$, $\tilde{E}(\hat{\gamma}_{i,c})$ and $\tilde{E}(\hat{\beta}_{i,c})$ in the math analysis, respectively.

Similar to what we have done in the mathematical analysis, if $\bar{\pi}_7$ is the highest posterior probability, we say its corresponding parameter combination is successful in BMA model selection, dangerous otherwise. Also, if $\bar{\gamma}_1$ and $\bar{\gamma}_2$ are bigger than 0.5 while $\bar{\gamma}_3$ is less than 0.5, it is a successful case of variable importance.

We keep the same 4 groups as defined in math analysis. For model selection, we calculate the Margin $m(\bar{\pi})$ and compare it with $m(\tilde{E}(\hat{\pi}))$ for each of the 4 groups. If the simulated

Margin is similar to the predicted Margin for a group, math analysis well predicts BMA in terms of selecting the correct HPM.

To better understand our prediction regarding variable importance, we obtain the odds ratio of each variable across the 4 groups. Specifically, we look at

$$OR_i = \frac{\bar{\gamma}_i/(1-\bar{\gamma}_i)}{\tilde{E}(\hat{\gamma}_i)/(1-\tilde{E}(\hat{\gamma}_i))}$$

$(i = 1, 2, 3)$ for each of the 4 groups. If OR_i is higher than 1 for some i, then on average $\tilde{E}(\hat{\gamma}_i)$ underestimates $\bar{\gamma}_i$. Our math analysis would predict X_i to be less important than it really is.

Finally, we will compare the Approximate Bias $D(\bar{\beta}_i)$ with $D(\tilde{E}(\hat{\beta}_i))$ to assess whether (3.5) gives a reasonable prediction of the BMA coefficient estimation.

5.1 Results of Simulation Study

Table 5.1 shows the confusion matrix of the math prediction and simulation study. The worst classifications happen among the cases when we predict BMA to be dangerous in only one respect (the SD and DS groups). When our formulae predict a case as SD, simulation will mostly classify such case as DD. Thus, $\tilde{E}(\hat{\pi}_k)$ is rather optimistic in predicting posterior probability of the true model when there is confusion in the variable importance measures. Similarly, the simulation classifies about half of the DS cases as DD. The formulas are most accurate at predicting DD cases, where the simulation agrees in 98.48% of those cases.

