

IS MODEL AVERAGING THE SOLUTION FOR ADDRESSING MODEL  
UNCERTAINTY? METHODOLOGICAL INSIGHTS, TOOLS FOR  
ASSESSMENT, AND CONSIDERATIONS FOR PRACTICAL USE

by

Katharine Michelle Banner

A dissertation submitted in partial fulfillment  
of the requirements for the degree

of

Doctor of Philosophy

in

Statistics

MONTANA STATE UNIVERSITY  
Bozeman, Montana

November, 2016

©COPYRIGHT

by

Katharine Michelle Banner

2016

All Rights Reserved

## DEDICATION

I dedicate this dissertation to my mom, Valerie Banner, who is the strongest person I know. Her courage in the face of overwhelming adversity gave me the strength to finish this work. This is for you, Mom, I love you to the moon and back!

## ACKNOWLEDGEMENTS

First and foremost, I would like to acknowledge my advisor, Dr. Megan Higgs, for her unwavering support and belief in me as a researcher. Her dedication to my learning and growth as a statistician is truly immeasurable. She set a standard I can only hope to achieve with students I may work with in the future. I cannot thank her enough for the countless hours she spent editing my work, pushing me to be my best, teaching me her unique approach to problems, and providing thoughtful advice. Thank you, Megan, you have been an amazing advisor to me both in academia, and in life.

Thank you to all of my family for being supportive of me in my pursuit of this degree. A very special thank you goes to my amazing parents who dedicated their lives to raising me, and who were there for me regardless of the hour to provide a listening ear and any support they could. Mom and Dad, I could not have done this without your unconditional love. Thank you.

I would also like to thank all of my friends (human and canine alike) who have offered their support through these last four years. I am indebted to all of you for providing your friendship, reassurance, and kind hearts. A special thanks goes to Liz Arnold for exchanging writing, helping solve problems, and being a devoted friend throughout this whole process.

Last, but surely not least, thank you (in no particular order) to: Dr. Steve Cherry, Dr. Jenny Green, Dr. Mark Greenwood, Dr. John Borkowski, Dr. Jim Robison-Cox, Dr. Lillian Lin, and all of the faculty members who have taught me and also offered their expertise to strengthen this work.

## TABLE OF CONTENTS

|                                                                                                                                                     |    |
|-----------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 1. INTRODUCTION .....                                                                                                                               | 1  |
| 1.1 Current State of Model Averaging.....                                                                                                           | 4  |
| 1.2 Specific Aims for This Research.....                                                                                                            | 6  |
| 2. CONSIDERATIONS FOR ASSESSING MODEL AVERAGING<br>OF REGRESSION COEFFICIENTS .....                                                                 | 10 |
| Contribution of Authors and Co-Authors .....                                                                                                        | 10 |
| Manuscript Information Page .....                                                                                                                   | 11 |
| Abstract .....                                                                                                                                      | 12 |
| 2.1 Introduction .....                                                                                                                              | 12 |
| 2.2 Model Averaging Within Multimodel Inference .....                                                                                               | 16 |
| 2.2.1 Model Averaging in Regression.....                                                                                                            | 19 |
| 2.2.2 Model Averaging of Predictions.....                                                                                                           | 22 |
| 2.2.3 Model Averaging of <i><b>Partial</b></i><br>Regression Coefficients .....                                                                     | 24 |
| 2.2.3.1 Two Components of the Model<br>Averaged Posterior Distribution .....                                                                        | 28 |
| 2.2.4 Graphical Tools for<br>Assessing Model Averaging .....                                                                                        | 30 |
| 2.3 Example 1: Haul-out Behavior of Weddell Seals.....                                                                                              | 32 |
| 2.4 Example 2: When Prediction Leads to Explanation .....                                                                                           | 43 |
| 2.5 Discussion .....                                                                                                                                | 46 |
| 2.5.1 Conclusion.....                                                                                                                               | 50 |
| 2.6 Acknowledgements.....                                                                                                                           | 51 |
| 3. INVESTIGATING THE POSTERIOR VARIANCE OF PARTIAL<br>REGRESSION COEFFICIENTS RESULTING FROM THREE<br>COMMON METHODS FOR MULTIMODEL INFERENCE ..... | 52 |
| Contribution of Authors and Co-Authors .....                                                                                                        | 52 |
| Manuscript Information Page .....                                                                                                                   | 53 |
| Abstract .....                                                                                                                                      | 54 |
| 3.1 Introduction .....                                                                                                                              | 54 |
| 3.2 Statistical Preliminaries .....                                                                                                                 | 58 |
| 3.2.1 Variance of Model-Averaged<br>Regression Coefficients .....                                                                                   | 59 |
| 3.2.1.1 Further Insight Into<br>the Model Averaged Posterior .....                                                                                  | 61 |

## TABLE OF CONTENTS - CONTINUED

|                                                                                         |     |
|-----------------------------------------------------------------------------------------|-----|
| 3.3 Simulation Study.....                                                               | 65  |
| 3.3.1 Ratios for Comparing Variances.....                                               | 67  |
| 3.3.1.1 All Models Versus Subset of Models.....                                         | 67  |
| 3.3.1.2 Subset of Models Versus Top Model.....                                          | 68  |
| 3.3.1.3 All Models Versus Top Model.....                                                | 69  |
| 3.3.2 Models.....                                                                       | 69  |
| 3.3.3 Simulation Study Experimental Design.....                                         | 71  |
| 3.3.3.1 Factor 1: Correlation.....                                                      | 72  |
| 3.3.3.2 Factor 2: Signal to Noise.....                                                  | 72  |
| 3.3.3.3 Factor 3: Truth.....                                                            | 73  |
| 3.3.3.4 Implementation.....                                                             | 74  |
| 3.4 Results.....                                                                        | 75  |
| 3.4.1 Analytical Results.....                                                           | 76  |
| 3.4.2 All Models Compared to Subset of Models.....                                      | 81  |
| 3.4.2.1 Comparing Correlations.....                                                     | 81  |
| 3.4.2.2 Comparing Signal to Noise Ratios.....                                           | 88  |
| 3.4.2.3 Comparing Truth in the Model Set.....                                           | 92  |
| 3.4.3 Subset of Models Compared to Top Model.....                                       | 100 |
| 3.4.3.1 Comparing Correlations.....                                                     | 101 |
| 3.4.3.2 Comparing Levels                                                                |     |
| Within Other Factors.....                                                               | 103 |
| 3.4.4 All Models Compared to Top Model.....                                             | 104 |
| 3.4.4.1 Comparison of Correlations.....                                                 | 104 |
| 3.4.4.2 Comparing Signal to Noise.....                                                  | 106 |
| 3.4.4.3 Comparing Truth in the Model Set.....                                           | 107 |
| 3.5 Discussion.....                                                                     | 108 |
| 3.5.1 All Models Versus Subset of Models.....                                           | 110 |
| 3.5.2 Subset of Models Versus Top Model.....                                            | 112 |
| 3.5.3 All Models Versus Top Model.....                                                  | 114 |
| 3.5.4 Potential Extensions.....                                                         | 116 |
| 3.6 Conclusion.....                                                                     | 117 |
| 4. THE MODEL AVERAGED POSTERiors PLOT MAPP PACK-<br>AGE FOR R STATISTICAL SOFTWARE..... | 119 |
| Contribution of Authors and Co-Authors.....                                             | 119 |
| Manuscript Information Page.....                                                        | 120 |
| 4.1 About this Chapter.....                                                             | 121 |
| Abstract.....                                                                           | 121 |
| 4.2 Motivation for the Package.....                                                     | 122 |

## TABLE OF CONTENTS - CONTINUED

|                                                                                                                      |     |
|----------------------------------------------------------------------------------------------------------------------|-----|
| 4.3 Model Averaging of Partial Regression Coefficients.....                                                          | 122 |
| 4.4 Installing <b>MApp</b> .....                                                                                     | 124 |
| 4.4.1 GitHub Install .....                                                                                           | 124 |
| 4.4.2 Local Disk Install.....                                                                                        | 125 |
| 4.5 <b>MApp</b> Plotting Functions .....                                                                             | 125 |
| 4.6 <b>MApp</b> Package Data .....                                                                                   | 126 |
| 4.7 The <b>MApp_bms</b> Function .....                                                                               | 128 |
| 4.7.1 Working with Small Model Sets .....                                                                            | 129 |
| 4.7.1.1 Reading the MAP Plot.....                                                                                    | 130 |
| 4.7.1.2 Using the MAP Plot for<br>Justifying Model Choices .....                                                     | 134 |
| 4.7.2 Working with Large Model Sets .....                                                                            | 136 |
| 4.8 <b>MApp</b> for IC-Based MA.....                                                                                 | 138 |
| 4.8.1 Using <b>MApp_IC</b> .....                                                                                     | 139 |
| 4.8.2 Using <b>MApp_IC_gen</b> .....                                                                                 | 143 |
| 4.9 <b>MApp</b> for Self-Programmed Samplers .....                                                                   | 147 |
| 4.9.1 <b>MApp_MCMC</b> for Matricies.....                                                                            | 147 |
| 4.9.2 Using <b>MApp_list</b> .....                                                                                   | 150 |
| 4.10 Conclusion .....                                                                                                | 152 |
| 5. CONCLUSION AND FUTURE WORK.....                                                                                   | 154 |
| REFERENCES CITED.....                                                                                                | 161 |
| APPENDICES .....                                                                                                     | 166 |
| APPENDIX A: Prior Specifications Continued .....                                                                     | 167 |
| APPENDIX B: Considerations for Assessing Model Averaging of<br>Regression Coefficients Supplementary Materials ..... | 171 |
| APPENDIX C: Simulation Study R code .....                                                                            | 186 |

## LIST OF TABLES

| Table                                                                    | Page |
|--------------------------------------------------------------------------|------|
| 2.1 Table of Posterior SDs .....                                         | 42   |
| 4.1 Brain Weight of Mammals Data .....                                   | 129  |
| 4.2 Posterior SDs: <b>brainData</b> .....                                | 133  |
| 4.3 Body Fat Data .....                                                  | 136  |
| 4.4 Posterior SDs: Body Fat Study .....                                  | 137  |
| 4.5 PMPs from IC Approximation .....                                     | 139  |
| 4.6 MA Coefficients from IC Approximation .....                          | 140  |
| 4.7 SDs from IC Approximation .....                                      | 140  |
| 4.8 Table of IC Weights for <b>brainData</b> Example .....               | 144  |
| 4.9 Estimated Coefficients for <b>brainData</b> Example .....            | 145  |
| 4.10 Estimated SEs for Coefficients<br>in <b>brainData</b> Example ..... | 145  |
| 4.11 Example Form of Posterior Draws .....                               | 148  |
| 4.12 Table of Posterior SDs .....                                        | 150  |

## LIST OF FIGURES

| Figure                                                                         | Page |
|--------------------------------------------------------------------------------|------|
| 1.1 Model Averaging in the Literature.....                                     | 7    |
| 2.1 MA Flow Chart .....                                                        | 15   |
| 2.2 Cut-and-coded Scatterplot Example .....                                    | 27   |
| 2.3 Seals Hauled Out in Big Razorback.....                                     | 33   |
| 2.4 MAP Plot: Seal Example .....                                               | 37   |
| 2.5 MAP Plot for {Time at max} .....                                           | 41   |
| 2.6 MAP Plot from Body Fat Example .....                                       | 45   |
| 3.1 Model Averaged Posterior Scenarios, $\mu_{cts} = 3$ .....                  | 63   |
| 3.2 Model Averaged Posterior Scenarios, $\mu_{cts} = 0.15$ .....               | 64   |
| 3.3 Variance Comparisons of Interest .....                                     | 66   |
| 3.4 Realizations of Simulated Data .....                                       | 71   |
| 3.5 Analytical Results Comparing $Var_{MA}/Var_{cts}$ .....                    | 79   |
| 3.6 Comparing Correlations, Reference Scenario .....                           | 82   |
| 3.7 Reference Scenario, $cor-35 = 0.0$ .....                                   | 84   |
| 3.8 Reference Scenario, $cor-35 = 0.9$ .....                                   | 85   |
| 3.9 Comparing Correlations; $sig-y = 1$ , $truth = YES$ .....                  | 86   |
| 3.10 Comparing Correlations;<br>$sig-y = 2.5$ , $truth = NO$ -taper4 .....     | 87   |
| 3.11 Comparing Correlations;<br>$sig-y = 2.5$ , $truth = NO$ 5.....            | 88   |
| 3.12 Comparing Signal to Noise Ratios;<br>$truth = YES$ , $cor-35 = 0$ .....   | 89   |
| 3.13 Comparing Signal to Noise Ratios;<br>$truth = YES$ , $cor-35 = 0.9$ ..... | 91   |
| 3.14 Comparing Signal to Noise Ratios;<br>$truth = NO$ 5, $cor-35 = 0.9$ ..... | 92   |

## LIST OF FIGURES - CONTINUED

| Figure                                                                                                          | Page |
|-----------------------------------------------------------------------------------------------------------------|------|
| 3.15 Comparing Signal to Noise Ratios;<br><i>truth</i> = NO-taper4, <i>cor-35</i> = 0.9 .....                   | 93   |
| 3.16 Comparing <i>truth</i> = YES to <i>truth</i> = NO5;<br><i>sig-y</i> = 2.5, <i>cor-35</i> = 0.0 .....       | 94   |
| 3.17 Comparing <i>truth</i> = YES to <i>truth</i> = NO5;<br><i>sig-y</i> = 2.5, <i>cor-35</i> = 0.9 .....       | 95   |
| 3.18 Comparing <i>truth</i> = YES to <i>truth</i> = NO-taper4;<br><i>sig-y</i> = 2.5, <i>cor-35</i> = 0.0 ..... | 97   |
| 3.19 Comparing <i>truth</i> = YES to <i>truth</i> = NO-taper4;<br><i>sig-y</i> = 2.5, <i>cor-35</i> = 0.9 ..... | 98   |
| 3.20 Comparing <i>truth</i> = NO5 to <i>truth</i> = N04;<br><i>sig-y</i> = 2.5, <i>cor-35</i> = 0.0 .....       | 99   |
| 3.21 Comparing <i>truth</i> = NO5 to <i>truth</i> = N04;<br><i>sig-y</i> = 2.5, <i>cor-35</i> = 0.9 .....       | 100  |
| 3.22 Comparing <i>truth</i> = YES to <i>truth</i> = N0-taper4;<br><i>sig-y</i> = 1, <i>cor-35</i> = 0.9 .....   | 101  |
| 3.23 Comparing Correlations; <i>sig-y</i> = 2.5, <i>truth</i> = YES .....                                       | 102  |
| 3.24 $r_{MAc}$ vs. $r_{MAt}$ Comparison, <i>cor-35</i> = 0.0 .....                                              | 105  |
| 3.25 $r_{MAc}$ vs. $r_{MAt}$ Comparison, <i>cor-35</i> = 0.3 .....                                              | 106  |
| 3.26 $r_{MAc}$ vs. $r_{MAt}$ Comparison, <i>cor-35</i> = 0.6 .....                                              | 107  |
| 3.27 $r_{MAc}$ vs. $r_{MAt}$ Comparison, <i>cor-35</i> = 0.7 .....                                              | 108  |
| 3.28 $r_{MAc}$ vs. $r_{MAt}$ Comparison, <i>cor-35</i> = 0.8 .....                                              | 109  |
| 3.29 $r_{MAc}$ vs. $r_{MAt}$ Comparison, <i>cor-35</i> = 0.9 .....                                              | 110  |
| 3.30 MAP Plot: Reference Scenario .....                                                                         | 113  |
| 3.31 Difference in SDs ( $SD_{MA}$ - $SD_{top}$ ) .....                                                         | 115  |
| 4.1 Example R Documentation for <b>MAPP_bms</b> .....                                                           | 127  |
| 4.2 MAP Plot: Brain Weights of Mammals Study .....                                                              | 131  |

## LIST OF FIGURES - CONTINUED

| Figure |                                                              | Page |
|--------|--------------------------------------------------------------|------|
| 4.3    | MAP Plot: Body Fat Study.....                                | 138  |
| 4.4    | MAP Plot for IC Approximation.....                           | 141  |
| 4.5    | MAP Plot from <code>MApp_IC</code> .....                     | 142  |
| 4.6    | MAP Plot: <code>MApp_IC_gen</code> .....                     | 146  |
| 4.7    | MAP Plot: <code>MApp_MCMC</code> .....                       | 148  |
| 4.8    | MAP Plot: <code>MApp_MCMC</code> , ‘‘common3’’.....          | 149  |
| 4.9    | MAP Plot: <code>MApp_list</code> .....                       | 151  |
| 4.10   | Help File for the Package .....                              | 152  |
| 5.1    | Comparison of SDs: <i>g</i> -prior vs. AIC Weights.....      | 158  |
| C.1    | Example of <code>pres_d_plot</code> , Plotted X1:X5 .....    | 209  |
| C.2    | Example of <code>pres_d_plot</code> , Plotted X5:X1 .....    | 210  |
| C.3    | Example of <code>pres_y</code> , Plotted X1:X5.....          | 211  |
| C.4    | Example of <code>pres_trt_comp</code> .....                  | 212  |
| C.5    | Example of <code>pres_y_comp</code> .....                    | 213  |
| C.6    | Example of <code>final_bean_d</code> : Comparing Ratios..... | 214  |
| C.7    | Example of <code>final_bean_d</code> : Comparing SDs .....   | 215  |

## ABSTRACT

Model averaging (MA) was developed as a way to combine predictions from many models, with the goal of reducing bias and incorporating model uncertainty into final predictive inferences. A new flavor of MA, focused on averaging partial regression coefficients over multiple models, has gained traction in fields such as Ecology, Biology, and Political Science, with motivation stemming from the concern that inferences based on a single model are too “naive” (i.e., do not fairly reflect sources of substantial uncertainty). However, coefficients appearing in multiple models do not necessarily hold the same interpretation across models, and averaging over them has the potential to result in inferences that are difficult to interpret. A gap exists between the theoretical development of MA and its current use in practice, potentially leaving well-intentioned researchers with unclear inferences, or difficulties justifying decisions to use (or not use) MA. Furthermore, it is questionable whether the perceived benefit of accounting for an additional source of uncertainty is realized in terms of increased variance for quantities of interest.

In this work, we revisit relevant foundations of regression modeling, suggest more explicit notation and graphical tools, and discuss how individual model results are combined to obtain a MA result, with the goal of helping researchers make informed decisions about MA. We present a new package for R Statistical Software providing plotting functions for visualizing components going into the MA posterior distribution. This package is meant to be used to assess the implicit assumptions made by using MA for regression coefficients, complete with guidelines for use and examples. We also design and conduct a simulation study to investigate how the variance for a partial regression coefficient of interest is different for three different approaches used within multimodel inference (MA using all models, MA using a subset of models, and conditioning inferences on one model). We assess whether the perceived benefit of accounting for model uncertainty is actually realized when more models are used for final inference, with the goal of helping researchers weigh tradeoffs between using variants of MA in place of one well thought out model.

## CHAPTER ONE

## INTRODUCTION

The purpose of a statistical model is to approximate a complex system adequately enough to ensure inferences drawn from the model are useful for understanding (or making predictions about) relationships of interest. Statistical modeling is not generally designed to uncover absolute truth, and statisticians often cite George Box’s famous assertion, “all models are wrong but some are useful,” when explaining this concept to novice researchers. Despite acknowledging more than one model may be useful, inferences are typically conditioned on a single model and as a result, uncertainty inherent in choosing that model (model uncertainty) is ignored. Most commonly, a set of models is considered and the data are used twice: once for model selection, and *again* for estimating parameters within the selected model. This “double-dipping” of the data has received criticisms dating back to the late seventies (e.g., Hodges, 1987; Leamer, 1978). Model combination was mentioned by Barnard (1963) as a potential way to address model uncertainty and similar ideas gained traction in the economics literature in the 1970s.

Hooten and Hobbs (2015) define *multimodel inference* to be any analysis considering a set of models, regardless of whether the final inferences are based on a single model (i.e., model selection) or a combination of many models (i.e., model averaging). Formally conducting multimodel inference is most naturally approached from a Bayesian perspective (Hodges, 1987) where one can incorporate the model set (the collection of models being considered) as an additional random variable. Then,

prior knowledge about models can be expressed by assigning probabilities to each model in the model set such that they sum to one, thus defining the probability mass function for the model set. Priors must also be specified for parameters in each model, after which, a fully Bayesian inference can be conducted for all models simultaneously. The results are in the form of a posterior distribution for the model set defined by posterior model probabilities (PMPs) as well as the joint posterior distribution for the parameters for each model in the model set. The set of posterior model probabilities is typically used to measure the degree of model uncertainty, and can be used to compare models by means of Bayes Factors (e.g., Kass and Raftery, 1995), or as weights in a weighted average to combine quantities of interest across models, a method called model averaging.

In the 1970s and 1980s, methods simultaneously considering large model sets and attempting to account for model uncertainty were far too computationally complex to be implemented with the computing power available. The early 1990s marked a breakthrough in computational roadblocks, and with the growth and newfound feasibility of Markov chain Monte Carlo (MCMC) methods (see Robert and Casella, 2011, for a short history of MCMC in statistics), came an explosion of work in the form of algorithms for implementing Bayesian model selection and variable selection techniques, and subsequently, model averaging (e.g., George and McCulloch, 1993, 1997; Geweke, 1996; Green, 1995; Kuo and Mallick, 1998). After the algorithms were developed, a substantial amount of work went into improving computational efficiency (e.g., Clyde et al., 1996; Clyde, 1999; Clyde et al., 2011) and considering the choice of priors for the model set and the parameters in the Bayesian implementation (e.g., Chipman, 1996; Chipman et al., 2001; Feldkircher and Zeugner, 2009; Link and Barker, 2006). Hoeting et al. (1999a) wrote a hallmark paper providing the foundations for model averaging in generalized linear models, Clyde and George

(2004) provide a history of model uncertainty in the context of Bayesian inference, and Hooten and Hobbs (2015) provide a recent overview of Bayesian model selection methods commonly used in Ecology, touching briefly on model averaging.

Although the Bayesian framework may be the most natural way to think about multimodel inference, there are common alternatives to implementing a fully Bayesian approach. To avoid the computational challenges that can arise in a fully Bayesian implementation of multimodel inference, Kass and Raftery (1995) illustrate the approximation of posterior model probabilities from the Bayesian Information Criterion (BIC) (Kass and Wasserman, 1995; Schwarz, 1978). Burnham and Anderson (2002) advocate an approximation using AIC, though unlike the BIC approximation, the AIC approximation is not based on any relationship to Bayes Factors. The use of AIC and BIC approximations are particularly attractive to researchers because AIC and BIC are easily obtainable for linear, generalized linear, and non-linear regression models. The difference between AIC and BIC approximations lies in the theoretical foundation and also in the prior specifications on the model set. If using BIC weights, the user must explicitly place prior uncertainty on all models in the model set (often accomplished by assigning equal weights to all models). When AIC weights are used, a prior is implicitly placed on the model set with higher probabilities assigned to more complex models, often unbeknown to the user. Link and Barker (2006) and Burnham and Anderson (2004) both provide summaries of the implications of such priors.

Hereafter, we use model averaging as a general term to refer to averaging quantities over models, using either the fully Bayesian implementation or the approximations.

## 1.1 Current State of Model Averaging

The methodology laying the groundwork for model averaging was developed in the context of prediction, and model averaging has primarily been studied and applied in this context. The motivating concept was driven by the basic principle of averaging; provided there are reasonable predictions from each individual model, averaging those predictions should decrease bias. We believe many of the algorithms for obtaining posterior model weights were developed with this goal in mind. However, the results provided by common and easily implementable software allow the method to be extended almost effortlessly to problems with more of an explanatory goal, and researchers are now commonly using model averaging to “account for model uncertainty” in inference for regression coefficients.

Extending the use of model averaging from prediction to obtaining “overall effects” of explanatory variables (over all models in the model set) has become more commonplace. This is accomplished by averaging regression coefficients over models to obtain a posterior distribution of a model-averaged regression coefficient associated with an explanatory variable. However, the interpretation of a *partial* regression coefficient is dependent on the other variables included in the model and their degree of collinearity with one another. Therefore, the interpretation of a model-averaged partial regression coefficient is often much more complicated than an “overall effect” of its associated variable, and in general, model-averaged partial regression coefficients are not always appropriate. Shmueli (2010) discusses the consequences of conflating prediction and explanation in a general context, and in her conclusion, she states, “Awareness of the distinction between explanatory and predictive modeling, and of the different scientific functions that each serve, is essential for the progress of scientific knowledge.” In general, discussions about the benefits of model averaging have

blurred the distinction between predictive goals and explanatory goals (interpretation of coefficients), leading researchers to believe model averaging is always an advantage for any inferential goal.

In their text geared towards biologists, Burnham and Anderson (2002) advocate accounting for model uncertainty when interest lies in coefficient estimates (not just predictions) by way of model averaging. Similarly, Montgomery and Nyhan (2010) caution political scientists against **not** using Bayesian model averaging: “[not using BMA] systematically understates the uncertainty of our results, generates fragile model specifications, and leads to the estimation of bloated models with too many control variables.” And finally, in their widely cited paper, *Bayesian Model Averaging: A Tutorial*, Hoeting et al. (1999a) suggest that model averaging can help researchers with model selection by *avoiding* it entirely, saying “model averaging [also] avoids the problem of having to defend the choice of model, thus simplifying presentation.” These three examples highlight some of the opinions purported in the literature. Statements like these present model averaging as a low hanging fruit for researchers, without cautioning about potential complications for inference.

Being able to say model uncertainty was accounted for in an analysis is particularly appealing to researchers working in fields where the uncertainty in model choice is potentially very large. In our experience working with biologists, there is increasing discomfort with using a single model for inference due to criticisms from colleagues and reviewers about not accounting for model uncertainty. Therefore, we see researchers being attracted to model averaging, regardless of their research objectives, and often without a clear indication as to why the method was chosen for a particular application. We see this as a problem, and we agree, in general, with Hocking (2003), who states,

A fundamental problem in all areas of science, and especially in statistics, is the gap between the development of new ideas and their implementation in practice. In the current computing environment, the problem becomes even more serious as new ideas are introduced into readily available statistical packages but not necessarily understood by practitioners.

The potential for this particular problem in the context of model averaging is evidenced by a literature search of peer-reviewed articles accessible to Montana State University (August, 2016) including “model averaging” in their titles. As expected, there is an increasing number of citations meeting these criteria over the last 45 years (see Figure 1.1). Software became available in the late 1990s - early 2000s, and we see a noticeable increase in publications per year over the last 10 years ('06 - Current). Methodology has not changed drastically in the last 10 years, so we feel it is fair to conclude most of the newer publications are likely applications of model averaging. We cannot claim all of these newer applied papers use model averaging in the context of explanatory inference, but we suspect many do. The prevalence of model averaging in applied statistical analyses is increasing (e.g., Figure 1.1), and we notice a gap between the development of the methodology and the use of model averaging in practice that should be addressed. Our observations and concerns are consistent with those raised in other recent publications (e.g., Cade, 2015; Fieberg and Johnson, 2015; Ver Hoef and Boveng, 2015) cautioning against using model averaging to make inferences about partial regression coefficients and suggesting alternative approaches to circumnavigate the convoluted interpretation of a model-averaged partial regression coefficient.

## 1.2 Specific Aims for This Research

Model averaging is becoming more prevalent in the applied sciences, and it is being used outside of its original context without considering any of the implicit assumptions that come with the extension of use. To compound this, easily accessible

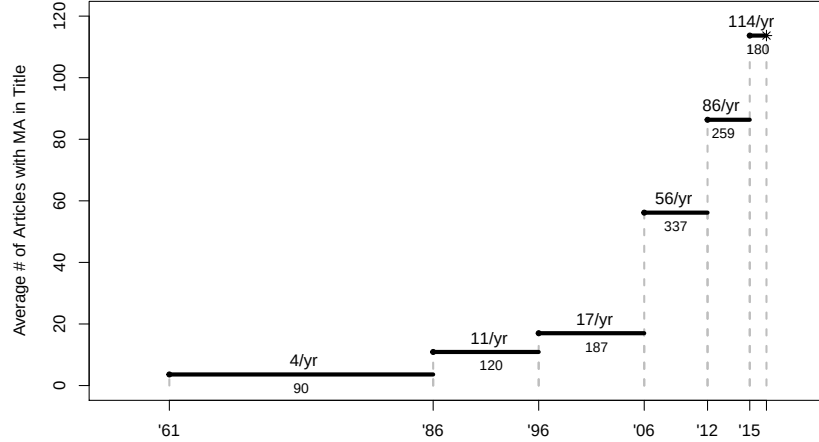


Figure 1.1: The average number of peer-reviewed journal articles published per year (rounded to the nearest integer) including model averaging in the title (1961-Current). Totals for the time periods are displayed below the line segments connecting start year to end year. The search was conducted on Montana State University’s Library catalog search in August, 2016.

software exists for implementing model averaging of regression coefficients, but there are no tools for assessing its appropriateness. This has created a gap that should be addressed so all researchers, regardless of statistical training, can make informed decisions about when to use model averaging.

Motivation for this research stemmed from our observations of model averaging being advocated, and used, outside of what we felt was its intended context. This inspired us to investigate the implicit assumptions required for model averaging partial regression coefficients, and also the motivations behind its use. The goals of this research are twofold: (1) to provide researchers much-needed tools and considerations for assessing and understanding the complex components model-averaged partial regression coefficients are comprised of; and, (2) to formally investigate the extent to which stated uncertainty in inferences increase as a result of model averaging.

We address these goals in the three subsequent chapters, each containing necessary background information and notation to be used therein.

Chapter 2 provides tools and considerations for assessing the appropriateness of model-averaged partial regression coefficients. We discuss model averaging in the common regression context and provide more insight into the difficult interpretation of the model-averaged regression coefficients. We use subscripted notation for partial regression coefficients and graphical tools for visualizing how the relationship between the mean response and a variable of interest may change across models. We provide researchers with necessary background information and tools to carefully consider whether model averaging is appropriate for their research objectives on a case-by-case basis. Using two examples, we introduce a tool we developed to aid in this process, the model-averaged posteriors plot (MAP plot). We developed an R (R Core Team, 2016) package to make this tool freely available to the public, which is the subject of Chapter 4.

It has been stated by many statisticians and researchers that model averaging should result in increased variance for quantities of interest because of its ability to account for model uncertainty in the analysis. Statements about how model uncertainty *should* propagate through an analysis to affect variances of quantities of interest have either been general, or specifically focused on prediction as the quantity of interest. However, common motivation for model averaging with regression coefficients is that the uncertainty surrounding parameters of interest would otherwise be too small. In Chapter 3 we investigate this claim using a simulation study in the linear regression context.

The software to create the MAP plot presented in Chapter 2 is part of an R (R Core Team, 2016) package we created for researchers who are considering model averaging of partial regression coefficients. This package is called **MApp**, model-

averaged posteriors plot, and includes functions to create the MAP plot for many different implementations of model averaging. Chapter 4 contains an extended vignette for this package, demonstrating the use of the functions in the package, and examples of the types of information that can be gleaned from the plots.

For computational convenience, and as a starting point for our investigations, we restricted our context to multiple linear regression problems with no higher order terms or interactions. Much, if not all of the work presented in this dissertation can be extended to the generalized linear model framework. Chapter 5 concludes our work, discusses ideas for extending this work beyond its restricted context, and presents directions for future research.

## CHAPTER TWO

### CONSIDERATIONS FOR ASSESSING MODEL AVERAGING OF REGRESSION COEFFICIENTS

#### Contribution of Authors and Co-Authors

Author: Katharine Banner

Contributions: Wrote the majority of the manuscript and R code.

Co-Author: Megan Higgs

Contributions: Provided feedback on statistical analysis and drafts of the manuscript and R code.

Manuscript Information Page

Katharine M. Banner and Megan D. Higgs

Ecological Applications

Status of Manuscript:

\_\_\_\_\_ Prepared for submission to a peer-reviewed journal

\_\_\_\_\_ Officially submitted to a peer-review journal

\_\_\_\_\_ Accepted by a peer-reviewed journal

  X   Published in a peer-reviewed journal

Publisher: Springer

Submitted: April 21, 2016; Published Online: August 8, 2016; <http://onlinelibrary.wiley.com/doi/10.1002/eap.1419/>

Permission: Copyright by the Ecological Society of America:

Banner, K. M. and Higgs M. D. (2016). Considerations for assessing model averaging of regression coefficients. *Ecological Applications*. <http://onlinelibrary.wiley.com/doi/10.1002/eap.1419/>

## Abstract

Model choice is usually an inevitable source of uncertainty in model-based statistical analyses. While the focus of model choice was traditionally on methods for choosing a single model, methods to formally account for multiple models within a single analysis are now accessible to many researchers. The specific technique of model averaging was developed to improve predictive ability by combining predictions from a set of models. However, it is now often used to average regression coefficients across multiple models with the ultimate goal of capturing a variable’s overall “effect.” This use of model averaging implicitly assumes the same parameter exists across models so that averaging is sensible. While this assumption may initially seem tenable, regression coefficients associated with particular explanatory variables may not hold equivalent interpretations across all of the models in which they appear, making explanatory inference about covariates challenging. Accessibility to easily implementable software, concerns about being criticized for ignoring model uncertainty, and the chance to avoid having to justify choice of a final model have all led to the increasing popularity of model averaging in practice. We see a gap between the theoretical development of model averaging and its current use in practice, potentially leaving well-intentioned researchers with unclear inferences or difficulties justifying reasons for using (or not using) model averaging. We attempt to narrow this gap by revisiting some relevant foundations of regression modeling, suggesting more explicit notation and graphical tools, and discussing how individual model results are combined to obtain a model-averaged result. Our goal is to help researchers make informed decisions about model averaging and to encourage question-focused modeling over method-focused modeling.

## 2.1 Introduction

In practice, there typically exists more than one reasonable model as a basis for statistical inference, and for over 50 years the potential implications of ignoring uncertainty in the process of choosing an inferential model have been discussed. Leamer (1978) cautioned that inferences conditional on one model may result in inflated precision for estimates and predictions. Similarly, Hodges (1987) explicitly laid out three types of uncertainty he argued should be addressed in any analysis:

uncertainty in structure (model), uncertainty in parameter estimates conditional on model, and uncertainty in measurement (inherent in data collection).

Algorithms for implementing Bayesian model selection and variable selection were developed in the late 1990s and the formal assessment of model uncertainty became possible (e.g., George and McCulloch 1993, 1997; Geweke 1996; Green 1995; Hoeting et al. 1999a; Kuo and Mallick 1998; Raftery et al. 1997a). A substantial amount of work went into improving algorithms' computational efficiency (e.g., Clyde et al. 1996; Clyde 1999; Clyde et al. 2011), and exploring their sensitivities to different prior specifications (e.g., Chipman 1996; Chipman et al. 2001; Feldkircher and Zeugner 2009; Link and Barker 2006). Relatively little work has gone into assessing the practical implications of incorporating model uncertainty into analyses, particularly for explanatory (rather than predictive) goals. Concerns raised previously have received little formal attention or are still unsettled within the statistical community, as evidenced by recent publications (Cade 2015; Fieberg and Johnson 2015; Hooten and Hobbs 2015; Ver Hoef and Boveng 2015).

Bayesian model averaging was originally developed as a method for improving out-of-sample predictions by combining predictions from multiple models with weights based on their posterior model probabilities. Algorithms for obtaining posterior model probabilities were developed with this predictive goal in mind. Software to implement model averaging has developed to the point where researchers can almost automatically obtain results regardless of their statistical backgrounds, and the potential benefits of model averaging are being advertised in a very broad sense. In general, discussions about the benefits of model averaging have blurred the distinction between predictive and explanatory inference, leading researchers to believe model averaging is always an advantage. This is concerning because the interpretation of partial regression coefficients can depend on other variables that have been included

in the model, so averaging regression coefficients across models may not be practically meaningful.

In their text geared towards biologists, Burnham and Anderson (2002) advocate for accounting for model uncertainty using AIC-based (non-Bayesian) model averaging in problems where the estimates of coefficients are of primary interest (not just for prediction). Similarly, Montgomery and Nyhan (2010) advocate for Bayesian model averaging for political scientists arguing, it can “help applied researchers to ensure that their estimates of effects of key independent variables are robust to a wide range of possible model specifications.” In their widely cited paper, *Bayesian Model Averaging: A Tutorial*, Hoeting et al. (1999a) suggest that model averaging for prediction can help researchers avoid having to defend a particular choice of model, with the benefit of a simplified presentation. While these are attractive qualities for predictive inference, we do not believe these statements were intended to extend to model averaging of regression coefficients.

The advertised allures of model averaging are strong and we see researchers being pulled toward the method regardless of their research objectives. Our observations and concerns are, in general, consistent with those recently raised by Cade (2015), who argues, the use of model averaging of regression coefficients may result in misleading inferences while leaving researchers “strangely, feeling satisfied that model uncertainty has been addressed.” Cade (2015) suggests using partial standard deviations to adjust the regression coefficients for their changing scales among models (due to multicollinearity).

There are many things a researcher must consider before deciding if model averaging is useful and appropriate for a particular problem (Fig. 2.1). After the decision to use model averaging has been made, there are additional considerations and decisions to be made about how model averaging will be implemented, such as

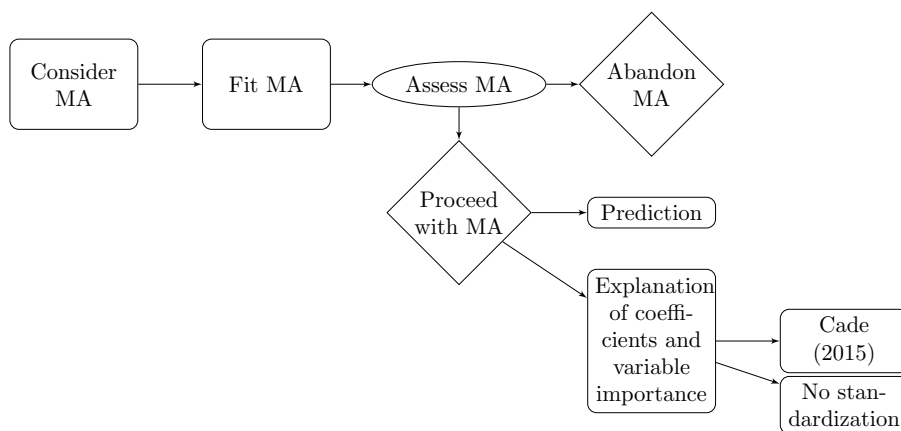


Figure 2.1: This flow chart depicts steps a researcher could take when considering model averaging. Much of the model averaging literature falls under the rectangular nodes. Articles related to algorithms and computational efficiency fall under ‘Fit MA’ (e.g., Barker and Link 2013; Clyde 2012, 1999; Clyde et al. 2011; Feldkircher and Zeugner 2009; Link and Barker 2006), articles related to predictive performance fall under ‘Prediction’ (e.g., Hoeting et al. 1999a,b; Raftery et al. 1997b), and literature related to using model averaging for explanatory inferences falls under ‘Explanation of coefficients and variable importance’ (e.g., Burnham and Anderson 2002; Cade 2015). Some of the literature raises concerns with particular aspects of model averaging and multimodel inference (e.g., Cade 2015; Fieberg and Johnson 2015; Ver Hoef and Boveng 2015). This paper falls under the ‘Assess’ node.

whether to employ the standardization methods described in Cade (2015). Much of the model averaging literature falls into categories represented with rectangles in Fig. 1 (e.g., Consider MA, Fit MA, etc.), with little falling into the Assess MA category. In this paper, we focus on the assessment step in the model averaging process.

Assessing the appropriateness of model averaging must be done on a case-by-case basis as it is difficult to understand the model-averaged result without understanding how the individual model results combine to create it. We look closer at the difference between model averaging predictions and regression coefficients, review foundations of linear regression, suggest helpful notation, and introduce graphical tools to help understand how results from individual models are combined to form model-averaged

results. We also highlight the importance of making modeling decisions in the context of the research questions, which we term question-focused modeling. These considerations help researchers make informed decisions about model averaging on a case-by-case basis, and provide a foundation for arguing against its use in cases where it is unnecessary or inappropriate.

We use two examples (Sections 2.3 & 2.4) with different analysis objectives to show how these considerations and graphical tools can be used in practice. In our first example, we demonstrate the practical importance of considering whether the added complexity of model averaging is worth the potential gains, particularly if inferences change very little compared to conditioning on one reasonable model. We highlight question-focused modeling and illustrate advantages it can have over model averaging. In our second example, we point out the often hidden secondary goals of analyses even when the stated goal is prediction. We use both examples to illustrate how we can assess the implications of the model averaging process relative to the problem at hand.

The information in this paper should be relevant to a broad spectrum of researchers, ranging from those with little statistical background to quantitative ecologists and applied statisticians. The information presented is meant to be foundational for some, a refresher for others, and for all, an aid for assessing when and why model averaging might be used as an effective research tool.

## 2.2 Model Averaging Within Multimodel Inference

Multimodel inference is an umbrella term for incorporating multiple models into a single analysis, including both model selection (selection of a model for inference from a clearly defined set of models) and model combination (e.g., model averaging) (Hooten and Hobbs, 2015). To conduct multimodel inference, a model set  $\mathcal{M} \equiv$

$\{M_1, M_2, \dots, M_J\}$  must be defined, where  $J$  is the total number of models considered. Ideally,  $\mathcal{M}$  is specified using sound science and expert knowledge prior to observing the data (e.g., Burnham and Anderson, 2002). However, for processes that are not well understood and do not have a pre-defined set of potential models, it is common for researchers to define  $\mathcal{M}$  as the set of all possible regression models made up from combinations of a set of potential input variables. The consideration of all first order combinations of quantitative input variables without any interactions is commonly referred to as *all subsets* regression and is the default in multiple software packages. Other strategies for choosing model sets are discussed elsewhere (e.g., Doherty et al., 2012).

Model averaging was originally developed in a Bayesian framework, where all unknowns are modeled with probability distributions and the hierarchy stemming from models and parameters within models is naturally incorporated. Knowledge about parameters *before* data are collected (prior) is combined with information from the data (through the likelihood) to form posterior distributions for all unknowns of interest, as opposed to focusing only on a likelihood function for inference. Including the model set  $\mathcal{M}$  as an unknown allows model uncertainty to be directly incorporated into an analysis. Prior distributions must be placed on  $\mathcal{M}$  and all of the parameters in each model. A prior for the discrete random variable  $\mathcal{M}$  is defined by assigning prior probabilities to each model in  $\mathcal{M}$ , such that  $p(\mathcal{M}) = \{Pr(M_1), Pr(M_2), \dots, Pr(M_J)\}$ , and  $\sum_{j=1}^J Pr(M_j) = 1$ . Similarly, we denote the joint prior distribution on the parameters in the  $j^{th}$  model as  $p(\boldsymbol{\theta}_j|M_j)$ , where  $\boldsymbol{\theta}_j$  is the parameter vector for the  $j^{th}$  model (including regression coefficients and any other parameters in model  $M_j$ , such as, but not limited to, (co)variance terms). The collection of posterior model probabilities  $\{Pr(M_j|\mathbf{y})$  for  $j = 1, 2, \dots, J\}$  defines the posterior distribution for  $\mathcal{M}$ , conditional on the observed data ( $\mathbf{y}$ ). The posterior model probability for an

individual model is connected to the prior model probability and the distribution  $p(\mathbf{y}|M_j)$  through Bayes' Theorem:

$$Pr(M_j|\mathbf{y}) = \frac{p(\mathbf{y}|M_j)Pr(M_j)}{\sum_{j=1}^J p(\mathbf{y}|M_j)Pr(M_j)} \quad (2.1)$$

$$\text{where } p(\mathbf{y}|M_j) = \int_{\boldsymbol{\theta}_j} p(\mathbf{y}|\boldsymbol{\theta}_j, M_j)p(\boldsymbol{\theta}_j|M_j)d\boldsymbol{\theta}_j. \quad (2.2)$$

Specifically, the likelihood function associated with model  $M_j$  is based on  $p(\mathbf{y}|\boldsymbol{\theta}_j, M_j)$ , and  $p(\mathbf{y}|M_j)$  is obtained by integrating over the posterior distribution of the parameters in  $M_j$ . The integration is typically intractable and therefore done computationally. Note in Eq. 2.2 that the prior distribution for the parameters is contributing to  $p(\mathbf{y}|M_j)$ , which in turn contributes to the posterior model probabilities.

To provide clear meaning to the posterior model probabilities, it is natural to assume “truth” is in the model set so that the posterior model probabilities represent the probability that a model is true, given the data and the priors. While this assumption is surely false in practice, we appeal to the rationale presented in Link and Barker (2006) and Barker and Link (2015) as to why it is reasonable to proceed. Collectively, the posterior model probabilities are also used to combine posteriors of quantities of interest from multiple models with the goal of incorporating model uncertainty into inferences, and are also often used to discuss the degree of uncertainty in the model set.

Let  $\phi$  be a quantity of interest, such as a prediction of a new observation or a regression coefficient. To generally define a posterior distribution for a model-averaged  $\phi$ , the marginal posterior distributions of  $\phi$  from individual models in  $\mathcal{M}$  are combined using posterior model probabilities to form the distribution commonly written as

$$p(\phi_{MA}|\mathbf{y}) = \sum_{j=1}^J Pr(M_j|\mathbf{y})p(\phi|M_j, \mathbf{y}). \quad (2.3)$$

Eq.2.3 is useful conceptually, but sometimes, as we will see in Eq. 9, the posterior distribution for a model-averaged  $\phi$  can be more complicated than it appears. For more details about the formal setup of Bayesian model averaging that is accessible to ecologists and environmental scientists we refer the reader to Link and Barker (2010).

In this paper, we present the fully Bayesian implementation of multimodel inference, but the concepts and discussion can be extended to the non-Bayesian approaches through a focus on combining point estimates and their standard errors, rather than on parameters and their posterior distributions. Formulas for the non-Bayesian context can be found elsewhere (e.g., Burnham and Anderson 2002; Link and Barker 2010). The broad use of AIC among ecologists has naturally led to its popularity for use in model averaging, and while AIC and BIC approximations initially seem a simpler option, practitioners should understand the implications of their choice for results (see Hooten and Hobbs 2015; Link and Barker 2006).

### 2.2.1 Model Averaging in Regression

In the context of linear regression, including generalized linear regression, model uncertainty typically enters through the variable selection process. For *all subsets* regression with  $p$  first-order input variables,  $\mathcal{M}$  contains  $J = 2^p$  elements, and a common motivating question is, which of  $p$  potential input variables should define the model(s) ultimately used for inference? Here, we follow convention and focus on the case where only first order terms of  $p$  quantitative input variables are considered. The issues we discuss in the context of this common convention are even more relevant and complicated when higher order terms such as interactions and polynomials are

considered. Thus, the challenges in implementing Bayesian model averaging with higher order terms often naturally leads to their exclusion, and we discuss concerns with this in Section 2.3.

Here, we use typical regression notation, defining  $\mathbf{y}$  as a  $(n \times 1)$  column vector of observations of a response variable and  $\mathbf{X}$  as a  $(n \times (p + 1))$  matrix with a column vector of 1s for the intercept and additional column vectors for the  $p$  potential input variables,  $\mathbf{X} = [\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p]$ . Let  $\beta_j$  be the  $(p + 1)$  column vector of regression coefficients associated with  $M_j$ . Assuming normal, independent errors with constant variance ( $\mathbf{y} \sim N_n(\boldsymbol{\mu}\{\mathbf{y}|\mathbf{X}\}, \sigma^2 I_{n \times n})$ ), we have our familiar multiple linear regression model. The mean of  $\mathbf{y}$  is modeled as a function of the first order terms of the  $p$  input variables and the intercept. More generally, it can be assumed that  $\mathbf{y}$  follows another distribution (such as Binomial or Poisson), and the mean of  $\mathbf{y}$  can be connected to the input variables through a link function  $g()$ ,

$$g(\boldsymbol{\mu}\{\mathbf{y}|\mathbf{X}\}) = \beta_{0,j} + \beta_{1,j}\mathbf{x}_1 + \beta_{2,j}\mathbf{x}_2 + \dots + \beta_{p,j}\mathbf{x}_p = \mathbf{X}\beta_j, \quad (2.4)$$

where  $g()$  is simply the identity function for multiple linear regression.

We include an index for model in our notation for regression coefficients to make it explicit that the meaning of the regression coefficient associated with input variable  $X_v$  in model  $M_j$  ( $\beta_{v,j}$ ) is model dependent. All models in  $\mathcal{M}$  can be defined by excluding certain input variables from the fullest model, which is equivalent to setting the  $\beta_{v,j}$ 's associated with those variables equal to 0.