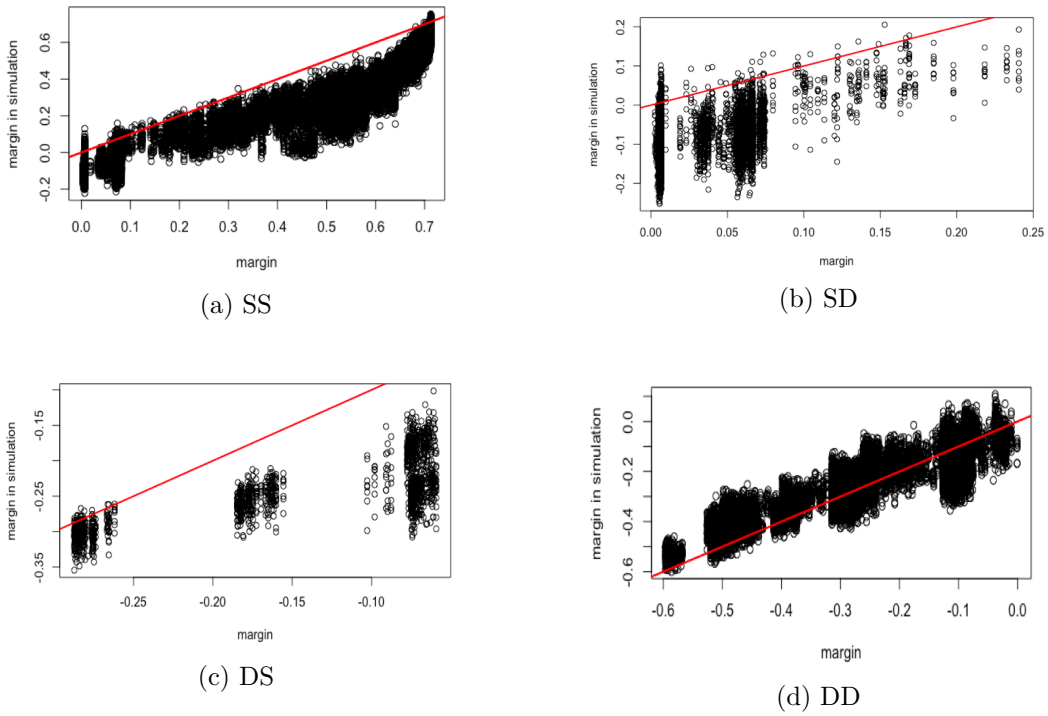
Table 5.1: Confusion Matrix of 4 Groups

		Simulation Result					
Math Prediction		**SS**	**SD**	**DS**	**DD**	**Total**	**Percentage Correct**
	SS	56288	1187	356	5809	63640	88.45%
	SD	88	601	17	6894	7600	7.91%
	DS	0	0	635	701	1336	47.53%
	DD	31	154	182	23857	24224	98.48%
	Total	56407	1942	1190	37261	96800	

Figure 5.1 shows the simulated Margin $m(\bar{\pi})$ for model posterior probabilities versus the predicted Margin $m(\tilde{E}(\hat{\pi}))$, with groups classified according to the math predictions. The scatter plots of SD and DS are more discrete due to fewer observations in these groups.

In general, the predicted Margin overestimates the simulated Margin except for group DD. This implies that when the formulae predict successful model selection (SS and SD groups)

Figure 5.1: Simulated Margin versus predicted Margin for each group



(a) SS

(b) SD

(c) DS

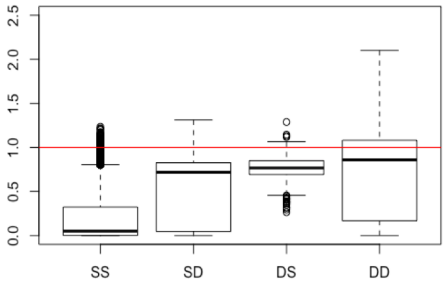
(d) DD

with a positive Margin, the simulation may indicate less certainty for the correct model, or even a preference for a different model, resulting in a negative margin. In SS, it is worthwhile to note that when the predicted margin is at least 0.1, the simulations rarely suggest (0.13%) that a different model is best. On the other hand, simulated margins for the SD group are usually below zero, indicating that the predicted failure of variable selection usually implies failure for model selection in the simulation. Most SD cases are classified as DD in simulations.

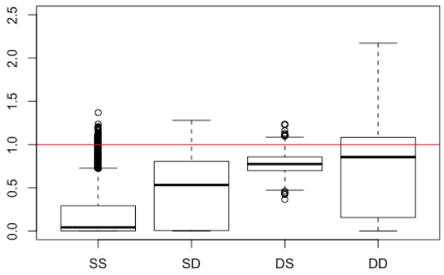
The predicted Margin also overestimates the simulated Margin in general in DS. When we see a negative Margin in DS, the simulated Margin is likely to be more negative. The DD group has the most accurate prediction on Margin, agreeing with the high correct classification rate in Table 5.1.

Moreover, Figure 5.2 shows the box plots of odds ratios. For all 4 groups, the odds ratio for each of X_1's and X_2's posterior probabilities is lower than 1, but is higher than 1 for X_3. In other words, on average, $\tilde{E}(\hat{\gamma}_i)$ overestimates $\bar{\gamma}_i$ for X_1 and X_2, but underestimates for X_3. The overestimation seems extremely severe for SS, but most predicted SS cases are also observed to be SS. We found in the SS group, the predicted $\tilde{E}(\hat{\gamma}_1)$ and $\tilde{E}(\hat{\gamma}_2)$ are very close to 1 too often, whereas the simulations suggest that they should be more moderate. The underestimation of $\tilde{E}(\hat{\gamma}_3)$ is not as severe by contrast, so it would seem that the real

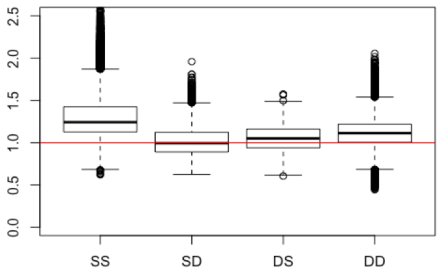
Figure 5.2: Odds Ratio boxplot for each of the 3 regression coefficients, separated by predicted groups