A considerable amount of work has been put into investigating the implications and limitations of different prior specifications for  $\mathcal{M}$  and  $\boldsymbol{\theta}_j$  in a regression context (e.g., Carlin and Chib 1995; Feldkircher and Zeugner 2009; George and McCulloch

1993, 1997; Geweke 1996; Kuo and Mallick 1998; Raftery et al. 1997a), and also the implementation of Markov chain Monte Carlo (MCMC) and other sampling schemes such as Bayesian adaptive sampling and reversible jump MCMC (e.g., Barker and Link 2013; Clyde et al. 2011; Green 1995; Link and Barker 2006, 2010). In depth discussions or critiques of the different methods available are beyond the scope of this paper. We focus on understanding how model averaging may be used when the default (or easier to use) priors are chosen for implementation.

One of the most common priors for  $\mathcal{M}$  is the discrete uniform prior, which places equal prior weight on each of the models considered,  $Pr(M_j) = 1/J$  for  $j = 1, 2, \dots, J$ . Assuming normal likelihoods, forms of the conjugate normal prior for regression parameters are commonly used because of their computational advantages and availability in R (R Core Team, 2016). One such prior is Zellner’s  $g$ -prior (Feldkircher and Zeugner, 2009; Zellner, 1984), which specifies a vague multivariate normal prior on the partial regression coefficients in each model, and an improper uniform prior on the standard deviation parameter. Specification of one hyper-parameter,  $g$ , allows the researcher to control how diffuse the prior will be, and the resulting posterior distributions for the regression coefficients from an individual model are multivariate  $t$  distributions (See Appendix S1 [Appendix A] for details).

Knowing the form of the posterior distributions under the  $g$ -prior greatly simplifies the implementation of multimodel inference from a computational perspective because it only requires approximating the first two moments, rather than the whole posterior distribution. However, the validity of the multivariate  $t$  form of the posterior distributions is contingent upon no severe violations of regression assumptions for *all* models in  $\mathcal{M}$ . In the *all subsets* setting, the size of  $\mathcal{M}$  grows exponentially with the number of input variables considered, which can make model checking prohibitive

even for moderately sized  $p$ . This poses a major limitation for multimodel inference, as recently pointed out by Ver Hoef and Boveng (2015).

### 2.2.2 Model Averaging of Predictions

The original motivation for the development of model averaging was to improve prediction, with the quantity of interest being a new observation at specified values of the input variables, denoted  $\phi_j = \tilde{y}_j$  for model  $M_j$  (consistent with Gelman et al. 2013). Although the values of predictions will change among models,  $\tilde{y}_j$  holds the same meaning for all models and can be directly compared among models, or combined in a weighted average to provide an average prediction over models. It has been demonstrated that model averaging can be advantageous under many out-of-sample prediction criteria (e.g., Hoeting et al. 1999a; Raftery et al. 1997a).

Within Bayesian inference, we obtain a posterior distribution of predictions from  $M_j$  for a particular set of input variables in  $\mathbf{X}^{new}$ . This posterior predictive distribution  $p(\tilde{y}|\mathbf{y})$  is defined by

$$p(\tilde{y}_j|\mathbf{y}, M_j) = \int_{\boldsymbol{\theta}_j} p(\tilde{y}_j|\boldsymbol{\theta}_j, M_j)p(\boldsymbol{\theta}_j|\mathbf{y}, M_j)d\boldsymbol{\theta}_j \quad (2.5)$$

and it describes the current knowledge about the prediction by combining uncertainty in the parameters of the model with variability coming from the distribution defining the likelihood (i.e., the two components on the right of Eq. 2.5).

We can think of the weighted average  $\tilde{y}_{MA} = \sum_{j=1}^J Pr(M_j|y)\tilde{y}_j$  as a model-averaged prediction in both the Bayesian and non-Bayesian context. In the Bayesian context,  $\tilde{y}_{MA}$  is an unknown of interest with a distribution. Specifically, its distribution is the mixture distribution of the posterior predictive distributions from the individual models, with weights equal to the posterior model probabilities (as in Eq. 2.3). In the non-Bayesian context, the point prediction of  $\tilde{y}_{MA}$  is constructed

using the point predictions from the individual models, and formulas are available to obtain the approximate standard error of a model-averaged point prediction (see Burnham and Anderson 2002; Link and Barker 2010).

This approach of directly averaging predictions works for linear and non-linear models, and even for combining predictions from models with different forms. However, for normal linear models, a shortcut using the regression coefficients has become the most common way to obtain model-averaged predictions because it bypasses the need to actually obtain predictions (or posterior predictive distributions) for the  $J$  models. This shortcut is particularly useful when there are many more models than there are input variables. For normal linear models, the posterior prediction from  $M_j$  can be written  $\tilde{y}_j = \mathbf{X}^{new} \boldsymbol{\beta}_j$ , leading to,

$$\begin{aligned} \tilde{y}_{MA} &= \sum_{j=1}^J Pr(M_j|y) \tilde{y}_j \\ &= \sum_{j=1}^J Pr(M_j|y) [\mathbf{X}^{new} \boldsymbol{\beta}_j], \end{aligned} \tag{2.6}$$

$$= \mathbf{X}^{new} \left[ \sum_{j=1}^J Pr(M_j|y) \boldsymbol{\beta}_j \right] = \mathbf{X}^{new} \boldsymbol{\beta}_{MA}. \tag{2.7}$$

Again, in a Bayesian setting  $\boldsymbol{\beta}_{MA}$  describes unknowns with distributions. In a non-Bayesian setting, carrying through point predictions and point estimates for regression parameters provides the same endpoint and illustrates the shortcut.

To better understand the equality in Eq. 2.7, we consider an example with two possible predictors, for which the *all subsets* model set is defined by four models: (intercept only,  $X_1$ ,  $X_2$ , or  $X_1$  and  $X_2$ ) where  $j = 1, 2, 3, 4$  respectively. The vectors of partial regression coefficients for each model can be written:  $\boldsymbol{\beta}_1 = (\beta_{0,1}, 0, 0)^T$ ,  $\boldsymbol{\beta}_2 = (\beta_{0,2}, \beta_{1,2}, 0)^T$ ,  $\boldsymbol{\beta}_3 = (\beta_{0,3}, 0, \beta_{2,3})^T$ , and  $\boldsymbol{\beta}_4 = (\beta_{0,4}, \beta_{1,4}, \beta_{2,4})^T$ . We rewrite Eqs.

2.6 and 2.7 to explicitly show the connection between model averaging of predictions and this shortcut through model-averaged regression coefficients (for this equation we denote  $Pr(M_j|\mathbf{y}) = w_j$ ),

$$\tilde{y}_{MA} = \mathbf{X}^{new} \begin{pmatrix} w_1\beta_{0,1} + w_2\beta_{0,2} + w_3\beta_{0,3} + w_4\beta_{0,4} \\ w_1 \times 0 + w_2\beta_{1,2} + w_3 \times 0 + w_4\beta_{1,4} \\ w_1 \times 0 + w_2 \times 0 + w_3\beta_{2,3} + w_4\beta_{2,4} \end{pmatrix} = \mathbf{X}^{new} \begin{pmatrix} \beta_{0,MA} \\ \beta_{1,MA} \\ \beta_{2,MA} \end{pmatrix}. \quad (2.8)$$

We suspect the current emphasis on interpreting model-averaged partial regression coefficients (as opposed to predictions) evolved from this practice. Model averaged regression coefficients are mixtures of regression coefficients from different models (e.g.,  $\beta_{1,MA} = w_2\beta_{1,2} + w_4\beta_{1,4}$  in Eq. 2.8), and while the natural desire is for an “overall effect,” we need to carefully think about what is actually represented by the weighted average. To build a foundation for this discussion, we revisit the properties of partial regression coefficients in the context of linear regression.

### 2.2.3 Model Averaging of ***Partial*** Regression Coefficients

A concept critical to understanding the potential challenges with interpreting model-averaged regression coefficients is that of *partial* regression coefficients. An appreciation of the meaning of the term *partial* is commonly overlooked, forgotten, or not given adequate attention in introductions to regression; it reflects the potential change in meaning of a regression coefficient associated with a particular variable when the other variables in the model change. To help illustrate this, and discuss the challenges in the context of model averaging, we use a subset of data from Sacher and Staffeldt (1974) also used as an example in *The Statistical Sleuth* (Ramsey and Schafer, 2013). These data are average values for brain weight (g), gestation length

(days), and body size (kg) for 96 species of mammals. Natural log transformations on average brain weight ( $lbrain$ ), body size ( $lbody$ ), and gestation length ( $lgest$ ) were performed as the relationships are approximately linear on the log-log scale.

In this example, primary interest is in the relationship between gestation length and mean brain weight. Given the available variables, and the common model averaging choice of not including higher order terms, we consider two models representing two distinctly different research goals: (1) we can investigate the overall relationship between gestation length and mean brain size using  $\mu\{lbrain|M_g\} = \beta_{0,g} + \beta_{1,g}lgest$ , and (2) we can investigate the relationship between gestation length and mean brain weight conditional on body size using  $\mu\{lbrain|M_{gb}\} = \beta_{0,gb} + \beta_{1,gb}lgest + \beta_{2,gb}lbody$ .

As one would suspect, body size and gestation length are collinear on the log-log scale ( $r = 0.85$ ). For those taught to view multicollinearity as having only negative consequences, this degree of collinearity may seem alarming. However, the more interesting of the two questions (from a biological point of view) is the second. In an observational study like this one, it would have been impossible for the researchers to experimentally control body size while manipulating gestation length of mammals. However, using regression, we can assess evidence of a relationship between gestation length and mean brain weight, after accounting for the strong (but not very interesting) relationship between body size and brain weight. We fully accept the larger posterior variances (or standard errors) that are a consequence of having less information in the data to estimate the relationship of interest conditional on body size; it is a small price to pay to be able to address the specific question of interest.

Returning to the two models, we do not expect the relationship between gestation length and mean brain weight *over all the mammal species* to be the same as the relationship between gestation length and mean brain weight *for mammal species with*

*the same (or similar) body sizes.* In other words, we do not expect the parameters  $\beta_{1,g}$  and  $\beta_{1,gb}$  to be equal. Note that with no interaction in the model, we are assuming a common  $\beta_{1,gb}$  applies to all body sizes (i.e., the relationship between gestation length and mean brain weight is the same across all body sizes). To gain intuition for the potential difference in meaning between these two parameters, we create arbitrary categories of body size and estimate the slope and intercept within each category (Fig. 2.2). This visual aid helps us compare the estimated relationships between gestation length and brain weight for groups of mammals with similar body sizes to the estimated relationship observed over all mammals (ignoring body size). As expected, the estimated relationship changes when we group by similar body sizes, though the separate slopes look very similar, which is consistent with the exclusion of the interaction term in the second model.

Using exploratory tools like Fig. 2.2 can be extremely informative when there are two collinear variables and we want to control for one of them while making inference about the other. Similar tools are partial residual (or added variable) plots and partial regression plots (presented in Cade, 2015), which can aid in understanding partial relationships captured in estimated coefficients for more than two explanatory variables. Although the scatterplot is limited to two explanatory variables, it provides a more intuitive way to visualize and illustrate the different information used to estimate the two different regression coefficients.

Conventional notation for regression coefficients does a poor job distinguishing between  $\beta_{1,g}$  and  $\beta_{1,gb}$  by labeling both parameters  $\beta_1$ , implying  $\beta_{1,g} = \beta_{1,gb}$ , which is rarely true. In general, the only case for which partial regression coefficients associated with a particular explanatory variable hold the same interpretation across models is when the explanatory variables are orthogonal. This case is only realized in balanced experimental designs and other settings where variables are constructed

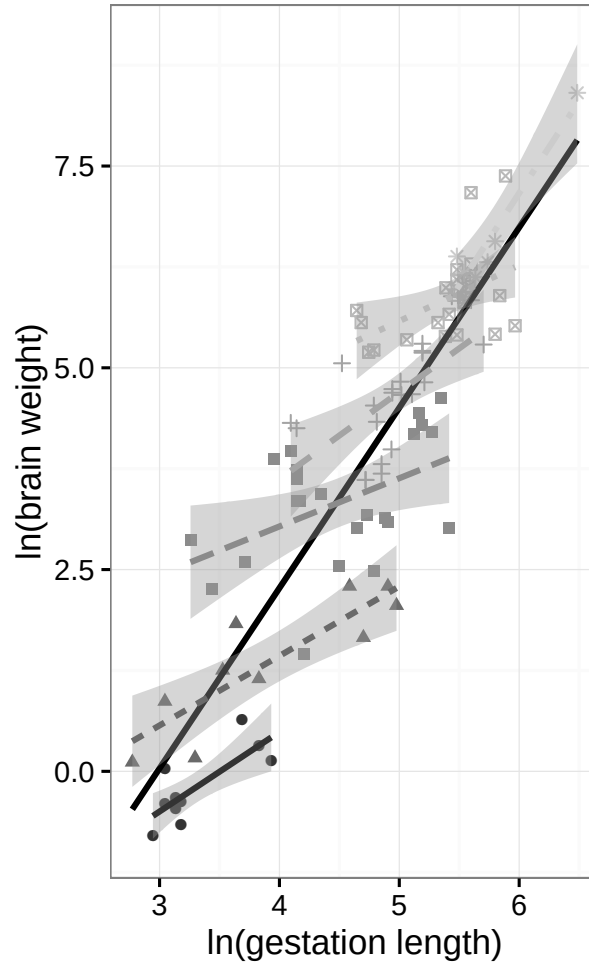


Figure 2.2: Interest lies in the relationship between  $lgest$  and  $lbrain$ , so we categorize the continuous variable we wish to account for,  $lbody$ , into subgroups of similar body size and create a scatterplot with points coded by these groups. We call this type of plot a *cut-and-coded scatterplot*. The corresponding lines represent the estimated relationships between  $lgest$  and  $lbrain$  conditional on the subgroups (we can think about the average slope across the body size subgroups as an approximate estimate of  $\beta_{1,gb}$ ). This figure was created using the `ggplot` function in the `ggplot2` R package (R Core Team, 2016; Wickham, 2009). Code is available in Appendix S2 [Appendix B]. A script file can be found in the Appendix DataS1 (Rcode-appendix-MAPPP.R) [online supplementary materials from Banner and Higgs (2016)].

to be orthogonal (e.g., principle components, orthogonal polynomials); it is very rare in observational studies. Therefore, we recommend explicit notation connecting a partial regression coefficient to the model it is defined in, as also suggested by others (e.g., Cade 2015; Hocking 2003; Jewell 2004).

Vague notation has also contributed to the common mistake of misinterpreting partial regression coefficients as “effects” of the explanatory variable they precede, regardless of which variables are in the model. We believe it also contributes to the attractiveness of capturing an “overall effect” through model averaging. Often, users fail to consider that (1) the meaning of the relationship associated with a partial regression coefficient is not exchangeable across models, and (2) the word “effect” implies a causal relationship between the input variable and the response, which is often inappropriate for observational studies without paying careful attention to principles and methods of causal modeling. We prefer interpreting partial regression coefficients as the linear relationship between the input variable and the mean response, after accounting for the other variables in the model.

### 2.2.3.1 Two Components of the Model Averaged Posterior Distribution

The total probability of the posterior distribution of a model-averaged regression coefficient is made up of two components, one coming from the models where the coefficient is set to zero, and the other coming from the models where the coefficient is estimated. We explicitly include both components by differentiating between the subset of models *including* the  $v^{th}$  input variable ( $\mathcal{M}_v = \{M_j | \beta_{v,j} \neq 0\}$ ), and its complement, the subset of models *excluding* the  $v^{th}$  input variable ( $\overline{\mathcal{M}}_v = \{M_j | \beta_{v,j} = 0\}$ ). For the remainder of the paper, we will refer to the posterior distribution conditional on  $\mathcal{M}_v$  as the *continuous component*, and the part conditional on  $\overline{\mathcal{M}}_v$

as the *zero component* (describing a point mass at zero). The notation may seem cumbersome, but it is important to formally represent both pieces because inferences depend on which components are used. This property of the model-averaged posterior distribution is not intuitively obvious, and is rarely discussed in detail in the Bayesian model averaging literature, with the exception of Hoeting et al. (1999a) who explicitly show it graphically.

The posterior distribution of the model-averaged partial regression coefficient associated with the input variable  $\mathbf{x}_v$  is a mixture distribution, which can be written as

$$p(\beta_{v,MA}|\mathbf{y}) = \begin{cases} \sum_{j \in \overline{\mathcal{M}}_v} Pr(M_j|\mathbf{y}) & \text{for } \beta_{v,MA} = 0 \\ \sum_{j \in \mathcal{M}_v} Pr(M_j|\mathbf{y})p(\beta_{v,j}|\mathbf{y}, M_j) & \text{for } \beta_{v,MA} \neq 0. \end{cases} \quad (2.9)$$

It is most typical to see reference to the two parts separately, and for different reasons; the continuous component is used for summarizing “overall effects,” and the probability associated with the zero component is used for measuring variable “importance.” The complement of the probability associated with the zero component is the posterior probability that the variable is included, termed the posterior inclusion probability (PIP). The PIP for input variable  $\mathbf{x}_v$  is  $PIP_v = 1 - \sum_{j \in \overline{\mathcal{M}}_v} Pr(M_j|\mathbf{y})$ , and a variable is often referred to as “important” if its PIP is greater than or equal to 0.5 (threshold suggested in the AIC-based context, see Burnham 2015; Burnham and Anderson 2002).

The zero component can be ignored without any practical consequences when using model averaging for prediction, and so it is often forgotten or omitted when the model averaging is extended to regression coefficients. When model averaging regression coefficients, it is important to be aware of the point mass at zero, which is

part of the posterior distribution and therefore is included when performing Bayesian model averaging. The inclusion of the zeros leads to shrinkage of regression coefficients toward zero. However, a common version of the AIC-based method for model averaging regression coefficients conditions on  $\mathcal{M}_v$ , and therefore ignores the zero component in constructing the average (see Burnham and Anderson 2002; Cade 2015).

To illustrate the difference, suppose we are interested in three models with the following forms:  $M_1 : \beta_{0,1}$ ,  $M_2 : \beta_{0,2} + \beta_{1,2}\mathbf{x}_1$ , and  $M_3 : \beta_{0,3} + \beta_{1,3}\mathbf{x}_1 + \beta_{2,3}\mathbf{x}_2$ . Let the posterior distribution for  $\mathcal{M}$  be  $p(\mathcal{M}|\mathbf{y}) = \{0.34, 0.33, 0.33\}$ . To model average the partial regression coefficients associated with  $\mathbf{x}_1$  over all models ( $\mathcal{M}$ ), we use both components to obtain  $\beta_{1,MA} = (0.34 \times 0) + (0.33 \times \beta_{1,2}) + (0.33 \times \beta_{1,3})$ . When users choose to average over only the continuous component, the weights are normalized over the models in  $\mathcal{M}_v$  to obtain  $\beta_{1,MA_{\mathcal{M}_v}} = (0.5 \times \beta_{1,2}) + (0.5 \times \beta_{1,3})$ . Again, this latter type of weighting is atypical in Bayesian model averaging, but common in AIC-based averaging, and therefore researchers should be aware of the differences for inference.

#### 2.2.4 Graphical Tools for Assessing Model Averaging

In practice, regression analysis should include both exploratory data analysis and diagnostic plots to assess appropriateness of the proposed model(s) (e.g., scatterplots, various plots of residuals, posterior predictive checks). Similar practices are also important, and perhaps even more so, for model averaging because we must assess the appropriateness of many individual regression models, and also consider whether it is appropriate to combine them. However, what we have observed in practice suggests less exploratory data analysis and model checking is actually done in the context of model averaging, which in part can be attributed to prohibitively large or at least inconveniently large  $\mathcal{M}$  (Ver Hoef and Boveng, 2015). Hoeting et al.

(1999a) recommend checking the usual diagnostics for the fullest model in the model set before proceeding with model averaging for predictions, which is similar to common advice when considering a set of nested models for multiple linear regression. The method for checking assumptions for many models simultaneously is inherently context and situation dependent, and therefore we do not provide broad suggestions here. However, in our next two points, we do help with assessing whether it is appropriate to combine results from individual models, assuming assumptions are not severely violated.

First, we have observed that choosing model averaging as an inferential tool may lead to a lack of consideration of higher order terms such as interactions or polynomials. This seems to be done out of convention or ease of implementation and is often not justified by plots of the data or subject matter expertise. Many of the usual regression plotting tools could help researchers identify interactions and higher order terms to include in an analysis. Including such terms complicates the model averaging process and the researcher will need to restrict the *all subsets* model set by forcing interactions and higher order terms into models together (Chipman, 1996). We provide an example of forcing variables into a model in Section 2.3, but we do not discuss the implications for the prior model probabilities when the model set is manipulated in this way because it is beyond the scope of this paper. We refer the reader to Chipman (1996) and Clyde and George (2004).

Second, we see the need for an accessible graphical tool to help researchers efficiently compare results from individual models and understand how they are combined into model-averaged posterior distributions. Such a tool allows researchers to assess the implicit assumption that the parameters (or estimates) being averaged have a common (enough) meaning across models, and to understand the weights given to each. Software for model averaging makes it relatively easy to obtain the posterior

for the model-averaged partial regression coefficient, but the ability to visualize how individual posteriors are combined to create it is absent. Another common and slightly more informative visualization tool, displays how the signs of the posterior means of partial regression coefficients associated with each input variable change across the top models (commonly called an image plot). However, the individual posterior distributions could have means with the same sign, and still be noticeably different. A comparison of the entire posterior distribution is much more desirable.

We developed the *model-averaged posteriors plot* (MAP plot hereafter), to provide a visual summary of all components going into, and resulting from, the averaging of partial regression coefficients across models. The plot is designed to be used in the process of deciding whether model averaging is an appropriate tool for inference, not as a way to display the results of model averaging at the end of an analysis. An effective way to assess whether model averaging is potentially appropriate for a particular case is to actually implement it and then use the MAP plot to carefully digest and critically evaluate the potential usefulness of model averaging for the problem at hand. This plot is not meant to replace the usual regression diagnostics and modeling decisions mentioned above, but it can help identify models that should be checked. For example, if model averaging makes sense for a particular example, and the model-averaged result is essentially a combination of five models, the researcher could easily check diagnostics for those five models. In the following sections, we use the MAP plot for assessing model averaging in two examples with different analysis objectives.

### 2.3 Example 1: Haul-out Behavior of Weddell Seals

In 2010, researchers studying Weddell Seal (*Leptonychotes weddellii*) haul-out behavior in Erebus Bay, Antarctica set up cameras to take photos of a portion of Big

Razorback haul-out site (Fig. 2.3). Photos were taken every forty-five minutes for the purpose of aligning the time-of-day for data collection to the time-of-day most seals were expected on the ice. One of the peak haul-out times is between the hours of 8 a.m. and 8 p.m., so we use only the data collected during these convenient times. Other explanatory variables thought to be related to the number of seals on the ice at a given time were also collected at each time point. The specific goal of this analysis is to make inference about the time-of-day associated with maximum mean seal counts, after accounting for wind speed ( $m/sec$ ), temperature ( $^{\circ}C$ ), and tide height ( $m$ ). Note that interest in the maximum necessitates a quadratic term for time-of-day.



Figure 2.3: A digital image used for enumerating Weddell seals on the fast ice in the Big Razorback haul-out site, Erebus Bay, Antarctica. This image was taken at 17:30 hrs on October 18, 2010. Weddell seal imagery obtained under authority of permit NMFS Permit No. 1032-1917-02.

For this data set and question of interest, we have already identified a useful model (up to consideration of interactions). However, suppose we are asked by reviewers to use model averaging to incorporate model uncertainty into the analysis (not an unrealistic scenario given our personal experiences). In an attempt to satisfy the reviewers, but still keep the question of interest in sight, we force the linear and quadratic components of time-of-day into all models in  $\mathcal{M}$ . It is possible that interactions should be investigated to allow the relationship between time-of-day and maximum expected seal count to depend on the values of the other explanatory variables, but we follow the common implementation of model averaging and do not include interactions in our model set. For this analysis, models are indexed by the explanatory variables they include. We define  $\mathcal{M} \equiv \{M_T, M_{Tw}, M_{Tt}, M_{Td}, M_{Twt}, M_{Ttd}, M_{Twd}, M_{Ttwd}\}$ , where  $T$  is defined to include both the linear and quadratic components of time-of-day (*time* and *time*<sup>2</sup>), and others are defined as  $w$  = wind speed,  $t$  = temp, and  $d$  = tide.

Given the count data, we could implement a Poisson generalized linear model using reversible jump MCMC (RJMCMC) similar to Link and Barker (2006). However, in this case all seal counts are greater than zero, averaging 15.79 seals, and natural log transformed counts reveal that linearity, normality, and constant variance assumptions for normal linear regression are reasonable. Therefore, we expect Poisson log-linear regression and normal linear regression on the log-transformed counts to provide very similar results for this example, and they do. We choose to use the simpler and computationally faster alternative because the choice of method does not change the conclusions or the points we wish to make with this example. Thus, the models in  $\mathcal{M}$  are of the form  $\mu\{\log(\mathbf{y})\} = \mathbf{X}\boldsymbol{\beta}_j$ , where errors are assumed independent and normally distributed.

There is an obvious violation of independence through temporal autocorrelation among residuals within a day. However, the violation exists for every model in the model set and ignoring it for this exercise does not affect the comparisons we use to illustrate our points. Preliminary analyses revealed an AR(1) correlation structure was sufficient to account for the correlation within days. Incorporating an AR(1) correlation structure into the analysis would result in us recommending that the researchers start sampling about 40 minutes earlier and end sampling about 40 minutes later than if the correlation was left unaccounted for. A sampling effort will span multiple hours, so the practical implications of the difference in results are minimal given the objective of the analysis.

The quantity of interest is the time-of-day that maximizes the expected number of seals, which can be expressed as a function of the regression coefficients associated with *time* and *time*<sup>2</sup> using the formula for the vertex of a parabola:  $\{\text{Time at Max}\}_j = -\frac{\beta_{\text{time},j}}{2\beta_{\text{time}^2,j}}$  for  $M_j$ . Using Bayesian inference, it is straightforward to computationally obtain a posterior distribution for this quantity, and in a likelihood setting, the delta method is a common choice for obtaining an associated standard error for the point estimate.

We implement Bayesian model averaging with the `bms` function in the `BMS` R package (Feldkircher and Zeugner 2009; R Core Team 2016). We use the default unit information formulation of the *g*-prior ( $g = n = 769$ ) on the parameters and a discrete uniform prior on  $\mathcal{M}$ . The standard output returned by the `bms` function tabulates posterior means, standard deviations for the model-averaged partial regression coefficients, posterior inclusion probabilities for each explanatory variable, and the frequency with which the posterior means are positive across models (which gets at the *partial* regression coefficient idea, but requires a sign switch, see Section 2.2.2). The user can also easily extract posterior model probabilities and

other posterior summaries from the `bms` output for all models in  $\mathcal{M}$ . Based on the prior we use, we can appeal to analytical results providing the form of the posterior distributions as  $t$ -distributions using moments obtained directly from `bms` (details available in [Appendix B]). We display all relevant information about the individual models and the resulting model-averaged posterior distributions in a MAP plot (Fig. 2.4).

For this example, the MAP plot is composed of five panels, one for each of the explanatory variables considered. A column (panel) displays the distributions of partial regression coefficients associated with the same explanatory variable across models. A row (going across panels) provides the posterior distributions for the partial regression coefficients from a particular model, with the model averaged posterior distribution in the bottom row. The probability associated with the the point mass of zero in the model-averaged distribution is provided as text. All explanatory variables were standardized to have zero mean and unit standard deviation prior to analysis due to large differences in their ranges (Fig. 2.4). An alternative form of standardization proposed in Cade (2015) uses partial standard deviations to adjust for changing multicollinearity among explanatory variables. However, as explained in Section 2.2.2, we are focusing on the early stages of assessing the appropriateness of model averaging, rather than deciding how to present results of model averaging. Creating the MAP plot at this stage promotes investigation of the differences in results across models before making the decision to possibly implement the standardization in Cade (2015).

There are several observations gleaned from the information provided by the MAP plot (Fig. 2.4). First, we observe that only two of the eight models received non-negligible posterior mass ( $M_{Ttd}$  with posterior model probability 0.884 and  $M_{Ttwd}$  with posterior model probability 0.116). The remaining models had such small posterior

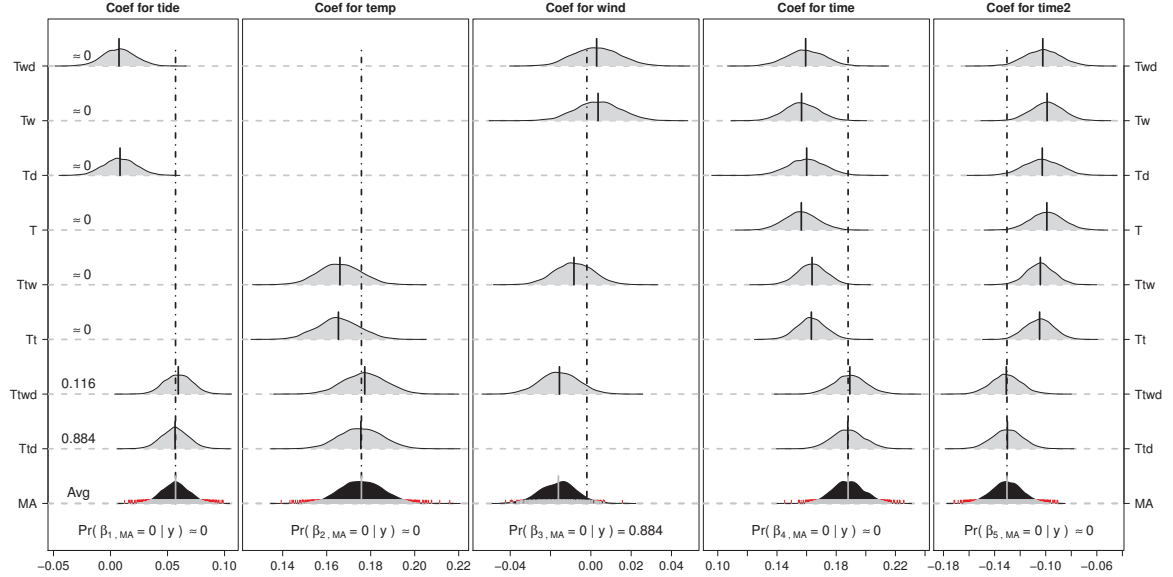


Figure 2.4: The MAP plot is split into five panels, one for each of the explanatory variables considered. Rows across panels display output from one model and are ordered by increasing magnitude of posterior model probability (printed along the leftmost  $y$ -axis). Probabilities given as  $\approx 0$  have a value less than 0.0001). The posterior distribution of the model-averaged partial regression coefficient associated with each explanatory variable is located at the bottom of the plot (continuous component is shown as a density plot and the zero-component is shown in text). Note that the point mass associated with the zero components associated with the coefficient *time* and *time*<sup>2</sup> are actually exactly zero because they appear in every model (although the label suggests otherwise). The vertical line shows the posterior mean of the model-averaged distribution. We created the **MApp** R (R Core Team, 2016) package that includes a MAP plot function to automatically construct the plot from different sources of model averaging output, along with a vignette (R packages used: `beanplot`, `BMS`, `dplyr`, `LearnBayes` Albert, 2014; `Feldkircher` and `Zeugner`, 2009; `Kampstra`, 2008; `Wickham` and `Francois`, 2015). It is available at <https://github.com/kbanner14/MApp-Rpackage>. Code to recreate the figure is available in Appendix S2 [Appendix B], a script file can also be found in Appendix Data S1 (Rcode-appendix-MAPP.R) [online supplementary materials from Banner and Higgs (2016)]

model probabilities that, in this case, model averaging is the result of a weighted average of only two models. The PMPs as visualized through the MAP plot is a useful tool for identifying when results of model averaging are a combination from a few models.

We also observe that the continuous component of the posterior distributions for the model-averaged coefficients associated with *tide* and *temp* are essentially the same as the model-averaged posterior including both the zero and continuous components. For example, the panel for *tide* shows that all models excluding *tide* have very small posterior model probabilities (given as  $\approx 0$ , with values  $\ll 0.0001$ ). Therefore, the probability associated with the point mass at zero for the model-averaged coefficients associated with *tide* is essentially zero, and the total posterior probability is almost entirely contained in the continuous component. Similar observations hold for *temp*. The panels for *time* and *time*<sup>2</sup> have posteriors for all models because both explanatory variables were forced into all models in  $\mathcal{M}$ , and therefore, the model-averaged posterior distributions for *time* and *time*<sup>2</sup> do not have zero components.

In contrast to the previous observations, we highlight the panel summarizing the partial regression coefficient for *wind*. One of the models excluding *wind* ( $M_{Ttd}$ ) has large posterior model probability (0.884) and the other models excluding *wind* ( $M_{Tt}$ ,  $M_T$ , and  $M_{Td}$ ) have small posterior model probabilities (given as  $\approx 0$ , with values  $\ll 0.0001$ ). The zero component of the model-averaged distribution associated with *wind* is the sum of posterior model probabilities from these models, totaling approximately 0.884, and the continuous component comes from just one model,  $M_{Ttwd}$ , with posterior model probability 0.116 ( $1 - 0.884 = 0.116$ ). Therefore, the model-averaged distribution associated with *wind* is essentially made up of one model contributing to the zero component, and one model contributing to the continuous component.

It is also clear from the plot that while the posterior model probability for the largest model  $M_{Ttw}$  is small, the posterior distributions for partial regression coefficients from  $M_{Ttw}$  are similar to their model-averaged counterparts. This is largely due to the fact that the addition of *wind* to the model with the highest posterior model probability,  $M_{Ttd}$ , did not appreciably change the posterior distributions of the partial regression coefficients for *temp*, *tide*, *time*, or *time*<sup>2</sup>. Observations such as these are meant to encourage readers to use the MAP plot to further investigate how information from individual models combines to create the model-averaged posterior, particularly in the presence of strong multicollinearity, and also to evaluate whether model averaging makes sense given the question of interest and desired inferences.

Now, we use the MAP plot to help assess averaging of the quantity of interest, time-of-day associated with maximum mean seal counts. We use posterior draws of  $\beta_{time,j}|\mathbf{y}, M_j$  and  $\beta_{time^2,j}|\mathbf{y}, M_j$ , to create posterior distributions for  $\{\text{Time at max}\}_j$  for all models, and then create the model-averaged distribution  $\{\text{Time at max}\}_{MA}$  (Fig. 2.5). Recall that inference for the quantity of interest could have been directly addressed using the model including all of the available covariates. Therefore, it is interesting to compare inferences drawn from model averaging to those drawn from the largest model, which is straightforward because the posterior distribution of  $\{\text{Time at max}\}_{MA}$  does not have a zero-component. We compare a posterior credible interval for  $\{\text{Time at max}\}_{MA}$  to that for  $\{\text{Time at max}\}_{Ttw}$ . After back-transforming, we have 95% posterior credible intervals of 16:13 hrs to 17:13 hrs for the model-averaged result, and 16:14 hrs to 17:14 hrs for the largest model. There is a direct interpretation for the posterior interval coming from the largest model: there is a 95% chance the time of day at which the maximum number of seals are on the ice is between 16:14 hrs and 17:14 hrs, after accounting for *tide*, *wind*, and *temp*. However, the interpretation

of the interval from the model-averaged result is less clear because we average over two models that do not account for the same covariates. In this example, we used model averaging to incorporate model uncertainty but ended up with a posterior interval no wider than what we would have obtained had we chosen a reasonable model based solely on the question of interest.

By definition, the variance of the model-averaged quantity is based on the entire posterior distribution, including both the zero and continuous components. It is not always reasonable to compare posterior variances from individual models to the posterior variance coming from only the continuous component. The posterior variance of the model-averaged quantity of interest can be larger than the variance based only on the continuous component when the zero component is large *and* the continuous component is far from zero. Similarly, it can be smaller when the zero component is large *and* the continuous component is near zero. The only coefficient in this example with this potential is that associated with *wind*, where the variance from the continuous component of the model-averaged distribution,  $\text{Var}_{cts}$ , is larger than the variance from the model-averaged distribution considering both the zero and continuous components,  $\text{Var}_{MA}$ , ( $\text{Var}_{MA} = 3.51 \times 10^{-5}$ ,  $\text{Var}_{cts} = 9.22 \times 10^{-5}$ ). For the other variables, it is reasonable to compare the individual model posterior variances to those of the continuous component of the model-averaged posterior because the probability associated with the zero component is essentially 0, and therefore the point mass at zero is not contributing to the variance. These comparisons can be easily made using Figs. 2.4 & 2.5, and the posterior standard deviations in Table 2.1. In this example there is little difference. This may seem unexpected, as discussions of multimodel inference lead us to expect that accounting for model uncertainty should always result in increased posterior variance (or decreased precision).

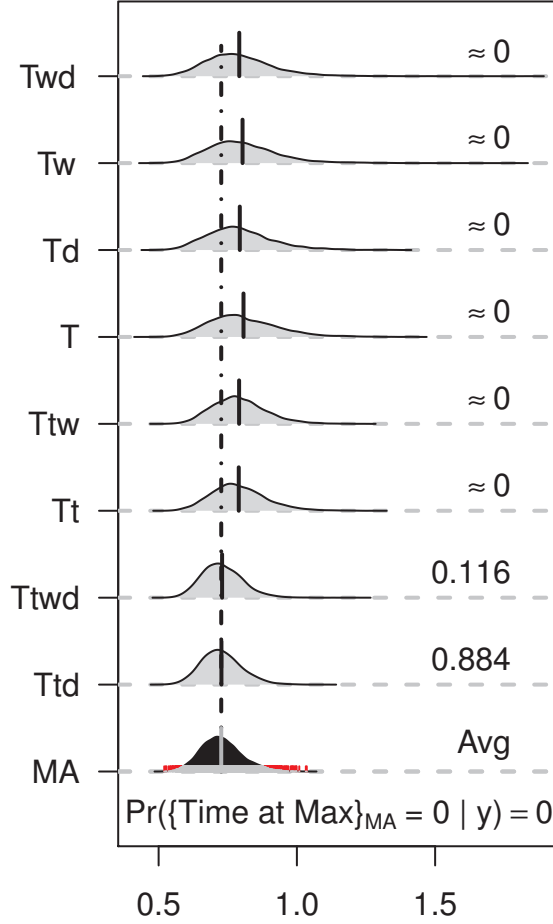


Figure 2.5: Posteriors for the  $\{\text{Time at max}\}$  parameter were created by finding the time-of-day corresponding to the maximum log total seals on the ice, which is a function of the partial regression coefficients associated with *time* and *time*<sup>2</sup> from Fig. 2.4. The MAP plot for  $\{\text{Time at max}\}$  shows the posterior distributions for the quantity of interest across the models in  $\mathcal{M}$  along with the model-averaged counterpart. Code to recreate this figure is available in Appendix S2 [Appendix B], a script file can also be found in Appendix DataS1 (Rcode-appendix-MAPP.R)[online supplementary materials from Banner and Higgs (2016)]

Table 2.1: Posterior standard deviations for the partial regression coefficients associated with each explanatory variable (each row) are shown for each model in the model set. The columns define models in the model set (columns Ttd - Twd). The posterior standard deviation coming from the continuous component of the model-averaged distribution is shown in the MA column. The row for {Time at Max} was appended to the results. This table is standard output from our plotting function, **MApp\_bms** which was programmed in R (R Core Team, 2016). Code is provided in Appendix S2 [Appendix B].

|                   | MA     | Ttd    | Ttwd   | Tt     | Ttw    | T      | Td     | Tw     | Twd    |
|-------------------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| Tide              | 0.0119 | 0.0119 | 0.0124 | —      | —      | —      | 0.0138 | —      | 0.014  |
| Temp              | 0.0097 | 0.0098 | 0.0098 | 0.0097 | 0.0096 | —      | —      | —      | —      |
| Wind              | 0.0095 | —      | 0.0096 | —      | 0.0096 | —      | —      | 0.0112 | 0.0113 |
| Time              | 0.0107 | 0.0107 | 0.0109 | 0.0096 | 0.0095 | 0.0111 | 0.0126 | 0.0113 | 0.0128 |
| Time <sup>2</sup> | 0.0119 | 0.0118 | 0.0121 | 0.0106 | 0.0107 | 0.0126 | 0.0139 | 0.0125 | 0.0141 |
| Time at max       | 0.0716 | 0.0717 | 0.0711 | 0.0971 | 0.0987 | 0.1236 | 0.1215 | 0.1243 | 0.1185 |
| PMP               | —      | 0.8837 | 0.1158 | 4e-04  | 0      | 0      | 0      | 0      | 0      |

From this example, we see that the incorporation of model uncertainty into an analysis does not always result in what we expect, and there is a need for more scrutiny on a case-by-case basis. This is important information for weighing whether the complexity of model averaging, along with its often added simplifying assumptions, is worth the supposed gains. When a research question is well defined, it can be addressed by building a question-driven model justified by the objective. This can be achieved by clearly defining the relationship(s) of interest, considering what variables should be accounted for (or controlled for) so that the relationship(s) of interest can be addressed, and possibly iterating on that model, as described by Ver Hoef and Boveng (2015). This process of building such a model can seem difficult because of the effort needed in justifying decisions and assumptions. Therefore, it is often tempting to elicit a more automatic approach to find the “best” model or combine results over multiple models. While model averaging may hold the allure of a decision-free model building technique (just use all of them), we must consider whether the added complexity is necessary. In this example, we believe model averaging was

unnecessary for the following reasons: (1) there was a well defined research question that could have been addressed directly by a single model (the largest model), and (2) inferences from the largest model were nearly identical to the model-averaged result.

#### 2.4 Example 2: When Prediction Leads to Explanation

In this section, we revisit an example from Hoeting et al. (1999a) using data from Penrose et al. (1985) and Johnson (1996). The objective is to predict percent body fat with age, weight, height, and ten body circumference measures. Training data are available for 251 adult males between the ages of 22 and 81 years. We refer to the variables as predictors rather than explanatory variables, because the analysis objective is prediction, rather than explanation using regression coefficients. We choose this example because these data appear multiple times in the model averaging literature (e.g., Burnham 2015; Burnham and Anderson 2002; Hoeting et al. 1999a; Zeugner 2011).

Even when the stated goal of model averaging is to improve predictive ability, it is often attractive to use the results to quantify the relative importance of predictors and interpret their “overall effects.” Both of these secondary goals typically lead to predictors being considered in isolation, outside the context of models created with collections of predictors. As previously discussed (Section 2.2.2), the appropriateness of model averaging for prediction does not imply it is appropriate for explaining “overall effects.” Careful examination of the individual model results using the MAP plot can help researchers decide if reporting conclusions focused on a variable in isolation may be appropriate.

Here, we keep the objective consistent with Hoeting et al. (1999a) and suppose the researchers have no *a priori* information about which variables are useful for predicting mean percent body fat, and no specific question of interest regarding any

of the variables beyond using first order combinations of the variables for prediction. We use the same *all subsets* model set with  $J = 2^{13} = 8192$  models, the same discrete uniform prior on  $\mathcal{M}$ , and a formulation of the  $g$ -prior for the regression coefficients that approximates the prior used in Hoeting et al. (1999a) (see Zeugner, 2011), and the `bms` function (Feldkircher and Zeugner, 2009).

The `bms` implementation of this example computes posterior model probabilities for all models in  $\mathcal{M}$ , and by default stores only the top 500. The results for this analysis indicate a substantial amount of model uncertainty; the ten models with the highest posterior model probabilities account for about 74% of the posterior probability, and the top 500 models capture about 99.8%. We create model-averaged distributions using only these 500 models. Using the MAP plot, we display results for three predictors (*weight*, *abdomen* circumference, and *wrist* circumference) using the ten models with highest probabilities (Fig. 2.6). The reader is encouraged to use the plot to investigate how information related to these variables in different models is combined to form the model-averaged results.

In Section 2.2.2, we introduced the PIP for the  $v^{th}$  variable ( $\text{PIP}_v$ ) as the sum of the posterior model probabilities over the models including that variable ( $\mathcal{M}_v$ ). PIPs are often interpreted as relative measures of overall “importance” for each explanatory variable over all models considered. See Cade (2015) for an alternative measure of relative importance and Burnham (2015) for a recent discussion about using AIC weights for this purpose. Rather than repeat arguments here, we tie our discussion of overall “importance” to the previously discussed implications of ignoring the *partial* in partial regression coefficients (Section 2.2.2).

The posterior model probabilities provide information about a collection of predictors working together to model the mean response, and each predictor occurs in isolation in only one model. Making statements about the “importance” of a single

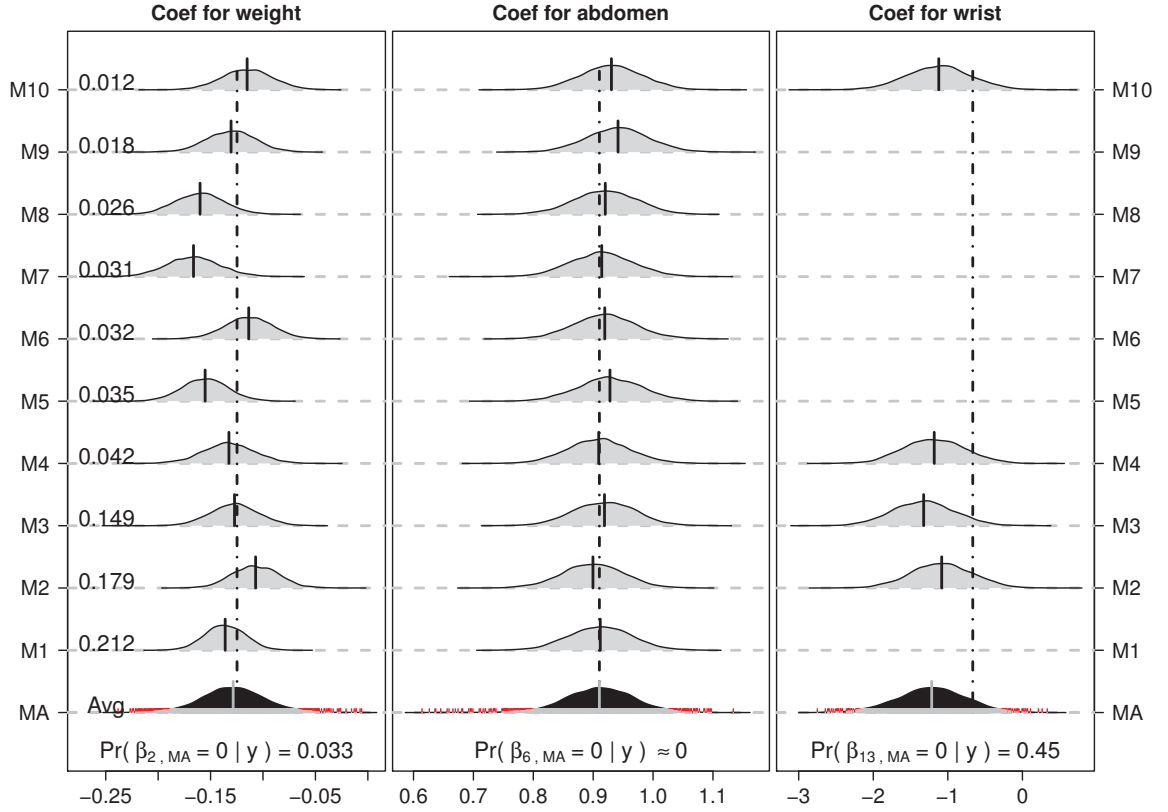


Figure 2.6: This MAP plot includes variables *weight*, *abdomen*, and *wrist* from the body fat example. The models with the two highest posterior model probabilities are those including *weight* and *abdomen*, and *weight*, *abdomen*, and *wrist*, represented in rows  $M_1$  and  $M_2$  respectively. For the other rows, not all variables included in the model are shown in the plot. This plot shows the posterior distributions for the partial regression coefficients associated with these three predictors for the ten models with the highest posterior model probabilities (note that the model-averaged distribution is an approximation, computed using the top 500 models which amount to 99.8% of the posterior model mass). Code to recreate this figure is available in Appendix S2 [Appendix B], a script file can also be found in Appendix DataS1 (Rcode-appendix-MAPP.R)[online supplementary materials from Banner and Higgs (2016)].

predictor with a posterior inclusion probability may imply (particularly to those with little background in regression or model averaging) that the predictor alone will be an effective predictor, regardless of what else accompanies it in the model. We show why such statements are potentially misleading by considering the PIP for *wrist* ( $\text{PIP}_{\text{wrist}} = 0.55$ ) which is large enough to be considered “important.” Now, consider how *wrist* performs by itself as compared to when it is coupled with *weight* and/or *abdomen*. The model with *wrist* alone does not even rank in the top 500 with respect to posterior model probability, but when *wrist* is combined with *weight* and *abdomen*, the resulting model ranks second (i.e., after accounting for *abdomen* and *weight*, *wrist* adds meaningful information for the prediction). However, one would not want to collect only information about wrist circumference to predict percent body fat. We bring up this point to promote discussion of the issues that arise when the importance of a variable is assessed in isolation, and to encourage careful thought about which quantities are chosen as a basis for conclusions.