(a) Odds Ratio for X_1: OR_1



(b) Odds Ratio for X_2: OR_2

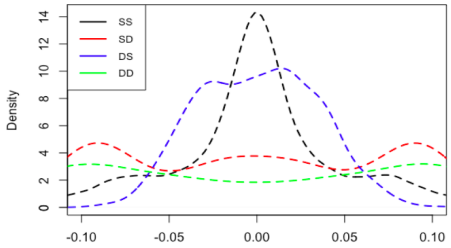


(c) Odds Ratio for X_3: OR_3

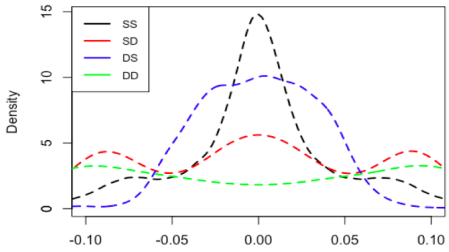
problem with identifying cases lies in overconfidence about the importance of real variables as opposed to misidentifying details of the unimportant variable.

Finally, we compare the parameter estimates. Figure 5.3 shows the density plots of Approximate Bias for the simulated coefficients. If we compare them with Figures 4.1(b), 4.1(c) and 4.1(d), the simulated Approximate Bias results agree with those of the predicted Approximate Bias for all 3 coefficients that group SS has the smallest Approximate Bias for β_i $(i = 1, 2, 3)$, followed by group DS. All bias estimates have more variability than their corresponding predicted values, perhaps partly because these bias estimates are based on only 100 estimates per case. In addition, scatter plots show $D(\tilde{E}(\hat{\beta}_i))$ predicts $D(\bar{\beta}_i)$ reasonably well (See Appendix B).

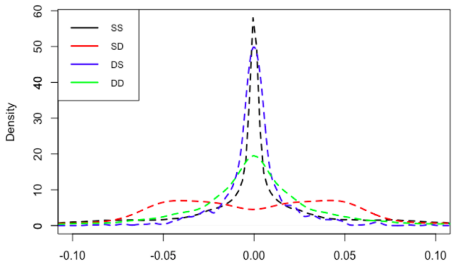
Figure 5.3: Parameter estimation of simulation studies



(a) Approximate Bias of simulated mean of β_1 coefficient: $D(\bar{\beta}_1)$



(b) Approximate Bias of simulated mean of β_2 coefficient: $D(\bar{\beta}_2)$



(c) Approximate Bias of simulated mean of β_3 coefficient: $D(\bar{\beta}_3)$

Chapter 6

Discussion and Conclusion

This paper aims to understand the BIC approximation to BMA in the presence of collinearity. We studied the 3 variable all subsets regression problem and derived the approximate expectation of model posterior probability ($\tilde{E}(\hat{\pi}_i)$), approximate expectation of variable importance ($\tilde{E}(\hat{\gamma}_i)$) and approximate expectation of coefficient estimators ($\tilde{E}(\hat{\beta}_i)$) to predict the BMA results for model selection, variable importance and coefficient estimation. These predictions were then compared with simulated data. We found that our approximate formulas for these quantities tend to be overoptimistic in predicting successful model selection and variable importance. In general, our formulae overestimate the posterior probabilities of the correct model and of the important variables. To be more specific, when we predict a parameter combination to be a dangerous case of model selection and/or a dangerous case in variable importance, it will most likely to be dangerous in both applications.

The prediction with regards to coefficient estimation is more accurate. We compared the Approximate Bias of each coefficient between the math prediction and the simulation. We found $D(\tilde{E}(\hat{\beta}_i))$ $(i = 1, 2, 3)$ predicts $D(\bar{\beta}_i)$ reasonably well.

In this study, we offered a way to systematically investigate BMA in the presence of multicollnearity. The math formulae are easy to compute for any number of variables and other combinations of variables in the true model. One could easily extend the math analysis to these cases in the hopes of discovering a pattern to when BMA could or should not be used. Thus, the entire body of the study is a proof of concept to show that there is potential for using theoretical analysis to identify when BMA can be successful in applications.