## 2.5 Discussion

In the last 20 years, it has become relatively easy to implement model averaging with automatic priors and closed form posteriors, and also through non-Bayesian approximations (e.g., AIC or BIC weights). This ease of implementation, coupled with the allures of addressing model uncertainty *and* avoiding having to justify choosing one model, have made model averaging an increasingly popular method with many researchers in the ecological and environmental sciences. We suggest a critical look at this method, especially when it is used with partial regression coefficients, before it becomes more commonplace.

In general, potential benefits of new methods are readily communicated, popularized, and propagated through the peer review process, while challenges

in using them are often left in the shadows. We believe statisticians have a responsibility to help researchers navigate the hard decisions about whether methods are appropriate or necessary for a particular problem. We are not the first to raise red flags suggesting a more careful look at multimodel inference and model averaging, and many of our concerns are consistent with those expressed elsewhere (e.g., Cade 2015; Fieberg and Johnson 2015; Thomas et al. 2007; Ver Hoef and Boveng 2015). In this paper, we offered researchers accessible information to help provide a basis for critically evaluating model averaging of partial regression coefficients on a case-by-case basis.

In Section 2.2.2 we provided notation for partial regression coefficients and the posterior distribution for model-averaged regression coefficients. Though subtle, these suggested changes for notation have the potential to remind the user of the implicit assumption underlying model averaging regression coefficients, and may prevent errant use of the method.

In Section 2.2.2, we briefly discussed question-focused modeling, which we believe is directly tied to discussions about the challenges of model averaging regression coefficients. In general, we find linear regression is a useful tool for three purposes: (1) estimating relationships between an explanatory variable and the mean response, often after accounting for other “control” variables, (2) estimating effects from experimental designs, and (3) building predictive models. It is typically easier in designed experiments to think about what variables should be “controlled for” or “accounted for” when making the treatment comparisons. However, this same thought exercise can, and should, extend to observational studies. Often, we can identify the relationship(s) of interest, think about what we would have controlled for in the design if random assignment had been possible, and then use regression as a way to adjust for the identified variables. This can be thought of as working toward

*ceteris paribus* (holding all else constant except the treatment assignment), even in observational studies. This describes question-focused modeling because the research question drives the choice of what variables to control for rather than allowing the analysis method to automatically choose them (method-focused modeling). Ver Hoef and Boveng (2015) and Fieberg and Johnson (2015) agree we have much to gain by carefully considering what variables are directly of interest and what variables we would like to account for. Automatic approaches like model averaging certainly hold an appeal, but model-based automatic approaches can be dangerous in the context of explanatory inference.

A common motivation for model averaging is to incorporate model uncertainty into the analysis, inferences, and ultimately management decisions. This implies an increase in posterior variance of predictions on average (with a decrease in bias), and it has been stated that the same should hold for the posterior variance of model-averaged partial regression coefficients (Leamer 1978; Raftery et al. 1997a). Beyond the scope of this paper, we are investigating the extent to which this may be the case in practice, and how this relates to the two components (the zero component and the continuous component) of the model-averaged posterior distribution. If the results are essentially the same for model averaging as compared to using one model, model averaging may not be a mistake, but it may be adding unnecessary complexity to an analysis for a perceived goal that is left unrealized.

When model averaging is used for prediction, it is natural to inquire which predictors are most useful for predicting the mean response. PIPs have been used as such a summary for individual predictors. However, a “good” model (as a collection of predictors) should be distinguished from a “good” predictor in isolation. We do not criticize the use of PIPs in general, but suggest careful thought about when they are appropriate and how they are interpreted when they are used for stating broad

conclusions. Although PIPs are a standard output from software packages designed for multimodel inference, we do not think they should be used simply because they are readily available. We deliberately use the posterior *exclusion* probabilities in the MAP plot because of their critical role in fully defining the posterior distribution of model-averaged regression coefficients.

Even though the goal of accounting for as many sources of uncertainty as possible is certainly worthy, we believe the complexity of the method used to achieve this goal, and the potential loss in the interpretability of results in the context of the question of interest, should be weighed against sheer sophistication and popularity of the method. The utility of a modeling approach should be measured by how well it serves the intended purpose of the research, by whether it is accessible for others to critique, and by the degree to which it facilitates practical interpretations and meaningful conclusions. With the examples in Sections 3 and 4, we demonstrated the utility of the MAP plot for all three of the aforementioned criteria. The MAP plot can help researchers understand the differences between the results from model averaging and those from individual models, assess the appropriateness and usefulness of model averaging for a particular problem, and find a starting place for justifying their choice of method to peers. We presented the MAP plot and discussion within a Bayesian framework, but the information contained in the plot and the ideas motivating it can directly be translated to model averaging performed using AIC. The main difference is displaying confidence intervals instead of posterior distributions, and such a plot can be used to elicit the same type of discussion and thought about model averaging (there is a function in our R package to create the MAP plot for AIC results `MApp-IC()`, see Appendix S1 [Appendix A] for details).

### 2.5.1 Conclusion

Negative attention directed towards *not* accounting for model uncertainty has led to researchers feeling as though they must account for it in their analyses. In such cases, model averaging may be the low hanging fruit (e.g., Burnham and Anderson 2002; Montgomery and Nyhan 2010), or may be suggested by peer reviewers. We have experienced this ourselves as statistical consultants through reviewer requests to conduct model selection or model averaging among competing models when it was at best, unnecessary, and at worst inappropriate. Reviewer comments have reflected both a lack of understanding of interpretation of partial regression coefficients and a loyalty to more automatic model selection techniques as a way to justify a model. This experience is echoed by other statisticians and scientists with solid statistical foundations (Cade 2015; Thomas et al. 2007, and the blog post: Brewer 2015).

Maintaining a focus on question-focused modeling in cases where it is appropriate can be difficult to justify to peers who may be expecting some form of multimodel inference in an analysis. With our tools for assessing model averaging, we shed light on the complexity of the posterior distribution of model-averaged partial regression coefficients and address a common misconception that model averaging is an easier, nearly automatic, alternative to question-focused modeling or other model selection methods. Crainiceanu et al. (2008) and Wilson and Reich (2014) describe methods that may fall closer to middle ground between multimodel inference and question-focused modeling by proposing algorithms to help identify a subset of models addressing a particular question of interest, and where the meanings of the partial regression coefficients of interest are similar enough across models to be reasonably averaged.

This paper is meant to provoke careful thought and continued discussions about model averaging, as well as provide common foundations to facilitate assessment

and discussion of model averaging on a case-by-case basis. We agree with Ver Hoef and Boveng (2015) that iterating on one model can often be better than considering many models, and we stress that adding complexity to analyses to “account for model uncertainty” should be seriously weighed against the cost of interpretation and the simplifying assumptions made in the process.

## 2.6 Acknowledgements

We would like to thank the subject matter editor, Dr. Kiona Ogle, as well as Dr. Brian S. Cade, Dr. Mevin Hooten, and two anonymous reviewers, all of whom provided invaluable feedback to strengthen and focus the manuscript. Additionally, we would like to thank Dr. Brian S. Cade for sharing his work throughout the process of developing this manuscript. The Weddell seal field research was supported by the National Science Foundation OPP-0635739 grant to R.A. Garrott, J.J. Rotella and D.B. Siniff. All of the camera images were obtained under authority of permit NMFS Permit No.1032-1917-02, and Dr. Robert Garrott put in a large effort obtaining the photos. Jay Rotella, Jesse DeVoe, and Michelle LaRue were largely involved in providing data sets and a wealth of information about their involvement in the Weddell Seal Project, and we thank them for their help and contributions.

## CHAPTER THREE

### INVESTIGATING THE POSTERIOR VARIANCE OF PARTIAL REGRESSION COEFFICIENTS RESULTING FROM THREE COMMON METHODS FOR MULTIMODEL INFERENCE

#### Contribution of Authors and Co-Authors

Author: Katharine Banner

Contributions: Wrote the majority of the manuscript and R code.

Co-Author: Megan Higgs

Contributions: Provided feedback on statistical analysis and drafts of the manuscript and R code.

Manuscript Information Page

Katharine M. Banner and Megan D. Higgs

Annals of Applied Statistics

Status of Manuscript:

☒ Prepared for submission to a peer-reviewed journal

☐ Officially submitted to a peer-review journal

☐ Accepted by a peer-reviewed journal

☐ Published in a peer-reviewed journal

Publisher: Institute of Mathematical Statistics.

Submitted:

## Abstract

The strategy of conditioning statistical inferences on one model is often criticized for ignoring uncertainty in the model selection process, and ultimately resulting in naïve final inferences. In this paper, we investigate the extent to which simultaneously incorporating multiple models into an analysis translates to changes in the variance for partial regression coefficients. Specifically, we compare three competing strategies within the multimodel inferential framework: (1) using all models in the model set and their posterior model probabilities (PMPs) as weights in model averaging (full model averaging), (2) using the subset of models including the variable of interest using re-normalized PMPs as weights in model averaging (model averaging over a subset of models); and (3) using a single model for final inference, specifically, the model with the largest PMP. We conduct a simulation study to investigate the assumed increase in posterior variance as a direct result of incorporating more models into an analysis. We found that the variance of model-averaged partial regression coefficients can either be larger or smaller than the variance of partial regression coefficients resulting from methods considering fewer models. In this paper, we identify specific situations in the context of multiple linear regression for both cases. In cases where the variance increases, we discuss whether it is worth the added complexity.

## 3.1 Introduction

Statistical modeling is an essential component to most quantitative scientific research. When building statistical models, researchers may identify a model thought to be useful for addressing their question(s) of interest (often by way of some model selection technique or criteria), and then condition their inferences on that model. When this approach is taken, one model is ‘selected’ for inference; however, even replicating the same study (same system and same study design) could result in a different model being ‘selected’. This strategy of conditioning inferences on one model is often criticized because it ignores uncertainty inherent in model selection, and as a consequence, resulting variances associated with parameters of interest may be understated (e.g., Buckland et al., 1997; Hodges, 1987; Leamer, 1978).

When an analysis considers many models, the resulting inference will be conditional on those models. However, considering different potential covariates and the spatial and temporal scales at which they are measured would result in a different set of models to consider, and yet another source of uncertainty (uncertainty in the model set). We focus on multimodel inference considering one set of models, and we are interested in how final inferences are different for different approaches within multimodel inference.

A popular multimodel inferential technique is model averaging, permitting inferences conditional on a set of models, rather than a single model. Starting with a model set (denoted  $\mathcal{M} = \{M_1, M_2, \dots, M_J\}$ ), model averaging simultaneously estimates parameters in each model *and* each models' posterior model probability (or model weight). Posterior model probabilities (or model weights) are then used to average quantities of interest over all models in  $\mathcal{M}$ . Model averaging was conceptualized under the assumption that a prediction was the quantity of interest. It is now being used to average partial regression coefficients over many models with the goal of assessing 'overall effects' while 'accounting for model uncertainty'. This extension of the method is typically motivated by the assumption that variances for *all* quantities of interest should be larger in the model-averaged result than in the result conditioning on a single model. However, challenges and implicit assumptions come with using model averaging to make inference about regression coefficients, and these are more complicated than they first appear (see Chapter 2).

Hoeting et al. (1999a) demonstrated predictive advantages for Bayesian model averaging (BMA) over other model selection techniques using simulation studies. Although they focused on prediction, they also compared  $p$ -values for regression coefficients (coming from a single model in  $\mathcal{M}$ ) to posterior inclusion probabilities (PIPs) for variables associated with the same regression coefficients (coming from

BMA). This was not a direct comparison of the variance associated with the coefficients of interest, but rather a comparison of inference about the ‘importance’ of the variables associated with the coefficients between the two methods. Hoeting et al. (1999a) concluded evidence for relationships between variables of interest and the mean response were overstated when inferences were conditioned on one model. In their discussion they state “BMA-based confidence intervals are better calibrated than single-model based confidence intervals; the latter tend to be too narrow,” suggesting inference based on one model overstates evidence for relationships, a claim not formally investigated in their paper. Raftery et al. (1997a) made similar conclusions based on comparisons made in the context of prediction, and Burnham and Anderson (2002) echo these conjectures in their textbook.

Inferences are sometimes called “naïve” if model uncertainty is not accounted for, and it is logical to think incorporating uncertainty from the model selection process into an analysis *should* lead to larger variances associated with quantities of interest. Thus, it is natural for researchers to feel pressure to use methods like model averaging in an effort to “account for model uncertainty” in their analyses. However, when model averaging is the tool used to account for model uncertainty, and partial regression coefficients are the quantities of interest, it is not always intuitive how the variance will change. This is largely due to the properties of the posterior distribution for the partial regression coefficients (discussed in Section 3.2). When model averaging is being considered, it is typical for final inference to be conditional on one of three subsets of models: (1) the whole model set, (2) the subset including the variable of interest (where the coefficient is estimated), or (3) the subset consisting of one “top” model typically as measured by the posterior model probability (PMP) (discussed in detail in Sections 3.3.1.1 - 3.3.1.3). We are interested in investigating how the

posterior variance of partial regression coefficients is affected by the subset of models used for final inference.

A lot of work has gone into adjusting variances or using methods to account for model uncertainty (e.g., Burnham and Anderson, 2002; Hoeting et al., 1999a), but little effort has been put into investigating when inferences from each approach will be most (or least) different in practice, especially in the context of inference for partial regression coefficients, rather than predictions. Incorporating model uncertainty into an analysis often comes with complicated inferences, and the perceived benefits of model averaging partial regression coefficients (less naïve inferences) are worthy of more investigation before model averaging becomes even more commonplace (e.g., Banner and Higgs, 2016; Cade, 2015; Ver Hoef and Boveng, 2015). To our knowledge, no one has investigated how the posterior variance of a partial regression coefficient changes according to the subset of models chosen for final inference.

In our informal investigations including observations from the model-averaged posteriors plot (presented in Chapter 2), we did not always observe the variance from model averaging to be larger than the variance from individual models (particularly models that would be chosen via AIC, PMP, or other model selection criteria). When we did observe differences, they were not so large that decisions based on inferences would change (e.g., Chapter 2, Section 2.3). Therefore, when the common phrase, “we incorporated model uncertainty into our inference” is used, it is not clear whether the perceived benefit of a more honest (larger) variance is realized. To help researchers gauge whether it is worth the complexity and increased difficulty in interpretation of regression coefficients, we are interested in learning more about the practical implications of including the additional uncertainty assumed to be realized by way of model averaging.

In this paper, we begin to investigate a rather large question: Is the change in posterior variance (when comparing three common strategies for incorporating model uncertainty into results) large enough to be practically meaningful for researchers? Furthermore, can we begin to identify cases where it either is, or is not practically meaningful? Addressing any pieces of these questions is both timely and relevant given the popularity of complex methods such as model averaging.

We restrict our context to modeling two relatively simple data generating models and employ *all subsets* linear regression. This simple context provides computational advantages in a fully Bayesian framework and provides a starting place for investigating the question of interest. Within this restricted context, we directly compare posterior variances resulting from three common approaches. First, we compare posterior variances from two approaches to model averaging: using all models in the model set, and using only the subset of models including the variable of interest. Second, we compare the posterior variance from the model with the largest PMP to that from the subset of models including the variable of interest. Finally, we compare the posterior variances from the model with the largest PMP to that from the full model-averaged distribution using all models.

### 3.2 Statistical Preliminaries

When methods for obtaining posterior model probabilities (necessary quantities for model averaging) were developed, many simulation studies focused on investigating the methods' ability to identify the data generating model as the model with highest posterior model probability (i.e., model selection; e.g., George and McCulloch, 1993; O'Hara and Sillanpää, 2009). To our knowledge, the properties of the variance of model-averaged posterior distributions for partial regression coefficients have not been formally investigated.

### 3.2.1 Variance of Model-Averaged Regression Coefficients

For our simulation study, we populate our model sets with linear regression models. We use typical regression notation, defining  $\mathbf{y}$  as a  $(n \times 1)$  column vector of observations of a response variable and  $\mathbf{X}$  as a  $(n \times (p+1))$  matrix with a column vector of 1s for the intercept and additional column vectors for the  $p$  potential input variables,  $\mathbf{X} = [\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p]$ . Let  $\beta_j$  be the  $(p+1)$  column vector of regression coefficients associated with  $M_j$ . Assuming normal, independent errors centered around a mean with constant variance ( $\mathbf{y} \sim N_n(\mu\{\mathbf{y}|\mathbf{X}\}, \sigma^2 I_{n \times n})$ ), we have our familiar multiple linear regression model. The mean of  $\mathbf{y}$  is modeled as a function of the first order terms of the  $p$  input variables and the intercept,

$$\mu\{\mathbf{y}|\mathbf{X}\} = \beta_{0,j} + \beta_{1,j}\mathbf{x}_1 + \beta_{2,j}\mathbf{x}_2 + \dots + \beta_{p,j}\mathbf{x}_p = \mathbf{X}\beta_j. \quad (3.1)$$

Excluding a variable from a linear regression model is equivalent to setting its associated partial regression coefficient to exactly zero. Therefore, the total probability of the posterior distribution of a model-averaged partial regression coefficient (over the entire model set) is made up of two components, one coming from the subset models where the coefficient is set to zero (i.e., variable excluded from the model), the other coming from the subset of models where the coefficient is estimated (i.e., variable included in the model). We denote the subset of models *including* the  $v^{th}$  input variable, as  $\mathcal{M}_v = \{M_j | \beta_{v,j} \neq 0\}$ , and its complement, the subset of models *excluding* the  $v^{th}$  input variable, as  $\overline{\mathcal{M}}_v = \{M_j | \beta_{v,j} = 0\}$ . This partition of  $\mathcal{M}$  into  $\{\mathcal{M}_v, \overline{\mathcal{M}}_v\}$  defines the model subsets associated with the two components of the model-averaged posterior distribution, the *continuous component* and the *zero component*, respectively. This is the main difference between using model averaging for predictions from linear regression models and using model averaging for

regression coefficients from linear regression models. When using model averaging for prediction, there is no point mass at zero because every model is capable of producing a directly interpretable prediction on the  $y$ -variable scale. However, the model-averaged posterior distribution for a partial regression coefficient includes a point mass at zero (the zero component) that cannot be ignored when considering how to summarize and interpret inferences. Therefore, the posterior distribution of the model-averaged partial regression coefficient associated with the input variable  $\mathbf{x}_v$  is a mixed distribution, which can be written as

$$p(\beta_{v,MA}|\mathbf{y}) = \begin{cases} \sum_{j \in \overline{\mathcal{M}}_v} \Pr(M_j|\mathbf{y}) & \text{for } \beta_{v,MA} = 0 \\ \sum_{j \in \mathcal{M}_v} \Pr(M_j|\mathbf{y})p(\beta_{v,j}|\mathbf{y}, M_j) & \text{for } \beta_{v,MA} \neq 0. \end{cases} \quad (3.2)$$

The quantities  $P(M_j|\mathbf{y})$  are termed posterior model probabilities (PMPs), which together, define the posterior distribution for the model set. PMPs are used to compute two important quantities: the posterior inclusion probability (PIP) and the posterior exclusion probability (PEP). The PIP is commonly used as a measure of variable “importance,” and is computed by taking the complement of the probability associated with the point mass at zero (i.e.,  $\text{PIP} = 1 - \sum_{j \in \overline{\mathcal{M}}} \Pr(M_j|\mathbf{y})$ ). Hereafter, we will refer to the probability associated with the point mass at zero as the posterior exclusion probability,  $\text{PEP} = \sum_{j \in \overline{\mathcal{M}}} \Pr(M_j|\mathbf{y})$ .

The variance of the posterior distribution for a model-averaged quantity of interest is provided in Hoeting et al. (1999a). Using slightly different notation to specify the quantity of interest as a model-averaged partial regression coefficient, we can write the variance of the posterior distribution for a model-averaged partial

regression coefficient, assuming conditional independence, as

$$\begin{aligned} \text{Var}(\beta_{v,MA}|\mathbf{y}) &= \left( \sum_{j=1}^J \left( \text{Var}(\beta_{v,j}|M_j, \mathbf{y}) + \text{E}(\beta_{v,j}|M_j, \mathbf{y})^2 \right) \text{Pr}(M_j|\mathbf{y}) \right) \\ &\quad - \left( \sum_{j=1}^J \text{E}(\beta_{v,j}|M_j, \mathbf{y}) \text{Pr}(M_j|\mathbf{y}) \right)^2. \end{aligned} \quad (3.3)$$

To obtain the posterior variance for the component of a model-averaged partial regression coefficient for variable  $\mathbf{x}_v$  considering only models where the coefficient is estimated, we first condition on that subset of models ( $\mathcal{M}_v$ ), and normalize the posterior model probabilities such that they sum to one,

$$\text{Pr}(M_j|\mathbf{y}, \mathcal{M}_v) = \frac{\text{Pr}(M_j|\mathbf{y})}{\sum_{j \in \mathcal{M}_v} \text{Pr}(M_j|\mathbf{y})}. \quad (3.4)$$

Then,

$$\begin{aligned} \text{Var}(\beta_{v,MA_{cts}}|\mathbf{y}) &= \left( \sum_{j \in \mathcal{M}} \left( \text{Var}(\beta_{v,j}|M_j, \mathbf{y}) + \text{E}(\beta_{v,j}|M_j, \mathbf{y})^2 \right) \text{Pr}(M_j|\mathbf{y}, \mathcal{M}_v) \right) \\ &\quad - \left( \sum_{j \in \mathcal{M}} \text{E}(\beta_{v,j}|M_j, \mathbf{y}) \text{Pr}(M_j|\mathbf{y}, \mathcal{M}_v) \right)^2. \end{aligned} \quad (3.5)$$

We abbreviate (3.3) and (3.5) as  $\text{Var}_{MA}$  and  $\text{Var}_{cts}$  respectively, and estimate them by plugging posterior means, variances, and model probabilities into (3.3), (3.4), and (3.5).

### 3.2.1.1 Further Insight Into the Model Averaged Posterior

To gain intuition for this complicated posterior distribution and its variance, we generate hypothetical examples of posterior distributions for model-averaged partial regression coefficients and illustrate how both components contribute to the total

model-averaged variance (Figures 3.1 & 3.2). For these examples, we assume the weighted average of posterior distributions from models including the variable of interest follows a normal distribution with mean  $\mu$  and standard deviation  $\sigma$ , and investigate different PEPs. Specifically, we define 18 scenarios using the combinations of levels of  $\mu = (0.15, 3)$ ,  $\sigma = (3, 1.5, 0.5)$ , and  $\text{PEP} = (0.1, 0.5, 0.9)$ . Results are split into two figures according to the levels of  $\mu$ : scenarios with  $\mu = 3$  (i.e., the weighted average of the posteriors from the individual models including the variable of interest is centered around 3) are displayed in Figure 3.1, and scenarios with  $\mu = 0.15$  are displayed in Figure 3.2. For both sets of results, PEPs are the same within a column (increasing in size from left to right) and standard deviations are the same within a row (decreasing in magnitude from top to bottom).

Three quantities are displayed in each plot: the total posterior variance of the model-averaged distribution ( $\text{Var}_{MA}$ ), the posterior variance of the continuous component of the model-averaged distribution ( $\text{Var}_{cts}$ ), and the ratio of these two variances ( $\text{ratio} = \frac{\text{Var}_{MA}}{\text{Var}_{cts}}$ ). Using the same logic used in the motivation for incorporating model uncertainty into analyses by way of multimodel inference, we may expect the variance of the method accounting for more models to be larger than the method accounting for fewer models (i.e., we may expect to see ratios larger than one). Contrary to this expectation, we actually observe the point mass at zero can either contribute extra variation to the model-averaged posterior distribution beyond that coming from the models including the variable (e.g., Figure 3.1, plots D, E, G, H, and I), or it can make the model-averaged posterior distribution less variable (e.g., Figure 3.1 plots B, C and F, and all plots in Figure 3.2).

In general, we notice when the mean of the continuous component is centered away from zero with relatively small variance, and the PEP is small to moderate, the ratio of variances tends to be larger than one. Similarly, when the component is

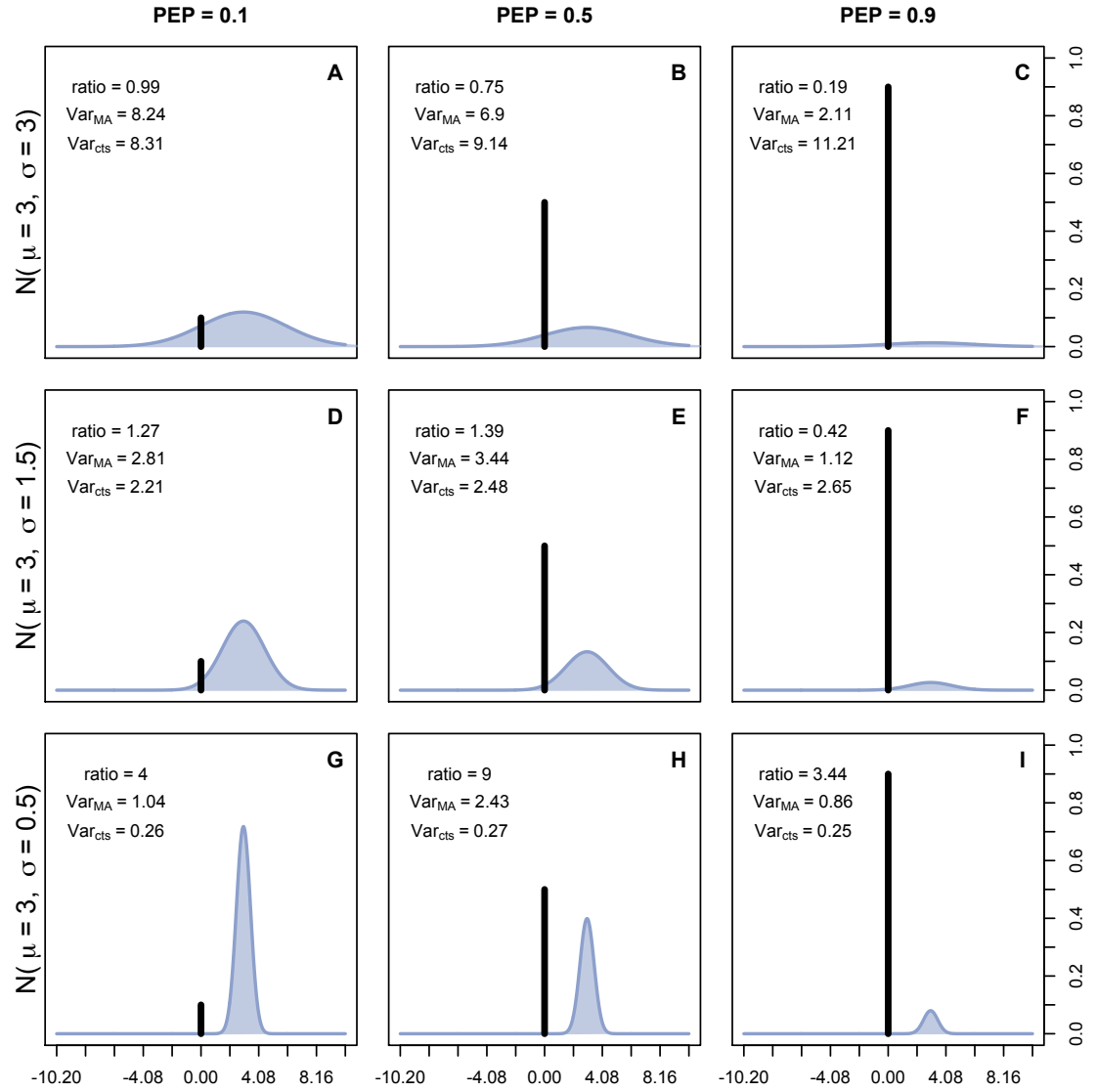


Figure 3.1: Nine scenarios investigating the variance of a mixed distribution comprised of a normal distribution centered at 3 units with standard deviation  $\sigma$  and a point mass at zero with probability equal to PEP.

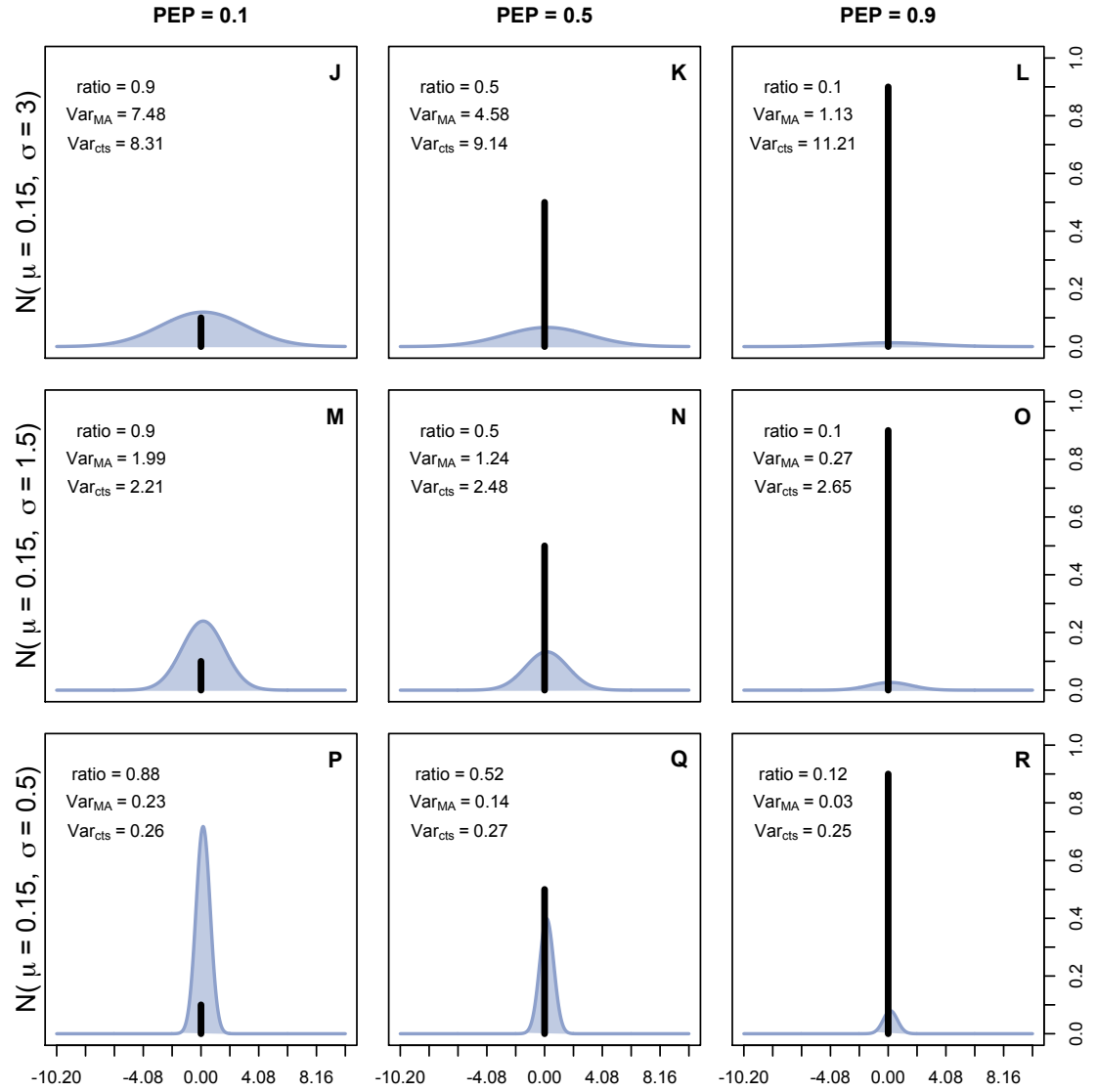


Figure 3.2: Nine scenarios investigating the variance of a mixed distribution comprised of a normal distribution centered at 0.15 units with standard deviation  $\sigma$  and a point mass at zero with probability equal to PEP.

centered away from zero with very small variance (e.g., plots G, H, and I in Figure 3.1), the models associated with the point mass at zero contribute a substantial amount of variability to the model-averaged distribution considering all models, even for large PEPs (e.g., plots H and I in Figure 3.1). Conversely, when the mean of the continuous component is centered near zero, the subset of models associated with the zero component actually end up reducing variation in the model-averaged posterior distribution as seen in all plots in Figure 3.2. This is interesting because the addition of the subset of models for which the coefficient is set to exactly zero, to the remaining models in  $\mathcal{M}$ , does not always result in increased variance, which is expected based on multimodel inference being purported as a tool for “incorporating additional uncertainty” inherent in model selection.

### 3.3 Simulation Study

As previously mentioned, we focus on three common subsets of models used for final inference: (1) using all models in the model set and all associated PMPs directly as weights for model averaging (full model averaging), (2) using the subset of models allowing the quantity of interest to be estimated by using re-normalized PMPs as weights in model averaging (model averaging over the models that make up the continuous component of the model-averaged posterior); and (3) using a single model, usually the model with the largest PMP.

When the quantity of interest is a partial regression coefficient and we are in an *all subsets* regression setting, the three subsets of models described above result in different distributions used for final inference: (1) the full model-averaged posterior (MA), comprised of the weighted average of the posteriors for the partial regression coefficients across all models in  $\mathcal{M}$ , including those where the coefficient is set to zero (resulting in a point mass at zero); (2) the re-normalized continuous component of

the full model-averaged distribution (*cts*), comprised of the weighted average of the posterior distributions for the subset of models in  $\mathcal{M}$  allowing the partial regression coefficient to be estimated (excluding the point mass at zero); and (3) the posterior distribution of the partial regression coefficient from the individual model with the largest PMP (*top*). These three choices of distributions for final inference and the three possible pairs defining comparisons we focus on in the simulation study are represented by the nodes and edges in Figure 3.3, respectively.

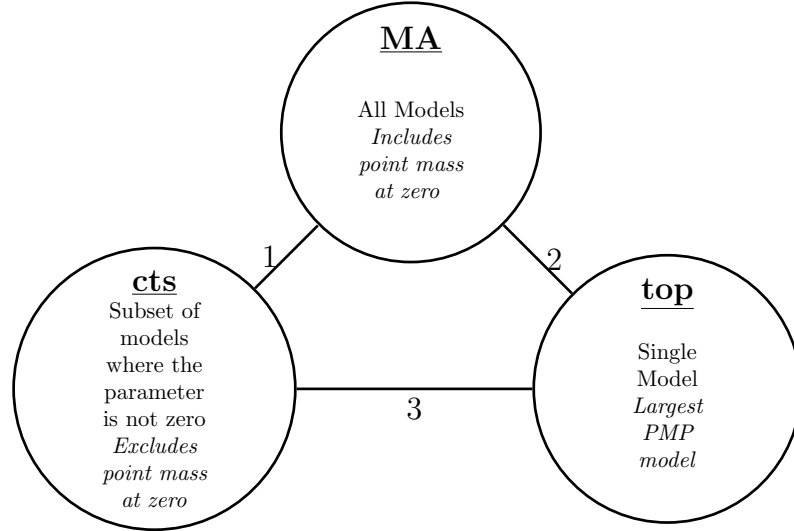


Figure 3.3: Three distributions commonly used for final inference based on different subsets of models in the model set are represented by the nodes. The three corresponding variance comparisons of interest are represented by the edges.

A very common summary for making inference is a range of plausible values for the quantity of interest, which is directly tied to the variance associated with that quantity (e.g., posterior intervals, confidence intervals). Therefore, to compare differences in inferences possibly resulting from choosing one of these three aforementioned methods, we will compare their variances (described in Sections 3.3.1.1 - 3.3.1.3). We will make these comparisons under different data generating

models, model sets, and correlations among the explanatory variables considered for the linear regression scenarios defined later in Section 3.3.2.

### 3.3.1 Ratios for Comparing Variances

Use of multimodel inference in practice is commonly motivated by the idea that the method incorporating the largest number of models will incorporate the most model uncertainty, and we generally expect incorporation of additional uncertainty to result in larger variances. For partial regression coefficients, it is not clear whether incorporating more models into an analysis (or “accounting for more model uncertainty”) always translates to larger variance and, when it does, if the increase is large enough to be practically meaningful. To investigate this, we make pairwise comparisons of the variances of the posterior distributions resulting from each of the three methods. We compute ratios such that values larger than one indicate the variance associated with the method incorporating more models is larger (i.e., we put the variance from the distribution including more models in the numerator). We use the ratios to make three comparisons in Figure 3.3:  $r_{MA,c} = Var_{MA}/Var_{cts}$  (Edge 1),  $r_{c,t} = Var_{cts}/Var_{top}$  (Edge 2), and  $r_{MA,t} = Var_{MA}/Var_{top}$  (Edge 3).

3.3.1.1 All Models Versus Subset of Models The difference between the full model-averaged distribution and its continuous component is the inclusion or exclusion of models for which the coefficient is set exactly to zero, corresponding to the inclusion or exclusion of the point mass at zero. The comparison of the variances from these two distributions (Edge 1 in Figure 3.3) helps describe how the point mass at zero coming from the collection of models excluding the variable of interest affects the model-averaged variance. The simulation study imparts insight into how the contribution to the variance, from the models informing the zero component, changes under different data generating scenarios (see Section 3.3.2).

In the Bayesian regression context, it is rare to see researchers condition inferences on the subset of models allowing the coefficient of interest to be estimated (i.e., condition inference on the models informing the continuous component of the model-averaged posterior distribution). However, this conditioning is actually quite popular, and commonly advocated for, when information criteria are used to obtain posterior model weights (e.g., Burnham and Anderson, 2002; Cade, 2015). Therefore, it is also interesting to compare these two methods of model averaging to see how much model uncertainty is “lost” when a particular subset of models is conditioned on for final inferences. We present results from comparing these two variances in Section 3.4.2.

3.3.1.2 Subset of Models Versus Top Model As previously described, model averaging of partial regression coefficients is sometimes conditioned on only those models containing the variable of interest (i.e., allowing the coefficient to be estimated). Therefore, it is interesting to investigate how inference from this method, which uses a subset of the models in the model set, compares to the method conditioning on the single model with the largest PMP. Using the ratio of variances defined by these two methods (Edge 2 in Figure 3.3), we compare variances under these two methods. This comparison provides insight into how different inference based on these two approaches can be different in practice. For example, when researchers choose to condition inferences on the subset of models estimating the coefficient, they may think they are “accounting for model uncertainty” (simply because they are averaging over models in their analyses), when, in fact, there could be little practical difference between their inferences and the simpler inference conditional on one model. That is, it is possible most of the “extra” uncertainty in the full model-averaged distribution comes from accounting for models excluding the

variable of interest (i.e., where the coefficient is set to exactly zero), and conditioning on the subset of models including the variable of interest often does not translate to increased variance for the coefficient of interest (as compared to just choosing one model). We present results from this comparison in Section 3.4.3.

3.3.1.3 All Models Versus Top Model Of the three methods discussed in this section, the two extremes are full model averaging over the entire set of models, and conditioning inference on one single model. Much of the work done in the context of prediction focuses on comparing these two methods, and they are often the methods being referred to when statements are made about comparing inferences conditional on one model and inferences incorporating model uncertainty. We directly compare the variances of the posterior distributions for partial regression coefficients resulting from these two methods (Edge 3 in Figure 3.3). This comparison also helps us form recommendations for researchers considering the two methods to help them weigh the tradeoff between simpler, more interpretable inferences, and incorporating model uncertainty into inference for quantities of interest. We use the simulation study to look at differences across a variety of scenarios which will be described in 3.3.2.

With these three comparisons (Sections 3.3.1.1 - 3.3.1.3), we learn about how the two components of the full model-averaged posterior distribution jointly contribute to its variance. With the simulation study, we begin investigating the claim that using more models in an analysis is synonymous with practically meaningful increases in variance for partial regression coefficients, and we can weigh trade-offs between “accounting for model uncertainty” and obtaining usable inferences.

### 3.3.2 Models

The form of the linear models we use in our simulation study come from an example in George and McCulloch (1993). The goal of George and McCulloch (1993)’s

example was to assess the performance of their stochastic search variable selection (SSVS) algorithm, which can be used to obtain posterior model probabilities for a set of models. In their paper, they investigate SSVS's ability to assign the data generating model the highest PMP when two of the input variables are highly collinear. Although we are not focusing on detecting the data generating model, this example is nice in that it has been discussed in the multimodel inference literature, it is relatively simple, and it is realistic enough for starting the investigation.

We use the general form of one of the simulation studies presented in George and McCulloch (1993) which considered five explanatory variables,  $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4, \mathbf{X}_5$  for both data generating models we consider. We keep the naming convention for the explanatory variables consistent with George and McCulloch (1993) and assign  $\mathbf{X}_4$  and  $\mathbf{X}_5$  to be the variables in the data generating models

$$\mathbf{y} = \mathbf{X}_4 + 1.2\mathbf{X}_5 + \boldsymbol{\epsilon}, \text{ and} \quad (3.6)$$

$$\mathbf{y} = \mathbf{X}_4 + 0.5\mathbf{X}_4^2 + 0.25\mathbf{X}_4^3 + 1.2\mathbf{X}_5 + \boldsymbol{\epsilon}, \quad (3.7)$$

where  $\boldsymbol{\epsilon} \sim N_{60}(\mathbf{0}, \sigma^2 \mathbf{I})$ . All explanatory variables are generated as independent and distributed as  $N_{60}(0, 1)$  random variables except for  $\mathbf{X}_3$  and  $\mathbf{X}_5$ , which are correlated. We define the correlation between  $\mathbf{X}_3$  and  $\mathbf{X}_5$  as  $\rho_{3,5}$ , and assume the variables are multivariate normal with mean equal to zero, standard deviation equal to one. Response and explanatory variables are all centered and scaled to have mean equal to zero and standard deviation equal to one, after the data are generated. Depending on condition (described in Sections 3.3.3.1 - 3.3.3.3), we use the right hand side of Equation 3.6 or Equation 3.7 to generate a response vector,  $\mathbf{y}$ , with 60 observations.

The relationships between  $X_4$  and  $y$  (plots A and C), and  $X_5$  and  $y$  (plots B and D) are shown for three realizations of data (shown in different colors), generated

using  $\sigma = 2.5$ , in Figure 3.4. Data generated according to Equation 3.6 are shown in plots A and B (top row) and data generated according to Equation 3.7 are shown in plots C and D (bottom row). The curvature from the tapering polynomial effect is visible in plot C.

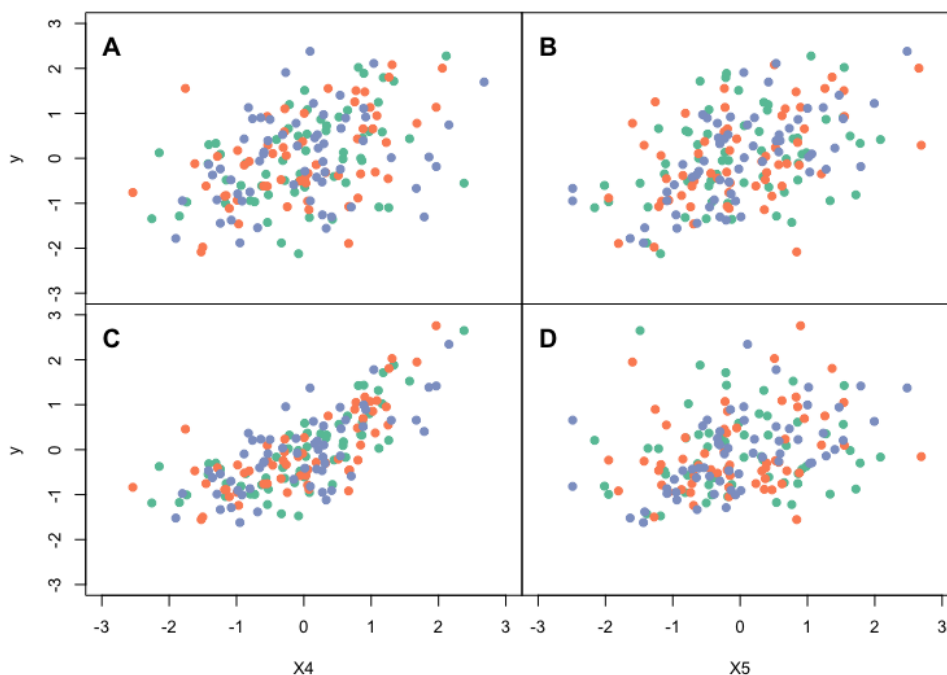


Figure 3.4: Three realizations of data are plotted within each plotting panel (coded by color) displaying the relationships between  $X_4$  and  $y$  (plots A and C), and  $X_5$  and  $y$  (Plots B and D). Data generated according to Equation 3.6 are shown in plots A and B (top row) and data generated according to Equation 3.7 are shown in plots C and D (bottom row).

### 3.3.3 Simulation Study Experimental Design

We are interested in comparing different variances often used when summarizing results from multimodel inference for partial regression coefficients: the full model-averaged variance, the variance from conditioning on the models where the coefficient is estimated, and the variance from the model with the largest posterior model probability. We are particularly interested in when and why these differences are

large enough to change inferences, or small enough not to change inferences. We investigate three factors that are potentially related to changes in the variance ratios described in Sections 3.3.1.1 - 3.3.1.3.

3.3.3.1 Factor 1: Correlation The probability associated with the point mass at zero (PEP) contributes to the posterior variances of the model-averaged partial regression coefficients (as shown in Figure 3.1). Similarly, results in George and McCulloch (1993) show that when  $\mathbf{X}_3$  and  $\mathbf{X}_5$  are highly correlated ( $\rho_{3,5} > 0.9$ ), the PEPs for both variables are close to 0.5, which we would expect to have a large impact on the model-averaged posterior distribution (particularly when its continuous component is centered away from zero with relatively small variance, as seen in Figure 3.1). Therefore, correlation between explanatory variables is tied to the PEPs for their associated partial regression coefficients, and thus to their full model-averaged posterior, though we are not aware of formal investigations of the change in variance. When variables are so highly correlated that one variable is essentially a proxy for the other (like the George and McCulloch (1993) case), researchers will often exclude one of the variables before selecting a model or building a model set. However, when variables are moderately correlated, researchers are less inclined to exclude variables prior to model selection. We have actually found many interesting questions that can only be addressed by multiple linear regression models including correlated variables. Therefore, we define a factor (*cor-35*) to capture six levels of correlation between  $X_3$  and  $X_5$  (0.0, 0.3, 0.6, 0.7, 0.8, 0.9) to get a sense for how moderate to high correlation among explanatory variables relates to resulting PEPs, and in turn to the different variance comparisons of interest in our investigation.

3.3.3.2 Factor 2: Signal to Noise We allow for two signal to noise ratios to reflect different data generating processes (systems of interest) and data collection

mechanisms having the potential to affect PEPs and variance ratios. The coefficients associated with  $X_4$  and  $X_5$  in the data generating model are 1 and 1.2, respectively. We manipulate the signal to noise ratio by forcing the value of the residual standard deviation to be equal to either 1 or 2.5. When the residual standard deviation (noise) is 2.5, the magnitude of the coefficients is small in comparison to the noise (low signal to noise ratio). Conversely, when the residual standard deviation is 1, the magnitude of the coefficients is large compared to the noise (high signal to noise). Therefore, we define *sig-y* as a factor with two levels: 1 and 2.5, corresponding directly to the signal to noise ratio (high and low, respectively). We expect higher signal to noise to dampen the effects of correlation on PEP, and in turn the variance comparisons of interest.

3.3.3.3 Factor 3: Truth Burnham and Anderson (2002) take issue with many simulations studies in the multimodel inferential context because they typically include the data generating model (the truth) in the model set. Burnham and Anderson (2002) argue that truth is infinitely complex and therefore, cannot be replicated by any model with a finite number of covariates; they also feel that results from simulation studies where truth is included in the model set are not realistic for problems in practice. We define the factor *truth* with three levels (YES, NO5, NO-taper4) indicating the presence of the data generating model in the model set (the YES and NO), and allowing for two different data generating models (the 4 and 5). For YES and NO5, the data are generated according to Equation 3.6, and the *all subsets* model set either contains models with  $X_5$  (including the data generating model - YES) or it does not (forcing the data generating model and other models containing  $X_5$  out of the model set - NO5). The level NO-taper4 uses Equation 3.7 as the data generating model including additional polynomial terms in  $X_4$  (a squared and cubed

term), with associated regression coefficients small in magnitude with respect to the two first order terms in the model. Thus, capturing unidentified (or unexpected) polynomial structure in the terms associated with the  $X_4$  variable, which we will call a tapering polynomial effect. Following typical convention for implementing model averaging, we use the *all subsets* model set defined by considering all five  $X$  variables with this data generating model (i.e., the same model set used in YES). Together, this data generating model and model set conducts model averaging without considering the tapering polynomial effects actually informing  $\mathbf{y}$ .