As with any study, there are certain limitations to our results. We restricted our analysis in the 3 variable all subsets regression scenario where the true model is in the form of (3.1). We also restricted sample size $n = 100$. More research is required to generalize our findings to the situations with more complex model and correlation structures.Throughout the analysis, we assumed the true model is known, which is rarely valid in reality. Researchers may

substitute the true model with the best predicted model in order to use the mathematical approach in this paper, or may investigate what would happen under a variety of plausible true models, given a particular observed correlation structure among explanatory variables. When there are many covariates or the candidate models are not linear regressions, the analytic approach may not be applicable. In addition, when it is believed that complex correlation structure exists, it might be more desirable to use the fully Bayesian approach of BMA and incorporate that information into priors instead of using the BIC version of BMA.

Finally, although we did not do it, we could consider developing a diagnostic for "safe" BMA usage by analyzing the correlation matrix for X. This relates to the observation that certain correlation structures tend to lead to greater likelihood of dangerous cases. We fixed r_{12} and r_{13} and let r_{23} vary within its range, but it would be better to study the boundary of the correlation matrix in 3 dimensions to obtain a better understanding of the correlation structure. Using correlation to identify these cases, rather than the regression parameter values, is important because we never can know the parameters in practice, but we can measure the correlation directly on $\boldsymbol{X}$.

Bibliography

Barbieri, Maria M., and James O. Berger. "Optimal Predictive Model Selection." The Annals of Statistics, vol. 32, no. 3, 2004, pp. 870-897.

Becker, Jan-Michael, et al. "How Collinearity Affects Mixture Regression Results." Marketing Letters, vol. 26, no. 4, 2015, pp. 643-659.

Brown, P. J., M. Vannucci, and T. Fearn. "Bayes Model Averaging with Selection of Regressors." Journal of the Royal Statistical Society. Series B (Statistical Methodology), vol. 64, no. 3, 2002, pp. 519-536.

Clyde, Merlise, and Edward I. George. "Model Uncertainty." Statistical Science, vol. 19, no. 1, 2004, pp. 81-94.

Daniell, P. J. "boundary Conditions for Correlation Coefficients." British Journal of Psychology. General Section, vol. 20, no. 2, 1929, pp. 190-194.

Daunizeau, Jean. "Semi-Analytical Approximations to Statistical Moments of Sigmoid and Softmax Mappings of Normal Variables." Brain and Spine Institute, 2017.

Draper, "Assessment and Propagation of Model Uncertainty." Mathematical Social Sciences, vol. 27, no. 1, 1994, pp. 116-117.

Elliot, Mark, et al. A Dictionary of Social Research Methods. Oxford University Press, Oxford, 2016.

FernÃąndez, Carmen, Eduardo Ley, and Mark F. J. Steel. "Benchmark Priors for Bayesian Model Averaging." Journal of Econometrics, vol. 100, no. 2, 2001, pp. 381-427.

Fragoso, Tiago M., Wesley Bertoli, and Francisco Louzada. "Bayesian Model Averaging: A Systematic Review and Conceptual Classification." International Statistical Review, vol. 86, no. 1, 2018, pp. 1-28.

Freckleton, Robert P. "Dealing with Collinearity in Behavioural and Ecological Data: Model Averaging and the Problems of Measurement Error." Behavioral Ecology and Sociobiology, vol. 65, no. 1, 2011, pp. 91-101.

Furnival, George M., and Robert W Wilson Jr. "Regression by Leaps and Bounds." Technometrics, vol. 42, no. 1, 2000, pp. 69.

Ghosh, Joyee, and Andrew E. Ghattas. "Bayesian Variable Selection Under Collinearity." The American Statistician, vol. 69, no. 3, 2015, pp. 165-173.

Hand, D. J. "Branch and Bound in Statistical Data Analysis." Journal of the Royal Statistical Society. Series D (the Statistician), vol. 30, no. 1, 1981, pp. 1-13.

Hoeting, Jennifer A., et al. "Bayesian Model Averaging: A Tutorial." Statistical Science, vol. 14, no. 4, 1999, pp. 382-401.

Kass, Robert E., and Larry Wasserman. "The Selection of Prior Distributions by Formal Rules." Journal of the American Statistical Association, vol. 91, no. 435, 1996, pp. 1343-1370.

Raftery, Adrian E., David Madigan, and Jennifer A. Hoeting. "Bayesian Model Averaging for Linear Regression Models." *Journal of the American Statistical Association*, vol. 92, no. 437, 1997, pp. 179-191.