We use all combinations of the levels of the three factors, defining a three factor factorial design with 36 treatments. We investigate all treatments via simulation and in Section 3.4 choose particular comparisons among treatments, or levels within a factor, to highlight the range of observed results.

3.3.3.4 Implementation We use the Bayesian implementation of model averaging described in Chapter 2, taking advantage of the closed form posteriors resulting from using the  $g$ -prior (Zellner, 1984). In this formulation, we place a unit information prior ( $g$ -prior with  $g = n$ ) on the regression parameters, and a uniform prior on the model set. This leads to posteriors in the form of  $t$ -distributions for the partial regression coefficients from each individual model, and a mixed distribution comprised of a weighted average of those  $t$ -distributions and a point mass at zero defining the model-averaged posterior (see Appendix A). We use the `bms` function in the **BMS** R package (Feldkircher and Zeugner, 2009; R Core Team, 2016) to obtain the analytical means (A.4) and variances (A.5) of the posterior distributions of the partial regression coefficients associated with each explanatory variable in each individual model. We also obtain posterior model probabilities for all models in  $\mathcal{M}$  and the posterior

variances of the model-averaged partial regression coefficients (also using the `bms` function).

For each of the 36 treatments defined in Section 3.3.2, we generate 300 realizations of data because we found that was a sufficient number of realizations to characterize the relationships we investigate. The explanatory variables are generated with the correlation structure specified in Section 3.3.2, however, we observed the range of actual correlations among the explanatory variables was between  $\pm 0.3$  of the specified value. To ensure the specified levels of *cor-35* are actually reflected in the simulated data, we set a tolerance of  $\pm 0.01$  from the specified value and only keep data realizations meeting this criterion. Within each level of correlation, the same random number generator seeds were set, effectively holding variance in the  $X$ s constant, to better describe differences among the levels of the other factors.

For each realization of data, we implement model averaging using the *all subsets* model set (the default in `bms`). For treatments excluding models containing  $X_5$  from the model set, we exclude  $X_5$  from the data frame passed to the `bms` function. We use output from the `bms` function to obtain PEPs for the partial regression coefficients associated with each explanatory variable included in each realization's  $\mathcal{M}$ , and to compute the desired ratios of variances using Equation 3.5 (defined in Section 3.3.1).

### 3.4 Results

The posterior exclusion probability (PEP) is the probability associated with the point mass at zero, coming from the sum of the posterior model probabilities (PMPs) over the subset of models for which the coefficient is set to exactly zero. Thus, the PEP is directly tied to the definition of the model-averaged posterior distribution for a partial regression coefficient. It is particularly interesting to think about the relationship between PEP and the ratio of variances comparing the variance of the

model-averaged distribution coming from all models in the model set to the variance of the distribution conditional on models where the coefficient is allowed to be estimated, because the subset of models excluded from the variance computed in the denominator defines the PEP. We begin by investigating this relationship analytically in Section 3.4.1.

### 3.4.1 Analytical Results

In this section, we investigate the theoretical relationship between the PEP and the ratio of variances comparing the variance of the full model-averaged distribution to the variance of only its continuous component (i.e., conditional on only models including the associated variable). In Sections 3.4.2 - 3.4.4, we use simulation to generalize these results to investigate different treatment comparisons of the three variances of interest.

The model-averaged posterior distribution for a partial regression coefficient is a mixed distribution comprised of a point mass at zero, coming from the subset of models where the coefficient is set to exactly zero, and a weighted average of continuous posterior distributions, coming from the subset of models where the coefficient was estimated. The probabilities associated with these two subsets of models are PEP and 1-PEP, respectively. Let  $W_1$  and  $W_2$  denote random variables with the two distributions described above, respectively. Assuming  $W_1$  and  $W_2$  each have a distribution with finite moments about the origin, we write the first two moments about the origin of  $W$  as

$$E(W) = [PEP \times E(W_1)] + [(1 - PEP) \times E(W_2)] \text{ and}$$

$$E(W^2) = [PEP \times E(W_1^2)] + [(1 - PEP) \times E(W_2^2)].$$

Using the identity,  $Var(W) = E(W^2) - (E(W))^2$ , and noting the expected value of  $W_1$  is zero, the variance of  $W$  can be written and simplified as follows

$$\begin{aligned} Var(W) &= [(PEP)E(W_1^2) + (1 - PEP)(E(W_2^2))] \\ &\quad - [(PEP)^2E(W_1)^2 + (1 - PEP)^2(E(W_2))^2] \\ &= [(1 - PEP)E(W_2^2)] - [(1 - PEP)^2(E(W_2))^2]. \end{aligned} \quad (3.8)$$

We use the mean and variance of  $W_2$  to write  $E(W_2^2) = Var(W_2) + (E(W_2))^2$ , and write the variance of  $W$  as a function of obtainable quantities,

$$\begin{aligned} Var(W) &= [(1 - PEP)(Var(W_2) + (E(W_2))^2)] - [(1 - PEP)^2(E(W_2))^2] \\ &= (1 - PEP) [Var(W_2) + (PEP)(E(W_2))^2]. \end{aligned} \quad (3.9)$$

Then, the ratio of the variance of the model-averaged distribution to the variance of its continuous component can be written as a function of the PEP,  $Var(W_2)$ , and  $E(W_2)$  as

$$\frac{Var_{MA}}{Var_{cts}} = \frac{(1 - PEP) [Var(W_2) + (PEP)(E(W_2))^2]}{Var(W_2)}. \quad (3.10)$$

It is interesting to investigate how the ratio of variances changes as a function of the PEP, given fixed values of  $E(W_2)$  and  $Var(W_2)$ , because the value of the PEP is directly tied to the subset of models excluded from  $Var_{cts}$ . That is, we learn something about how the subset of models making up the zero component contributes to the variance of the full model-averaged posterior distribution. We first investigate properties of the relationship between the ratio of variances and PEP by using (3.10), and then plot the relationship for specific values of  $E(W_2)$  and  $Var(W_2)$ .

Although the PEP may never be exactly zero or one in practice, we can think about the values of the ratio of variances as the PEP approaches these two endpoints by taking limits. The limit of the ratio of variances as PEP goes to zero is one, and the limit as PEP goes to one is zero. The ratio of variances is a second order polynomial in PEP and is continuous on the interval  $[0, 1)$  (there is a removable discontinuity at  $PEP = 1$  because  $Var[W_2]$  does not exist at  $PEP = 1$ ). By the intermediate value theorem, we know there are ratios of variances that are less than one, implying there are PEPs such that the variance of the continuous component of the model-averaged posterior is larger than the variance of the full model-averaged distribution. The first and second derivatives of (3.10) reveal a maximum at  $PEP = \frac{1}{2} \left(1 - \frac{Var(W_2)}{E^2(W_2)}\right)$ , and solving directly for the value(s) of PEP such that the ratio of variances is one results in  $PEP = 1 - \frac{Var(W_2)}{E^2(W_2)}$ . Therefore, the most interesting properties of the curve, namely the maximum value of the ratio of variances and the value(s) for which the variances are the same are both a function of the relationship between the expected value and the variance of the component of the model-averaged distribution made up by models allowing the coefficient to be estimated (as previously postulated from Figure 3.1).

We are interested in understanding how the variance of the full model-averaged posterior (including all models) is influenced by the subset of models defining the point mass at zero, particularly for partial regression coefficients of interest. We expect the posteriors for the partial regression coefficients of interest to be centered away from zero, so we fix the central location of  $W_2$  at 3 and use three different standard deviations (3, 1.5, 0.5) to investigate the general relationship between ratio of variances ( $Var_{MA}/Var_c$ ) and the PEP for different values of the ratio of  $E(W_2)$  to  $Var(W_2)$  values (Figure 3.5).

The variance of only the continuous component of the model-averaged posterior distribution is either equal to, or greater than, the variance of the model-averaged

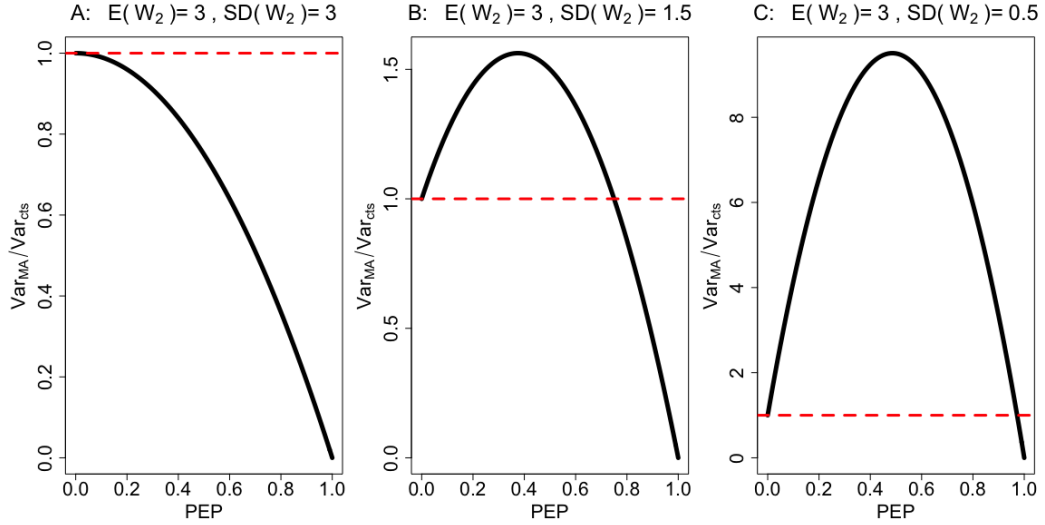


Figure 3.5: Analytical results for the ratio of variance comparing the model-averaged distribution to its continuous component. The three examples presented have continuous components with normal distributions with mean 3 and standard deviation equal to 3, 1.5, and 0.5 respectively (read left to right).

distribution when the mean and standard deviation of the continuous component are equal (plot A in Figure 3.5). This is not surprising because the PEP is the probability associated with the point mass at zero (coming from models where the coefficient is exactly zero), and when the variance of the continuous component (coming from models where the variable is included) is large in comparison to its mean, zero will likely be a plausible value (e.g., plots A, B and C in Figure 3.1). Therefore, as the PEP increases, the posterior distribution from full model averaging, considering all models, has larger mass associated with the point mass at zero and its variance is smaller in comparison to the variance of the continuous component, considering the subset of models for which the coefficient is estimated.

When the mean of the continuous component of the model-averaged posterior distribution is centered away from zero and twice the value of its standard deviation, the ratio of variances is larger than one for a range of values of PEPs,  $(0, 0.75)$ ,

achieving a maximum value of about 1.56 at a PEP of 0.375 (plot B in Figure 3.5). This indicates the variance from the full model-averaged distribution is, at most, about one and a half times larger than the variance of its continuous component. Furthermore, this maximum occurs when there is a substantial point mass at zero (i.e., the sum of the posterior model probabilities over models where the variable is excluded is sizable), *and* the posteriors coming from the subset of models for which the coefficient is estimated suggest it is different from zero. As the value of PEP increases from 0.375 to 1, the values of the ratio of variances decrease as a direct result of more of the full model-averaged posterior distribution being made up by the point mass at zero. Provided the corresponding continuous component is distributed such that it includes zero and the probability associated with a tight interval around zero is not tiny, the variance of the full model-averaged distribution will decrease in comparison to the variance of the continuous component as the PEP increases (e.g., plots D, E, and F in Figure 3.1).

In the last case (plot C in Figure 3.5), the weighted averaged of the posteriors for the partial regression coefficients coming from models where the corresponding variable is included is centered away from zero with very small variance and is very inconsistent with the point mass at zero. Therefore, even for very small PEP, the point mass at zero adds enough variability to the fully model-averaged posterior distribution (weighted average over *all* models) to make it more variable than its corresponding continuous component (as seen in Figure 3.1 plots H, I, and J). It is also interesting to note that the values of the ratios of variances in this example are much larger than they were in the previous example. The maximum value occurs at a PEP of 0.486 and is equal to 9.51, indicating the full model-averaged posterior distribution has a variance almost ten times larger than the variance of only its continuous component.

To summarize, we filled in the holes from the intuition we built with derived formulas, and investigated how the curve representing the relationship between PEP and the ratio of the variance of the full model-averaged distribution to the variance of its continuous component changes for three different values sets of values of  $E(W_2)$  and  $Var(W_2)$ . These analytical curves exist for the entire domain of PEP, and we expect to see similar shapes in the simulation results presented in Sections 3.4.2 - 3.4.4. However, there may be some combinations of PEP and  $Var_{MA}/Var_{cts}$  that would be unrealistic to observe in practice, and the simulation study reveals which parts of the curve are most likely under the different treatments (described in Sections 3.3.3.1 - 3.3.3.3).

#### 3.4.2 All Models Compared to Subset of Models

In this section, we investigate the simulation results for the ratio of variances  $Var_{MA}$  to  $Var_{cts}$ . This is an interesting ratio to investigate because we know what we may expect to see given the theoretical results in Section 3.4.1. We also get a sense for how much additional variance the subset of models defining the zero component contributes to the variance of the full model-averaged posterior distribution. We can make this comparison for every realization of data for all partial regression coefficients of interest. We organize results into sections by the factors described in Section 3.3.3.1 - 3.3.3.3 and highlight comparisons among particular levels adding the most information to our overall characterization of this ratio of variances. We start by investigating the differences among the levels of correlation within fixed levels of signal to noise and truth in the model set, which we use as a basis for comparisons among levels of the other factors and among the three ratios of variances.

3.4.2.1 Comparing Correlations We first compare results among the six levels of correlation within low signal to noise and truth in the model set. We refer to this

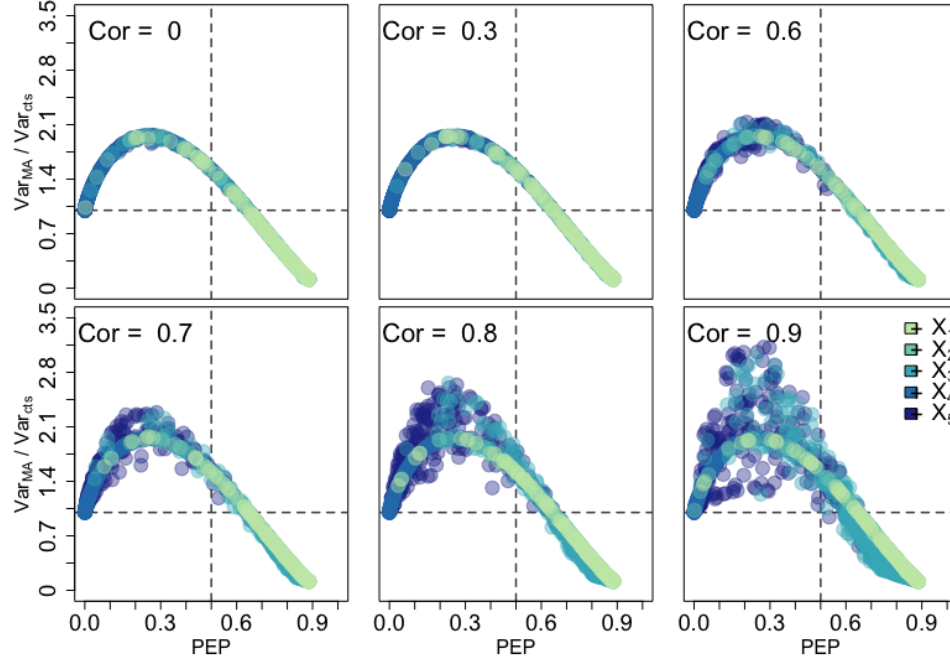


Figure 3.6: Results from the reference scenario investigating the relationship between PEP and  $Var_{MA} / Var_{cts}$  for all correlations and all variables of interest.

collection ( $cor-35 = \text{all}$ ,  $sig-y = 2.5$ , and  $truth = \text{YES}$ ) as the “reference scenario” because we will compare its results to those from other combinations of levels of signal to noise and truth in the model set. Results from three hundred realizations at each level of correlation for this reference scenario are shown in Figure 3.6. Each ‘o’ represents an ordered pair  $(PEP, Var_{MA} / Var_{cts})$  resulting from the model averaging process applied to one realization of data. There are five ‘o’s per realization (one associated with each explanatory variable), and the plotting characters are color coded by variable. We observe a very strong relationship between the ratio of variances and the PEP for all levels of correlation (Figure 3.6) resembling one of the analytical curves presented in Section 3.4.1. This relationship appears to be the same for the ratios of variances associated with all explanatory variables. However, it is difficult

to see the range of PEPs observed for each variable in each plot because they are superimposed.

Results for each variable are shown individually for the correlation values 0.0 and 0.9 in Figures 3.7 and 3.8, respectively (results from intermediate correlations can be created following the steps outlined in Appendix C). We chose to display these two correlations because the results for a correlation of 0.3 closely resemble those for 0.0 and the results for the correlations 0.6, 0.7 and 0.8 closely resemble those for 0.9 (with slightly more variation around the observed relationship for variables  $X_3$  and  $X_5$  as the correlation between  $X_3$  and  $X_5$  increases, as seen in Figure 3.6). In conclusion, the underlying relationship between PEP and the ratio of variances is the same for all variables for each level of correlation, and this relationship is not surprising given the analytical results in Section 3.4.1. They most closely relate to the case where the mean of the continuous component was centered away from zero and had relatively small standard deviation (about half the size of the mean). It appears there are similar contributions from the models associated with the continuous component and the models associated with zero component to the full model-averaged posterior distribution for each variable.

Something we were not able to capture in the analytical results is the variability around the relationship between PEP and the ratio of variances for the correlated variables ( $X_3, X_5$ ). This variability increases as the correlation between the two variables increases (Figure 3.6), a direct result of the correlation between these two variables inducing more uncertainty (i.e., PMPs are substantial for many models in the model set). The true data generating model includes  $X_4$  and  $X_5$ , so when  $X_3$  and  $X_5$  are highly correlated, the subset of models with relatively large posterior model probabilities becomes more variable (i.e., sometimes the model with the largest PMP will include  $X_3$  and  $X_4$ ,  $X_4$  and  $X_5$ , or all three variables). Variability in which model

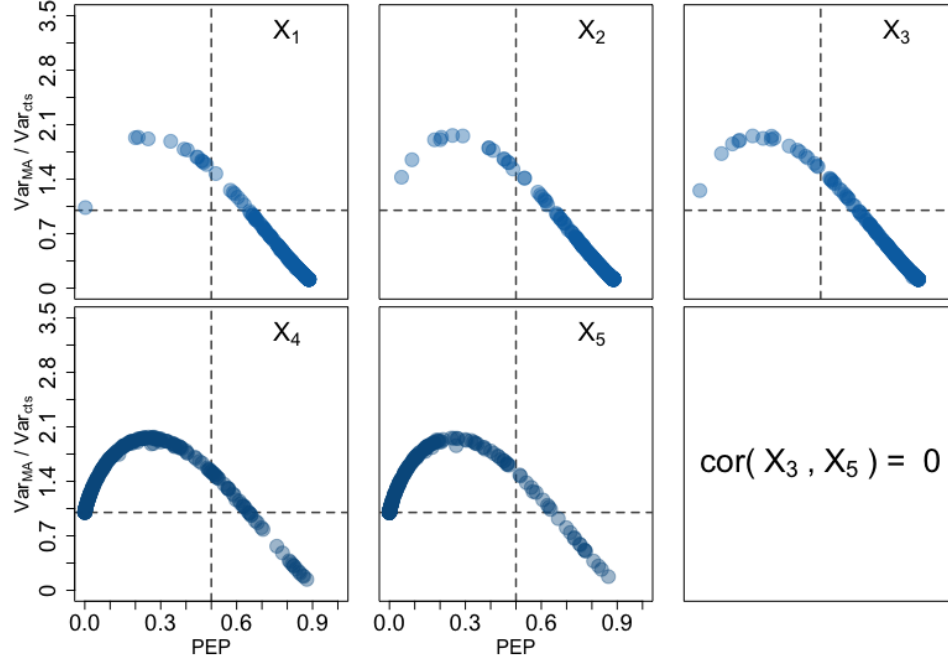


Figure 3.7: Results from the reference scenario investigating the relationship between PEP and  $Var_{MA} / Var_{cts}$  for correlation of  $X_3$  and  $X_5$  equal to 0.0, displayed for all variables of interest. Results associated with variables that are truly in the data generating model have slightly darker plotting characters than those that are not in the data generating model.

ends up with the largest PMP across realizations results in more options for the ratio of variances for fixed PEP (e.g., Figures 3.6 - 3.8).

When results are split up by explanatory variable (Figures 3.7 and 3.8), we can see the range of realistic PEPs for realizations of data from each explanatory variable. We observe more realizations with lower PEPs for the variables  $X_4$  and  $X_5$  than we do for the variables  $X_1$ ,  $X_2$ , and  $X_3$  because  $X_4$  and  $X_5$  are truly in the data generating model, and therefore, generally explain more variability in  $y$ . As the correlation between  $X_3$  and  $X_5$  increases, the density of realizations at lower PEPs increases for  $X_3$ . This is consistent with what we expect because when  $X_3$  is highly

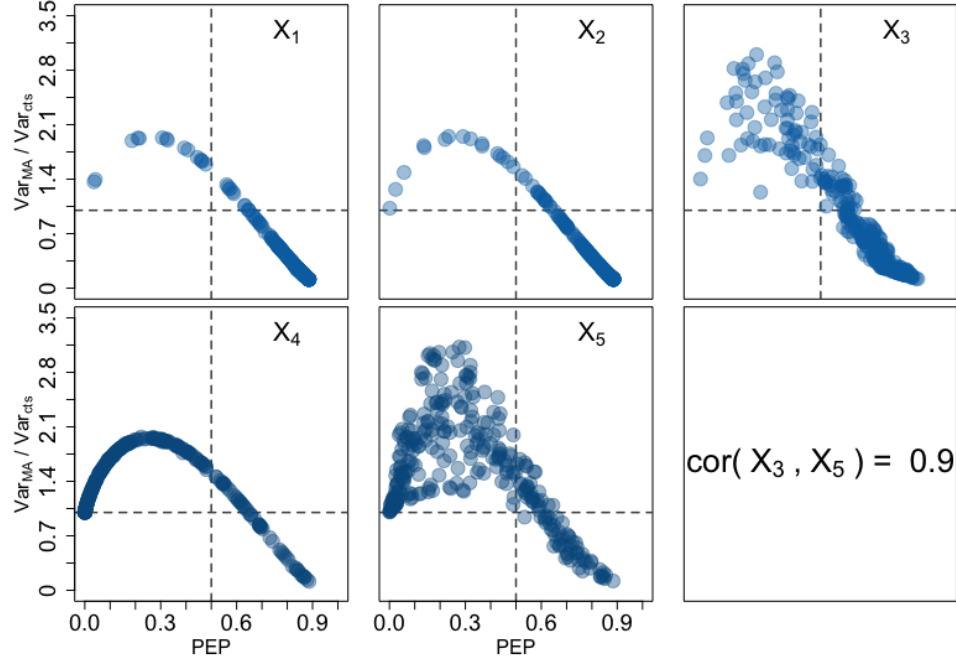


Figure 3.8: Results from the reference scenario investigating the relationship between PEP and  $Var_{MA} / Var_{cts}$  for correlation of  $X_3$  and  $X_5$  equal to 0.0, displayed for all variables of interest. Results associated with variables that are truly in the data generating model have slightly darker plotting characters than those that are not in the data generating model.

correlated with  $X_5$  (a variable in the true data generating model), models including  $X_3$  will tend to receive more of the posterior model mass for more realizations of data.

The results presented in Figures 3.6 - 3.8 highlight a range of values of PEP for each variable where the magnitude of the posterior variance coming from model averaging using all models will be higher than the variance coming from only the models allowing the coefficient to be estimated. There is greater potential for this multiplicative difference to be larger when the correlation between  $X_3$  and  $X_5$  is high. Therefore, it is not always true that full model averaging (using all of the models in the model set) results in a larger variance for a model-averaged partial regression coefficient than model averaging using the subset of models for which the coefficient

is allowed to be estimated. There is greater potential for variances from fully model-averaged posterior distributions to be larger than the variances from their continuous components when the partial regression coefficient of interest is associated with a variable that is correlated with a variable in the data generating model.

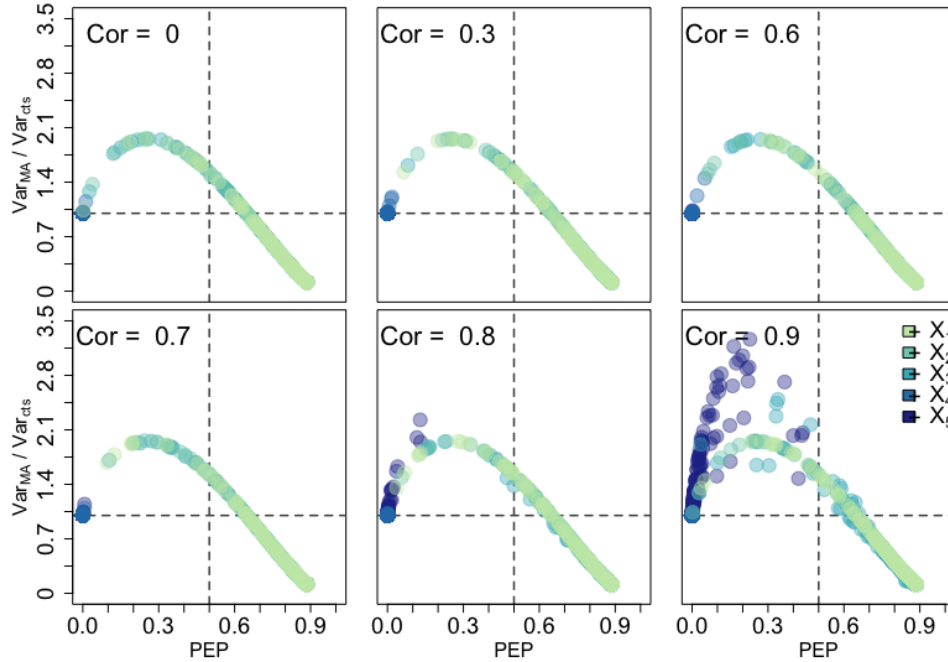


Figure 3.9: Results for each level of correlation from three hundred realizations of data simulated with high signal to noise, and used for model averaging where truth is included in the model set.

**Signal to Noise** Results across all levels of correlation for high signal to noise level and model averaging conducted with the data generating model in the model set are shown in Figure 3.9. The conclusions regarding comparisons among the correlations are not drastically different for low signal to noise as compared to high signal to noise (holding truth in the model set constant). That is, the correlation between  $X_3$  and  $X_5$  affects the variation around the underlying relationship between PEP and the ratio of variances similarly to the reference scenario. We note that the range of PEPs

observed for low signal to noise is more consistent with what we would expect given the variables in the data generating model, and we discuss this difference in more detail in Section 3.4.2.2.

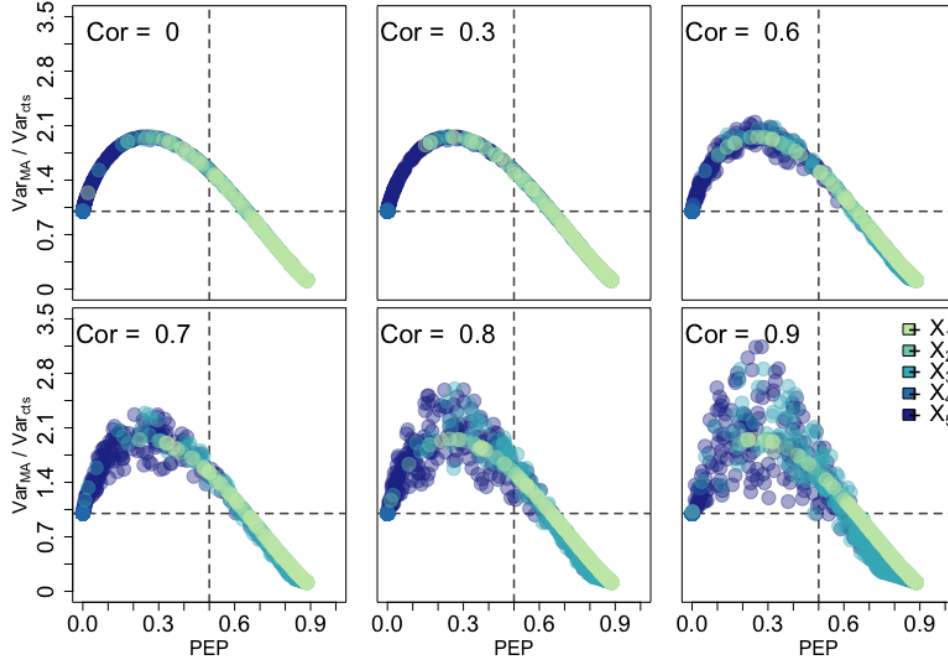


Figure 3.10: Results for each level of correlation from three hundred realizations of data simulated with low signal to noise ratio, and used for model averaging where truth is excluded from the model set by excluding true tapering polynomial terms associated with  $X_4$ .

**Truth in the Model Set** Holding signal to noise constant, conclusions regarding comparisons among the correlations are similar when we move from truth being included in the model set ( $\text{truth} = \text{YES}$ ) to the level of truth in the model set where tapering effects for  $X_4$  are in the data generating model (i.e., Equation 3.7), but excluded from model set ( $\text{truth} = \text{NO-taper4}$ ; see Figure 3.10). However, when  $X_5$  is included in the data generating model but excluded from the model set ( $\text{truth} = \text{NO5}$ ), we make different observations for comparisons among levels of correlation for fixed signal to noise levels. These observations indicate an interaction between correlation

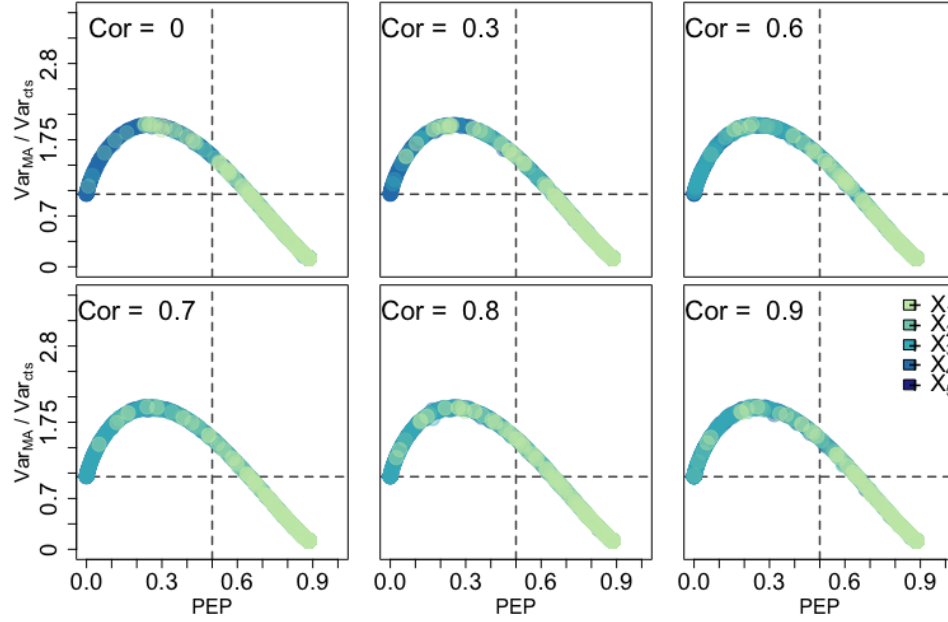


Figure 3.11: Results for each level of correlation from three hundred realizations of data simulated with low signal to noise, and used for model averaging where truth is excluded from the model set by excluding the variable  $X_5$  from models in the model set.

and truth in the model set (see Figure 3.11). We discuss broader comparisons between the relationship between the levels of truth and the model set for fixed levels of correlation in Section 3.4.2.3.

3.4.2.2 Comparing Signal to Noise Ratios In this section, we change the level of signal to noise to high ( $sig-y = 1$ ), and compare results to those obtained in the previous section when signal to noise was low. We highlight two of the correlations considered (0.0 and 0.9) because results from treatments with correlations equal to 0.3 are similar to the results when correlation is equal to 0.0, and the results when the correlation is equal to 0.6, 0.7, and 0.8 are similar to results when correlation is equal to 0.9 (Figure 3.12 and Figure 3.13). In both figures, the plotting character indicates

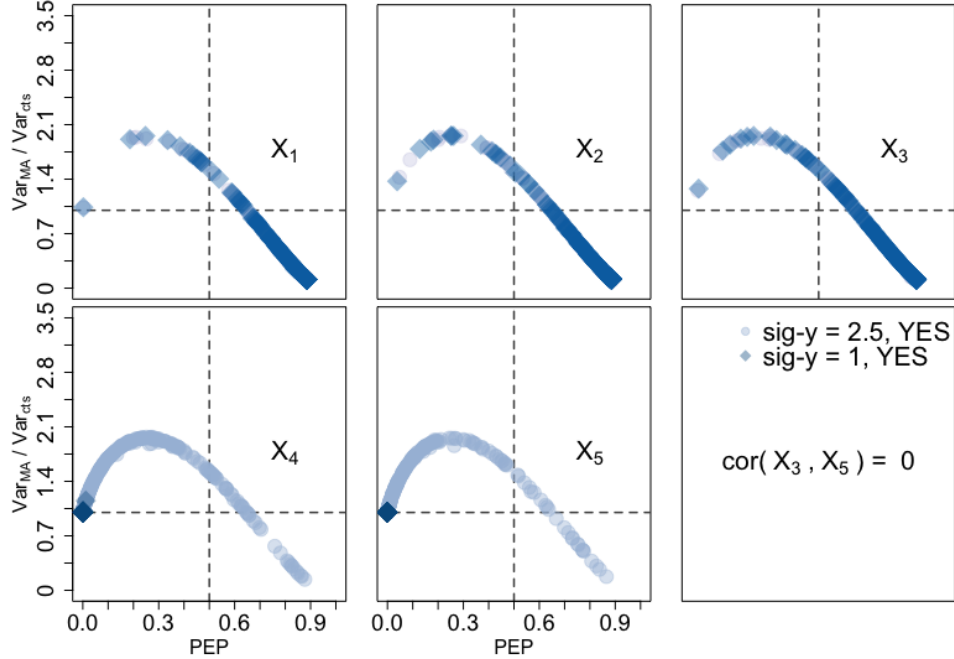


Figure 3.12: This figure facilitates the comparison of results from data generated with low signal to noise to those generated with high signal to noise. Truth in the model set is held constant at  $truth = YES$  (data generating model included in the model set used for model averaging) and correlation is held constant at  $cor-35 = 0.0$  for both sets of results. Plotting characters are coded such that ‘o’ represents low signal to noise ( $sig-y = 2.5$ ), and ‘◇’ represents high signal to noise ( $sig-y = 1$ ). Results associated with variables that are truly in the data generating model have slightly darker plotting characters than those that are not in the data generating model.

the level of signal to noise used in the data generating process (‘o’: low signal to noise,  $sig-y = 2.5$ ; ‘◇’: high signal to noise,  $sig-y = 1$ ). Results from intermediate correlations can be created following the steps outlined in Appendix C.

First, we focus on comparing the two signal to noise ratios when the correlation is equal to zero (Figure 3.12). For both levels of signal to noise, variables not in the data generating model ( $X_1$ ,  $X_2$ , and  $X_3$  top row of Figure 3.12) tend to have high PEPs, and variables in the data generating model ( $X_4$  and  $X_5$  bottom row of Figure 3.12) tend to have low PEPs. However, results coming from realizations

generated with high signal to noise ( $\diamond$ ) are more concentrated around the PEPs than we would expect (i.e., they are almost always 0 for  $X_4$  and  $X_5$  and are almost always greater than 0.5 for  $X_1$ ,  $X_2$ , and  $X_3$ ) than the results coming from realizations generated with low signal to noise ( $\circ$ ). Therefore, when the signal to noise ratio is high and correlation is low, variables in the data generating model (including the data generating model itself) tend to have large posterior model probabilities. For variables in the data generating model ( $X_4$  and  $X_5$ ), there is little difference between  $Var_{cts}$  and  $Var_{MA}$  because the PEP is essentially 0. For variables not in the data generating model ( $X_1$ ,  $X_2$ ,  $X_3$ ),  $Var_{cts}$  is larger than  $Var_{MA}$  for most realizations of data, and the associated PEPs are almost always greater than 0.5. Therefore, for cases where the signal to noise level is high, correlation among the data generating variables is low, and truth is in the model set, there are very few occasions when the inference made accounting for all models in the model set will result in larger variances than the inference made by conditioning on the subset of models where the coefficient is estimated.

When the correlation between  $X_3$  and  $X_5$  is high (refer to Figure 3.13), we see more variability in the ratios of variances for fixed PEPs for the posteriors for coefficients associated with  $X_3$  and  $X_5$  for both levels of signal to noise. However, when the signal to noise ratio is high, the ratio of variances for  $X_4$  is almost always equal to one because models with  $X_4$  consistently have high PMPs for almost all realizations of data.

In summary, for realizations of data generated with a large signal to noise ratio, we observe large posterior model probabilities for models including variables found in the data generating model, even for large correlations between  $X_3$  and  $X_5$ . This results in very few realizations with ratios of variances larger than one, implying  $Var_{MA} \approx Var_{cts}$ . Consequently, when studying systems with high signal to noise

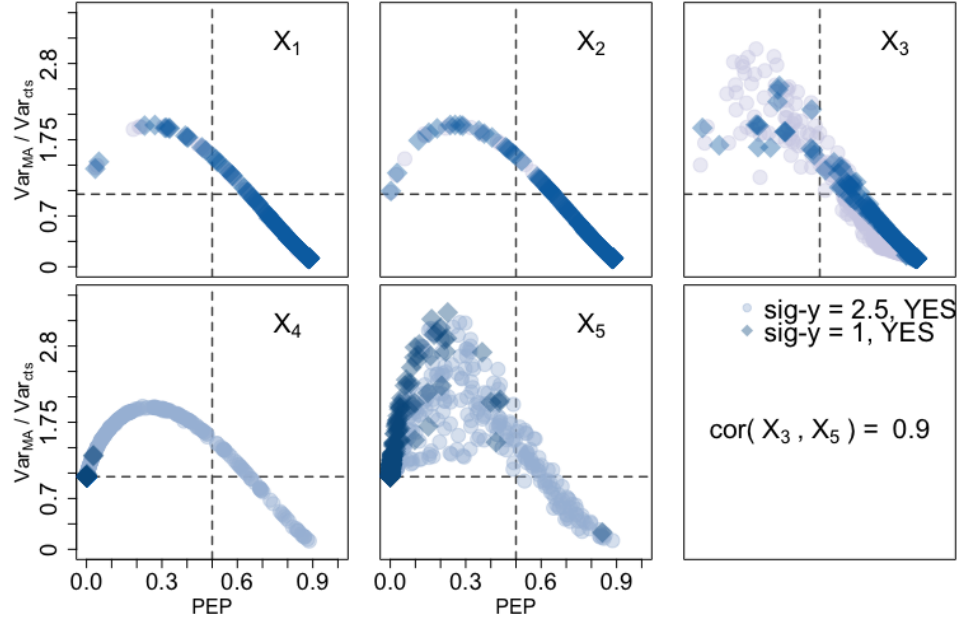


Figure 3.13: This figure facilitates the comparison of results from data generated with low signal to noise to those generated with high signal to noise. Truth in the model set is held constant at  $truth = YES$  (data generating model included in the model set used for model averaging) and correlation is held constant at  $cor-35 = 0.9$  for both sets of results. Plotting characters are coded such that ‘o’ represents low signal to noise ( $sig-y = 2.5$ ), and ‘◊’ represents high signal to noise ( $sig-y = 1$ ). Results associated with variables that are truly in the data generating model have slightly darker plotting characters than those that are not in the data generating model.

ratios, there will be little differences between inferences based on all models in the model set (full model averaging) and inferences based on only models including the explanatory variable (i.e., where the coefficient is estimated).

**Truth in the Model Set** The comparison between the levels of signal to noise for both of the other levels of truth in the model set are very similar to what we observed for the reference scenario (Figures 3.14 & 3.15). When the signal to noise ratio is higher, we observe large posterior model probabilities for models including variables found in the data generating model, even for large correlations between  $X_3$  and  $X_5$ .

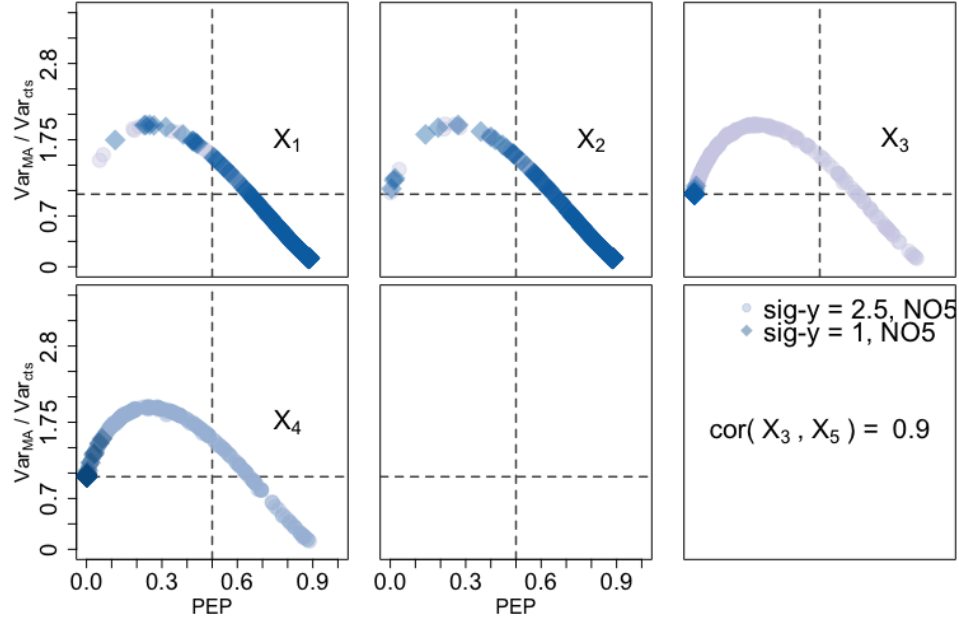


Figure 3.14: This figure facilitates the comparison of results from data generated with low signal to noise to those generated with high signal to noise. Truth in the model set is held constant at  $truth = NO5$  (data generated according to Equation 3.6 and models including  $X_5$  excluded from the model set used for model averaging) and correlation is held constant at  $cor-35 = 0.9$  for both sets of results. Plotting characters are coded such that ‘o’ represents low signal to noise ( $sig-y = 2.5$ ), and ‘◇’ represents high signal to noise ( $sig-y = 1$ ). Results associated with variables that are truly in the data generating model have slightly darker plotting characters than those that are not in the data generating model.

The increased PEPs for variables in the data generating model at high signal to noise (as compared to low signal to noise) are slightly less pronounced for the realizations coming from data generated according to  $truth = NO-taper4$  (Figure 3.15).

3.4.2.3 Comparing Truth in the Model Set The three levels of truth in the model set are YES, NO5, and NO-taper4: YES, corresponding to conducting model averaging with the data generating model in the model set (Equation 3.6 and model set defined by *all subsets* for the five  $X$  variables); NO5, corresponding to conducting

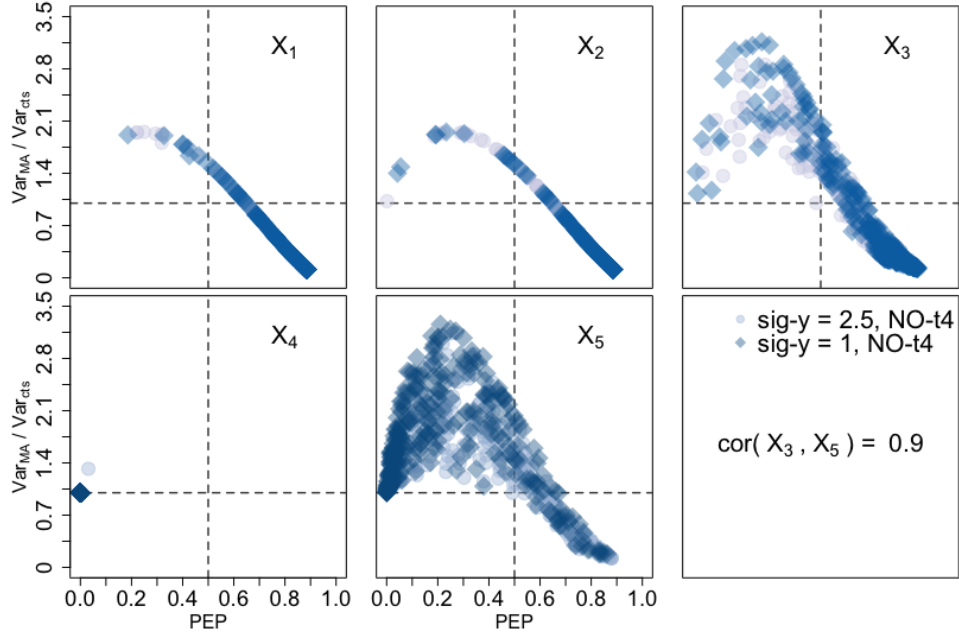


Figure 3.15: This figure facilitates the comparison of results from data generated with low signal to noise to those generated with high signal to noise. Truth in the model set is held constant at  $\text{truth} = \text{NO-taper4}$  (data generated according to Equation 3.7 and *all subsets* model set with no polynomial terms used for model averaging) and correlation is held constant at  $\text{cor-35} = 0.9$  for both sets of results. Plotting characters are coded such that ‘o’ represents low signal to noise ( $\text{sig-y} = 2.5$ ), and ‘◊’ represents high signal to noise ( $\text{sig-y} = 1$ ). Results associated with variables that are truly in the data generating model have slightly darker plotting characters than those that are not in the data generating model.

model averaging with the data generating model defined in Equation 3.6 (including  $X_4$  and  $X_5$ ), but defining the *all subsets* model set without  $X_5$  (excluding models including  $X_5$  from the model set used in YES); and NO-taper4, corresponding to a data generating model including tapering polynomial effects associated with  $X_4$  (Equation 3.7), and then conducting model averaging with the *all subsets* model set used in YES. In the following paragraphs, we make the three possible pairwise comparisons among the levels of this factor within fixed correlations and low signal to noise ratio.

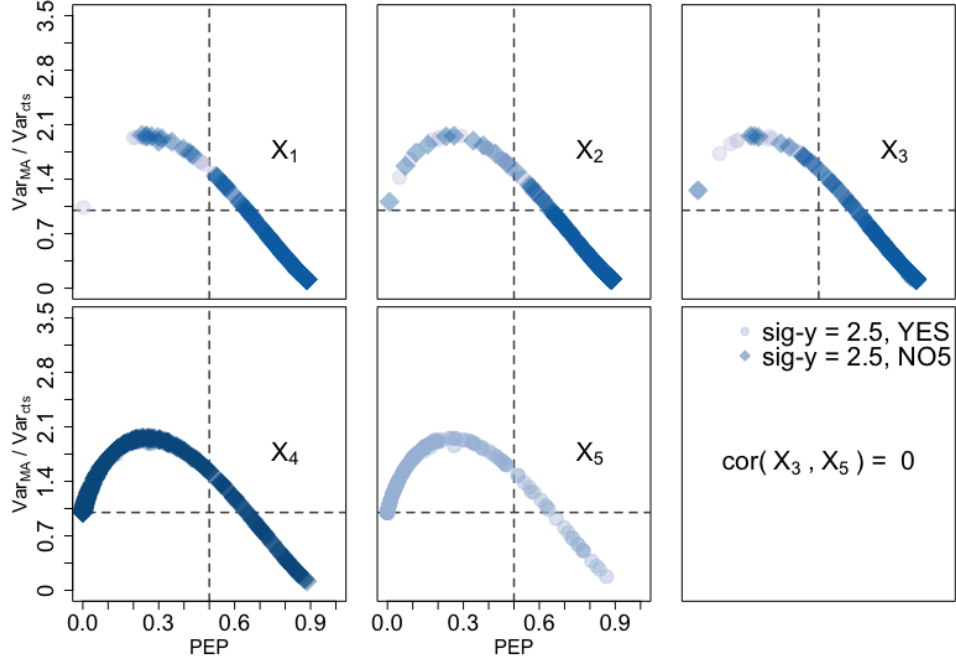


Figure 3.16: This figure facilitates the comparison of results from scenarios where model averaging is conducted with  $truth = YES$  to those conducted with  $truth = NO5$  (data generated according to Equation 3.6 and models including  $X_5$  excluded from model set). Signal to noise is held constant at low signal to noise ratio, and correlation is held constant at  $cor-35 = 0.0$  for both sets of results. Plotting characters are coded such that ‘o’ represents  $truth = YES$  and ‘ $\diamond$ ’ represents  $truth = NO5$ . Results associated with variables that are truly in the data generating model have slightly darker plotting characters than those that are not in the data generating model.