Raftery, Adrian E. "Bayes Factors and BIC: Comment on "A Critique of the Bayesian Information Criterion for Model Selection"." Sociological Methods and Research, vol. 27, no. 3, 1999, pp. 411.

Wasserman, Larry. "Bayesian Model Selection and Model Averaging." Journal of Mathematical Psychology, vol. 44, no. 1, 2000, pp. 92-107.

Appendix A

We give the box plot of the determinant of $\boldsymbol{X'X}$ for each group. We see the median of determinate is relatively higher for cases that are predicted to be successful in both model selection and variable importance compare to the cases that are predicted to be dangerous in both of the two aspects. However, such amount may not be significant in predicting BMA performance.

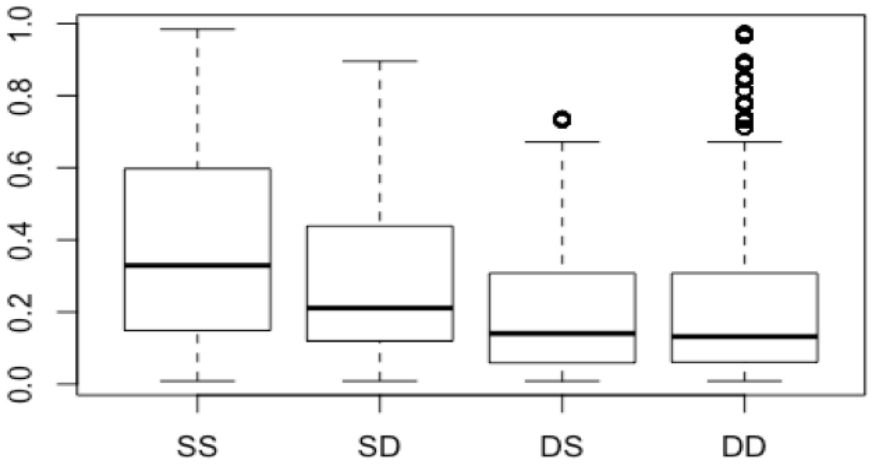


Figure A.1: Determinant of the Full Model Correlation Matrix Boxplot

We also give the histograms for the coefficient parameters observed in each of the predicted groups. In group SS, most β_1 and β_2 have values grater than 0.5. By contrast, β_1 and β_2 in group DS are mostly clustered between -0.4 and 0.4. The histograms of group SD and group DD are similar, while β_1 and β_2 have the highest densities around -0.2 to 0.2, they are also uniformly spread for all other tested values.

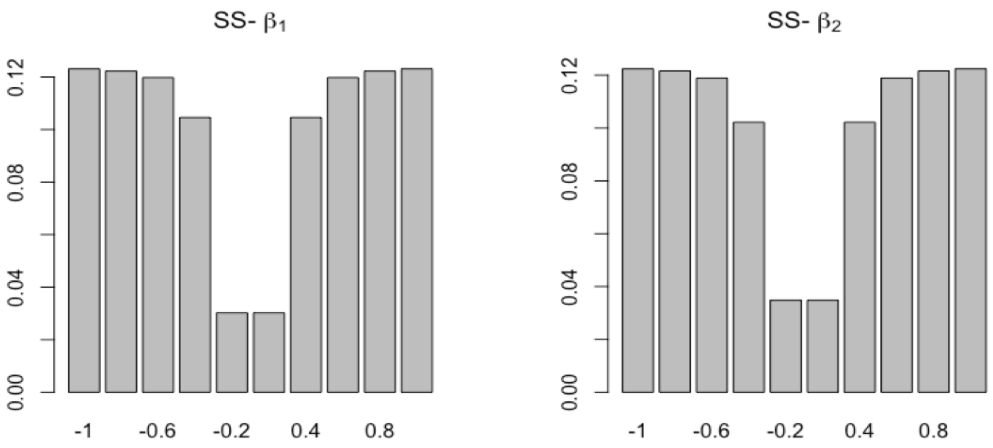


Figure A.2: SS β_1 and β_2 Histograms

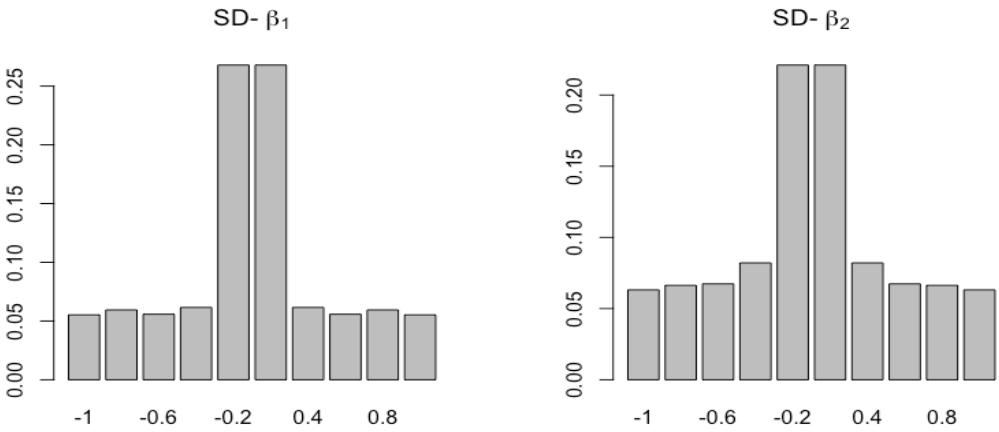


Figure A.3: SD β_1 and β_2 Histograms

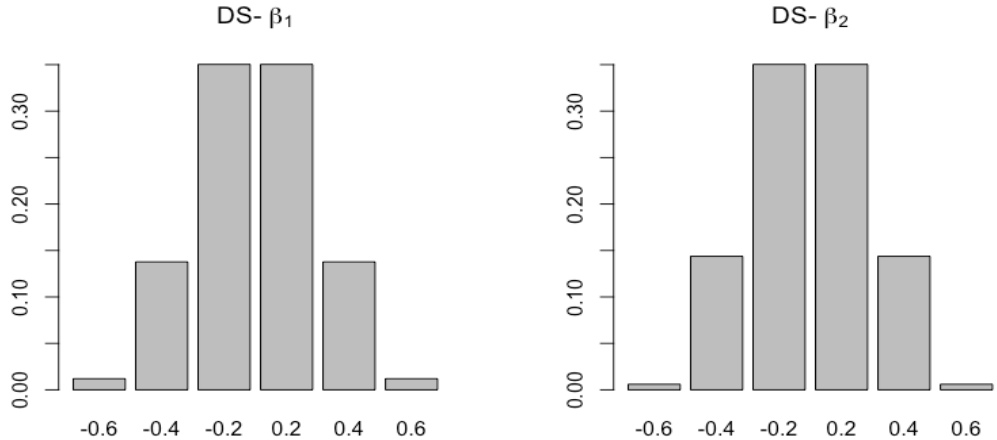


Figure A.4: DS β_1 and β_2 Histograms

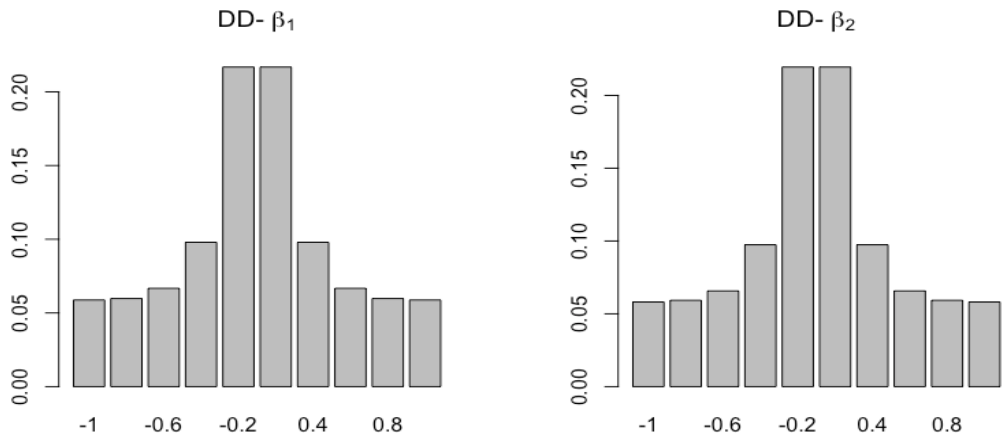


Figure A.5: DD β_1 and β_2 Histograms

Appendix B

We give the scatter plots for $D(\bar{\beta}_i)$ versus $D(\tilde{E}(\hat{\beta}_i))$ $(i = 1, 2, 3)$ for all 4 groups. The $D(\bar{\beta}_i)$ measures the Approximate Bias of each simulated mean of coefficient and $D(\tilde{E}(\hat{\beta}_i))$ measures the Approximate Bias of each of the predicted coefficient. We see that on average, $D(\tilde{E}(\hat{\beta}_i))$ well predicts $D(\bar{\beta}_i)$ for all the groups.