**Comparing YES to NO5** Here, we compare results from the reference scenario described in Section 3.4.2.1 (plotted as ‘o’), where truth is in the model set, to those obtained using a model set excluding models containing a data generating variable  $X_5$  (plotted as ‘ $\diamond$ ’). We investigate this comparison for all correlations and, again, highlight results from two correlations ( $cor-35 = 0, 0.9$ ) characterizing the range of results we observe (Figures 3.16 & 3.17 respectively). When the correlation is 0 (Figure 3.16), the results excluding truth from the model set are similar to the results observed in the reference scenario for all variables except  $X_5$ . When  $X_5$  is excluded

from the model set, its associated partial regression coefficients are not estimated, and therefore ratios of variances are not available, so we are focusing on the results for  $X_3$  (correlated with  $X_5$ ) and  $X_4$  (the other data generating variable).

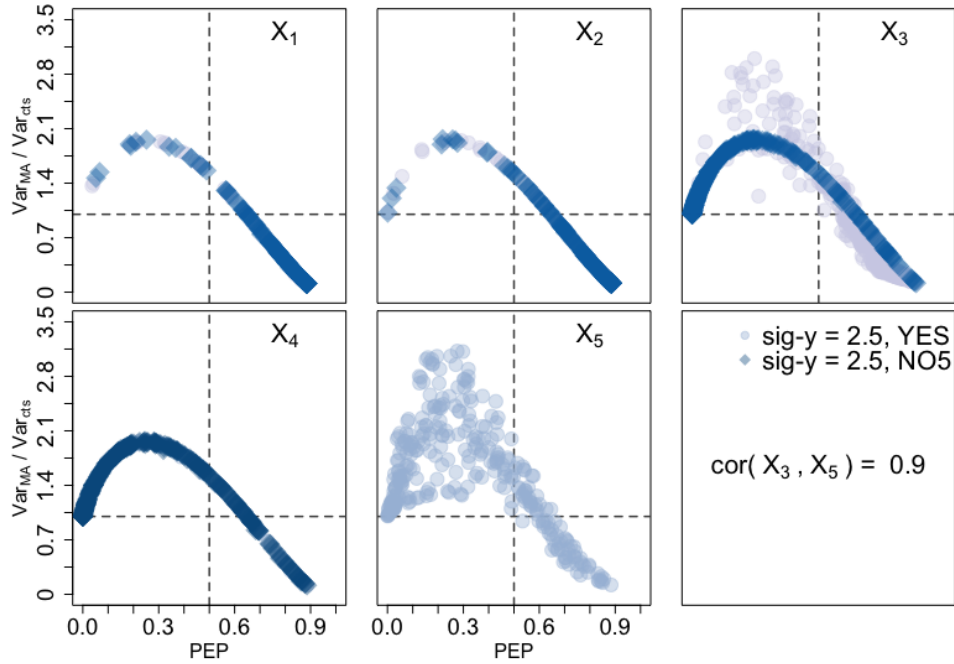


Figure 3.17: This figure facilitates the comparison of results from scenarios where model averaging is conducted with  $\text{truth} = \text{YES}$  to those conducted with  $\text{truth} = \text{NO5}$  (data generated according to Equation 3.6 and models including  $X_5$  excluded from model set). Signal to noise is held constant at low signal to noise ratio, and correlation is held constant at  $\text{cor-35} = 0.9$  for both sets of results. Plotting characters are coded such that ‘o’ represents  $\text{truth} = \text{YES}$  and ‘ $\diamond$ ’ represents  $\text{truth} = \text{NO5}$ . Results associated with variables that are truly in the data generating model have slightly darker plotting characters than those that are not in the data generating model.

When the correlation is 0.9, there are notable differences between results from the reference scenario and those obtained after excluding  $X_5$  (Figure 3.17). We note that unlike results from the reference scenario (‘o’), there is no increased variability around the relationship between the ratio of variances and the PEPs for any of the variables. Instead, we observe many realizations resulting in small PEPs for  $X_3$  and it

appears as though  $X_3$  has become a proxy for  $X_5$  resulting in high PMPs for models including  $X_3$  and  $X_4$  almost exclusively.

We see the potential for the variance of the partial regression coefficients associated with  $X_3$  and  $X_4$  to be larger when all models are considered as opposed to only the subset of models for which the coefficient is estimated. However, the maximum value of the ratio of variances is not affected by the correlation between  $X_3$  and  $X_5$  because all variables used to generate the models in the model set are generated as independent (but their correlation is not necessarily zero) when  $X_5$  is excluded from the model set.

**Comparing YES to NO-taper4:** Here, we compare results from the reference scenario to results obtained by generating data using a polynomial tapering effect of  $X_4$ , but then using the same model set as used in YES (does not include the polynomial terms; NO-taper4). We highlight the same correlations (0, and 0.9) for the same reason discussed in the previous sections. Results for the reference scenario are plotted with ‘o’s, and results from NO-taper4 are plotted with ‘◇’s (Figures 3.18 and 3.19). When the tapering effects are present in the data generating model, but not in the model set, the relationship between PEP and the ratio of variances is very similar to what we observed in the reference scenario, except for the variable  $X_4$ . When the data generating model has more information coming from polynomial terms in  $X_4$ , the results after model averaging reflect large PMPs for models including  $X_4$ , and the ratio of variances is almost always one with very little chance of being larger. That is, the variance associated with the partial regression coefficient for  $X_4$  coming from the method using all models (model averaging) is almost always equivalent to the variance coming from models conditioning on the subset of models allowing  $X_4$  to be estimated.

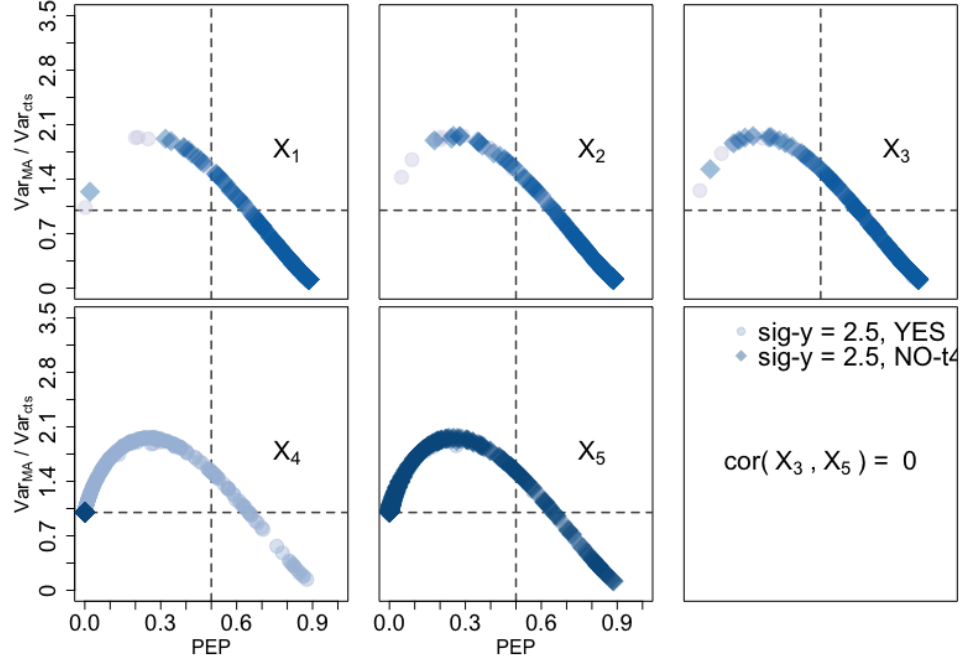


Figure 3.18: This figure facilitates the comparison of results from scenarios where model averaging is conducted with  $\text{truth} = \text{YES}$  to those conducted with  $\text{truth} = \text{NO-taper4}$  (data generated according to Equation 3.7 and *all subsets* model set is used). Signal to noise is held constant at low signal to noise ratio, and correlation is held constant at  $\text{cor-35} = 0.0$  for both sets of results. Plotting characters are coded such that ‘o’ represents  $\text{truth} = \text{YES}$  and ‘ $\diamond$ ’ represents  $\text{truth} = \text{NO-taper4}$ . Results associated with variables that are truly in the data generating model have slightly darker plotting characters than those that are not in the data generating model.

**Comparing NO5 to NO-taper4:** Here, we compare results obtained from model averaging after excluding  $X_5$  from the model set (NO5, plotted with ‘o’) to results obtained by excluding truth from the model set by excluding any higher order terms associated with  $X_4$  from the model set (NO-taper4, plotted with ‘ $\diamond$ ’), seen in Figures 3.20 and 3.21. For small correlations, we observe differences in the relationship between the ratio of variances and PEP for the variables  $X_4$  and  $X_5$  (Figure 3.20). When tapering effects are included in the data generating model, but excluded from models in the model set, almost all of the realizations result in PEPs for  $X_4$  very

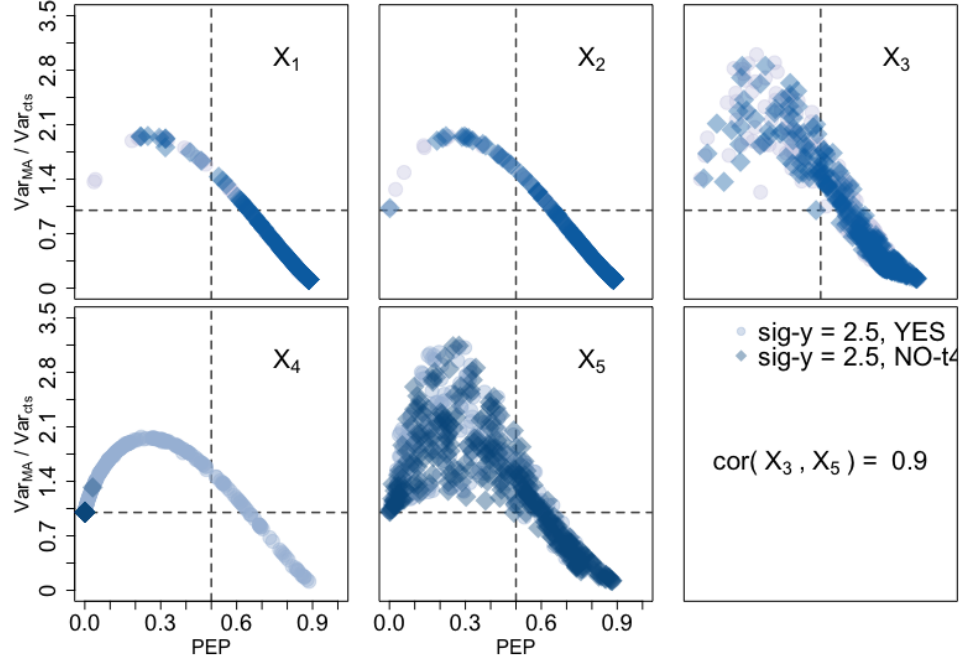


Figure 3.19: This figure facilitates the comparison of results from scenarios where model averaging is conducted with  $\text{truth} = \text{YES}$  to those conducted with  $\text{truth} = \text{NO-taper4}$  (data generated according to Equation 3.7 and *all subsets* model set is used). Signal to noise is held constant at low signal to noise ratio, and correlation is held constant at  $\text{cor-35} = 0.9$  for both sets of results. Plotting characters are coded such that ‘o’ represents  $\text{truth} = \text{YES}$  and ‘◇’ represents  $\text{truth} = \text{NO-taper4}$ . Results associated with variables that are truly in the data generating model have slightly darker plotting characters than those that are not in the data generating model.

close to one. When  $X_5$  is included in the data generating model, but is excluded from models in the model set, we see realizations resulting in PEPs for  $X_4$  covering almost the entire domain of PEP. As noted in the previous two sections, ratios of variances and PEPs are unavailable for  $X_5$  when it is excluded from models in the model set, so there are differences in results for  $X_5$  between the levels of truth in the model set NO-taper4 and NO5 because of this. When the correlation is large, we observe more variability around the relationship between the ratio of variances and PEP for  $X_3$  and  $X_5$  when tapering effects for  $X_4$  are in the data generating model, but excluded from

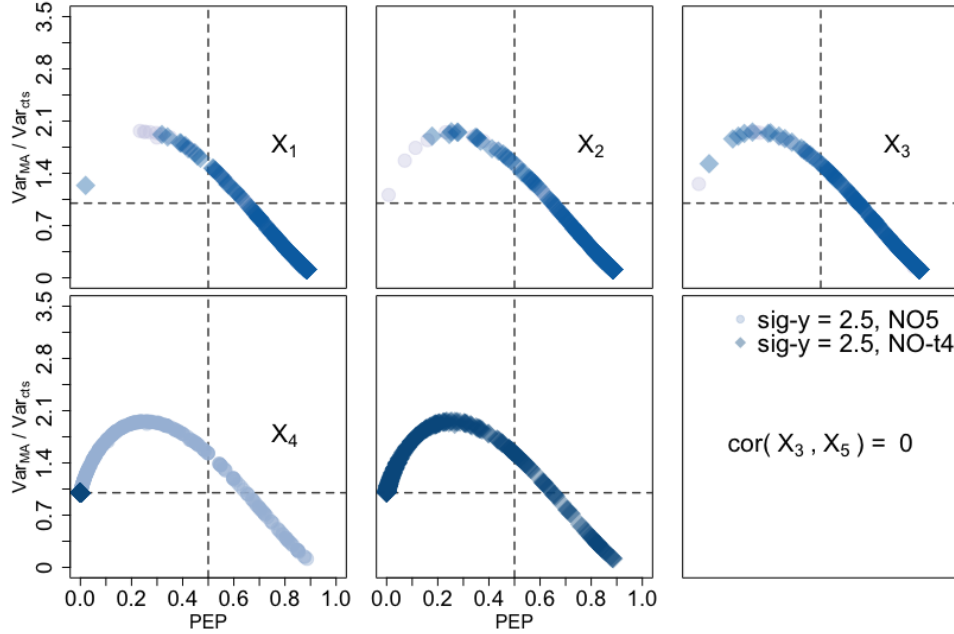


Figure 3.20: This figure facilitates the comparison of results from scenarios where model averaging is conducted with  $truth = NO5$  (data generated according to Equation 3.6 and models including  $X_5$  are excluded from the model set) to those conducted with  $truth = NO\text{-}taper4$  (data generated according to Equation 3.7 and *all subsets* model set is used). Signal to noise is held constant at low signal to noise ratio, and correlation is held constant at  $cor_{35} = 0.0$  for both sets of results. Plotting characters are coded such that ‘o’ represents  $truth = NO5$  and ‘ $\diamond$ ’ represents  $truth = NO\text{-}taper4$ . Results associated with variables that are truly in the data generating model have slightly darker plotting characters than those that are not in the data generating model.

the model set. However, the differences between the two levels of truth in the model set does not change for the larger correlation between  $X_3$  and  $X_5$  (Figure 3.21).

**Signal to Noise Ratio** Comparisons among the levels of truth in the model set for high signal to noise ratio are extremely similar to those shown in the previous three comparisons. The only difference is that realizations of data for all three levels of truth in the model set generated with high signal to noise tend to result in high PMPs for variables in the data generating model (or variables correlated with those in

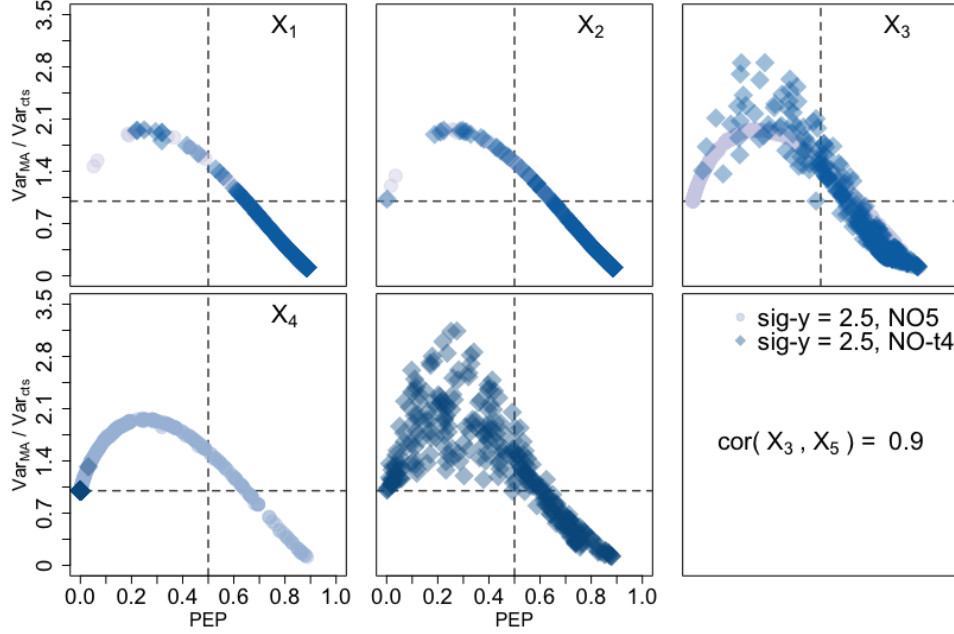


Figure 3.21: This figure facilitates the comparison of results from scenarios where model averaging is conducted with  $truth = NO5$  (data generated according to Equation 3.6 and models including  $X_5$  are excluded from the model set) to those conducted with  $truth = NO\text{-}taper4$  (data generated according to Equation 3.7 and *all subsets* model set is used). Signal to noise is held constant at low signal to noise ratio, and correlation is held constant at  $cor_{35} = 0.9$  for both sets of results. Plotting characters are coded such that ‘o’ represents  $truth = NO5$  and ‘ $\diamond$ ’ represents  $truth = NO\text{-}taper4$ . Results associated with variables that are truly in the data generating model have slightly darker plotting characters than those that are not in the data generating model.

the data generating model - in the case of NO5) more often than when we made these same comparisons fixing signal to noise to be low (Figure 3.22). This observation is similar to what we observed in Section 3.4.2.2; see Figure 3.22 for a comparison of YES to NO-taper4 for high signal to noise ratio.

### 3.4.3 Subset of Models Compared to Top Model

Model averaging for partial regression coefficients is sometimes conditioned on the subset of models for which the associated variable is included (i.e., the coefficient

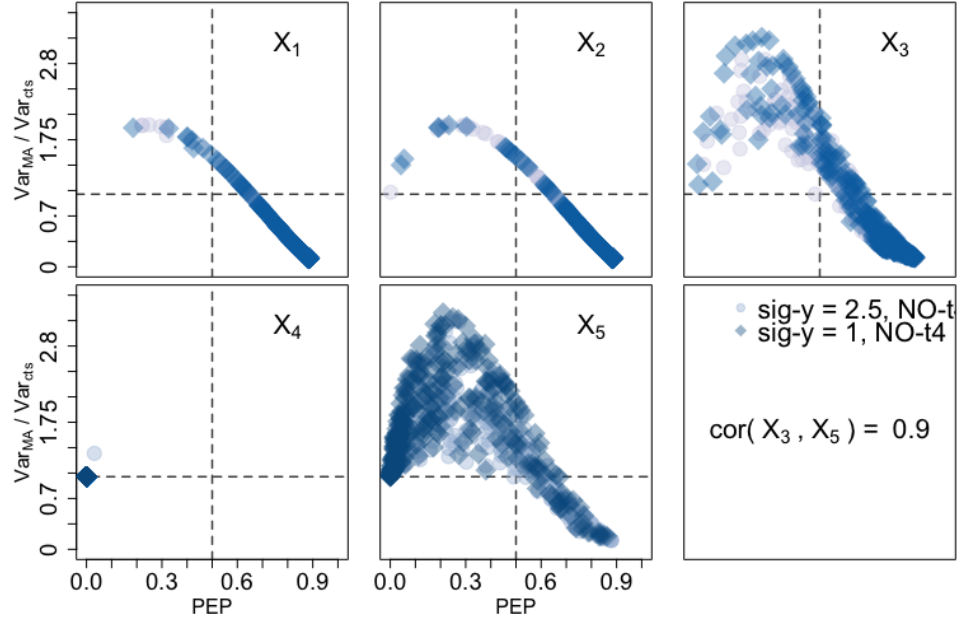


Figure 3.22: This figure facilitates the comparison of results from scenarios where model averaging is conducted with  $truth = YES$  to those conducted with  $truth = NO\text{-taper4}$  (data generated according to Equation 3.7 and *all subsets* model set is used). Signal to noise is held constant at high signal to noise ratio, and correlation is held constant at  $cor\text{-}35 = 0.9$  for both sets of results. Plotting characters are coded such that ‘o’ represents  $truth = YES$  and ‘◇’ represents  $truth = NO\text{-taper4}$ . Results associated with variables that are truly in the data generating model have slightly darker plotting characters than those that are not in the data generating model.

is not set exactly to zero). In this section, we compare the variance of the posterior for this type of model averaging ( $Var_{cts}$ ) to the variance of the posterior distribution for the same partial regression coefficient from the single model with the largest PMP ( $Var_{top}$ ).

3.4.3.1 Comparing Correlations Again, we start by comparing results among the levels of correlation with a low signal to noise ratio and model averaging was conducted with truth in the model set (treatments defining the reference scenario in

Section 4.2.1). The ratio of variances comparing  $Var_{cts}$  to  $Var_{top}$  is shown for all treatments in the reference scenario (Figure 3.23).

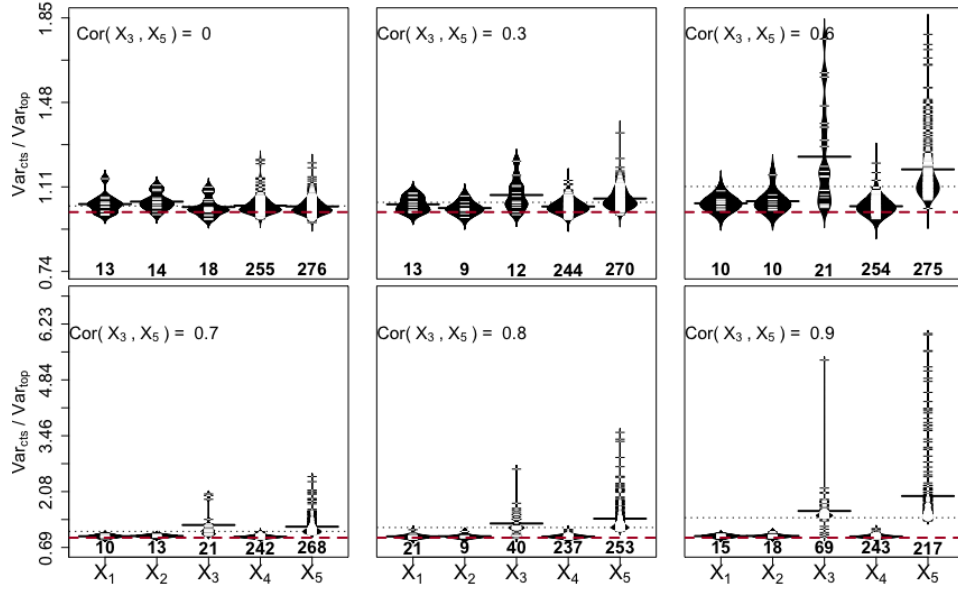


Figure 3.23: Comparison of posterior variances between the subset of models defining the continuous component of the model-averaged distribution and the single model with highest PMP. Note the dependence in ratios from the same realization is not reflected in these plots. Each violin plot contains the results from the same 300 realizations, just for a different variable. The numbers at the bottoms of the violin plots indicate the number of data realizations resulting in top models including the variable being summarized by the violin plot (note that the density functions are sensitive to the number of observations going into them).

We are interested in investigating when these two variances are similar and when they are different because they both correspond to methods used for final inference. Both methods investigated in this section exclude the subset of models setting partial regression coefficients to exactly zero, and because of this, it is not relevant to look at the relationship between the ratio of these method's variances and PEP. Therefore, we instead use violin plots (Kampstra, 2008) to display the distribution of ratio of variances for 300 realizations of simulated data, split up by each level of correlation

and each explanatory variable (Figure 3.23). Very few of the 300 realizations result in variables  $X_1$  or  $X_2$  being included in the top PMP model, so there are fewer observations in violin plots for those two variables.

All distributions of the ratios are centered close to one when the correlation between  $X_3$  and  $X_5$  is small, indicating little difference in variance for the two methods at low correlations. As the correlation between  $X_3$  and  $X_5$  increases, the distribution of the ratio of variances for coefficients associated with  $X_3$  and  $X_5$  is almost always strictly greater than 1 indicating more variability in the result conditional on the subset of models than the result conditional on one model.

### 3.4.3.2 Comparing Levels Within Other Factors

Results comparing levels within the other two factors were very similar to those presented in Figure 3.23. Similarly, for most other treatments, we observe the ratio of variances being very close to one for all realizations. See Appendix C for code to generate plots of results from other treatments. However, some treatments result in large multiplicative differences in  $Var_{cts}$  and  $Var_{top}$ , and the distributions of ratios of variances for these treatments are very similar to the two rightmost plots in the bottom row of Figure 3.23. These treatments are defined by combinations of the following levels of factors:  $sig-y = 1,2$   $truth = YES, NO$ -taper4, and  $cor-35 = 0.8, 0.9$ . Therefore, when we include two correlated variables in the model set, and one of those variables is in the data generating model, there is the potential for increased variance for inference conditional on the subset of models allowing the coefficient to be estimated as compared to the variance from the top model.

### 3.4.4 All Models Compared to Top Model

In this section we obtain results for the ratio of variances comparing the method using all models (full model averaging) to the method conditional on the model with the largest PMP ('MA to top'). The results we observe in this section are very similar to those observed in Section 3.4.2, so we compare the results from this ratio of variance ('MA to top') to the first ratio of variances we investigated ('MA to cts'). We use this comparison to highlight similarities between this ratio of variances (MA to top) and the ratio of variances comparing 'MA to cts' and exploit these similarities for brevity.

3.4.4.1 Comparison of Correlations Graphical comparisons of the results between these two ratios of variances are shown for the reference scenario (low signal to noise, truth in the model set, and all correlations) in Figures 3.24 - 3.29. When the correlation is small ( $cor-35 = 0.0, 0.3$ ), very few realizations result in top PMP models including  $X_1$ ,  $X_2$ , or  $X_3$ , resulting in very few ratios of variances comparing the method of full model averaging to the method using the top PMP model for those variables. We also note, no realizations result in top PMP models with variables having associated PEP greater than 0.6.

Therefore, in practice, the top PMP model tends not to include partial regression coefficients for variables whose PEP is larger than 0.6. The ratio of variances comparing full model averaging (using all models) to the top PMP model tends to be higher than the ratio of variances comparing the two components of the model-averaged posterior distribution. This is expected after investigating the ratio of variances between model averaging conditional on the subset of models for which the coefficient is estimated and the largest PMP model because the variance of this conditional model-averaged result ( $Var_{cts}$ ) was consistently larger than the variance from the model with the highest PMP. We also note the ratio comparing 'MA to top'

exhibits a similar relationship with PEP than the ratio comparing ‘MA to cts’, but it is shifted slightly higher on the  $y$ -axis (ratio of variances axis). As the correlation between  $X_3$  and  $X_5$  increases, this vertical shift becomes larger.

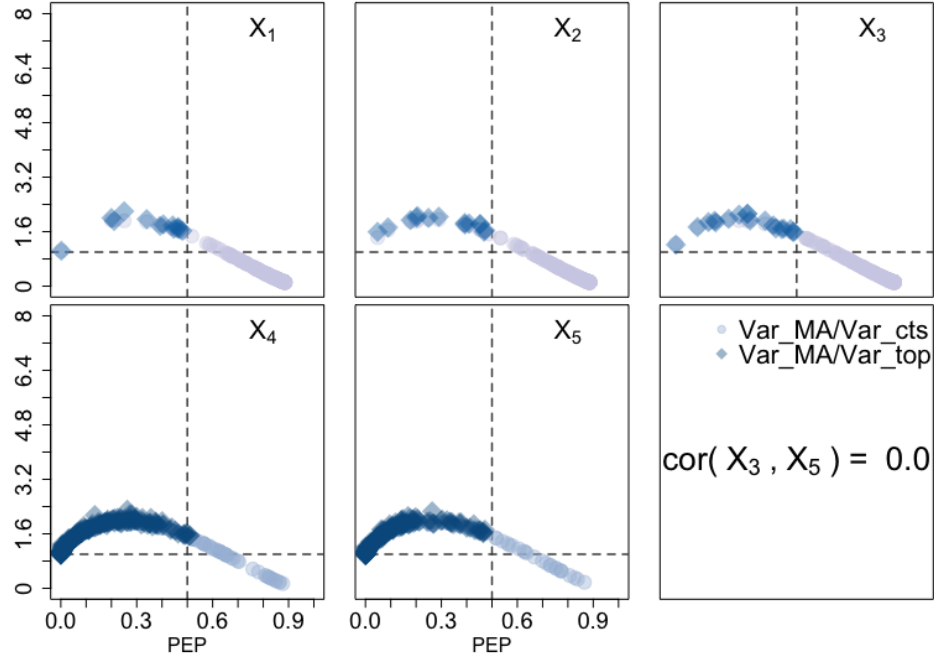


Figure 3.24: Comparison of results from two ratios of variances generated according to the reference scenario with a correlation of 0.0 (low signal to noise and truth in the model set). Plotting characters are coded such that ‘o’ represents ‘MA to cts’ and ‘◊’ represents ‘MA to top’. Results associated with variables that are truly in the data generating model have slightly darker plotting characters than those that are not in the data generating model.

Comparing these two ratios for each level of correlation and fixed levels of signal to noise and truth in the model set shows potential for the variance of fully model-averaged partial regression coefficients to be much larger (up to 8 times larger) than their corresponding variances associated with the single model with the largest PMP. However, this only appears to be the case when there is large correlation among

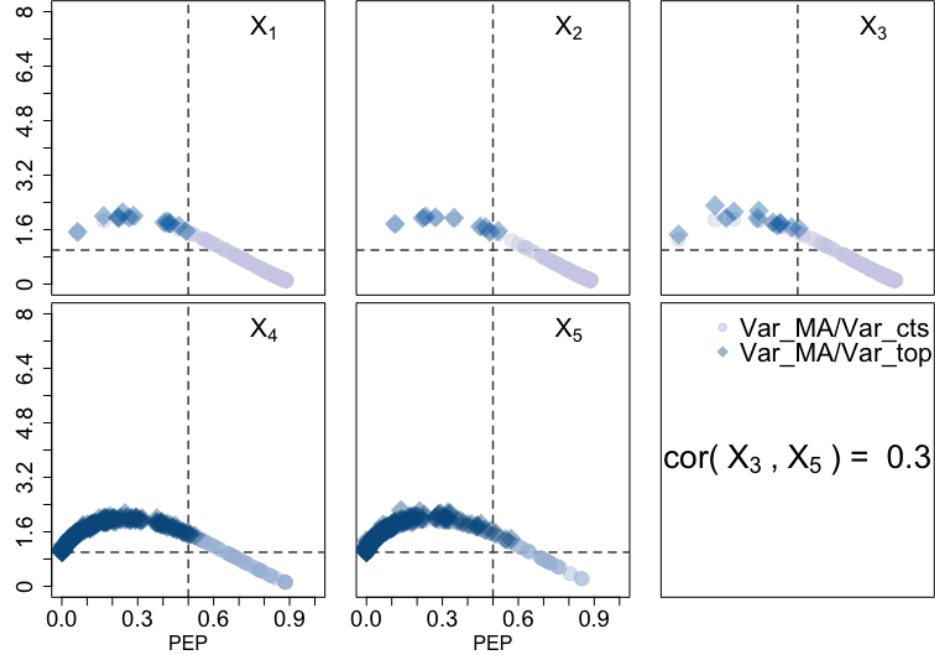


Figure 3.25: Comparison of results from two ratios of variances generated according to the reference scenario with a correlation of 0.3 (low signal to noise and truth in the model set). Plotting characters are coded such that ‘o’ represents ‘MA to cts’ and ‘◇’ represents ‘MA to top’. Results associated with variables that are truly in the data generating model have slightly darker plotting characters than those that are not in the data generating model.

explanatory variables, which makes inference for model-averaged partial regression coefficients difficult, resulting in a trade-off.

3.4.4.2 Comparing Signal to Noise The relationship between the two ratios of variances we observed in the previous section (slight vertical shift and smaller range of PEP for ‘MA to top’ versus ‘MA to cts’) holds for comparing the two levels of signal to noise within fixed levels of truth in the model set. That is, based on the results presented in Section 3.4.2.2, we know what we will see for this ratio of variances. Our conclusions remain the same; when the signal to noise is high, we have less of a chance

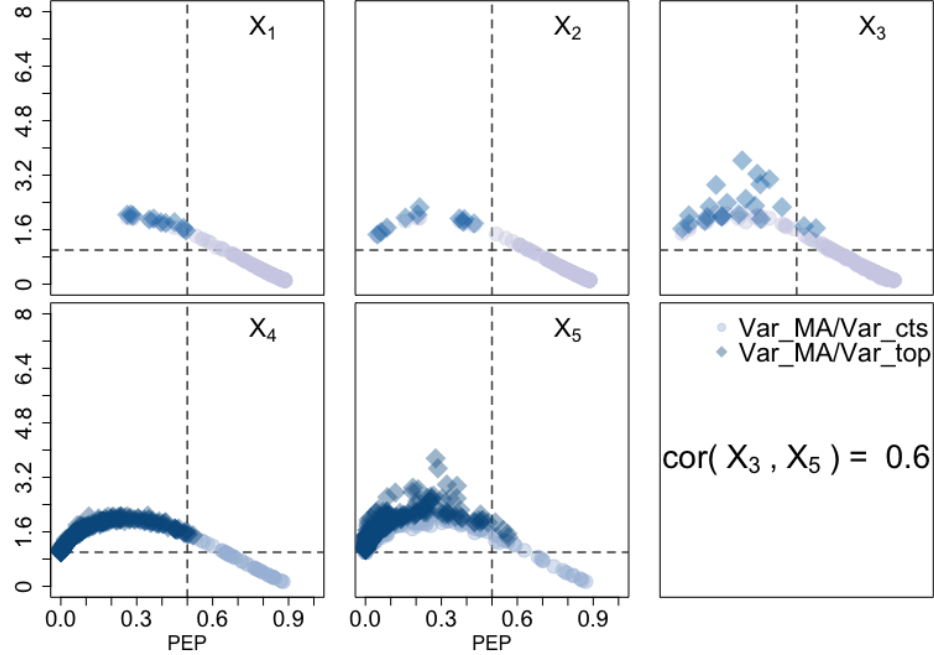


Figure 3.26: Comparison of results from two ratios of variances generated according to the reference scenario with a correlation of 0.6 (low signal to noise and truth in the model set). Plotting characters are coded such that ‘o’ represents ‘MA to cts’ and ‘◊’ represents ‘MA to top’. Results associated with variables that are truly in the data generating model have slightly darker plotting characters than those that are not in the data generating model.

of seeing a much larger variance coming from the method accounting for all models when compared to the variance coming from the top model, for all correlations and levels of truth in the model set.

3.4.4.3 Comparing Truth in the Model Set The relationships between the three levels of truth in the model set for fixed signal to noise and correlation behave similarly to those highlighted above. When we exclude  $X_5$  from models in the model set, we observe results nearly identical to those presented in Section 3.3.3.3, but with the vertical shift along the ratio of variance axis described in Section 3.4.4.1. Our

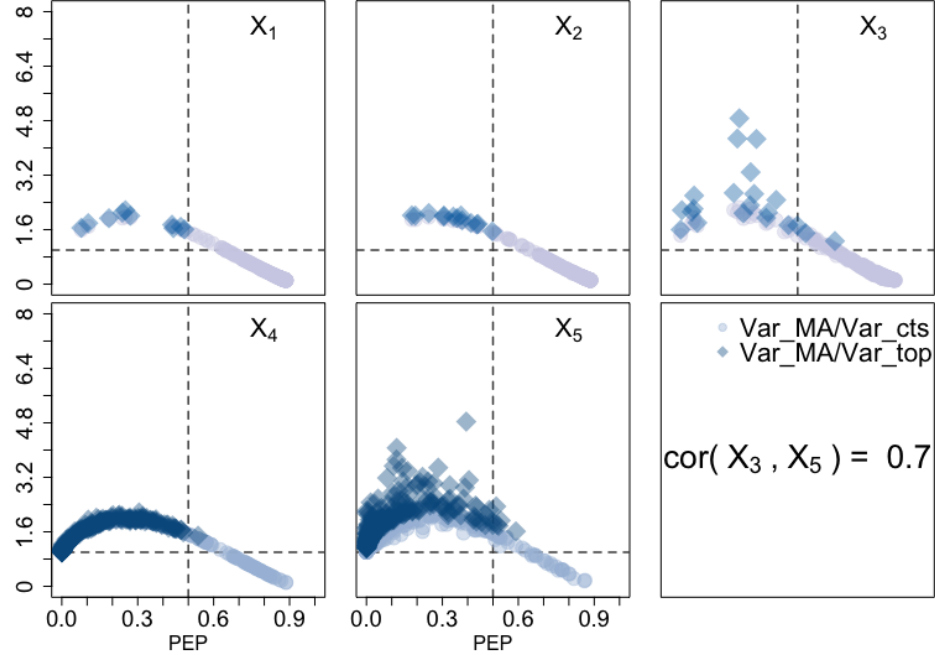


Figure 3.27: Comparison of results from two ratios of variances generated according to the reference scenario with a correlation of 0.7 (low signal to noise and truth in the model set). Plotting characters are coded such that ‘o’ represents ‘MA to cts’ and ‘◊’ represents ‘MA to top’. Results associated with variables that are truly in the data generating model have slightly darker plotting characters than those that are not in the data generating model.

conclusions remain the same; when truth is in the model set, we see potential for the variance coming from the method conditional on all models to be much larger than the variance coming from the method conditional on the top PMP model, especially when the correlation between  $X_3$  and  $X_5$  is high and the signal to noise ratio is low.

### 3.5 Discussion

The goal of this paper is to compare variances among methods for conducting multimodel inference and better understand implications for inference. We presented results investigating ratios of variances for partial regression coefficients from three

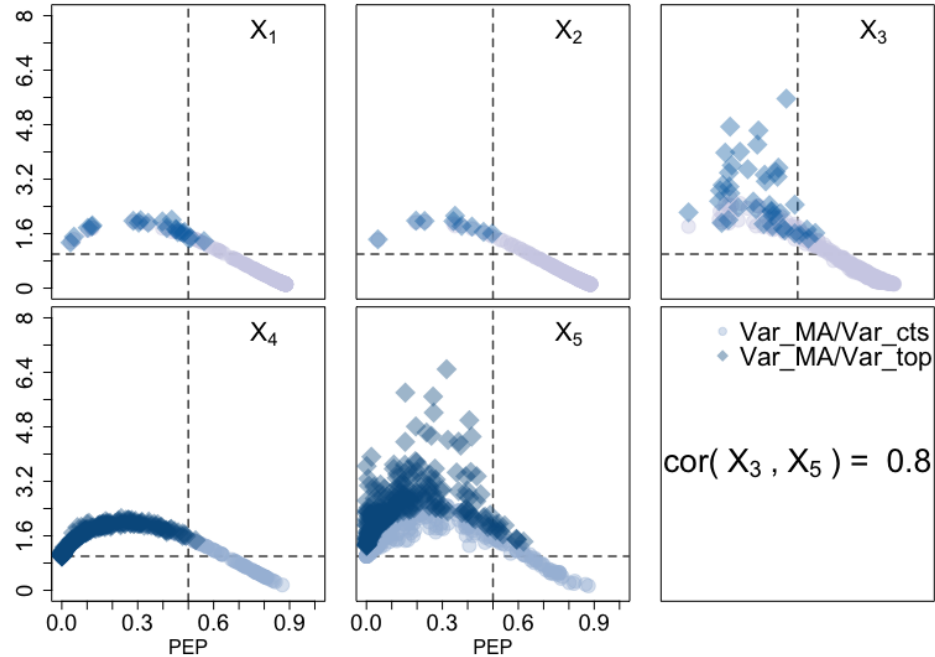


Figure 3.28: Comparison of results from two ratios of variances generated according to the reference scenario with a correlation of 0.8 (low signal to noise and truth in the model set). Plotting characters are coded such that ‘o’ represents ‘MA to cts’ and ‘◊’ represents ‘MA to top’. Results associated with variables that are truly in the data generating model have slightly darker plotting characters than those that are not in the data generating model.

common methods of multimodel inference. We studied a range of data generating models and model sets by varying correlations among explanatory variables, signal to noise ratios, and which variables were included in the data generating model and the model set. In most cases, methods accounting for more models resulted in larger variances, as expected based on the motivation behind multimodel inference and model averaging. That is, model averaging is often used by researchers who would like to ensure their inferences “account for model uncertainty.” We compared posterior variances because they are used to summarize uncertainty in quantities of interest

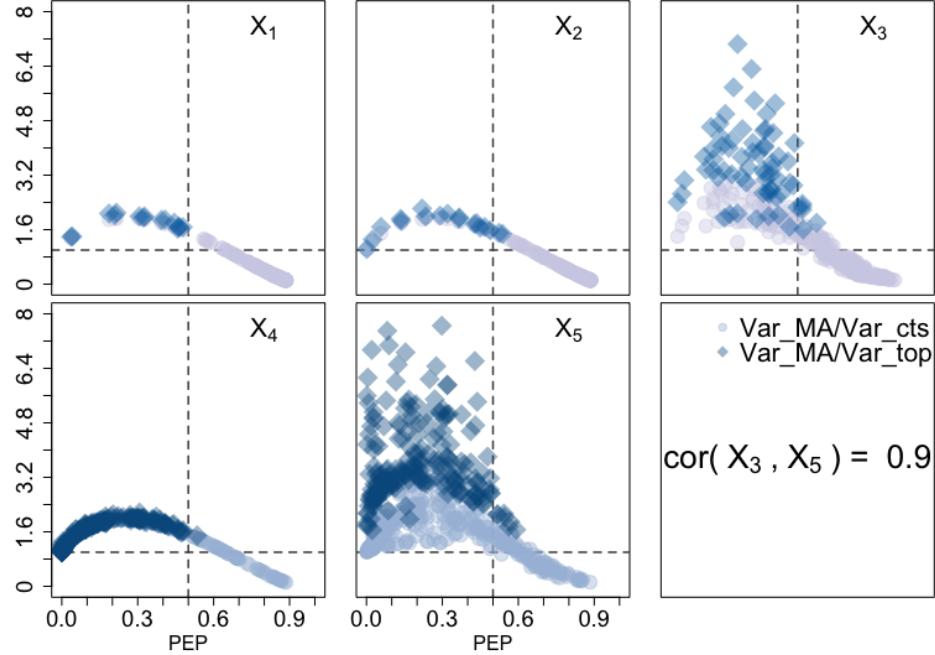


Figure 3.29: Comparison of results from two ratios of variances generated according to the reference scenario with a correlation of 0.9 (low signal to noise and truth in the model set). Plotting characters are coded such that ‘o’ represents ‘MA to cts’ and ‘◊’ represents ‘MA to top’. Results associated with variables that are truly in the data generating model have slightly darker plotting characters than those that are not in the data generating model.

and thus are an integral part of inference. In the following sections, we provide more discussion surrounding our observations and conclusions.

### 3.5.1 All Models Versus Subset of Models

We made this comparison to better understand how the two components of the model-averaged posterior distribution for a model-averaged partial regression coefficient contribute to its posterior variance. We directly assessed how including the subset of models for which the coefficient is not estimated affected the full model-averaged variance by comparing the variance of the full model-averaged posterior

distribution (considering all models) to the variance from the distribution considering only those models where the coefficient was estimated. Both components of the posterior distribution exist for all partial regression coefficients considered in the model set, so we were able to make this comparison for all realizations of data.

We observed many PEPs for which the method including fewer models had smaller variance than the method including all models (much like the results observed in Section 3.4.1). In general, we observed variables not included in the data generating model tended to result in model-averaged posteriors of their partial regression coefficients with high probabilities at the point mass of zero (high PEPs), and thus smaller full model-averaged variance than the variance coming from its continuous component (conditional on the subset of models where the coefficient was estimated). This is contrary to what we would have expected based on the idea that methods accounting for more models result in larger variances. In contrast, partial regression coefficients associated with variables in the data generating model tended to have larger full model-averaged variance, and when there was correlation between explanatory variables, the multiplicative increase had the potential to be even larger for a given PEP. This observation was consistent with the belief that methods accounting for more models will result in larger variances.

Our observations held for almost all of the treatment comparisons made; we learned that the addition of the subset models for which the coefficient is set to exactly zero to the subset of models for which the coefficient is estimated can lead to either an increase or a decrease in posterior variance of model averaging considering all models, and that the direction of this change is context dependent.

### 3.5.2 Subset of Models Versus Top Model

When comparing the variance from model averaging using only the models for which the coefficient was estimated to that of a single model, we observed differences for the treatments for higher correlations between  $X_3$  and  $X_5$  for both signal to noise ratios, and for two levels of truth in the model set (namely when the data generating model was included in the model set, and when the model set failed to include variables related to tapering effects that were included in the data generating model). We saw that  $Var_{cts}$  was strictly larger than  $Var_{top}$  for the partial regression coefficients associated with  $X_3$  and  $X_5$  when their correlation was high, and that the two variances were very similar otherwise. Here, we explain why we only observed a large difference between  $Var_{cts}$  and  $Var_{top}$  for variables that are highly correlated.

The component of the model-averaged posterior distribution conditional on models allowing the coefficient associated with  $X_5$  to be estimated is constructed using information from the posterior distributions from models where  $X_3$  and  $X_5$  are included in isolation (i.e.,  $X_3$  is included and  $X_5$  is excluded, or visa-versa) *and* the posterior distributions associated with models where  $X_3$  and  $X_5$  are included together. As the correlation increases between the variables and both are included in the model, the posterior variance of their associated partial regression coefficients increases. These models also tend to receive a non-negligible portion of the posterior model mass and therefore contribute substantially to the posterior variance of the component of the model-averaged posterior distribution conditional on models including  $X_5$ . It is unlikely that the model with the largest posterior model probability will be a model where both  $X_3$  and  $X_5$  are included, so the ratio of the variances will consistently be larger than one. We illustrate this with Figure 3.30.

Results from model averaging for one of the realizations of data generated under the reference treatment condition ( $sig - y = 2.5$ ,  $truth = \text{YES}$ , and  $cor - 35 =$

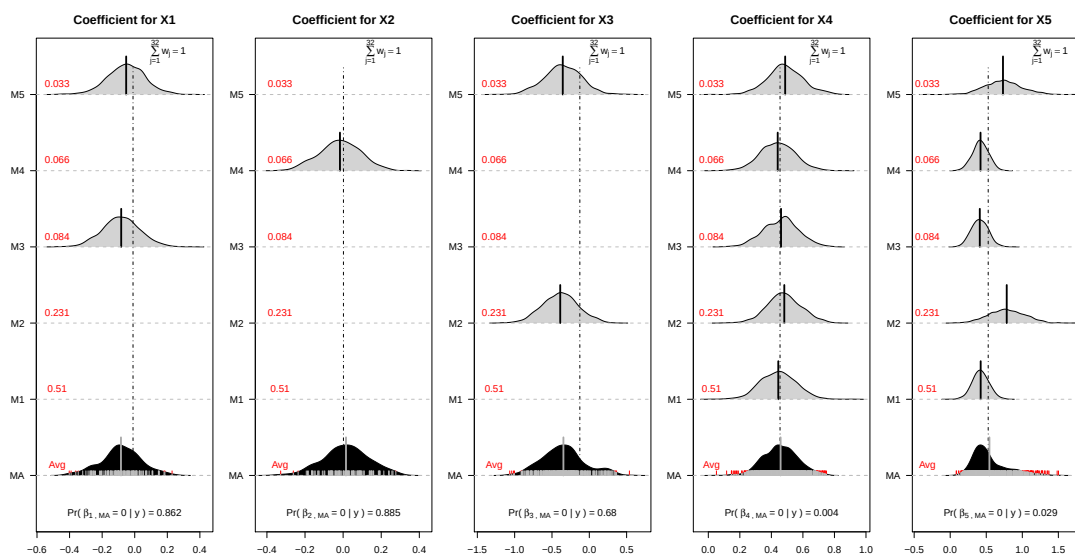


Figure 3.30: Results from model averaging for one of the realizations of data generated under the treatment conditions;  $sig - y = 2.5$ ,  $truth = \text{YES}$ , and  $cor - 35 = 0.9$ .

0.9) are shown in Figure 3.30. The model with the largest PMP is labeled M1, and the posterior distributions for the partial regression coefficient for the variables in this model can be seen by looking across that row (labeled M1). In this realization of data, the top PMP model includes  $X_4$  and  $X_5$ , has posterior model probability equal to 0.51, and the coefficient associated with  $X_5$  has a noticeably smaller posterior variance than the component of the model-averaged posterior distribution conditional on models allowing the coefficient associated with  $X_5$  to be estimated (seen in the bottom MA row). Also note, M2 (the model with the second highest PMP) includes  $X_3$ ,  $X_4$ , and  $X_5$ , has a posterior model probability of 0.231, and the posterior for the partial regression coefficient associated with  $X_5$  in this model has a much larger variance than that from M1. The contribution from the posterior for the partial regression coefficient associated with  $X_5$  coming from M2 to the model-averaged posterior distribution for the partial regression coefficient associated with  $X_5$  contributes extra variation to the model-averaged result (e.g., long right tail in the continuous component of the

model-averaged posterior distribution for the coefficient associated with  $X_5$ ). This extra variation culminates in a much larger variance in the continuous component of the model-averaged posterior distribution for the coefficient associated with  $X_5$  than its corresponding posterior coming from M1.

Comparing variances from the method using the subset of models including the variable of interest to the model with the largest PMP reveals potential for the method accounting for more models to have larger variance, particularly when there is correlation among explanatory variables.

### 3.5.3 All Models Versus Top Model

We observed large multiplicative differences in the posterior variances for partial regression coefficients (top model versus MA) for highly correlated variables. However, large multiplicative differences do not necessarily mean the two variances will be different for practical purposes (such as inference). The difference in the posterior standard deviations ( $SD_{MA} - SD_{top}$ ) for the treatment condition resulting in the highest ratios of variances is shown in Figure 3.31. Even though the differences are mostly greater than zero (indicating higher posterior standard deviation in the model-averaged result), the actual difference is at most around 0.20 standard deviation units (Figure 3.31). The explanatory variables were standardized to have mean zero and standard deviation one, so the units are standard deviations of the particular explanatory variables. Therefore, the resulting posterior interval for a partial regression coefficient of interest (for data like these) will be at most 0.40 standard deviations wider if we “account for model uncertainty by using model averaging.”

One could argue that when variables are highly correlated with other explanatory variables, model averaging of partial regression coefficients is inappropriate (Chapter

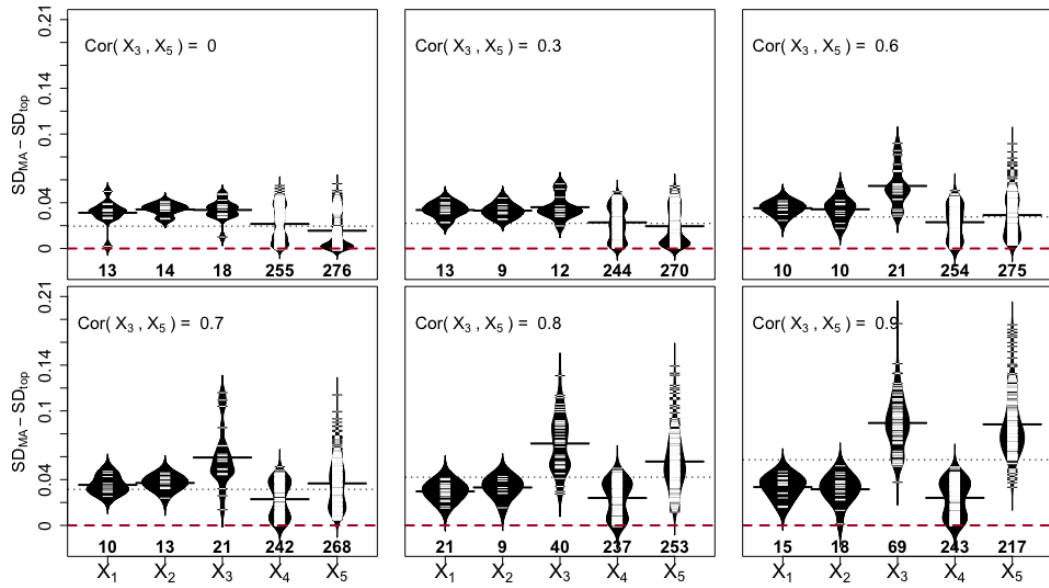


Figure 3.31: The difference in the posterior standard deviations ( $SD_{MA} - SD_{top}$ ) for the treatment condition resulting in the highest ratios of variances. The numbers at the bottoms of the violin plots indicate the number of data realizations resulting in top models including the variable being summarized by the violin plot (note that the density functions are sensitive to the number of observations going into them).

2). Let  $X_4$  and  $X_5$  be variables of interest. Then we could think of  $X_4$  as a variable of interest that is uncorrelated with other explanatory variables and  $X_5$  as a variable of interest that is correlated with other variables in the model (namely  $X_3$ ). So, model averaging the partial regression coefficient for  $X_4$  may actually make sense, since its interpretation stays relatively constant regardless of what other variables enter the model. However, model averaging  $X_5$  is arguably inappropriate as its interpretation will depend on whether or not  $X_3$  is in the model. When model averaging for a partial regression coefficient is arguably ‘appropriate’, there is little difference in the resulting inference made from the model-averaged result as compared to the top individual model, and the only case for which model averaging may largely affect the inference made is when the interpretation for a model-averaged partial regression coefficient is unclear (see Chapter 2 or Banner and Higgs (2016) for more on this discussion).

#### 3.5.4 Potential Extensions

We investigated results for specific combinations of our three factors of interest in a linear regression setting and we expect these results to also hold for generalized linear models. A generalized linear model transforms the response using a link function to make the relationship between the mean transformed response and the explanatory variables linear. We expect to see similar results on the transformed scale, however back-transforming the response to the linear scale could make the differences in variances more drastic after the transformation. Executing a simulation study like the one we conducted using a fully Bayesian implementation of multimodel inference for generalized linear regression models is prohibitive because of the computational expense.

A common alternative to fully Bayesian implementation is to approximate PMPs with weights using information criteria such as AIC and BIC, focusing on

point estimates and standard errors, rather than posterior distributions. These approximations to model averaging are discussed in detail in Burnham and Anderson (2002)’s text. Formulae analogous to (3.3), (3.4), and (3.5) exist for this approach to model averaging, and these approximations are potentially useful for investigating differences in variances from different types of mulitmodel inference, further discussed in Chapter 5.

### 3.6 Conclusion

When explanatory variables are approximately uncorrelated, there is little difference between the variance of the model-averaged result and the variance of the posterior from the top individual model. As the correlation increases between two of the explanatory variables, the model-averaged result tends to have larger variance than the individual result. However, when there is large correlation between explanatory variables, the model-averaged result is a weighted average of partial regression coefficients with different interpretations and model averaging may not make sense. In summary, when it is reasonable to model average partial regression coefficients there is little difference between the model-averaged posterior variance and the corresponding variance from an individual model. Thus, the motivating benefit of model averaging may not be realized. In other words, in cases where the variance of the model-averaged quantity tends to be larger than the variance from an individual model the interpretation of the model-averaged regression coefficient is challenging.

In this work we found little evidence suggesting large differences in variances for conditioning inferences on different subsets of models of different sizes. Following a suggestion made by Burnham and Anderson (2002), it may be reasonable for researchers to use the model-averaged posterior variance with an estimated slope coefficient that is conditional on one model as a way to “account for model

uncertainty” and also keep the straightforward interpretation of the partial regression coefficient, but in the cases we investigated here, such an adjustment is likely unnecessary from a practical perspective.

## CHAPTER FOUR

### THE MODEL AVERAGED POSTERiors PLOT MAPP PACKAGE FOR R STATISTICAL SOFTWARE

#### Contribution of Authors and Co-Authors

Author: Katharine Banner

Contributions: Wrote R code, vignette, and developed the package.

Co-Author: Megan Higgs

Contributions: Provided feedback on drafts of the vignette.

Manuscript Information Page

Katharine M. Banner and Megan D. Higgs

Journal of Statistical Software

Status of Manuscript:

☒ Prepared for submission to a peer-reviewed journal

☐ Officially submitted to a peer-review journal

☐ Accepted by a peer-reviewed journal

☐ Published in a peer-reviewed journal

Publisher: The Foundation for Open Access Statistics

Submitted:

## 4.1 About this Chapter

R statistical software (R Core Team, 2016) is designed for statistical computation and graphics and is popular among statisticians and scientific researchers alike. The software is free and open source, meaning anyone can download the software for personal use, and anyone can contribute packages to be used with the software. To make the model averaged posteriors plot accessible to as many users of model averaging as possible, we wrote the **MApp** package in the R programming language (R Core Team, 2016). **MApp** was developed to support work presented in Chapter 2 (Banner and Higgs, 2016), which discusses model averaging of regression coefficients for linear regression models. We provide a brief summary of model averaging in 4.2, and then introduce the functions in **MApp** and discuss our intentions for their use.

## Abstract

For any averaging to make sense, it is important to ensure the quantities being averaged have similar meanings across all models being considered. When considering model averaging partial regression coefficients, this becomes very important, as partial regression coefficients associated with variables of interest often have different interpretations depending on which other explanatory variables are included in the model. There are strong allures for using model averaging as it is purported as a tool capable of “accounting for model uncertainty,” and providing information about “overall effects” of “important” variables. Currently, model averaging for regression coefficients is not being used in conjunction with tools for assessing the implicit assumptions for averaging, solely because the tools are unavailable. We developed the model averaged posteriors plot (MAP plot) and **MApp** package as a tool for assessing the appropriateness of model-averaged partial regression coefficients. The MAP plot clearly displays all components going into the model-averaged results and facilitates assessment of the appropriateness of model-averaged partial regression coefficients on a case-by-case basis. Functions for multiple implementations of model averaging are available in **MApp**; this paper provides examples and discussions to help researchers come to informed decisions about when (and when not) to use model-averaged regression coefficients for final inference.

## 4.2 Motivation for the Package

Model averaging is advertised as a tool for “accounting for model uncertainty” in analyses. To conduct model averaging, a set of models is considered ( $\mathcal{M} = \{M_1, M_2, \dots, M_J\}$ ). When considering model averaging in the linear regression framework, all first order combinations of the explanatory variables, or *all subsets*, is a common model set to consider. For example, consider  $p$  potential explanatory variables, the *all subsets* model set has size  $J = 2^p$  models, and is defined by models where each of the  $p$  variables is either in or out of a particular model.

Once the model set is defined, parameters of interest and posterior model probabilities, or weights based on information criteria such as AIC or BIC, are calculated for all models in  $\mathcal{M}$ . The probabilities (or, more generally, weights) are used in a weighted average to combine information about quantities of interest across all models considered. Chapter 2 Section 2.2.2 and Hoeting et al. (1999a) provide more information about Bayesian model averaging, and Burnham and Anderson (2002) describe model averaging based on information criteria.

## 4.3 Model Averaging of Partial Regression Coefficients

Model averaging was developed in the context of prediction and its use has recently been extended to partial regression coefficients (e.g., as discussed in Banner and Higgs, 2016; Cade, 2015). Inference drawn from partial regression coefficients typically results in explanations of “overall effects” for each of the coefficients of interest, presenting a unique question: what does “overall effect” mean in the context of the research question? In the presence of multicollinearity, the interpretations of partial regression coefficients are not necessarily constant across models, making “overall effects” difficult to understand. Similarly, the posterior distributions for

partial regression coefficients take careful thought. Variables associated with partial regression coefficients can be included or excluded from a linear regression model. Excluding a variable from a linear regression model is equivalent to setting its associated partial regression coefficient to exactly zero. Therefore, the total probability of the posterior distribution of a model-averaged partial regression coefficient is made up of two components, one coming from the models where the coefficient is set to zero (i.e., variable excluded from the model), and the other coming from the models where the coefficient is estimated (i.e., variable included in the model).

The **MApp** package provides the user with graphical tools for understanding all of the components going into the posterior for the model-averaged partial regression coefficients so its appropriateness may be assessed on a case-by-case basis. Banner and Higgs (2016) contains a detailed discussion of potential issues with, and considerations for assessing, common inferences drawn from model averaging partial regression coefficients.

There are many approaches for conducting model averaging including: self-programmed reversible jump Markov chain Monte Carlo (RJMCMC) samplers; implementations in WinBUGS (Lunn et al., 2000), described in Link and Barker (2006); and methods based on information criteria, described in Burnham and Anderson (2002). To accommodate as many users of model averaging as possible, we designed our plotting functions to work with posterior draws from existing software in R (R Core Team, 2016) (such as **BMS**), self-coded reversible jump MCMC samplers, as well as AIC-based and BIC-based model averaging.

## 4.4 Installing MApp

Currently, **MApp** must be installed from GitHub or loaded from a local disk. The package, **devtools** (Wickham and Chang, 2016) is necessary for both installations. To install **devtools**, run `install.packages("devtools")` in the R console.

### 4.4.1 GitHub Install

GitHub (<https://github.com>) is an online service for hosting Git repositories. Git (<https://git-scm.com>) is a version control software used primarily by developers for maintaining code. You do not need to be a Git user or have a registered GitHub account to use **MApp**. To install **MApp**, follow these two steps:

1. Install Version 0.1 of **MApp** from GitHub.

```
# install the package
library(devtools)
install_github("kbanner14/MApp-Rpackage", subdir = "MApp")
# load the package
library(MApp)
```

2. To update **MApp** when new versions are available, unload the package with `unload(inst("MApp"))` and repeat Step 1. Note that both functions, `inst` and `unload`, are in the **devtools** package.

**MApp** uses functions from the following packages: **LearnBayes** (Albert, 2014), **beanplot** (Kampstra, 2008), **magrittr** (Bache and Wickham, 2014), **dplyr** (Wickham and Francois, 2015), and **BMS** (Feldkircher and Zeugner, 2009). These packages will be loaded automatically with a call to `install_github`. Some users have experienced issues with the `install_github` install; in the event issues arise, another option for installing **MApp** is to load it from a local directory.

#### 4.4.2 Local Disk Install

Follow these four steps for this type of installation.

1. Download all files in the **MApp-Rpackage/** repository from the *Download Zip* button on the right hand side of the screen here: <https://github.com/kbanner14/MApp-Rpackage>. Git users may also clone the repository.
2. Set your working directory to the location of the **MApp-Rpackage-master/MApp** repository on your computer.
3. Run the following lines of code to load the package and its documentation.

```
# load the package
devtools::load_all(".")
# load the R documentation files
devtools::document(".")
```

4. Install package dependencies using `install.packages("packagename")`. The names of the required packages are: **LearnBayes** (Albert, 2014), **beanplot** (Kampstra, 2008), **magrittr** (Bache and Wickham, 2014), **dplyr** (Wickham and Francois, 2015), and **BMS** (Feldkircher and Zeugner, 2009).

#### 4.5 MApp Plotting Functions

The **MApp** package provides plotting functions to help researchers dissect and understand how results from individual models are combined to form the common inferences drawn from model averaging of partial regression coefficients. There are five major plotting functions:

- **MApp\_bms** works with **bma** objects obtained from the **bms** function in the **BMS** package. Details about the **bms** function can be found in the **BMS** tutorial here: <http://bms.zeugner.eu/tutorials/bms.pdf>. The **MApp\_bms** function returns the MAP plot and a table of posterior standard deviations.

- **MApp\_IC** works with a data frame and conducts AIC-, AICc-, or BIC-based model averaging for *all subsets* regression. This function returns the MAP plot and the model-averaged results.
- **MApp\_IC\_gen** works with results from model averaging based on information criteria. This function requires the user to have the results first, but provides more flexibility than **MApp\_IC** when returning the MAP plot.
- **MApp\_MCMC** works with default output from the implementation of model averaging using the **OpenBUGS** or self-programmed RJMCMC samplers.
- **MApp\_list** works with draws from self-programmed RJMCMC samplers that store results from each model in a list. This function returns the MAP plot and a table of posterior standard deviations.

Documentation exists for all functions in **MApp**. Use the `?` or `help` to view the documentation (Figure 4.1).

```
? MApp_bms
```

#### 4.6 MApp Package Data

- **bfat** A dataset containing percent body fat measurements, and 13 other body size measurements for 251 men. These data were originally collected by Johnson (1996), and were used by Hoeting et al. (1999a) in their hallmark paper, *Bayesian Model Averaging, A Tutorial*.
- **brainData** A dataset containing the log brain weight  $\log(g)$ , log body weight  $\log(kg)$ , log gestation length  $\log(days)$ , and litter size ( $\#$  offspring) for 96 species of mammals. These data were modified from **case0902** in the R package **Sleuth2** (Ramsey et al., 2012).

## Model Averaged Posteriors Plot (MAP plot)

### Description

For `bma` objects obtained using the `bms` function in the `BMS` package. Creates a graphical display showing all relevant output from the model averaging procedure. Used to graphically compare posterior distributions of partial regression coefficients from individual models in the model set to the posterior distributions resulting from the model averaged partial regression coefficients. Also used to compare posterior standard deviations of the continuous piece of the model averaged distribution to corresponding standard deviations from individual posterior distributions in tabular form.

### Usage

```
MApp_bms(x, plot_wind, num_sims = 1000, max_display = NULL,
  mod_names = NULL, include_coef = NULL, ...)
```

### Arguments

- |                           |                                                                                                                                                                                                                                     |
|---------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <code>x</code>            | an object of class <code>bma</code> from the <code>BMS::bms()</code> function                                                                                                                                                       |
| <code>plot_wind</code>    | A vector of length 2 specifying the number of rows and columns to partition the plotting window into.                                                                                                                               |
| <code>max_display</code>  | An integer specifying the number of models to display in the plotting window. Can also take on the argument "common3" to show the top PMP model, the full model, and the model averaged distribution.                               |
| <code>mod_names</code>    | A vector specifying the names of the models. If left blank, models will be named $M_1$ , $M_2$ , ..., $M_J$ , where $M_1$ will have the largest posterior model probability and $M_J$ has the smallest posterior model probability. |
| <code>include_coef</code> | A vector specifying which partial regression coefficients to display.                                                                                                                                                               |

### Value

The `MApp_bms` plot and a table of posterior standard deviations for the top `max_display` individual models specified, along with the posterior standard deviation of the model averaged parameter.

Figure 4.1: Example R documentation for the function, `MApp_bms`.

#### 4.7 The `MApp_bms` Function

`MApp_bms` is designed to be used with the `bms` function in the `BMS` package (Feldkircher and Zeugner, 2009). The `bms` function executes a popular implementation of model averaging by placing Zellner’s  $g$ -prior (Zellner, 1984) on the regression coefficients, a user-specified prior on the model set, and a user-specified value for the hyper-parameter,  $g$ . This implementation results in analytical posterior distributions for the regression coefficients in the form of  $t$  distributions, allowing for quick computations. Detailed information about the  $g$ -prior implementation and its resulting posteriors is provided in Appendix A.

The function `bms` takes on a data frame and prior information and returns model-averaged results (posterior model probabilities (PMPs), and means and standard deviations of the model-averaged distributions for each partial regression coefficient). The function also saves useful information about the models considered, namely the central moments for the posterior distributions of the partial regression coefficients. We use this information to obtain the  $t$  distributions for all partial regression coefficients in each model considered. Specifically, we use analytical results (Appendix A) to find the appropriate degrees of freedom, and the central moments from `bms` to compute means and variances for each  $t$  distribution. `MApp_bms` uses this information to obtain a specified number of Monte Carlo draws from posterior distributions of the partial regression coefficients for each individual model. Then, using the posterior model probabilities, `MApp_bms` samples proportionally from all individual posteriors to create a model-averaged posterior distribution for each partial regression coefficient of interest. Therefore, `MApp_bms` obtains draws from all distributions going into the model-averaged result and displays this information using density and rug plots, or beanplots, from the `beanplot` package (Kampstra, 2008) using a Gaussian smoothing

kernel with default bandwidth `nrd0`. The bandwidth is set according to Silverman’s ‘rule of thumb’ (see the text Silverman B. W. (1986). *Density Estimation*, page 48, eqn (3.31)).

#### 4.7.1 Working with Small Model Sets

In this example we use `brainData`, which is comprised of average values of log gestation length `log(days)`, litter size (number of offspring), log body weight `log(kg)` and log brain weight `log(g)` for 96 species of mammals. The number of animals going into the average values for each species were not provided.

```
data(brainData)
head(brainData)
```

We suppose the researchers are interested in the relationship between log brain weight and log gestation length for all 96 mammal species. This relationship can be addressed directly with a simple linear regression model regressing log brain weight on log gestation length. Now we assume the researchers are asked to use model averaging to “account for model uncertainty.” We show how the MAP plot can be used to understand the results of model averaging the partial regression coefficients associated with log gestation length and to potentially defend the researchers’ choice to a simpler model to address their question of interest.

Table 4.1: First five rows of `brainData`.

| lbrain | lbody | lgest | litter |
|--------|-------|-------|--------|
| 2.86   | 1.25  | 3.26  | 1.00   |
| 1.25   | -0.07 | 3.53  | 4.60   |
| 1.15   | -1.90 | 3.83  | 3.00   |
| 0.13   | -3.02 | 3.93  | 1.50   |
| 0.31   | -2.75 | 3.83  | 1.50   |

We can use the `bms` function from the `BMS` package to obtain a `bma` object.

```
library(BMS)
brain <- bms(brainData, g = "UIP",
             mprior = "uniform",
             user.int = F)
class(brain)
```

```
## [1] "bma"
```

We use `bms` to place Zellner's  $g$ -prior on the coefficients, and we specify a discrete uniform prior on the  $2^3 = 8$  models in the *all subsets* model set. `MApp_bms` takes on two arguments that do not have default values: `x` and `plot_wind`. For this example, we specify `x = bma-object`, and `plot_wind = c(1,3)`, which tells the function to split the plotting window up into one row with three columns. If we know which models go with which posterior model probabilities, we can add specific model names in place of the default  $M_1, M_2, \dots, M_8$  using the `mod_names` argument.

```
map_brain <- MApp_bms(x = brain, plot_wind = c(1,3),
                     mod_names = c("BGL", "BG", "BL", "B", "G", "GL", "L", "Null"))
```

The model-averaged posteriors plot displays both components of the model averaged posterior distribution. The continuous component is represented by the black density and rug plot (beanplot), and the zero component is represented by text providing the probability of the point mass at zero at the bottom of the panels. The dashed line extending up from the beanplot in the bottom row of each panel shows the average of the model-averaged posterior distribution, which incorporates the zero component; the larger the point mass at zero, the more the line is pulled towards zero.

4.7.1.1 Reading the MAP Plot Figure 4.2 has three panels, one for each of the explanatory variables considered (log gestation length (*lgest*), log body size (*lbody*), and litter size (*litter*)). The posterior distributions of all partial regression coefficients for log body size, along with the posterior distribution of the model averaged partial

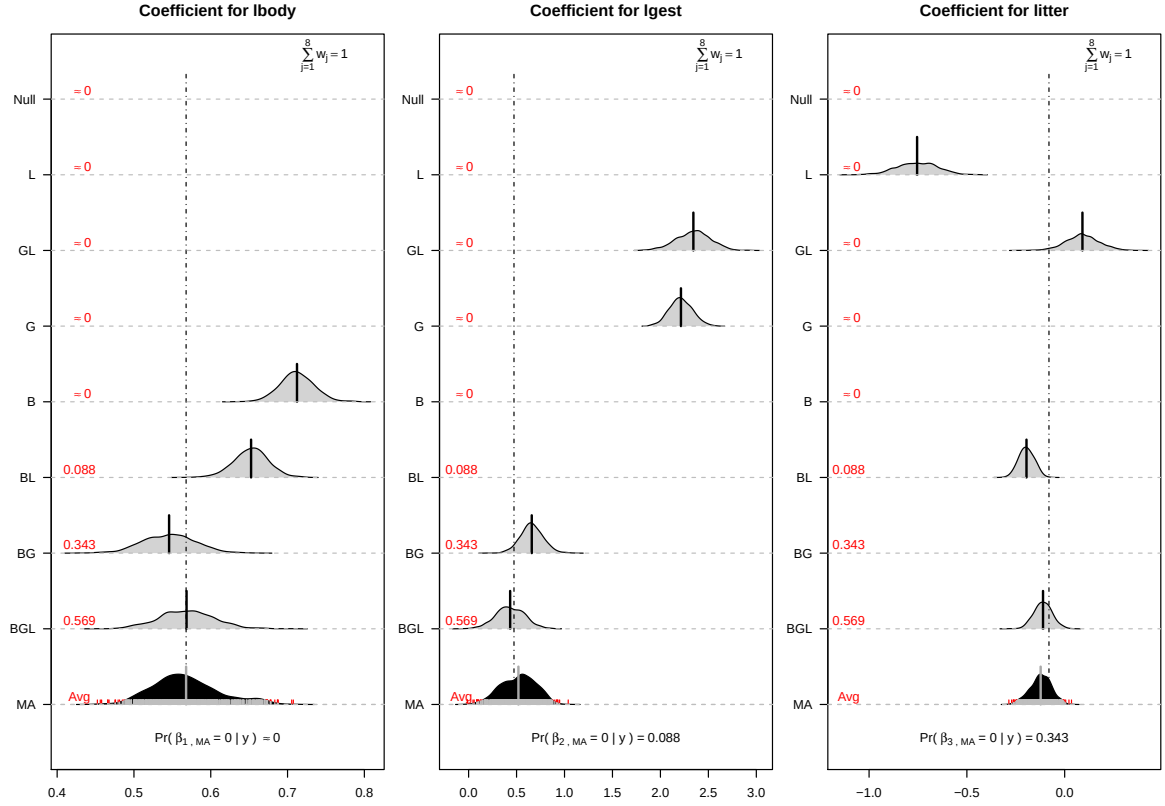


Figure 4.2: MAP plot for the partial regression coefficients associated with each explanatory variable in the **brainData** dataset.

regression coefficient associated with log body size (black beanplot and text) are shown in the first (leftmost) panel of Figure 4.2. In this case, all models excluding log body size had such small posterior model probabilities that their sum is arbitrarily close to zero. That is, the point mass at zero is so small that it is essentially negligible and the posterior distribution for the model averaged partial regression coefficient from  $\mathcal{M}$  (all models) is essentially equivalent to the posterior distribution resulting from using only the subset of models including log body size. The posterior distribution of the model averaged partial regression coefficient for log body size conditional on the models including log body size gives essentially the same results as the model-averaged posterior distribution considering all models.

To contrast this, we highlight the panel summarizing the partial regression coefficient for litter (far right). One of the models excluding litter (Model BG) has a relatively large posterior model probability of 0.343, and the other two models excluding litter (Models G and B) have negligible posterior model probabilities. Therefore, the zero component of the posterior distribution of the model-averaged partial regression coefficient associated with litter is the sum over the posterior model probabilities for models G, B and GB, which is approximately 0.343. The mean line is pulled away from the center of the continuous component (black beanplot) toward zero, showing the effect of the relatively large point mass at zero. Conditioning on models including litter, the model-averaged posterior distribution is essentially a weighted average of the model with all three explanatory variables (with weight 0.569) and the model with log body and litter (with weight 0.088). Therefore, in this case we have one model contributing to the zero component and two models contributing (very unequally) to the continuous component of the model-averaged distribution. This type of information is important for assessing the appropriateness of a model-averaged regression coefficient, and it is all easily discerned from the MAP plot.

Each row across panels displays the posterior distributions of partial regression coefficients associated with explanatory variables considered in a particular model; the model name is shown in the margin and its posterior model probability is shown on the  $y$ -axis. For example, the second row in Figure 4.2 is associated with the model containing only litter (Model L). The posterior model probability for this model is so small that it is 0.000 when rounded to three decimal places (reflected by a label of  $\approx 0$  in the MAP plot). The rows are ordered by increasing posterior model probability to facilitate easy comparisons between posteriors for partial regression coefficients from models with large PMPs (i.e., models that contribute largely to the model-averaged

posterior distribution) and their corresponding model-averaged posterior distribution (displayed in the bottom row). The posterior distribution for the coefficient associated with *litter* in Model L is shown in the panel for *litter*, and the panels for *lgest* and *lbody* do not display posterior distributions because the partial regression coefficients associated with those explanatory variables were set explicitly to zero in Model L. All other rows are read in the same fashion. Comparing rows facilitates comparisons of posteriors for coefficients and model probabilities across models.

The MAP plot also returns a table for comparing the posterior standard deviations of the model-averaged results to those from individual models (Table 4.2). Here, we view this table by calling the stored results (`map_brain`).

```
map_brain
```

Table 4.2: Posterior standard deviations for partial regression coefficients coming from individual models, their corresponding model-averaged posteriors (MA), and the subset of models including the associated explanatory variable (MA\_cts).

|        | MA     | MA_cts | BGL    | BG     | BL     | B      | G      | GL     | L      | Null |
|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|------|
| lbody  | 0.0438 | 0.0438 | 0.0349 | 0.0351 | 0.0226 | 0.0216 | —      | —      | —      | —    |
| lgest  | 0.2217 | 0.1743 | 0.1508 | 0.1197 | —      | —      | 0.1211 | 0.1695 | —      | —    |
| litter | 0.0721 | 0.0544 | 0.0472 | —      | 0.0366 | —      | —      | 0.0829 | 0.1008 | —    |
| PMP    | —      | —      | 0.569  | 0.3433 | 0.0877 | 0      | 0      | 0      | 0      | 0    |

A common motivation for model averaging is to “account for model uncertainty,” which is often equated to increased variances for quantities of interest. However, it is unclear whether this translation of model uncertainty into increased variances is realized for partial regression coefficients. In Chapter 3, we investigated this via a simulation study. We also include a table of posterior standard deviations with MAP to facilitate comparisons for particular examples.

#### 4.7.1.2 Using the MAP Plot for Justifying Model Choices

We return to the original question of interest stated at the beginning of the example. The researchers were specifically interested in the relationship between mean brain weight and gestation length over all mammals (ignoring body size and litter size). The regression coefficient representing this relationship comes from the model with log gestation length as the only explanatory variable (Model G, third row in Figure 4.2). We observe two things in Figure 4.2: (1) the model of interest has such a small posterior model probability that it is essentially ignored in the model averaging, and (2) the posterior distribution for the partial regression coefficient associated with log gestation length in the model of interest is clearly shifted to the right compared to the posterior distributions of the partial regression coefficients with large enough posterior model probabilities to contribute largely to the model averaging. That is, the posteriors contributing to the model-averaged result for log gestation length and the posterior for the partial regression coefficient addressing the question of interest tell very different stories.

We easily see from the MAP plot that the total posterior probability for models excluding log gestation length is about 0.088. The component of the model-averaged posterior distribution made up by the subset of models for which log gestation length is included is essentially a weighted average of the posterior distributions from models GBL and GB:  $p(\beta_{3.MA}|y, \mathcal{M}_L^*) = 0.569p(\beta_{3.gbl}|y) + 0.343p(\beta_{3.bl}|y)$ . Although we can see in Figure 4.2 that the two posterior distributions contributing to the continuous component of the model averaged distribution are similar, the result is a mixture of two coefficients with different meanings: one conditional on log body size alone, and one conditional on both log body size and litter size. As we mentioned above, neither of these directly addresses the relationship of interest.

We have seen researchers interpret model-averaged coefficients as “overall effects” of their associated explanatory variables. However, this is technically incorrect and potentially very misleading. Not only is the interpretation difficult, but the model-averaged distribution for the coefficient associated with log gestation length is essentially a weighted average of posterior distributions from only *two* of the *eight* models. While information like this is extremely important, it is not typically reported or discussed and the MAP plot tells a clear visual story to convey this information.

This example, though simple, illustrates the complexities of model averaging regression coefficients, and demonstrates that model averaging leads to tough decisions of its own, just like defending the choice of one model. Such decisions should be accompanied by a thorough description of the main components going into the model averaging. Whether the researcher chooses to condition inference on one model or use an approach like model averaging, the decisions and justifications along the way should be driven by the research question and the intended use of the results. When a specific question is well defined, an appropriate way to address that question can be to build a regression model driven by the question itself, rather than automatically considering a model set like all subsets. Building such a model requires careful thought and effort to justify, making it tempting to elicit some sort of automatic approach to find the “best” model or combine results over multiple models. Something we often lose sight of when building models is that the “best” model, whether it be based on AIC,  $p$ -values, or some other model selection criteria, may not be consistent with addressing the original question. So, while model averaging may hold the allure of a decision-free model building technique (just use all of them), we must consider whether the result of the procedure leaves us with the ability to address the relationships we set out to investigate in the first place. We urge researchers to think hard about what research questions they are investigating and whether

“accounting for model uncertainty” is worth the potential loss in interpretation and ability to address the question(s) of interest. We present the MAP plot is a tool to help achieve this goal.

#### 4.7.2 Working with Large Model Sets

We will use the `bfat` data which contain 13 body fat measurements from 251 men to demonstrate the flexibility of our plotting function for problems with a larger set of potential predictors. First, we will take a look at these data and create a `bma` object, `bfat_bms`, using the `bms` function. Again, we chose a uniform prior on the model space and Zellner’s  $g$ -prior (Zellner, 1984) with  $g = n$ .

```
data(bfat)
head(bfat)
```

Table 4.3: The first five rows of the `bfat` data set.

| bf.brozek | age | weight | height | neck  | chest  | abdomen | hip    | thigh | knee  | ankle | ex.bicep | forearm | wrist |
|-----------|-----|--------|--------|-------|--------|---------|--------|-------|-------|-------|----------|---------|-------|
| 12.60     | 23  | 154.25 | 67.75  | 36.20 | 93.10  | 85.20   | 94.50  | 59.00 | 37.30 | 21.90 | 32.00    | 27.40   | 17.10 |
| 6.90      | 22  | 173.25 | 72.25  | 38.50 | 93.60  | 83.00   | 98.70  | 58.70 | 37.30 | 23.40 | 30.50    | 28.90   | 18.20 |
| 24.60     | 22  | 154.00 | 66.25  | 34.00 | 95.80  | 87.90   | 99.20  | 59.60 | 38.90 | 24.00 | 28.80    | 25.20   | 16.60 |
| 10.90     | 26  | 184.75 | 72.25  | 37.40 | 101.80 | 86.40   | 101.20 | 60.10 | 37.30 | 22.80 | 32.40    | 29.40   | 18.20 |
| 27.80     | 24  | 184.25 | 71.25  | 34.40 | 97.30  | 100.00  | 101.90 | 63.20 | 42.20 | 24.00 | 32.20    | 27.70   | 17.70 |

```
# create bma object using bfat data
bfat_bms <- bms(bfat, mprior = "uniform", g = "UIP", user.int = F)
```

If we consider all potential first-order combinations of the predictors, there are  $2^{13} = 8192$  potential models in the model set. By default, a `bma` object generated by `bms` will store weights and posterior summary measures for the top 500 models considered. Using the closed form posterior results for model averaging with a  $g$ -prior on the regression coefficients, `MApp.bms` will obtain a specified number of draws from the analytical posterior distributions for all models stored in the `bma` object and will represent the model-averaged distribution by sampling from these individual posteriors with normalized probabilities to retain proportionality from

model averaging including all models. The proportion of posterior model mass accounted for by the top 500 models should be checked to ensure the approximate model-averaged distribution created by `MApp_bms` is reasonable. In this case, the top 500 models accounted for approximately 97% of the posterior mass and we can be confident that the approximate model-averaged distributions will be very similar to the model-averaged distributions when all models are considered.

```
# see how much posterior model mass is accounted for by top 500 models
sum(pmp.bma(bfat_bms)[,1])
```

```
## [1] 0.9749476
```

Plotting 500 models simultaneously would make the MAP plot impossible to read, so in `MApp_bms` we built in functionality to display only a subset of those models using the `max_display` argument. We can also specify which coefficient to include in the model-averaged posteriors plot with the `include_coef` argument. If `max_display` and `include_coef` are left unspecified, the function will display the 20 models with the largest PMP.

```
map_bfat <- MApp_bms(bfat_bms, plot_wind = c(1,3),
  include_coef = c(13, 2, 6),
  max_display = 10,
  num_sims = 1000)
```

```
map_bfat
```

Table 4.4: Posterior standard deviations for partial regression coefficients associated with wrist, weight, and abdomen circumference.

|         | MA     | MA_cts | M1     | M2     | M3     | M4     | M5     | M6     | M7     | M8     | M9     | M10    |
|---------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| wrist   | 0.6496 | 0.4597 | 0.4271 | 0.4218 | 0.4052 | —      | 0.4264 | 0.4541 | 0.4241 | —      | 0.4057 | 0.4274 |
| weight  | 0.0395 | 0.0324 | 0.0226 | 0.0229 | 0.0241 | 0.0199 | 0.0251 | 0.028  | 0.0253 | 0.0234 | 0.0292 | 0.0254 |
| abdomen | 0.0657 | 0.0657 | 0.0528 | 0.0533 | 0.051  | 0.0545 | 0.0527 | 0.0651 | 0.0518 | 0.0528 | 0.0533 | 0.0538 |
| PMP     | —      | —      | 0.1636 | 0.0788 | 0.0463 | 0.0376 | 0.034  | 0.0246 | 0.0235 | 0.0206 | 0.0195 | 0.0164 |

Figure 4.3 and Table 4.4 provide the MAP plot and posterior standard deviations for three variables in the data set. To view all of the explanatory variables, the user

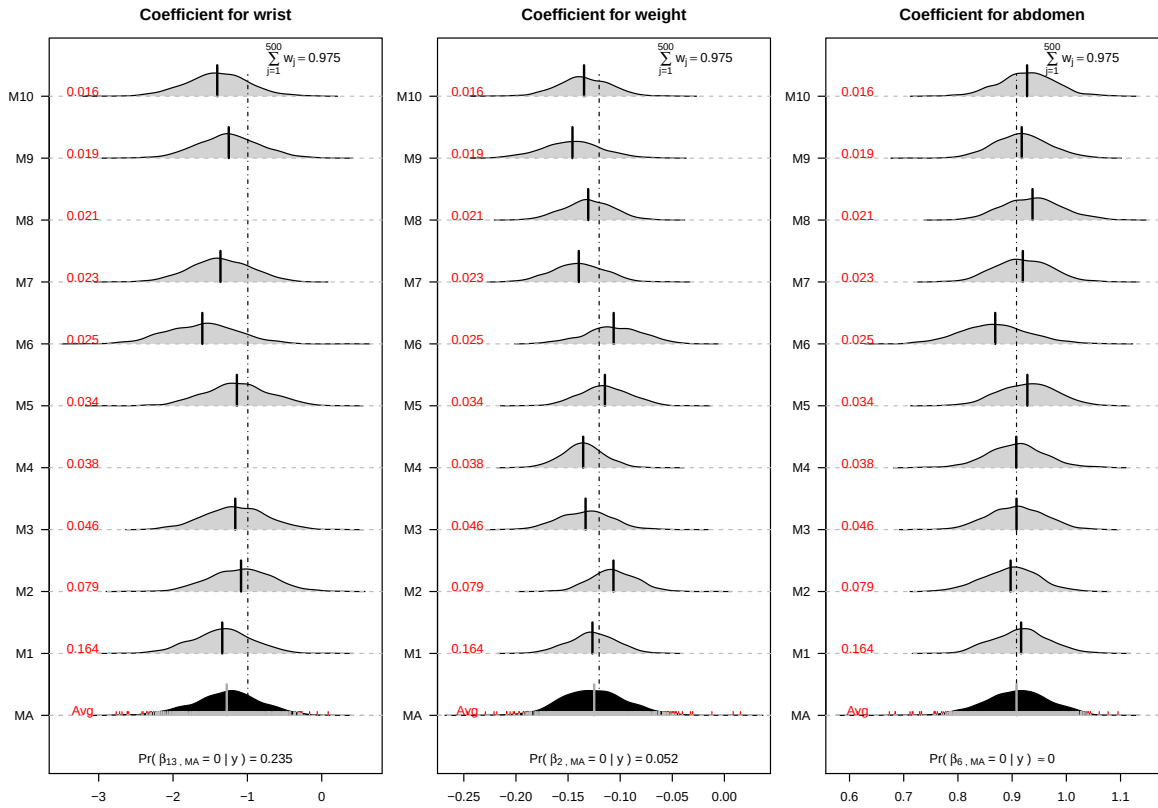


Figure 4.3: MAP plot for partial regression coefficients associated with three of the explanatory variables in `bfat`.

can remove the `include_coef` argument from the function call and the MAP plot will be made for all explanatory variables being considered.

#### 4.8 MApp for IC-Based MA

It is also common to obtain approximate posterior model probabilities (model weights) from certain information criterion (most commonly: BIC, AICc, or AIC). A version of the MAP plot can be created for comparing confidence intervals for partial regression coefficients across the models in the model set.

#### 4.8.1 Using `MApp_IC`

`MApp_IC` generates results from approximations to posterior model probabilities (PMPs) based on two user specifications: (1) the type of information criterion, and (2) the set of models to be used in the averaging (all models, or only those with the explanatory variable included). The default for `MApp_IC` conducts BIC-based model averaging (`type = "BIC"`), placing a discrete uniform prior on the model set (although this may be changed by providing an argument for `mod_prior`), and considers all models in the model set in the averaging (`w_plus = FALSE`). Here, we make the MAP plot for the `brainData` with the default specifications.

```
MApp_IC(brainData, plot_wind = c(1,3),
        type = "BIC", w_plus = FALSE)
```

```
## Subset of models used in MA: All models in the model set
## Type of information criteria used in MAP plot: BIC
```

Table 4.5: Approximate posterior model probabilities resulting from AICc-based and BIC-based model averaging.

|    | Model   | AICc_pmp | BIC_pmp | AICc    | BIC     |
|----|---------|----------|---------|---------|---------|
| 1  | MA_AICc |          |         |         |         |
| 2  | MA_BIC  |          |         |         |         |
| 3  | M1      | 0.895    | 0.368   | 136.310 | 148.465 |
| 4  | M2      | 0.088    | 0.318   | 140.947 | 150.765 |
| 5  | M3      | 0.017    | 0.259   | 144.271 | 154.089 |
| 6  | M4      | 0.000    | 0.055   | 171.447 | 178.880 |
| 7  | M5      | 0.000    | 0.000   | 274.682 | 282.114 |
| 8  | M6      | 0.000    | 0.000   | 275.600 | 285.418 |
| 9  | M7      | 0.000    | 0.000   | 381.491 | 388.924 |
| 10 | M8      | 0.000    | 0.000   | 424.408 | 429.407 |

The function returns messages providing the type of model averaging used and other useful output, including AICc-based MA results.

Sometimes researchers choose to condition their results on the subset of models including the variables of interest (i.e., models where associated coefficients are not set

Table 4.6: Estimated model-averaged coefficients resulting from AICc-based and BIC-based model averaging, along with their corresponding maximum likelihood estimates from each individual model. If an estimate is zero its corresponding variable is not included in the model.

|    | Model   | Est_Int | Est_lbody | Est_lgest | Est_litter |
|----|---------|---------|-----------|-----------|------------|
| 1  | MA_AICc | 0.746   | 0.574     | 0.452     | -0.102     |
| 2  | MA_BIC  | 1.039   | 0.597     | 0.375     | -0.092     |
| 3  | M1      | 0.823   | 0.575     | 0.440     | -0.110     |
| 4  | M2      | -0.457  | 0.551     | 0.668     | 0.000      |
| 5  | M3      | 2.917   | 0.658     | 0.000     | -0.197     |
| 6  | M4      | 2.332   | 0.719     | 0.000     | 0.000      |
| 7  | M5      | -6.665  | 0.000     | 2.234     | 0.000      |
| 8  | M6      | -7.528  | 0.000     | 2.371     | 0.094      |
| 9  | M7      | 5.622   | 0.000     | 0.000     | -0.761     |
| 10 | M8      | 3.865   | 0.000     | 0.000     | 0.000      |

Table 4.7: Estimated standard errors for model-averaged coefficients resulting from AICc-based and BIC-based model averaging, along with their corresponding maximum likelihood estimates from each individual model. If an estimate is zero its corresponding variable is not included in the model.

|    | Model   | SE_Int | SE_lbody | SE_lgest | SE_litter |
|----|---------|--------|----------|----------|-----------|
| 1  | MA_AICc | 0.746  | 0.034    | 0.152    | 0.049     |
| 2  | MA_BIC  | 1.314  | 0.056    | 0.273    | 0.080     |
| 3  | M1      | 0.662  | 0.033    | 0.137    | 0.042     |
| 4  | M2      | 0.458  | 0.032    | 0.109    | 0.000     |
| 5  | M3      | 0.119  | 0.020    | 0.000    | 0.034     |
| 6  | M4      | 0.073  | 0.020    | 0.000    | 0.000     |
| 7  | M5      | 0.562  | 0.000    | 0.117    | 0.000     |
| 8  | M6      | 0.960  | 0.000    | 0.170    | 0.084     |
| 9  | M7      | 0.293  | 0.000    | 0.000    | 0.101     |
| 10 | M8      | 0.222  | 0.000    | 0.000    | 0.000     |

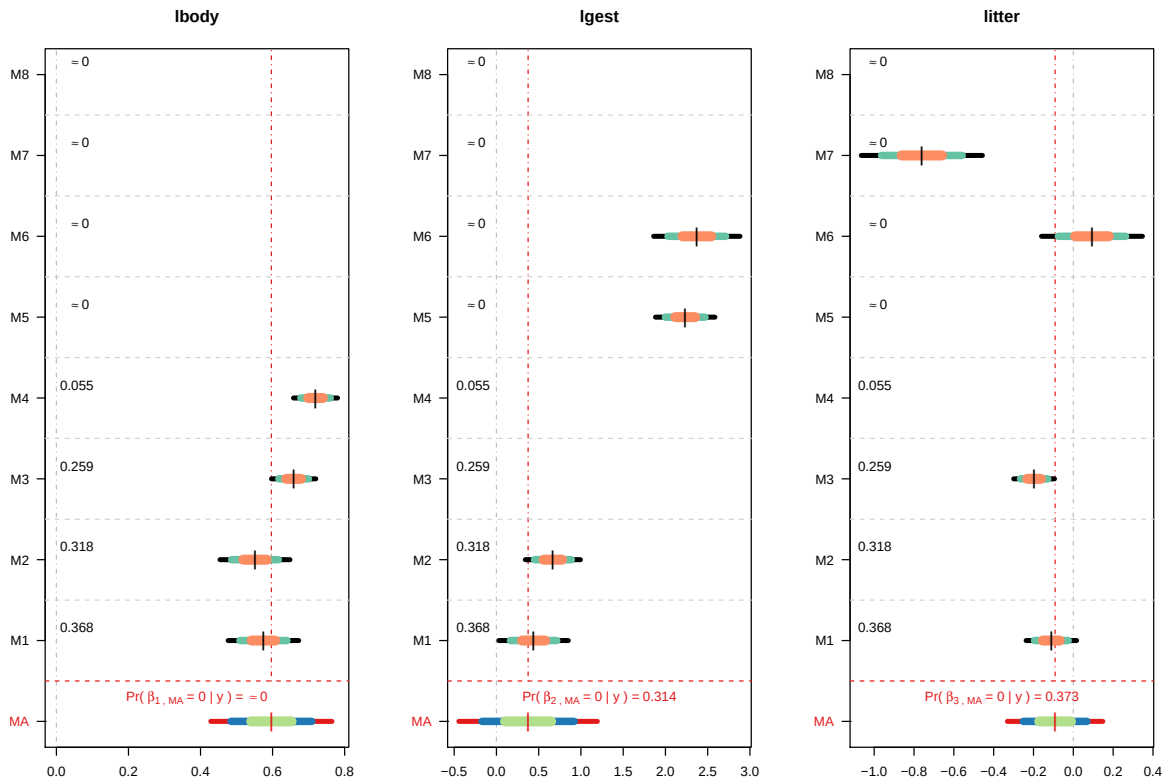


Figure 4.4: MAP plot created using `MAP_IC` with BIC-based model averaging, using all models in the model set. Intervals are created using  $\pm 1, 2$ , and  $3$  SEs from the estimated mean. Intervals for individual models are represented by orange, teal, and black segments, respectively. Intervals for the model-averaged results are represented by green, blue, and red segments, respectively.

to 0), as described in Burnham and Anderson (2002). Setting the argument `w_plus = TRUE` tells `MAPP_IC` to conduct model averaging on this restricted model set.

```
MApp_IC(brainData, plot_wind = c(1,3),
        type = "BIC", w_plus = TRUE)
```

```
## Subset of models used in MA: Models where coefficient is not set to 0
## Type of information criteria used in MAP plot: BIC
```

The message now indicates that model averaging was conducted conditional on only the models for which the explanatory variables were included. Comparing Figures 4.4 and 4.5, we note that the model averaging conditional on the subset

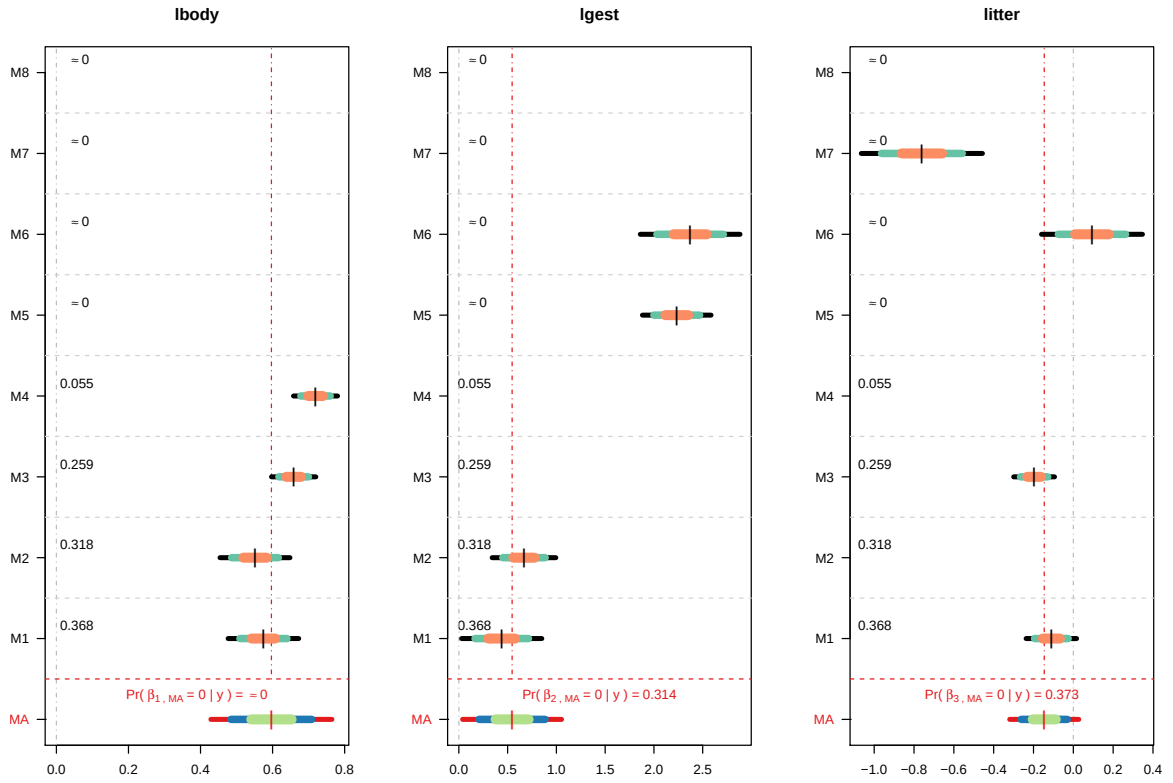


Figure 4.5: MAP plot for BIC-based model averaging and conditioning on the subset of models for which the variable associated with each coefficient of interest is included (Mapp\_IC with w\_plus = TRUE). Intervals are created using  $\pm 1, 2$ , and 3 SEs from the estimated mean. Intervals for individual models are represented by orange, teal, and black segments, respectively. Intervals for the model-averaged results are represented by green, blue, and red segments, respectively.

of models for which the coefficient is included results in estimates for the model-averaged partial regression coefficients associated with log gestation length and litter size that are farther from zero than the estimates from all models. We also observe slightly smaller standard errors for the model-averaged partial regression coefficients associated with log gestation length and litter size when model averaging is conducted using only the subset of models where those variables are included. This is to be expected because the models for which the coefficient associated with these explanatory variables is set to zero are excluded from the averaging. We also note

that there is little difference in Figures 4.4 and 4.5 in the panel displaying partial regression coefficients associated with log body weight. Again, this is not surprising because the models for which the coefficient associated with log body weight is set exactly to zero have posterior model probabilities so small that they are essentially negligible.