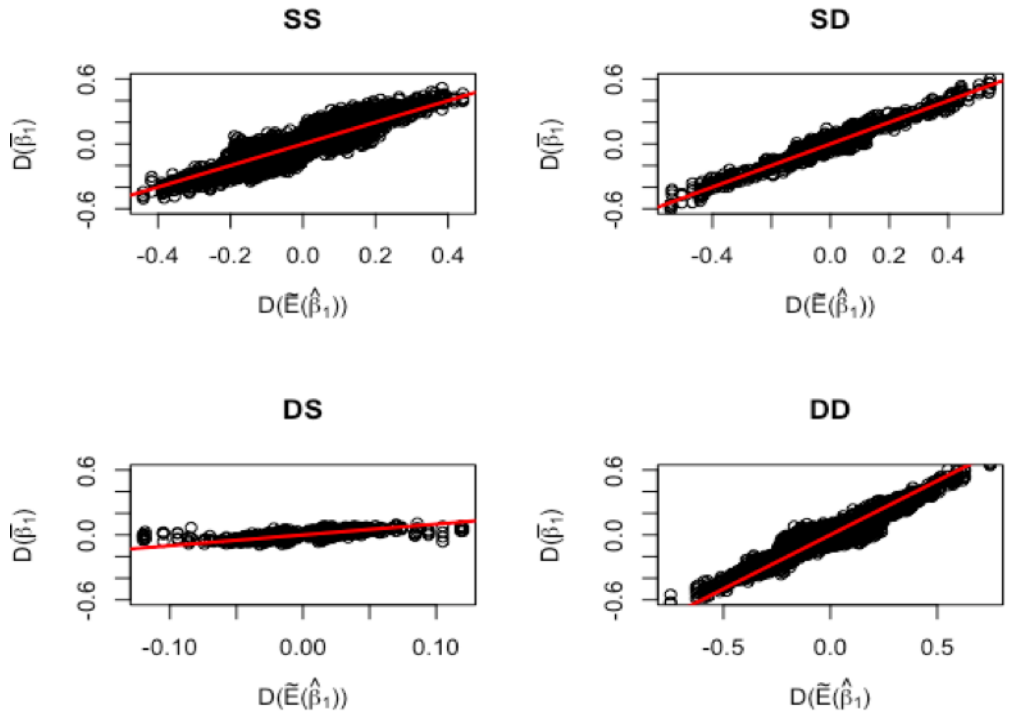


Figure B.1: Simulated Approximate Bias $(D(\bar{\beta}_1))$ versus predicted Approximate Bias $(D(\tilde{E}(\hat{\beta}_1)))$ of β_1 scatter plot