The `MApp_IC` function works best with modest numbers of explanatory variables because it uses `bms` to define an *all subsets* model set. When 9 explanatory variables are considered, the *all subsets* model set contains 512 models, and by default `bms` will only store results from 500 models. Therefore, if this function is used with data containing 9 or more explanatory variables, it is important for the user to make sure the sum of the PMPs for the top 500 models (estimated using `bms`) is large enough that conditioning on those 500 models is appropriate for IC-based model averaging. A less automatic, but more flexible function is `MApp_IC_gen` (demonstrated in Section 4.8.2).

#### 4.8.2 Using `MApp_IC_gen`

Results from IC-based model averaging can be used as input arguments to `MApp_IC_gen` to create the MAP plot. This function requires the user to input four arguments: (1) `x_coef`, a matrix of maximum likelihood estimates of the coefficients of interest for each model considered, (2) `x_se`, a matrix of standard errors for those estimates, (3) `pmp`, a vector (length =  $J$ ) of model weights, and (4) `inmat`, a matrix of zeros and ones corresponding to the variables included in each model (all matrices will be  $J$  rows and  $p$  columns). The user can choose not to include all  $J$  models by using a subset of the models in the four inputs. In this case the user should ensure the subset of models being used makes up enough of the posterior model mass that

results from model averaging conditional on the subset of models are similar to results conditional on all models.

We use the `braindata` to demonstrate this function, and we use the `bma` object created by the `bms` function to quickly create `inmat`. However, `inmat` can be specified in any way as long as each row represents a model and the columns represent all explanatory variables of interest.

```
post_means <- t(brain$topmod$betas())
post_means[post_means != 0] <- 1
inmat <- post_means
```

We use another function in this package, `IC_approx`, to obtain the posterior model probabilities associated with all models considered. This function computes AICc-based and BIC-based posterior model probabilities (model weights) for each model in the model set specified by `inmat`.

```
IC_approx <- approx_pmp(inmat, Xmat = brainData[,-1], Yvec = brainData[,1])
```

```
## Subset of models used in MA: All models in the model set
```

Results from using `IC_approx` are shown in Tables 4.8 - 4.10.

Table 4.8: Posterior model probabilities calculated based using two types of information criteria: AICc, and BIC.

| Model   | AICc_pmp | BIC_pmp | AICc    | BIC     |
|---------|----------|---------|---------|---------|
| MA_AICc | NA       | NA      | NA      | NA      |
| MA_BIC  | NA       | NA      | NA      | NA      |
| M1      | 0.895    | 0.368   | 136.310 | 148.465 |
| M2      | 0.088    | 0.318   | 140.947 | 150.765 |
| M3      | 0.017    | 0.259   | 144.271 | 154.089 |
| M4      | 0.000    | 0.055   | 171.447 | 178.880 |
| M5      | 0.000    | 0.000   | 274.682 | 282.114 |
| M6      | 0.000    | 0.000   | 275.600 | 285.418 |
| M7      | 0.000    | 0.000   | 381.491 | 388.924 |
| M8      | 0.000    | 0.000   | 424.408 | 429.407 |

Table 4.9: Maximum likelihood estimates for all coefficients from each model considered and their model-averaged counterparts according to AICc-based and BIC-based model averaging. If an estimate is zero its corresponding variable is not included in the model.

| Model   | Est_Int | Est_lbody | Est_lgest | Est_litter |
|---------|---------|-----------|-----------|------------|
| MA_AICc | 0.746   | 0.574     | 0.452     | -0.102     |
| MA_BIC  | 1.039   | 0.597     | 0.375     | -0.092     |
| M1      | 0.823   | 0.575     | 0.440     | -0.110     |
| M2      | -0.457  | 0.551     | 0.668     | 0.000      |
| M3      | 2.917   | 0.658     | 0.000     | -0.197     |
| M4      | 2.332   | 0.719     | 0.000     | 0.000      |
| M5      | -6.665  | 0.000     | 2.234     | 0.000      |
| M6      | -7.528  | 0.000     | 2.371     | 0.094      |
| M7      | 5.622   | 0.000     | 0.000     | -0.761     |
| M8      | 3.865   | 0.000     | 0.000     | 0.000      |

Table 4.10: Standard errors obtained from maximum likelihood methods for all coefficients from each model considered and their model-averaged counterparts according to AICc-based and BIC-based model averaging. If an estimate is zero its corresponding variable is not included in the model.

| Model   | SE_Int | SE_lbody | SE_lgest | SE_litter |
|---------|--------|----------|----------|-----------|
| MA_AICc | 0.746  | 0.034    | 0.152    | 0.049     |
| MA_BIC  | 1.314  | 0.056    | 0.273    | 0.080     |
| M1      | 0.662  | 0.033    | 0.137    | 0.042     |
| M2      | 0.458  | 0.032    | 0.109    | 0.000     |
| M3      | 0.119  | 0.020    | 0.000    | 0.034     |
| M4      | 0.073  | 0.020    | 0.000    | 0.000     |
| M5      | 0.562  | 0.000    | 0.117    | 0.000     |
| M6      | 0.960  | 0.000    | 0.170    | 0.084     |
| M7      | 0.293  | 0.000    | 0.000    | 0.101     |
| M8      | 0.222  | 0.000    | 0.000    | 0.000     |

We use these results with `MApp_IC_gen` to create the MAP plot. It is important to note that the four pieces of information required to use `MApp_IC_gen` need not be obtained from `IC_approx`, in fact, they can be obtained from any software providing IC-based PMPs (model weights) along with estimated coefficients and standard errors from each model being considered.

```
# BIC weights
x_coef <- IC_approx[c(2:10), c(7:9)]
x_se <- IC_approx[c(2:10), c(11:13)]
pmp <- as.numeric(as.character(IC_approx[c(3:10), 2]))
MApp_IC_gen(x_coef = x_coef, x_se = x_se, pmp = pmp, inmat = inmat)
```

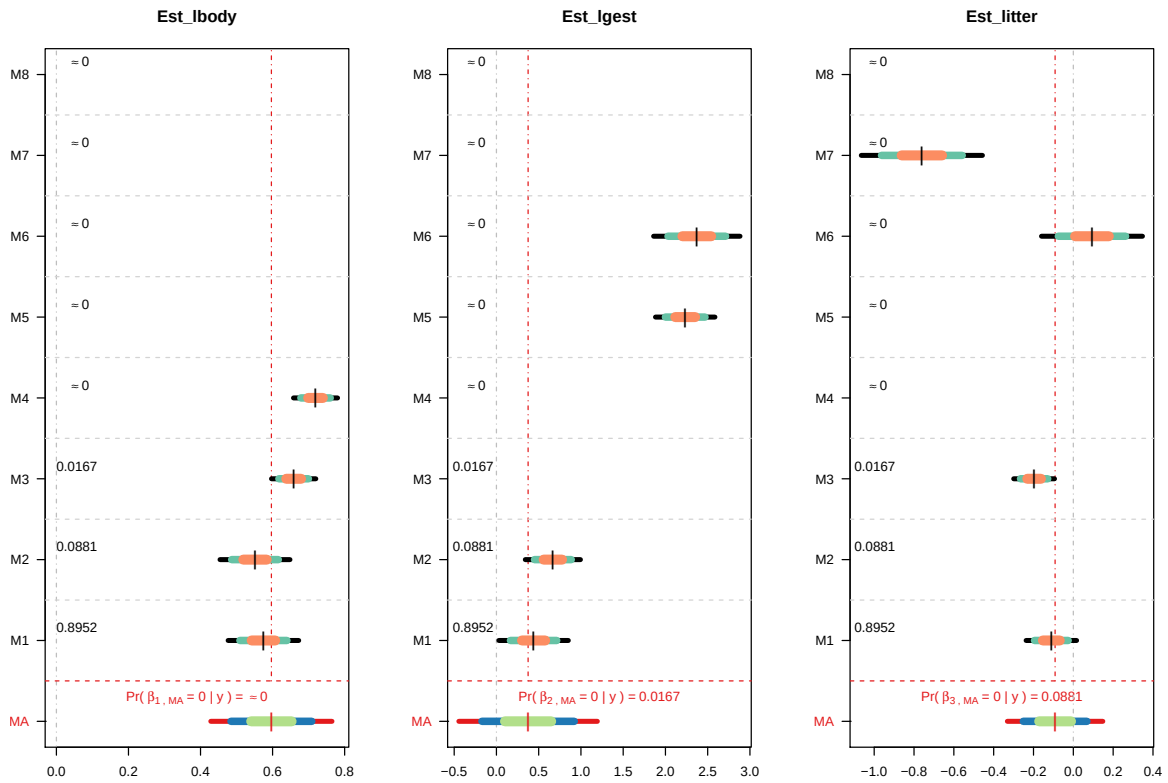


Figure 4.6: MAP plot for IC-based model averaging using `MApp_IC_gen`. Intervals are created using  $\pm 1, 2$ , and  $3$  SEs from the estimated mean. Intervals for individual models are represented by orange, teal, and black segments, respectively. Intervals for the model-averaged results are represented by green, blue, and red segments, respectively.

## 4.9 MApp for Self-Programmed Samplers

Another way to conduct Bayesian model averaging is to use **BUGS**, **OpenBUGS**, or a self-programmed sampler (e.g., using reversible jump MCMC). In this case, the user will have posterior draws for partial regression coefficients from each of the models  $\mathcal{M}$ . Collections of posterior draws are usually stored one of two ways: (1) in a list (length is usually equal to the cardinality of the model set, unless certain models are never visited), or (2) in a matrix with posterior draws for quantities of interest including model as a variable in the matrix. The **MApp** package has functions for both classes of posterior draws.

### 4.9.1 MApp\_MCMC for Matrices

Here we show how the MAP plot can be made using results similar to the posterior draws that would be obtained from implementing model averaging with **WinBUGS** (Lunn et al., 2000) or **R2OpenBugs** (Sturtz et al., 2005). Typically draws are extracted using **coda** (Plummer et al., 2006), and they will be a **mcmc.list** object with length equal to the number of chains. This list can be converted to a matrix with a column for the model, and columns for the estimates of partial regression coefficients by using the **as.matrix** function in the **coda** package and ‘rbind-ing’ the matrices together. Here, we use fake data considering 3 variables and 7 models. The data are of the form shown in Table 4.11.

```
head(mcmc_mat)
```

To create the MAP plot using **MApp\_MCMC**, with the default options, use the following (see Figure 4.7).

```
mcmc_beans <- MApp_MCMC(mcmc_mat, plot_wind = c(1,3))
```

Table 4.11: Example of the form of posterior draws in a matrix.

| model | X1      | X2     | X3     |
|-------|---------|--------|--------|
| M3    | -0.5044 | 0.0000 | 6.5181 |
| M2    | -1.8689 | 3.2709 | 0.0000 |
| M5    | -1.4742 | 0.0000 | 0.0000 |
| M7    | -0.0000 | 0.0000 | 5.9674 |
| M3    | 0.1373  | 0.0000 | 6.0382 |

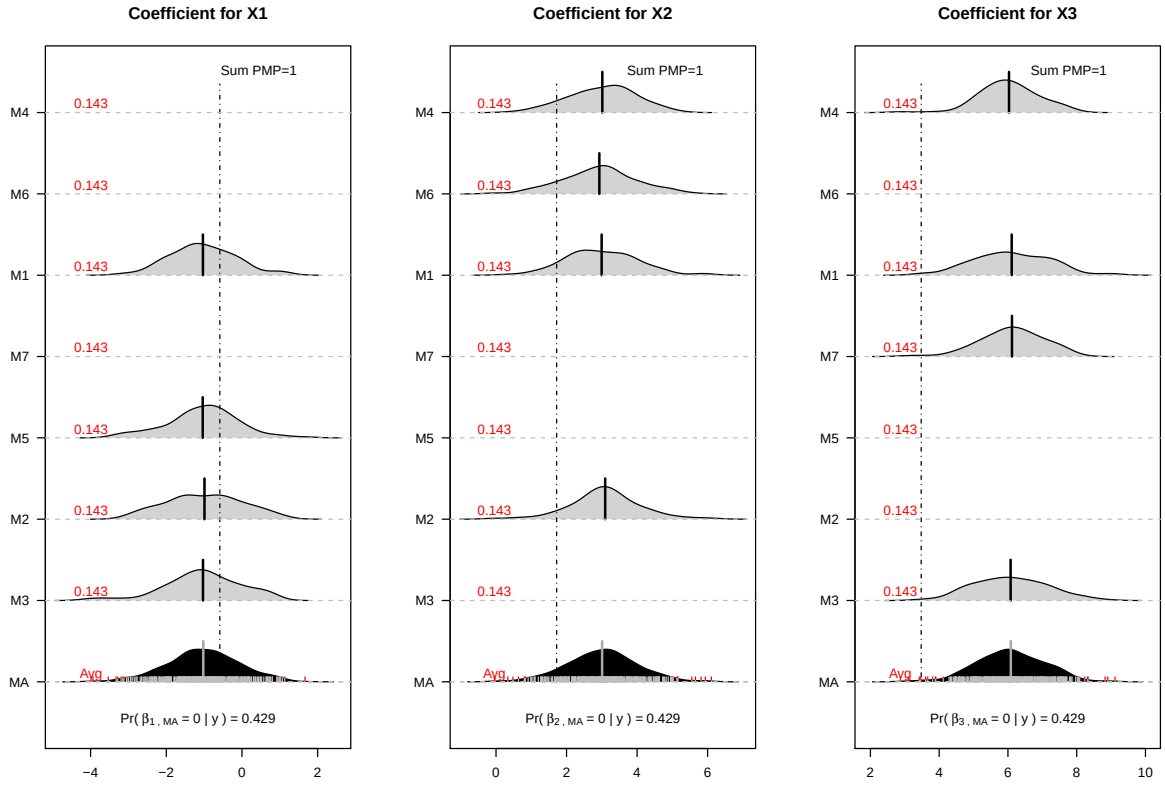


Figure 4.7: MAP plot created using MApp\_MCMC.

Another option for `max_display` (common to `MApp_bms`, `MApp_MCMC`, and `MApp_list`) is `"common3"`, which displays the top PMP model, the model including all explanatory variables, and the model-averaged result (see Figure 4.8).

```
ex_mcmc <- MApp_MCMC(mcmc_mat, plot_wind = c(1,3), max_display = "common3")
```

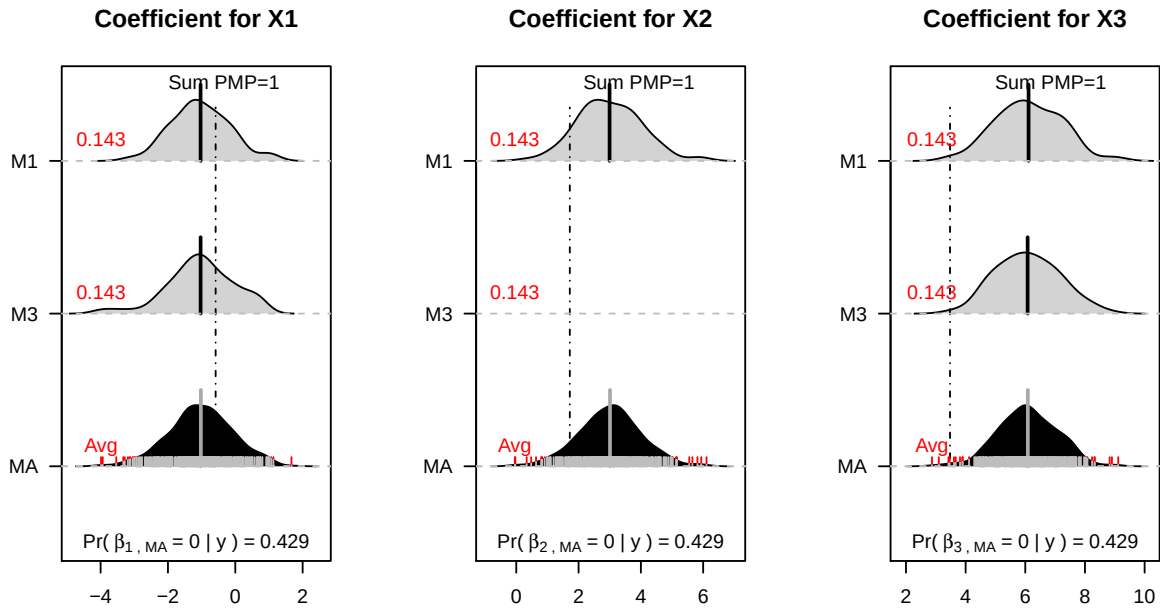


Figure 4.8: MAP plot created using `MApp_MCMC` and the `"common3"` argument.

In this example, the posterior model probabilities are all equal, resulting in equal posterior exclusion probabilities (point masses at zero) for the partial regression coefficients associated with each explanatory variable. The overall mean for the model-averaged posterior distribution gets pulled toward zero in all three panels, but we see the most shrinkage toward zero in the panel for  $X_3$ . This is a result of the individual model posteriors for the coefficients associated with  $X_3$  being distributed similarly; i.e., all being centered away from zero with relatively small variance.

The table of posterior standard deviations shows how a relatively large posterior exclusion probability can affect the posterior variance of the model-averaged distribution. We observe that the posterior variance of the continuous component

of the model-averaged distribution is similar to the posterior variances from the corresponding individual models (Table 4.12). When considering the zero component, the model averaged posterior distribution has larger variance for the partial regression coefficients associated with  $X_2$  and  $X_3$ , and smaller variance for the partial regression coefficients associated with  $X_1$  (Table 4.12). This is a result of the location of the continuous component relative to zero. We discuss relationships like this in detail, and investigate how they change for different linear regression scenarios using simulation in Chapter 3

*# look at the table of posterior standard deviations for all models*  
mcmc\_beans

Table 4.12: Posterior standard deviations from individual models and their model-averaged counterpart.

|     | MA     | MA_cts | M3     | M2     | M5     | M7     | M1     | M6     | M4     |
|-----|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| X1  | 0.8793 | 0.9518 | 1.0079 | 0.9991 | 0.9418 | —      | 0.8651 | —      | —      |
| X2  | 1.6736 | 1.0084 | —      | 1.0032 | —      | —      | 1.0294 | 1.0503 | 0.9565 |
| X3  | 3.1055 | 0.9944 | 1.0734 | —      | —      | 0.924  | 1.116  | —      | 0.8541 |
| PMP | —      | —      | 0.1429 | 0.1429 | 0.1429 | 0.1429 | 0.1429 | 0.1429 | 0.1429 |

#### 4.9.2 Using `MApp_list`

Another common format for posterior draws is a list with one element for each model, in which case, `MApp_list` can be used. For this example, the posterior draws from the model with the highest PMP are in the first element of the list, and posterior draws from the model with the smallest PMP are in the last element of the list.

```
MApp_list(mcmc_list, plot_wind = c(1,5), max_display = 10)
```

```
## Table of posterior standard deviations:
##           compare individual models to the MA posterior.
##
##
##           (press enter to display table)
```

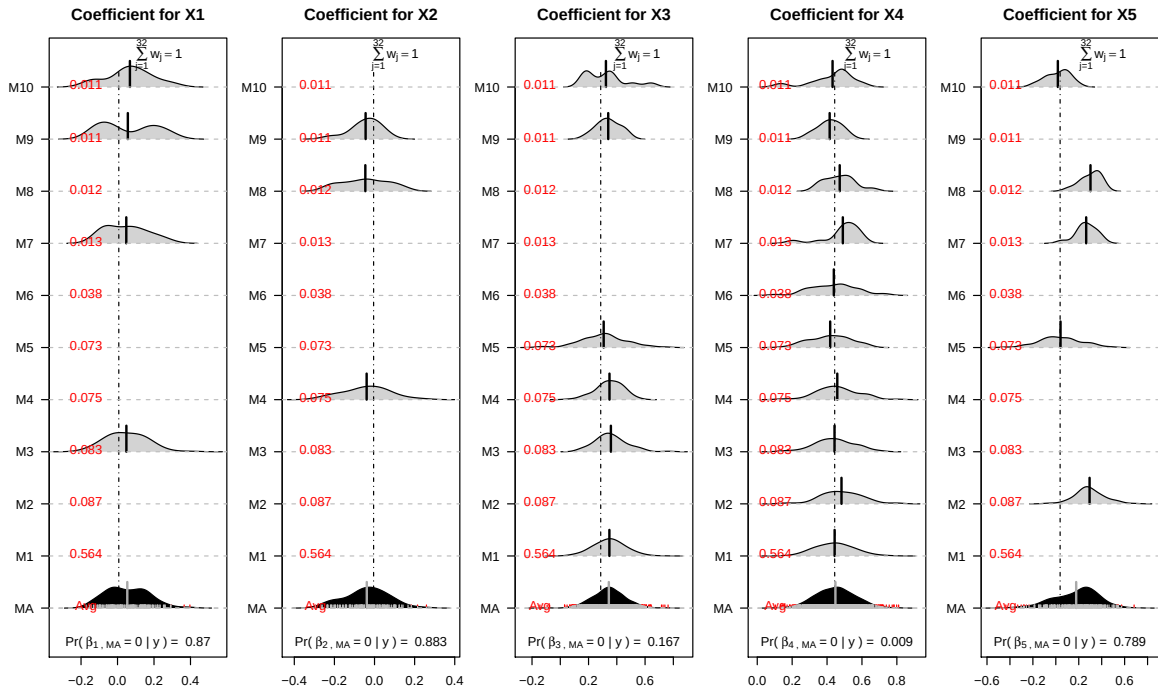


Figure 4.9: An example MAP plot produced using MApp\_list.

| ##     | MA       | MA_cts   | M1       | M2       | M3       | M4       | M5       |
|--------|----------|----------|----------|----------|----------|----------|----------|
| ## X1  | "0.0485" | "0.1198" | "---"    | "---"    | "0.1283" | "---"    | "---"    |
| ## X2  | "0.0394" | "0.1142" | "---"    | "---"    | "---"    | "0.1083" | "---"    |
| ## X3  | "0.1647" | "0.1116" | "0.1061" | "---"    | "0.099"  | "0.0899" | "0.1443" |
| ## X4  | "0.1215" | "0.1146" | "0.1118" | "0.1274" | "0.1137" | "0.1083" | "0.1116" |
| ## X5  | "0.1116" | "0.2034" | "---"    | "0.1161" | "---"    | "---"    | "0.1615" |
| ## PMP | "---"    | "---"    | "0.5636" | "0.0875" | "0.0835" | "0.0746" | "0.0726" |
| ##     | M6       | M7       | M8       | M9       | M10      |          |          |
| ## X1  | "---"    | "0.0918" | "---"    | "0.0907" | "0.0645" |          |          |
| ## X2  | "---"    | "---"    | "0.1389" | "0.1152" | "---"    |          |          |
| ## X3  | "---"    | "---"    | "---"    | "0.106"  | "0.23"   |          |          |
| ## X4  | "0.1111" | "0.1272" | "0.1527" | "0.0883" | "0.1147" |          |          |
| ## X5  | "---"    | "0.1539" | "0.1273" | "---"    | "0.1724" |          |          |
| ## PMP | "0.0378" | "0.0129" | "0.0119" | "0.0109" | "0.0109" |          |          |

Unlike the MAP plot, the number of draws from each individual model is proportional to the posterior model probability, and some beanplots may be drawn with very few observations.

## 4.10 Conclusion

In this chapter, we highlighted the plotting functions in **MApp**, however, many of the other functions in this package are useful for studying properties of model-averaged posteriors for partial regression coefficients under the *g*-prior implementation of Bayesian model averaging. Use `help(package = "MApp")` to see a list of all functions in the package and their descriptions (Figure 4.10).

### Documentation for package ‘MApp’ version 0.1

- [DESCRIPTION file](#).

### Help Pages

|                                |                                                                                                    |
|--------------------------------|----------------------------------------------------------------------------------------------------|
| <a href="#">aicc</a>           | Compute AICc for a model                                                                           |
| <a href="#">approx_pmp</a>     | Compute individual and model averaged results using the AIC and BIC approximations.                |
| <a href="#">bfat</a>           | Body fat data for 251 men                                                                          |
| <a href="#">bms_post_sim</a>   | Simulate draws from posterior distributions of partial regression coefficients.                    |
| <a href="#">brainData</a>      | Brain weight, body weight, gestation length and litter size                                        |
| <a href="#">cov_fun</a>        | Compute analytical posterior variance covariance matrix of partial regression coefficients         |
| <a href="#">est_MA</a>         | Compute model averaged estimates for partial regression coefficients.                              |
| <a href="#">MApost_bms_sim</a> | Generate individual and MA posterior distributions                                                 |
| <a href="#">MApp_bms</a>       | Model Averaged Posteriors Plot (MAP plot)                                                          |
| <a href="#">MApp_IC</a>        | Model Averaged Posteriors Plot for MA Using Information Criteria                                   |
| <a href="#">MApp_IC_gen</a>    | Model Averaged Posteriors Plot                                                                     |
| <a href="#">MApp_list</a>      | Model Averaged Posteriors Plot                                                                     |
| <a href="#">MApp_MCMC</a>      | Model Averaged Posteriors Plot                                                                     |
| <a href="#">mean_fun</a>       | Compute analytical posterior means                                                                 |
| <a href="#">pmp_AIC</a>        | Compute approximate posterior model probabilities using the AIC approximation for a set of models. |
| <a href="#">pmp_BIC</a>        | Compute approximate posterior model probabilities using the BIC approximation for a set of models. |
| <a href="#">se_MA</a>          | Compute the standard error for model averaged partial regression coefficients.                     |
| <a href="#">vec_fun</a>        | Convert a list of posterior samples into a long data frame.                                        |

Figure 4.10: Help for the **MApp** package provides a list of functions and their descriptions.

Although the functions in this package display the results from model averaging, we do not intend for the plots to be used for that purpose. We encourage researchers to use the plotting functions in **MApp** as tools for assessing the appropriateness of model averaging partial regression coefficients, and to defend their choice of not using the method to reviewers.

## CHAPTER FIVE

## CONCLUSION AND FUTURE WORK

This body of work was meant to provoke thought and discussion, as well as provide usable tools for aiding researchers in making informed decisions about when, and more importantly, when not, to use model averaging for inference with partial regression coefficients. We began by highlighting complexities and implicit assumptions arising when model averaging is used in the context of explaining regression coefficients, rather than for its original purpose of combining predictions. We focused on the complicated interpretation of a model-averaged partial regression coefficient when collinearity among explanatory variables is present in an analysis, and suggested using explicit notation to make it clear coefficients associated with the same variable do not necessarily hold the same interpretation across models, and therefore cannot always be sensibly averaged to describe an “overall effect” of their associated variable.

We also provided two graphical tools, cut-and-coded scatterplots and the model-averaged posteriors plot (MAP plot), to aid in identifying when a partial regression coefficient of interest is sensitive to other variables in a model. We developed the MAP plot as a graphical assessment tool for visualizing posterior distributions for partial regression coefficients from individual models along with their model-averaged counterparts to facilitate easy comparisons and allow the user to see how results from individual models combine to form the model-averaged posterior distribution. We demonstrated the use of these graphical tools in the context of multiple examples and showed how they can help explain or justify why model averaging may be an

added complexity that is at best unnecessary, and at worst incorrect for a particular question of interest.

We discussed alternatives to model averaging such as using question-focused modeling to build statistical models for addressing very specific questions of interest, and then iterating on, or expanding, that model (e.g., as described in Gelman et al., 2013; Ver Hoef and Boveng, 2015). We provided references to other work (e.g., Crainiceanu et al., 2008; Wilson and Reich, 2014) that presents methodology for identifying subsets of models addressing a particular question of interest where the meanings of the partial regression coefficients of interest are similar enough across models to be reasonably averaged. We weighed trade-offs between the complexity of inferences drawn from model averaging and the perceived gain of accounting for another source of uncertainty (model uncertainty) in an analysis.

Cade (2015) has similar concerns as we do about using model averaging with partial regression coefficients and proposes a potential “fix” to standardize regression coefficients twice, once before model averaging, and also after the model averaging of partial regression coefficients. He says this provides well-defined units for the model-averaged posterior. We considered this standardization method as a potential option for the MAP plot and ultimately decided against it because we are not convinced such standardization is appropriate. We plan to conduct a deeper investigation of Cade (2015)’s methods in future work.

Another large piece of this work focused on investigating whether multimodel inferential techniques, commonly used to ensure variability due to model selection is properly accounted for in an analysis (e.g., model averaging), actually result in this claimed benefit. We compared three commonly used approaches for multimodel inference: (1) model averaging using all models in the model set, (2) model averaging using only the subset of models including the variable of interest, and (3) conditioning

inferences on one single model. Our goal in making these comparisons was to better understand how accounting for more models in an analysis directly translates to increases or decreases in the posterior variance of partial regression coefficients.

We designed and executed a simulation study comparing posterior variances from the three methods outlined above for three factors related to the posterior variance of a partial regression coefficient; specifically, correlation among explanatory variables, the signal to noise ratio, and the combination of the data generating model and the model set used. The motivation behind conducting multimodel inference is described by stating that inferences drawn from methods accounting for fewer models will fail to incorporate model uncertainty (something that could have been accounted for in the analysis), and result in understated variances for quantities of interest. We investigated this claim using a simulation study and found, as the correlation increases among explanatory variables (and one of them is in the data generating model), the method using more models does tend to have larger posterior variance than the method using fewer models. However, the two methods using more models also used model averaging, and when there is large correlation among explanatory variables, obtaining model-averaged partial regression coefficients results in confusing inferences (discussed in Chapter 2). We also found cases where the method using fewer models resulted in larger posterior variance than the method using more models, so it is not fair to claim that conditioning inferences on a single model always results in naïve (liberal) inferences.

We also looked at the magnitude of the difference in posterior standard deviations between model averaging using all models and the method using one single model (top posterior model probability model). In the most extreme cases (largest ratios comparing the variances of the two methods), we found the difference in posterior standard deviations was 0.20 standard deviation units, implying a maximum

increase in width for a posterior interval of around 0.40 standard deviation units. These results were context dependent (i.e., for relatively simple scenarios in a linear regression setting), but there is a push for researchers to use these types of methods to “account for model uncertainty” in their analyses, when the actual difference in final inference may not be large enough to be practically meaningful.

In the simulation study, we restricted our context to multiple linear regression with independent normal errors. We expect our findings to extend to generalized linear models, but we did not formally investigate this claim. Currently, it is still too computationally expensive to conduct a simulation study like the one we executed using a fully Bayesian approach. However, model weights based on information criteria constitute easily obtainable approximations to posterior model probabilities in the generalized linear model context, and could potentially be used to conduct a similar study to the one we conducted, as long as one considers the prior on the model set induced by the choice of criterion.

Using one of the data generating models used in the simulation study (Equation 3.6), low signal to noise ratio, and an *all subsets* model set that included the data generating model (see Chapter 3, Section 3.3.3), we compared posterior standard deviations for partial regression coefficients associated with each explanatory variable to those obtained from using AIC weights for all correlations between  $X_3$  and  $X_5$ . We did this for 50 realizations of data for each correlation considered (Figure 5.1). It is interesting to note that the standard deviations for the variables *not* in the data generating model ( $X_1$ ,  $X_2$ , and  $X_3$ ) tend to be larger for the AIC method than they do for the fully Bayesian implementation. This preliminary comparison suggests information criteria-based approximations to posterior model probabilities and their analogous formulas for obtaining point estimates and standard errors for partial regression coefficients could be used to quickly conduct simulation studies like

the one we did, but instead, for generalized linear models. This is an avenue we will continue to explore.

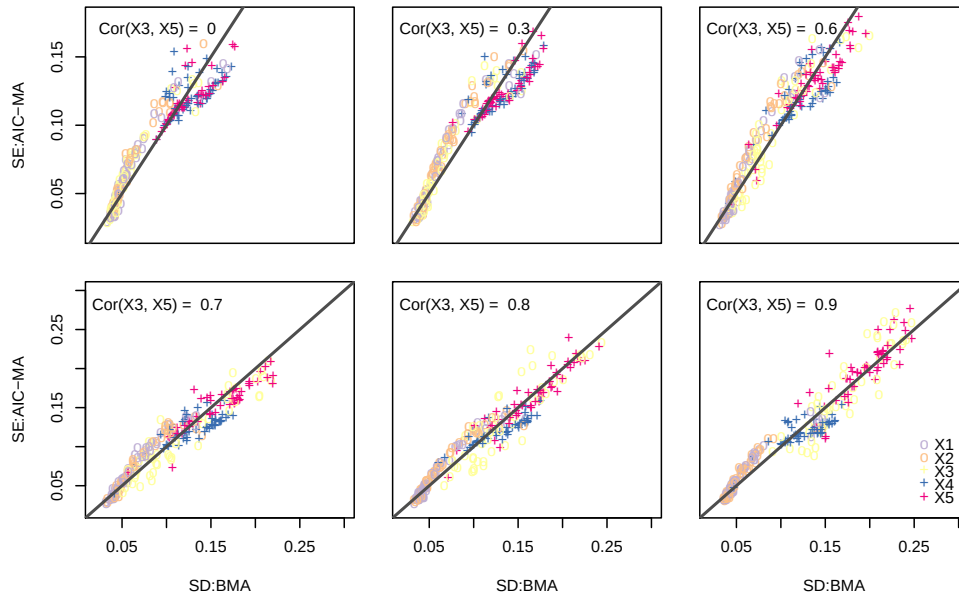


Figure 5.1: This graph displays a comparison of posterior standard deviations for partial regression coefficients obtained via model averaging using the  $g$ -prior to the standard errors of the same partial regression coefficients obtained using the AIC approximation.

Motivated by the assumption that inferences accounting for model uncertainty will always be less precise than inferences conditional on one model, Burnham and Anderson (2002) advocate the use of model averaging and use frequentist sampling theory to attempt to derive an *unconditional* (i.e., *not* conditional on a single model) sampling variance for use with estimates of regression coefficients (see Burnham and Anderson (2002), Section 4.3.2, pgs. 158-162). They do this by incorporating an additional variance component in the variance estimator to capture variability in the estimated regression coefficients among the models considered. However, they do not end up with a result they are comfortable recommending for widespread use. These authors conclude by suggesting researchers use the estimated sampling variance for

the model-averaged partial regression coefficient proposed in Buckland et al. (1997) as the unconditional sampling variance for a partial regression coefficient of interest. Their reasoning for this suggestion has a theoretical basis, but some of the steps were heuristic and they suggest more work be put into finding an *unconditional* variance estimator. We did not investigate or suggest adjustments to variance estimates conditional on one model, but we found model averaging can result in a smaller variance than alternative methods considering fewer models. Therefore, we agree with Burnham and Anderson (2002) that their *unconditional* sampling variance estimator deserves more attention, and it may not be the case that the variance should always increase when model averaging is conducted, especially when the quantity of interest is a partial regression coefficient.

Although we restricted the context of Banner and Higgs (2016) and our simulation study to focus on linear regression models, we created an R (R Core Team, 2016) package (**MApp**) for making the MAP plot providing functions for a variety of implementations of model averaging. We created a function to work directly with objects obtained from a flexible package designed for implementing model averaging according to many variations of the *g*-prior (BMS R package Feldkircher and Zeugner, 2009; Zellner, 1984, see Appendix A for details about the *g*-prior). We also created plotting functions to work with model-averaged results obtained from self-programmed reversible jump Markov chain Monte Carlo samplers (RJMCMC) and AIC and BIC approximations, which are methods that can accommodate generalized linear models. We created a detailed vignette for this package describing how to use all of the functions, and providing tips for using the MAP plot as an assessment tool for researchers considering model averaging for partial regression coefficients (available on RPubS: <http://www.rpubs.com/kbanner14/203078>). Our code is available to anyone on GitHub (<https://github.com/kbanner14/MApp-Rpackage>), and we plan

to publish the package to the Comprehensive R Archive Network (CRAN). We also plan to extend the package to work more directly with some of the other packages designed for conducting model averaging such as *Bayesian Model Averaging* BMA (Raftery et al., 2014) and *Bayesian Adaptive Sampling* BAS (Clyde, 2012).

Statisticians are in a position of great responsibility when it comes to communicating how to appropriately use methodology. In these exciting times where computational roadblocks are becoming rarer, and complex methodology is becoming more accessible to a wide range of researchers, we need to be sure to provide researchers who may use these complex methods such as model averaging with appropriate assessment tools and information regarding appropriate use. We hope to provide researchers who may feel pressured to use more complicated and sophisticated analyses, even when they might not be needed, with tools and arguments to defend their more parsimonious choices. In this work we discussed issues with a new use of model averaging we have seen gaining traction as a way to “address model uncertainty” and make inferences about the “overall effects” of “important” variables. We started to address all of the phrases in quotations in this work and we provide ideas for future work, all in the hopes of encouraging careful consideration before the use of this popular method.

## Bibliography

- Albert, J. (2014). *LearnBayes: Functions for Learning Bayesian Inference*. R package version 2.15. <http://CRAN.R-project.org/package=LearnBayes>.
- Amini, S. M. and Parmeter, C. F. (2011). Bayesian model averaging in r. *Journal of economic and social measurement*.
- Bache, S. M. and Wickham, H. (2014). *magrittr: A Forward-Pipe Operator for R*. R package version 1.5. <https://CRAN.R-project.org/package=magrittr>.
- Banner, K. M. and Higgs, M. D. (2016). Considerations for assessing model averaging of regression coefficients. *Ecological Applications*.
- Barker, R. J. and Link, W. A. (2013). Bayesian multimodel inference by rjmc: A gibbs sampling approach. *The American Statistician*, 67(3):150–156.
- Barker, R. J. and Link, W. A. (2015). Truth, models, model sets, aic, and multimodel inference: A bayesian perspective. *The Journal of Wildlife Management*, 79(5):730–738.
- Barnard, G. A. (1963). New methods of quality control. *Journal of the Royal Statistical Society Series A*, 126:255.
- Brewer, M. (2015). Ten top tips for reviewing statistics: A guide for ecologists. [https://methodsblog.wordpress.com/2015/06/03/reviewing\\_statistics/](https://methodsblog.wordpress.com/2015/06/03/reviewing_statistics/).
- Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997). Model selection: An integral part of inference. *Biometrics*, 53(2):603–618.
- Burnham, K. P. (2015). Multimodel inference: Understanding aic relative variable importance values. <http://warnercnr.colostate.edu/~kenb/pdfs/KenB/AICRelativeVariableImportanceWeights-Burnham.pdf>.
- Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference A Practical Information-Theoretic Approach*. Springer, New York, NY.
- Burnham, K. P. and Anderson, D. R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, 33(2):261–304.
- Cade, B. S. (2015). Model averaging and muddled multimodel inference. *Ecology*, 96(6):2370–2382.
- Carlin, B. P. and Chib, S. (1995). Bayesian model choice via markov chain monte carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(3):473–484.

- Chipman, H. (1996). Bayesian variable selection with related predictors\*. *The Canadian Journal of Statistics*, 24(1):17–36.
- Chipman, H., George, E. I., and McCulloch, R. E. (2001). The practical implementation of bayesian model selection. *IMS Lecture Notes - Monograph Series*, 38:67–134.
- Clyde, M. (2012). *BAS: Bayesian Adaptive Sampling for Bayesian Model Averaging*. R package version 1.0.
- Clyde, M., Desimone, H., and Parmigiani, G. (1996). Prediction via orthoganalized model mixing. *Journal of the American Statistical Association*, 91:1197–1208.
- Clyde, M. and George, E. I. (2004). Model uncertainty. *Statistical Science*, 19(1):81–94.
- Clyde, M. A. (1999). Bayesian model averaging and model search strategies. *Bayesian Statistics*, 6:157–185.
- Clyde, M. A., Ghosh, J., and Littman, M. L. (2011). Bayesian adaptive sampling for variable selection and model averaging. *Journal of Computational and Graphical Statistics*, 20(1):80–101.
- Crainiceanu, C. M., Dominici, F., and Parmigiani, G. (2008). Adjustment uncertainty in effect estimation. *Biometrika*, 95(3):635–651.
- Doherty, P. F., White, G. C., and Burnham, K. P. (2012). Comparison of model building and selection strategies. *Journal of Ornithology*, 125(Suppl 2):S317–S323.
- Feldkircher, M. and Zeugner, S. (2009). Benchmark priors revisited: On adaptive shrinkage and the supermodel effect in bayesian model averaging. *International Monetary Fund (IMF) Working Paper*, pages 1–39. Written within the IMF in the Finance Department.
- Fieberg, J. and Johnson, D. H. (2015). Mmi: Multimodel inference or models with management implications. *The Journal of Wildlife Management*, 79(5):708–718.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman and Hall CRC Press, Boca Raton, FL, third edition.
- George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.
- George, E. I. and McCulloch, R. E. (1997). Approaches for bayesian variable selection. *Statistica Sinica*, 7:339–373.

- Geweke, J. (1996). Variable selection and model comparison in regression. *Bayesian Statistics*, 5:609–620.
- Green, P. J. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732.
- Hocking, R. R. (2003). *Methods and Applications of Linear Models*. John Wiley & Sons, Hoboken, NJ.
- Hodges, J. S. (1987). Uncertainty, policy analysis and statistics. *Statistical Science*, 2(3):259–291.
- Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999a). Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–401.
- Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999b). Rejoinder: Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):412–417.
- Hooten, M. B. and Hobbs, N. T. (2015). A guide to bayesian model selection for ecologists. *Ecological Monographs*, 85(1):3–28.
- Jewell, N. B. (2004). *Texts in Statistical Science: Statistics for Epidemiology*. Chapman & Hall CRC Press, Boca Raton, FL.
- Johnson, R. W. (1996). Fitting percentage of body fat to simple body measurements. *Journal of Statistics Education*, 4(1). <http://www.amstat.org/publications/jse/v4n1/datasets.johnson.html>.
- Kampstra, P. (2008). Beanplot: A boxplot alternative for visual comparison of distributions. *Journal of Statistical Software, Code Snippets*, 28(1):1–9.
- Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.
- Kass, R. E. and Wasserman, L. (1995). A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of the American Statistical Association*, 90(431):928–934.
- Kuo, L. and Mallick, B. (1998). Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, 60(1):65–81.
- Leamer, E. E. (1978). *Specification Searches*. Wiley, New York.
- Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). Mixtures of g priors for bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423.

- Link, W. A. and Barker, R. J. (2006). Model weights and the foundations of multimodel inference. *Ecology*, 87(10):2626–35.
- Link, W. A. and Barker, R. J. (2010). *Bayesian Inference With Ecological Applications*. Elsevier, London, NW1 7BY, UK.
- Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). Winbugs – a bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10:325–337.
- Montgomery, J. M. and Nyhan, B. (2010). Bayesian model averaging: Theoretical developments and practical applications. *Political Analysis*, 18(2):245–270.
- O’Hara, R. B. and Sillanpää, M. J. (2009). A review of bayesian variable selection methods: What, how and which. *Bayesian Analysis*, 4(1):85–118.
- Penrose, K., Nelson, A., and Fisher, A. (1985). Generalized body composition prediction equation for men using simple measurement techniques (abstract). *Medicine and Science in Sports Exercise*, 17(2):189.
- Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). Coda: Convergence diagnosis and output analysis for mcmc. *R News*, 6(1):7–11.
- R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Raftery, A., Hoeting, J., Volinsky, C., Painter, I., and Yeung, K. Y. (2014). *Bayesian Model Averaging*. R package version 3.17.1. <http://CRAN.R-project.org/package=BMA>.
- Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997a). Bayesian Model Averaging for Linear Regression Models. *Journal of the American Statistical Association*, 92(437):179.
- Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997b). Bayesian Model Averaging for Linear Regression Models. *Journal of the American Statistical Association*, 92(437):179.
- Ramsey, F., Schafer, D., Schafer, D. W., Sifneos, J., and Turlach, B. A. (2012). *Sleuth2: Data sets from Ramsey and Shafer’s “Statistical Sleuth (2nd ed)”*. R package version 1.0-7.
- Ramsey, F. L. and Schafer, D. W. (2013). *The Statistical Sleuth: A course in Methods of Data Analysis*. Brooks/Cole, Cengage Learning, Boston, MA, third edition.
- Robert, C. and Casella, G. (2011). A short history of markov chain monte carlo: Subjective recollections from incomplete data. *Statistical Science*, 0(00):1–14.

- Sacher, G. A. and Staffeldt, E. (1974). Relation of gestation time to brain weight for placental mammals: Implications for the theory of vertebrate growth. *American Naturalist*, 108:593–613.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.
- Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3):289–310.
- Sturtz, S., Ligges, U., and Gelman, A. (2005). R2winbugs: A package for running winbugs from r. *Journal of Statistical Software*, 12(3):1–16.
- Thomas, D. C., Jerrett, M., Kuenzli, N., Louis, T. A., Dominici, F., Zeger, S., Schwarz, J., Burnett, R. T., Krewski, D., and Bates, D. (2007). Bayesian model averaging in time-series studies of air pollution and mortality. *Journal of Toxicology and Environmental Health*, 70:311–315.
- Ver Hoef, J. M. and Boveng, P. L. (2015). Iterating on a single model is a viable alternative to multimodel inference. *Journal of Wildlife Management*, 79:719–729.
- Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. Springer, New York, NY. <http://had.co.nz/ggplot2/book>.
- Wickham, H. and Chang, W. (2016). *devtools: Tools to Make Developing R Packages Easier*. R package version 1.11.1.
- Wickham, H. and Francois, R. (2015). *dplyr: A Grammar of Data Manipulation*. R package version 0.4.3. <https://CRAN.R-project.org/package=dplyr>.
- Wilson, A. and Reich, B. J. (2014). Confounder selection via penalized credible regions. *Biometrics*, 70:852–861.
- Zellner, A. (1984). *Posterior odds ratios for regression hypotheses: General considerations and some specific results*. Univ. Chicago Press, Chicago, IL.
- Zeugner, S. (2011). Replicating the body fat example from “bayesian model averaging: A tutorial” (1999) with bms in r. <https://modelaveraging.wordpress.com/2011/02/07/replicating-the-body-fat-example-from-bayesian-model-averaging-a-tutorial-1999-with-bms-in-r/>.