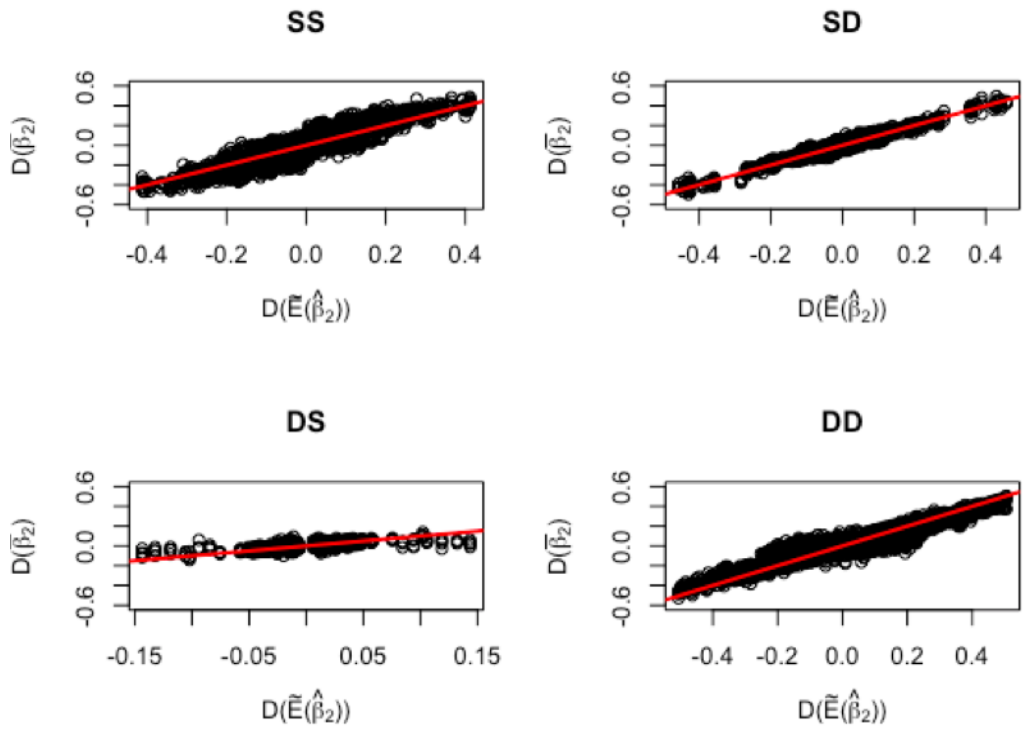


Figure B.2: Simulated Approximate Bias ($D(\bar{\beta}_2)$) versus predicted Approximate Bias ($D(\tilde{E}(\hat{\beta}_2))$) of β_2 scatter plot

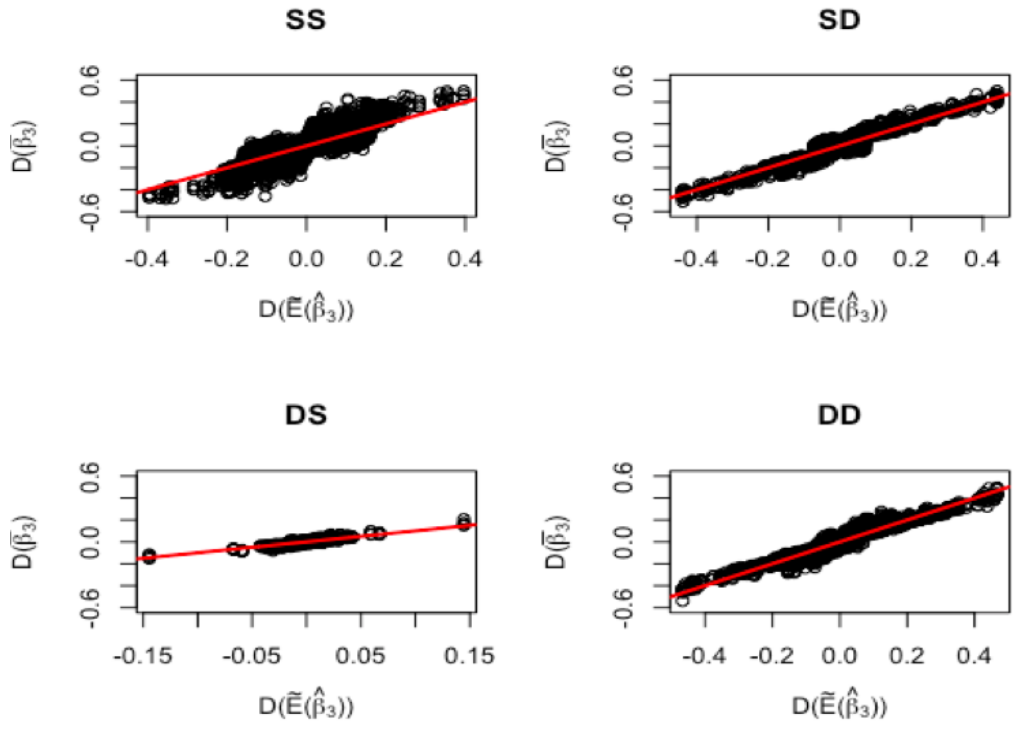


Figure B.3: Simulated Approximate Bias ($D(\bar{\beta}_3)$) versus predicted Approximate Bias ($D(\tilde{E}(\hat{\beta}_3))$) of β_3 scatter plot