## APPENDICES

APPENDIX A

PRIOR SPECIFICATIONS CONTINUED

Here we provide further details regarding prior specification when choosing to implement Bayesian multimodel inference. Difficulties arise in specifying prior distributions for  $\mathcal{M}$  and  $\boldsymbol{\theta}_j|M_j$  (especially when  $J = 2^p$  is large). Depending on prior specifications, computing the integral to obtain the marginal of  $\mathbf{y}$  given  $M_j$  in Equation 2 can be intractable, making MCMC or other sampling techniques necessary. Thus, a large portion of research concerning Bayesian multimodel inference in the last 20 years has focused on different prior specifications and implementation of MCMC and other sampling schemes such as importance sampling (e.g., Carlin and Chib (1995); Chipman et al. (2001); Clyde (2012, 1999); Feldkircher and Zeugner (2009); George and McCulloch (1993, 1997); Geweke (1996); Kuo and Mallick (1998); Raftery et al. (1997a))

We do not mean to re-describe or critique the different priors available. We are interested in investigating how these methods are, or may be, used by practitioners who will likely choose the default or easier-to-use priors. Here, we describe the priors we chose to use in our examples in more detail. The discrete uniform prior for the model space assumes that all models are independent and assigns equal prior probability to each of the models considered ( $Pr(M_j) = 1/J$  for  $j = 1, 2, \dots, J$ ). The independence assumption can be naive, but despite its naivety, it is commonly used in practice (see Chipman (1996); Chipman et al. (2001) for more information and George and McCulloch (1993); Hoeting et al. (1999a) for a general description of all “independence priors”).

Zellner’s  $g$ -prior (Feldkircher and Zeugner, 2009; Zellner, 1984) is commonly used to automatically assign priors to the parameters in each model (the variance and the intercept are excluded in the assignment and receive their own common reference priors (A.2)-(A.3)). Among the packages available in R, variations of the  $g$ -prior are a popular choice (e.g., BMS (Feldkircher and Zeugner, 2009), BMA (Raftery et al.,

2014) and BAS (Clyde, 2012)). We refer the reader to Amini and Parmeter (2011) for a summary and comparison of the three packages (Amini and Parmeter (2011) make a mistake in their comparison by reporting that BMA is only capable of using AIC and BIC approximations, which is not true).

The  $g$ -prior is

$$p(\boldsymbol{\theta}_j|g, M_j) \sim \text{MVN}\left(0, g\sigma^2(\mathbf{X}'_j\mathbf{X}_j)^{-1}\right) \quad (\text{A.1})$$

$$p(\beta_0) \propto 1 \quad (\text{A.2})$$

$$p(\sigma) \propto \frac{1}{\sigma}. \quad (\text{A.3})$$

The only hyper-parameter requiring specification is  $g$ . In the prior,  $g$  is multiplied by the inverse of the Fisher information matrix in a given model  $(\mathbf{X}'_j\mathbf{X}_j)^{-1}$ , which is related to the variance of the error in the regression  $(\sigma^2(\mathbf{X}'_j\mathbf{X}_j)^{-1})$ . Larger values of  $g$  reflect more uncertainty in how many variables should be included. When  $g$  is set equal to the total sample size, the prior variance is set by the amount of observed Fisher information contained in one observation  $(n\sigma^2(\mathbf{X}'_j\mathbf{X}_j)^{-1})$  and this formulation is called the unit information prior. This approach was recommended by Kass and Wasserman (1995), and more details can be found there and in Feldkircher and Zeugner (2009); Liang et al. (2008).

We can gain insight into the  $g$ -prior by looking at the resulting closed form posterior distributions of the partial regression coefficients. As noted in Section 2.1, the posterior distributions of the partial regression coefficients coming from model  $M_j$  follow multivariate  $t$ -distributions  $MV_t(\boldsymbol{\theta}_j|g, M_j, \mathbf{y}, \mathbf{x})$  with  $n - p_j - 1$  degrees of freedom,

$$E[\boldsymbol{\theta}_j|g, M_j, \mathbf{y}, \mathbf{x}] = \frac{g}{1+g} \hat{\boldsymbol{\theta}}_j, \text{ and} \quad (\text{A.4})$$

$$Cov[\boldsymbol{\theta}_j|g, M_j, \mathbf{y}, \mathbf{x}] = \frac{(\mathbf{y} - \bar{y})^T (\mathbf{y} - \bar{y})}{N - 3} \left( \frac{g}{1+g} \right) \left( 1 - \frac{g}{1+g} R_j^2 \right) (X_j^T X_j)^{-1}. \quad (\text{A.5})$$

In Equation A.4,  $\hat{\boldsymbol{\theta}}_j$  is the vector of maximum likelihood estimates of the partial regression coefficients, and  $R_j^2$  is the coefficient of determination from the ordinary least squares fit of  $M_j$ . Therefore, the influence of the hyper-parameter,  $g$ , comes into the posterior as a shrinkage estimator  $\left(\frac{g}{1+g}\right)$ , which approaches one as  $g \rightarrow \infty$ .

Exploiting the closed form posterior distributions using the  $g$ -prior allows enumeration of analytical posterior summaries for up to  $p = 30$  (depending on the software package used). When  $p$  is large, most R (R Core Team, 2016) software packages employ MCMC sampling to search through the model space until convergence is achieved. Typically, this will result in a set containing far fewer than  $2^p$  elements, and that set will capture all but a negligible amount of the posterior model mass. For the smaller set, results from the MCMC approximation can be compared to the analytical results posterior distributions (A.4, A.5) to ensure convergence.

APPENDIX B

CONSIDERATIONS FOR ASSESSING MODEL AVERAGING OF REGRESSION  
COEFFICIENTS SUPPLEMENTARY MATERIALS

## B.1 Setup

This document provides instructions for reproducing all analyses in *Model Averaging of Regression Coefficients: Considerations for Assessing Model Averaging Regression Coefficients*. Some of the necessary functions and data files can be loaded by loading the **MApp** (Model Averaged Posteriors Plot) package from GitHub. To use **MApp** follow these instructions:

1. Install the package dependencies: **LearnBayes**, **beanplot**, **dplyr**, and **BMS**.  
Run `install.packages(c("LearnBayes", "beanplot", "dplyr", "BMS"))` in your R console to install these packages.
2. Install **devtools** and **roxygen2**: run  
`install.packages(c("devtools", "roxygen2"))` in your R console.
3. Download the ZIP file of this repository: <https://github.com/kbanner14/MApp-Rpackage> (click Download ZIP in upper right hand corner).
4. Set your working directory to the location of the **MApp-Rpackage-master** Directory on your computer. The path in quotes ("PATH/TO") should point to where **MApp-Rpackage-master** is saved.
5. Load packages

```
# load plotting package and BMS
# if you don't have ggplot2 installed run the commented line below
# install.packages("ggplot2")
library(ggplot2)
library(BMS)

# load MApp - includes data for brain weight example and example 3
devtools::load_all("PATH/TO/MApp-Rpackage-master/MApp/.")
```

6. Load Weddell seal data using `read.csv()`. Note that the path in quotes ("PATH/TO") should point to where `seal-count-data.csv` is saved. *The Weddell*

seal field research was supported by the National Science Foundation OPP-0635739 grant to R.A. Garrott, J.J. Rotella and D.B. Siniff. All of the camera images were obtained under authority of permit NMFS Permit No.1032-1917-02. If you are uncomfortable setting the path to the .csv file, use option 2 in the code chunk.

### B.1.1 Create Fig. 2

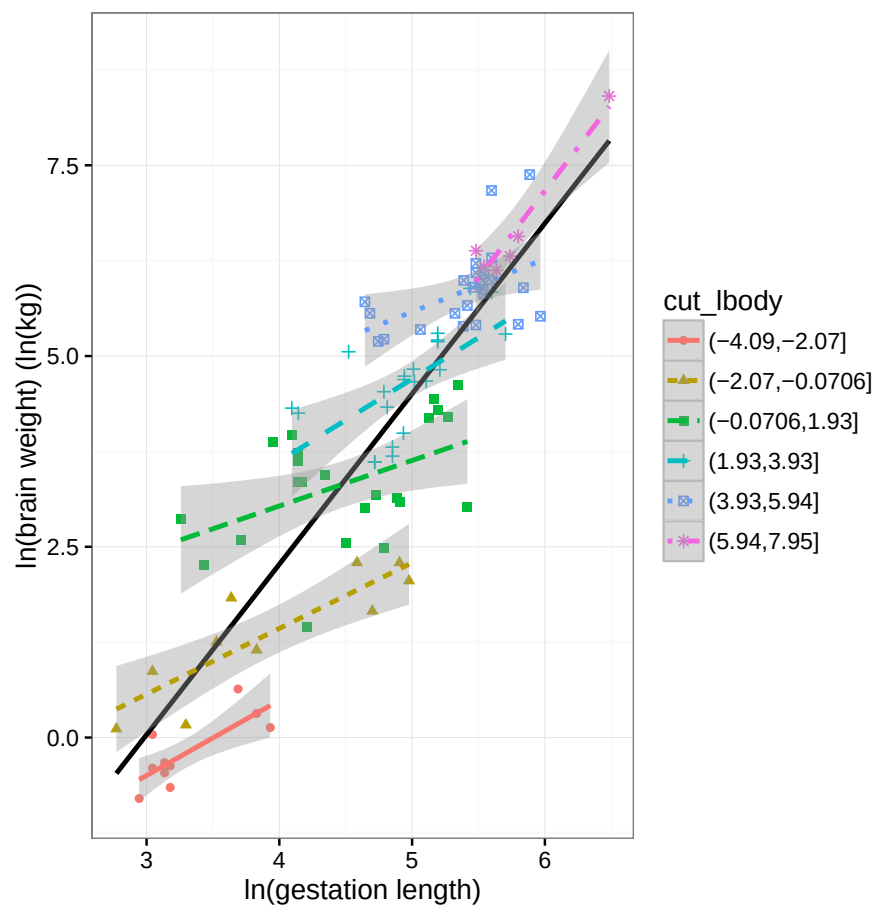
```
# look at the brain weight data
head(brainData)

##      lbrain      lbody    lgest litter
## 1 2.8622009  1.25276297 3.258097    1.0
## 2 1.2527630 -0.07257069 3.526361    4.6
## 3 1.1474025 -1.89711998 3.828641    3.0
## 4 0.1310283 -3.01593498 3.931826    1.5
## 5 0.3148107 -2.74887220 3.828641    1.5
## 6 3.0910425  0.74193734 4.905275    1.0

# create groups for body size
cut_lbody <- cut(brainData$lbody, breaks = 6)

# initialize plot using ggplot()
ggplot(data = brainData, aes(y = lbrain, x = lgest)) +
  # add points by fixed body size groups
  geom_point(aes(colour = cut_lbody, shape = cut_lbody)) +
  # add regression line for overall (marginal) relationship
  geom_smooth(method = "lm", colour = 1, se = FALSE) +
  # add regression line for relationship conditional
  # on fixed body size groups
  geom_smooth(method = "lm", aes(group=cut_lbody,
                                colour = cut_lbody,
                                lty = cut_lbody)) +

  #add labels
  xlab("ln(gestation length)") +
  ylab("ln(brain weight) (ln(kg))") +
  #make background white
  theme_bw()
```



## B.2 Create Fig. 3b

The code below will create a pairs plot, coding by different tide groups to visualize the seal data.

```
# create the tide groups
seal_sub$cut_tide2 <- cut(seal_sub$tide, breaks = 4)

# create printer friendly color scheme for the variable cut_tide2
seal_sub$colors_sub <- with(seal_sub,
  ifelse(cut_tide2 == "(-1.75,-0.556]",
    "#66c2a5",
    ifelse(cut_tide2 == "(-0.556,0.63]",
      "#fc8d62",
      ifelse(cut_tide2 == "(0.63,1.82]",
        "#8da0cb", "#e78ac3"))))

# create the scatterplot matrix
pairs(seal_sub[,c(1,6,3,4,5)], lower.panel = NULL, col = seal_sub$colors_sub)
```

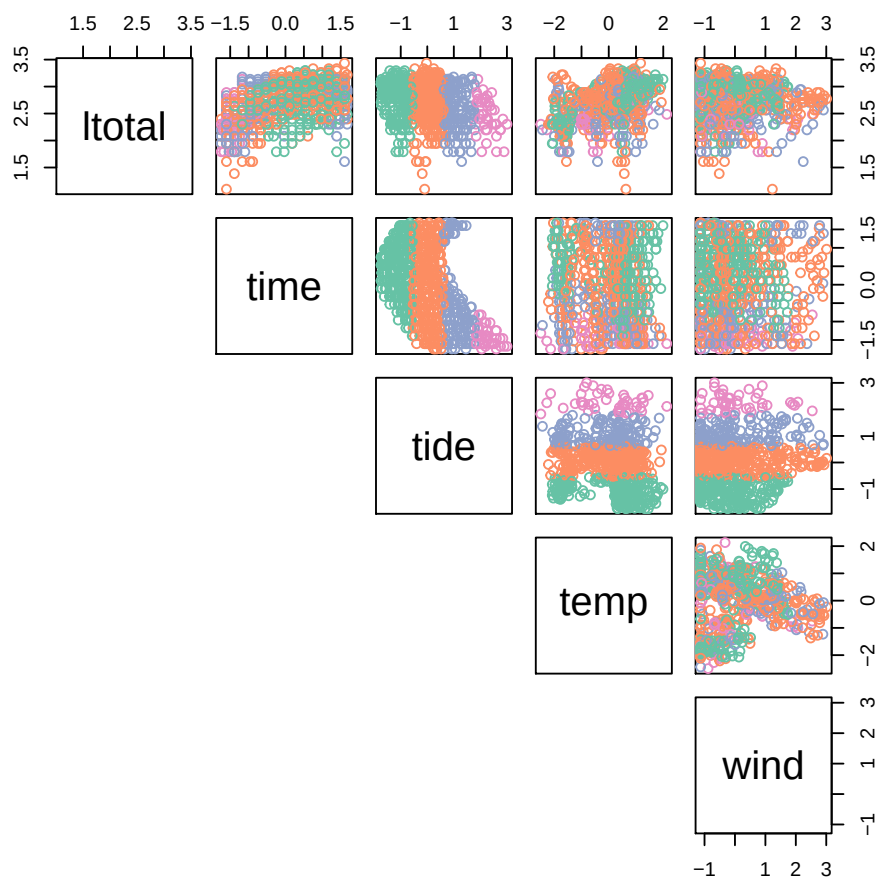
## B.3 Recreate Example 1

The seal data were standardized (centered and scaled). The means and sds used in the standardization are shown for each variable in the table below.

|      | tide  | temp   | wind  | time  | time2  |
|------|-------|--------|-------|-------|--------|
| mean | 0.360 | -9.745 | 6.458 | 6.120 | 49.758 |
| sd   | 0.207 | 4.996  | 5.725 | 3.510 | 44.047 |

To recreate Example 1, use the following code chunks. Note that the `MApp_bms` function in our package provides some extra annotations, namely the one showing the amount of the posterior model mass used in the model averaging (upper left corner). Note also that  $w_j$  is shorthand for the posterior model probability from model  $M_j$ .

```
# obtain the bma object for these data, considering all subsets of
# tide, temp, and wind after forcing time of day and
# time of day squared into the model.
sub_bms <- bms(seal_sub[,c(1,3:7)], mprior = "uniform",
  g = "UIP", user.int = F,
```

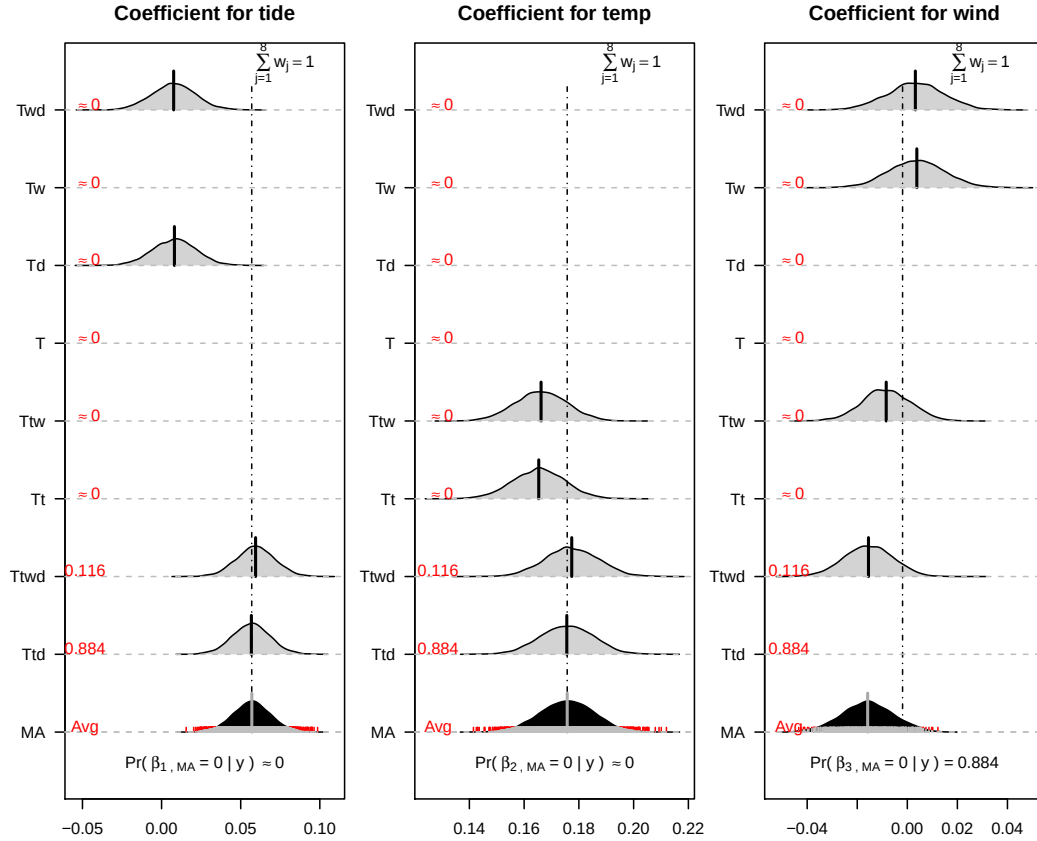


```

fixed.reg = c("time", "time2"))

# create the MApp plot to visualize the results from bms (Fig. 4)
mapp_seal <- MApp_bms(sub_bms, plot_wind = c(1,3), num_sims = 5000,
  mod_names = c("Ttd", "Ttwd", "Tt", "Ttw",
    "T", "Td", "Tw", "Twd"))

```

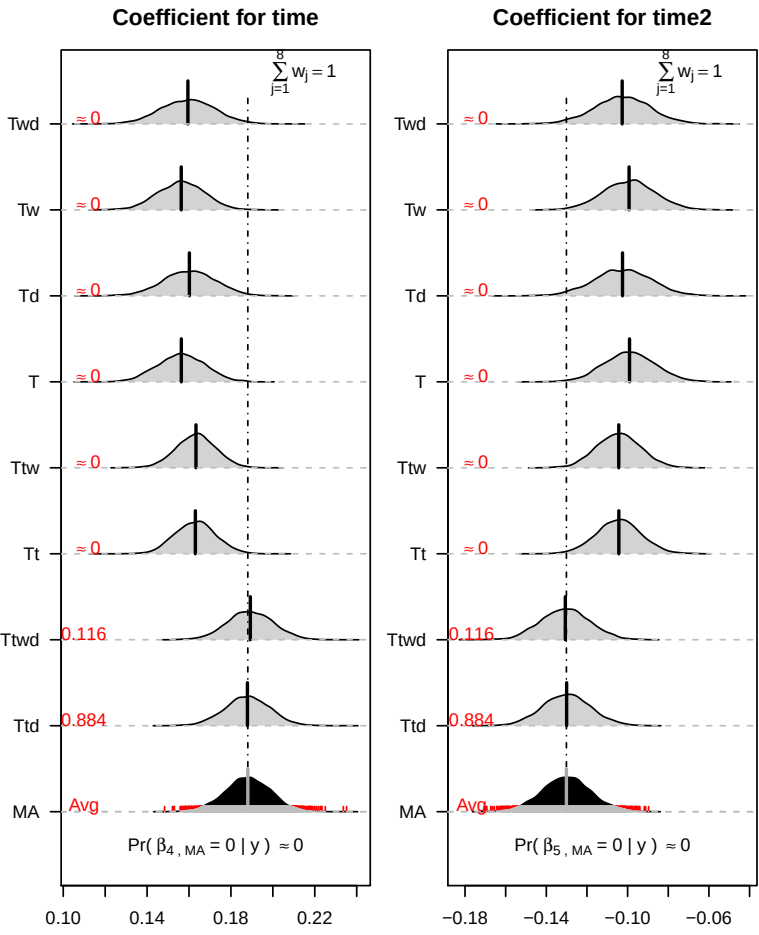


```

# typing mapp_seal will create Table 1
mapp_seal

```

|       | MA     | MA_cts | Ttd    | Ttwd   | Tt     | Ttw    | T      | Td     | Tw     | Twd    |
|-------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| tide  | 0.012  | 0.0119 | 0.0119 | 0.0122 | —      | —      | —      | 0.0138 | —      | 0.0139 |
| temp  | 0.0098 | 0.0098 | 0.0098 | 0.0098 | 0.0095 | 0.0094 | —      | —      | —      | —      |
| wind  | 0.0062 | 0.0101 | —      | 0.0097 | —      | 0.0097 | —      | —      | 0.0112 | 0.0113 |
| time  | 0.0107 | 0.0107 | 0.0108 | 0.0109 | 0.0096 | 0.0094 | 0.0112 | 0.0125 | 0.0112 | 0.0128 |
| time2 | 0.0119 | 0.0119 | 0.0118 | 0.012  | 0.0111 | 0.0108 | 0.0126 | 0.0142 | 0.0127 | 0.0139 |
| PMP   | —      | —      | 0.8837 | 0.1158 | 4e-04  | 0      | 0      | 0      | 0      | 0      |



## B.3.1 Create Fig. 3

```

# model names (for plotting)
mod_names <- c("Ttd", "Ttwd", "Tt", "Ttw",
              "T", "Td", "Tw", "Twd")

# set up objects needed for obtaining posterior draws
x <- sub_bms

# obtain 5000 draws
num_sims = 5000

# only interested in time-at-max
include_coef <- c(6)

# grab what we need from bms
results <- data.frame(coef(x))
PIP <- results[order(results$Idx), ][, 1]
Yvec <- x$X.data[,1]
Xmat <- x$X.data[, -1]

# Create input matrix to work in simulation function.
# Rows ordered w.r.t posterior model probability
post.means <- x$topmod$betas()
inmat <- post.means
inmat[inmat != 0] <- 1
inmat <- t(inmat)
weights <- sort(pmp.bma(x)[,1], decreasing = T)

g <- x$gprior.info$g

# obtain posterior draws according to the analytical posteriors
mcmc_list <- bms_post_sim(input_mat = inmat, Xmat = Xmat,
                        Yvec = Yvec, num_sims = num_sims, g)

# create the time at max variable using draws from \beta_time and \beta_time2
max_posts <- lapply(mcmc_list, function(x){
  x <- x[,c(4:5)]
  max <- -x[,1]/(2*x[,2])
  return(max)
})

# give them names
names(max_posts) <- mod_names

# append draws from time-at-max posterior to mcmc_list
for (i in 1:8) {
  mcmc_list[[i]]$Time_at_max <- max_posts[[i]]
}

#number of models

```

```

K <- length(mcmc_list)

# display all models
max_display <- K

# number of reg. parameters
p <- dim(mcmc_list[[1]])[2]

# assign model names
names(mcmc_list) <- mod_names

#grab variable names
var_names <- names(mcmc_list[[1]])

# create MA posterior for all coefficients simultaneously
# generate samps per model using multinomial distribution
samps <- rmultinom(num_sims, 1, weights)
samps_mod <- apply(samps, 1, sum)
names(samps_mod) <- mod_names

index_samp <- which(samps_mod != 0 )
MA_post <- data.frame()
for (m in index_samp) {
  mod_samp <- mod_names[m]
  draws <- apply(mcmc_list[[mod_samp]], 2, sample, size = samps_mod[m])
  MA_post <- rbind(MA_post, draws)
}

# reshape list into data frame for plotting
coef_frames <- data.frame()
for (i in 1:K) {
  Post.Vec <- mcmc_list[[i]]$Time_at_max
  Model <- rep(mod_names[i], length(Post.Vec))
  coef_frames <- rbind(coef_frames, data.frame(Post.Vec, Model))
}

# append model averaged result with coef_frames
app <- data.frame(MA_post[,6], rep("MA", num_sims))
names(app) <- c("Post.Vec", "Model")
coef_frames <- rbind(app, coef_frames)

# make MAP plot
par(mfrow = c(1,1), las = 1)

# Initilize plot
allplot <- beanplot::beanplot(Post.Vec ~ Model,
                              names = levels(coef_frames$Model),
                              data = coef_frames,
                              what = c(0, 0, 0, 0),
                              border = 1,
                              horizontal = T, bw = "nrd0", log = "",

```

```

                                cex.axis = 1)

# Add horizontal lines
modLines <- seq(1:(max_display + 1))
abline(h = c(1, modLines), lty = 2, col = "gray")

# Add top model beans
beanplot::beanplot(Post.Vec ~ Model, side = "second",
  show.names = F,
  data = coef_frames,
  add = T,
  subset = Model != "MA",
  what = c(0, 1, 1, 0),
  border = 1,
  ll = 0.12,
  horizontal = T,
  col = c("lightgray"), bw = "nrd0", log = "")

# Add MA posterior mean line
MA.mean <- mean(subset(coef_frames, Model == "MA")$Post.Vec, na.rm = T)
segments(MA.mean, 1, MA.mean, (max_display + 1.35), col = "black", lty = 4)

# Add MA bean
beanplot::beanplot(Post.Vec ~ Model, side = "second",
  show.names = F,
  data = coef_frames,
  add = T,
  subset = Model == "MA",
  what = c(0, 1, 1, 1),
  border = 1,
  ll = 0.12,
  horizontal = T,
  col = c("black", "gray", "red", "darkgrey"),
  bw = "nrd0", log = "")

# Add model weights. display 0.000 for ~0's to convey rounding
disp.weights <- round(weights[1:max_display], 3)
disp.weights <- ifelse(disp.weights == 0, "< 0.000",
  as.character(disp.weights))

MaxMin <- numeric(2)
MaxMin[1] <- min(coef_frames[, 1], na.rm = T)
MaxMin[2] <- max(coef_frames[, 1], na.rm = T)

# Add text to the plot
text(rep(MaxMin[2], max_display + 1),
  c(1:(max_display + 1) + 0.25),
  c("Avg", disp.weights),
  col = "black",
  cex = 1,
  pos = 2)

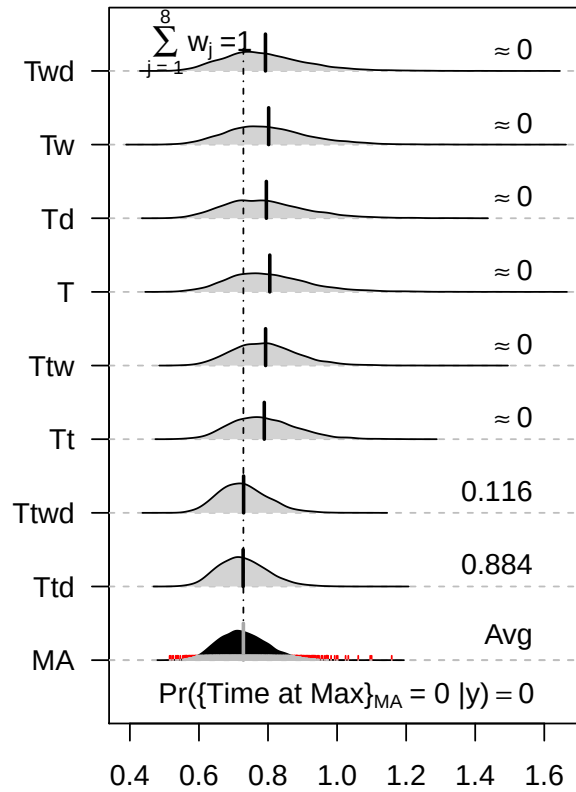
```

```

# Add zero component - in this case it has probability 0
text(mean(MaxMin),
     0.1,
     expression("Pr({Time at Max}[MA] ~ \"|y)\" == 0),
     col = "black",
     cex = 1,
     pos = 3)

text(.4, (max_display + 1.2),
     expression(sum(w[j], j = "j = 1", 8) ~ "=1"),
     col = "black",
     cex = 1,
     pos = 4)

```



### B.3.2 Posterior Credible Intervals

We can easily create posterior intervals for the time at expected max seals for the model averaged result. Note that these posterior credible intervals are based off of Monte Carlo simulations, so the results you get may differ slightly from those

presented in this document. We note that the natural log function of a positive number,  $\log(y)$  where  $y \geq 0$  is monotonic, and the value of  $x$  that maximizes  $\log(y) = ax^2 + bx + c$  is the same value that maximizes  $y = \exp(ax^2 + bx + c)$ .

```
ci_ma <- quantile(MA_post$Time_at_max, c(0.025, 0.975))

# back-transform
# x_std = (x- mu_x)/sd_x implies x_std*sd_x + mu_x
# mu_x = 6.120, sd_x = 3.510
ci_ma_backt <- ci_ma*3.51 + 6.12
# convert back to military time (time is hours from 8)
# add 8 and convert fraction of hour into minutes
mt_ma <- ci_ma_backt + 8

# split up hrs and minutes
hrs_ma <- round(mt_ma)
# convert fractions of hrs into minutes
mins_ma <- mt_ma %% 1
# create the interval
mt_ma <- paste(hrs_ma, round(mins_ma*60), sep = ":")
mt_ma

## [1] "16:14" "17:14"
```

If we condition inference on the MA model, there is a 95% chance that the true time when we can expect the maximum seals on the ice was between 16:14 hrs and 17:14 hrs. Now we can compare this to the model with all potential covariates (the largest model).

```
# posterior interval for max seals from largest model
ci_full <- quantile(mcmc_list[[2]]$Time_at_max, c(0.025, 0.975))
ci_full_backt <- ci_full*3.51 + 6.12
# convert back to military time (time is hours from 8)
# add 8 and convert fraction of hour into minutes
mt_full <- ci_full_backt + 8

# split up hrs from fractions of hrs
hrs_full <- round(mt_full)
# convert fractions of hrs into mins
mins_full <- mt_full %% 1
# create the interval
mt_full <- paste(hrs_full, round(mins_full*60), sep = ":")
mt_full

## [1] "16:15" "17:14"
```

If we condition inference on the largest model, there is a 95% chance that the true time when we can expect the maximum seals on the ice was between 16:15 hrs and 17:14 hrs. Now we can compare this to the model with all potential covariates (the largest model).

## B.4 Example 2

To create Fig. 5, we first obtain posterior weights and summaries of the posterior distributions of model averaged partial regression coefficients using the `bms()` function.

- `mprior = "uniform"` places a uniform prior on the model set
- `g = 1700` approximates the prior used in Hoeting et. al. (1999) see Zeugner (2011) for details.

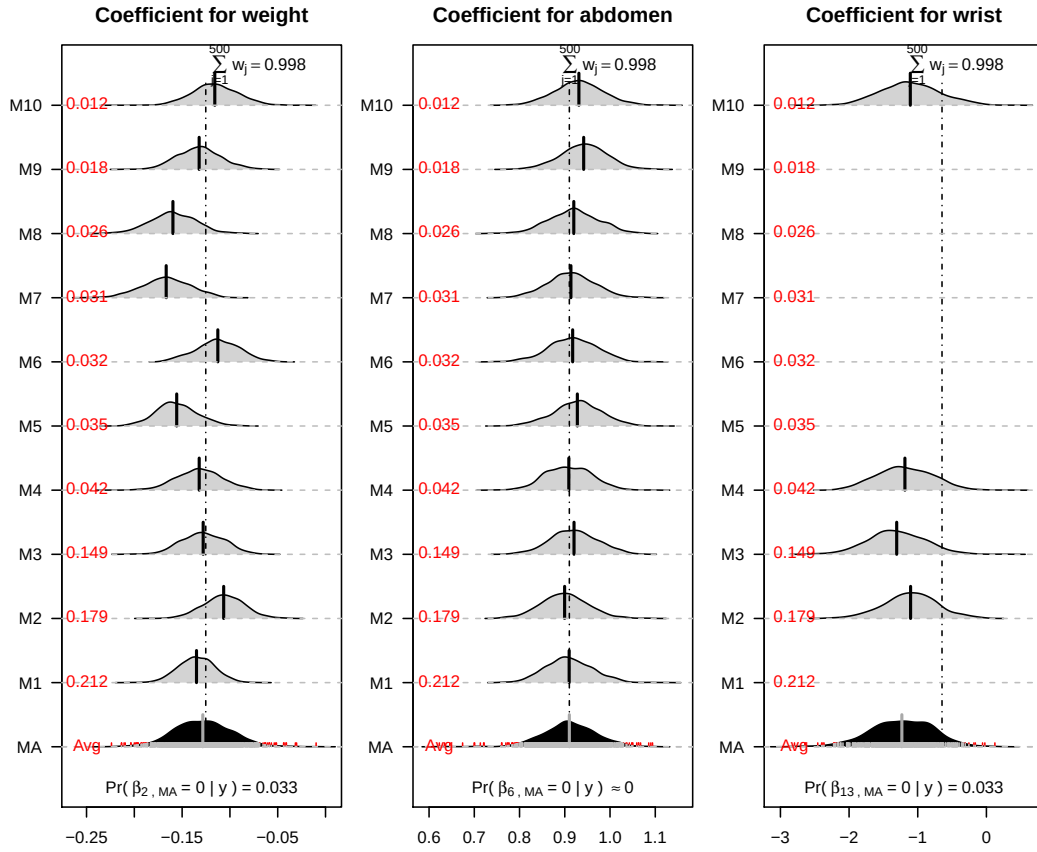
To create Fig. 5, we use the `MApp_bms()` function.

- `include_coef` specifies which coefficients to include in the display
- `max_display` tells the plotting function how many models to display on the plot
- `plot_wind` tells the function how to split the plotting window

```
# create bma object
fat_bms <- bms(bfat, mprior = "uniform",
              g = 1700, user.int = F)

# create MApp
mapp_fat <- MApp_bms(fat_bms, plot_wind = c(1,3),
                    mod_names = NULL,
                    max_display = 10,
                    include_coef = c(2,6,13),
                    num_sims = 1000)
```

```
# display comparison of cts. components of posterior distributions
# for partial regression coefficients
mapp_fat
```



|         | MA     | MA_cts | M1     | M2     | M3     | M4     | M5     |
|---------|--------|--------|--------|--------|--------|--------|--------|
| weight  | 0.0353 | 0.0297 | 0.0187 | 0.0214 | 0.0225 | 0.0239 | 0.0216 |
| abdomen | 0.0597 | 0.0597 | 0.0514 | 0.0518 | 0.0517 | 0.051  | 0.0527 |
| wrist   | 0.697  | 0.4523 | —      | 0.3895 | 0.416  | 0.4052 | —      |
| PMP     | —      | —      | 0.2125 | 0.1789 | 0.1494 | 0.0416 | 0.0349 |

|         | MA     | MA_cts | M6     | M7     | M8     | M9     | M10    |
|---------|--------|--------|--------|--------|--------|--------|--------|
| weight  | 0.0353 | 0.0297 | 0.0226 | 0.025  | 0.0241 | 0.0227 | 0.024  |
| abdomen | 0.0597 | 0.0597 | 0.0527 | 0.0517 | 0.0549 | 0.0509 | 0.0539 |
| wrist   | 0.697  | 0.4523 | —      | —      | —      | —      | 0.4327 |
| PMP     | —      | —      | 0.0318 | 0.0309 | 0.0259 | 0.0183 | 0.0121 |

APPENDIX C

SIMULATION STUDY R CODE

This appendix contains the R code used for simulating and displaying the results discussed in Chapter 3.

### C.1 Simulation Code

The code provided in this Section and Section C.2 is also available on Github: <https://github.com/kbanner14/MAsim>. The functions in the MAsim package can be loaded using `devtools::load_all("PATH/TO/MAsim/.")`. Once the package is loaded, the user can generate the simulated results using the code in Section C.3.

```
# must load MApp funtions first https://github.com/kbanner14/MApp-Rpackage/MApp/

# general function for simulation study
data_sim <- function(n = 60, p = 5, betas = c(0,0,0,1,1.2),
                    sig_y = 2.5, sig_x = rep(1, 5), cor_x = 0,
                    cor_vars = c(3,5), tol = 0.05, truth4 = F,
                    b42 = 0.5, b43 = 0.25, ...){
  # for now we are just going to assume that two of the variables are
# correlated X3 and X5, can change this with cor_vars
  try(if(length(betas) != p)
    stop("dimensions of betas and number of parameters (p) must agree"))

  # draw iid normal x's, then use cor matrix (R) and Cholesky
# factorization to get desired correlation structure in x's

  var_x <- sig_x^2
  cov <- diag(var_x)
  R <- diag(p)
  R[cor_vars[1], cor_vars[2]] <- cor_x
  R[cor_vars[2], cor_vars[1]] <- cor_x
  U <- t(chol(R))

  # simulate the x matrix
  Xmat <- LearnBayes::rmnorm(n, mean = betas, varcov = cov)
  Xmat <- t(U %*% t(Xmat))
  c <- abs(cor_x - cor(Xmat[,cor_vars[1]], Xmat[, cor_vars[2]]))

  while(c >= tol) {
    Xmat <- LearnBayes::rmnorm(n, mean = betas, varcov = cov)
    Xmat <- t(U %*% t(Xmat))
    c <- abs(cor_x - cor(Xmat[,cor_vars[1]], Xmat[, cor_vars[2]]))
  }
}
```

```

}

dimnames(Xmat)[[2]] <- paste0("X", 1:p)

# generate y
epsilon <- rnorm(n, 0, sig_y)
# check for truth 4 scenario
if(truth4 == F){
  Y <- Xmat%*%matrix(betas) + epsilon
} else {
  X4sq <- Xmat[,4]^2
  X4cu <- Xmat[,4]^3
  Xmat <- cbind(Xmat, X4sq, X4cu)
  betas <- c(betas, b42, b43)
  Y <- Xmat%*%matrix(betas) + epsilon
  Xmat <- Xmat[,1:5]
}

# make the dat frame
dat <- scale(data.frame(Y,Xmat), center = T, scale = T)
dat <- data.frame(dat)
return(dat)
}

#~~~~~#
# sim function #
#~~~~~#
sim_d2 <- function(nsim, n = 60, p = 5, betas = c(0,0,0,1,1.2),
  sig_y = 2.5, sig_x = rep(1,5), cor_x = 0,
  cor_vars = c(3,5), tol = 0.05, seed_num = NULL, case = 1,
  truth4 = F,
  ...){

  dat <- as.list(1:nsims)

  for(i in 1:nsims){
    dat[[i]] <- data_sim(n = n, p = p, betas = betas,
      sig_y = sig_y, sig_x = sig_x, cor_x = cor_x,
      cor_vars = cor_vars, tol = tol, truth4 = truth4)
  }
  if (case == 1) {
    sum_dat <- lapply(dat, FUN = function(x){
      dat_bms <- BMS::bms(x, mprior = "uniform", user.int = F, g = "UIP")
      var_top <- dat_bms$topmod$betas2()[,1] - (dat_bms$topmod$betas()[,1])^2
      vars <- var_betaMA(dat_bms)
      var_ma <- vars[,1]
      var_cts <- vars[,2]
      names(var_top) <- dimnames(vars)[[1]]
      names(var_cts) <- dimnames(vars)[[1]]
      names(var_ma) <- dimnames(vars)[[1]]
      d <- vars[,1]/vars[,2]
    })
  }
}

```

```

    names(d) <- dimnames(vars)[[1]]
    PIP <- coef(dat_bms)[order(coef(dat_bms)[,5]), ][,1]
    return(c(d, PIP, var_ma, var_cts, var_top))
  }
)
sum_dat <- do.call(rbind, sum_dat)
out <- as.list(1:2)
out[[1]] <- dat
d <- tidyr::gather(data.frame(sum_dat[,1:p]), "X1:X5")
names(d) <- c("X", "d")
d$PIP <- tidyr::gather(data.frame(sum_dat[, (p+1):(2*p)]),
  "X1:X5")[,2]
d$var_ma <- tidyr::gather(data.frame(sum_dat[, (2*p+1):(3*p)]),
  "X1:X5")[,2]
d$var_cts <- tidyr::gather(data.frame(sum_dat[, (3*p+1):(4*p)]),
  "X1:X5")[,2]
d$var_top <- tidyr::gather(data.frame(sum_dat[, (4*p+1):(5*p)]),
  "X1:X5")[,2]
d$var_top[d$var_top == 0] <- NA
d$d_topMA <- d$var_ma/d$var_top
d$d_topCts <- d$var_cts/d$var_top
d$diff_MAT <- sqrt(d$var_ma) - sqrt(d$var_top)
d$diff_ct <- sqrt(d$var_cts) - sqrt(d$var_top)

out[[2]] <- d
names(out) <- c("dat", "summary")
} else {
  sum_dat <- lapply(dat, FUN = function(x){
    dat_bms <- BMS::bms(x[,c(1,2:5)], mprior = "uniform",
      user.int = F, g = "UIP")
    var_top <- dat_bms$topmod$betas2()[,1] - (dat_bms$topmod$betas()[,1])^2
    vars <- var_betaMA(dat_bms)
    var_ma <- vars[,1]
    var_cts <- vars[,2]
    names(var_top) <- dimnames(vars)[[1]]
    names(var_cts) <- dimnames(vars)[[1]]
    names(var_ma) <- dimnames(vars)[[1]]
    d <- vars[,1]/vars[,2]
    names(d) <- dimnames(vars)[[1]]
    PIP <- coef(dat_bms)[order(coef(dat_bms)[,5]), ][,1]
    return(c(d, PIP, var_ma, var_cts, var_top))
  })
sum_dat <- do.call(rbind, sum_dat)
out <- as.list(1:2)
out[[1]] <- dat
p <- p-1
d <- tidyr::gather(data.frame(sum_dat[,1:p]), "X1:X4")
names(d) <- c("X", "d")
d$PIP <- tidyr::gather(data.frame(sum_dat[, (p+1):(2*p)]),
  "X1:X4")[,2]

```

```

d$var_ma <- tidyr::gather(data.frame(sum_dat[, (2*p+1):(3*p)]),
                          "X1:X4")[,2]
d$var_cts <- tidyr::gather(data.frame(sum_dat[, (3*p+1):(4*p)]),
                          "X1:X4")[,2]
d$var_top <- tidyr::gather(data.frame(sum_dat[, (4*p+1):(5*p)]),
                          "X1:X4")[,2]
d$var_top[d$var_top == 0] <- NA
d$d_topMA <- d$var_ma/d$var_top
d$d_topCts <- d$var_cts/d$var_top
d$diff_MAt <- sqrt(d$var_ma) - sqrt(d$var_top)
d$diff_ct <- sqrt(d$var_cts) - sqrt(d$var_top)
out[[2]] <- d
names(out) <- c("dat", "summary")
}
return(out)
}

var_betaMA <- function(x){
  # function to take a bma object created using bms and return the variance of
  # the posterior distribution for the MA coefficient according to the formula
  # provided in Hoeting et. al 1999

  # calculate posterior means for each individual model delta_hat_k in Hoeting
  mean_betas <- x$topmod$betas()

  # I think we want the intercept only model.
  # # forget the intercept only model
  # idx <- which(apply(mean_betas, 2, sum) != 0)
  mean_betas <- t(mean_betas)

  inmat <- mean_betas
  inmat[which(inmat != 0)] <- 1

  # calculate the variances of the coefficients in each individual model
  # (Var[delta/D,M_k])
  var_betas <- x$topmod$betas2() - x$topmod$betas()^2
  var_betas <- t(var_betas)

  # calculate posterior model weights for the top models
  weights <- matrix(sort(BMS::pmp.bma(x)[,1], decreasing = T))

  # Calculate the model averaged posterior estimates squared.
  Ebeta_all2 <- (t(weights)%*%mean_betas)^2
  Ebeta_indiv2 <- mean_betas^2
  term1 <- t(weights)%*%(var_betas + Ebeta_indiv2)
  var_est <- term1 - Ebeta_all2
  dimnames(var_est)[[2]] <- dimnames(coef(x)[order(coef(x)[,"Idx"]), ])[[1]]
  dimnames(var_est)[[1]] <- "var_MA"
  var_est <- t(var_est)
  var_est <- data.frame(var_est)
  var_est$sd_MA <- sqrt(var_est$var_MA)

```

```

var_cts_norm <- numeric(dim(mean_betas)[2])
for(i in 1:length(var_cts_norm)){
  idx <- which(inmat[,i] == 1)
  # posterior means from the individual models that include variable
  mb <- mean_betas[idx,i]
  # PMP's for models that include variable
  wt <- weights[idx]
  # normalize those weights
  wt_re <- wt/sum(wt)
  # posterior variances
  vb <- var_betas[idx,i]
  # individual means squared for first term
  # (var(beta/m_k, y) + E(beta/mk, y)^2)
  Eb_ind2 <- mb^2
  #
  Eb_cts2 <- (t(wt_re)%*%mb)^2
  t1 <- t(wt_re)%*%(vb + Eb_ind2)
  var_cts_norm[i] <- t1 - Eb_cts2
}

var_est$var_cts_norm <- var_cts_norm
var_est$sd_cts_norm <- sqrt(var_cts_norm)
var_est <- var_est[,c(1,3,2,4)]

return(var_est)
}

# function to obtain the approximate MA posterior means and sd's, and also
# the number of draws taken from each model considered.

sum_stat <- function(dat, num_sims = 1000, mod_names = NULL,...){

  # function to track useful summary statistics for evaluating
  # MCMC error in our simulation process for the UIP g-prior formulation of
  # BMA. Input a data frame and return mean_MA, sd_cts, sd_MA, and total
  # samples.
  # This version uses the multinomial distribution to sample from individual
  # posteriors

  # generate BMA object and save important info
  x <- bms(dat, mprior = "uniform", g = "UIP", user.int = F)

  # create input matrix for models visited
  post_means <- x$topmod$betas()
  inmat <- post_means
  inmat[inmat != 0] <- 1

  # save data and dimensionality
  Yvec <- x$X.data[,1]

```

```

Xmat <- x$X.data[,-1]
K <- dim(inmat)[1]

# find posterior model probabilities
weights <- matrix(sort(pmp.bma(x)[,1], decreasing = T))

# generate individual and MA posterior distributions for coefficients
# Save the number of samples per model going into the MA posterior
MA_posts <- MApp::MApost_bms_sim(x, num_draws = num_sims, mod_names = mod_names)
posts <- MA_posts[[1]]
mod_samps <- MA_posts[[2]]

# track informative summary statistics
MA_Qsum <- apply(posts, 2, function(x){
  SD_MA <- sd(x)
  mean_MA <- mean(x)
  x <- x[x != 0]
  cts_SD <- sd(x)
  return(c(mean_MA, cts_SD, SD_MA))
})

# sort the info to match original sampler (not multinomial)
MA_Qsum <- data.frame(t(MA_Qsum))
names(MA_Qsum) <- c("mean_MA", "SD_cts", "SD_MA")
MA_Qsum$tot_samp <- sum(mod_samps)

# return data frame and samps per model
return(list(MA_Qsum[,c(4,1,2,3)], mod_samps))
}

```

## C.2 Plotting Functions

```

process_fun <- function(df_vec, y){
  xs <- unique(df_vec[[2]][,1])
  nsims <- dim(df_vec[[2]])[1]/length(xs)
  colors <- rep(c('#c7e9b4', '#7fcdbb',
                  '#41b6c4', '#2c7fb8',
                  '#253494')[1:length(xs)],
                each = nsims)

  idx_d <- which(attributes(df_vec)$names == "summary")
  max_y <- sapply(df_vec[idx_d], function(x){
    return(max(x[,y], na.rm = T))
  })
  max_y <- max(max_y) + sd(max_y)/2
  for(i in idx_d){

```

```

    pep <- 1 - df_vec[[i]][,"PIP"]
    df_vec[[i]]$pep <- pep
    df_vec[[i]]$colors <- colors
    df_vec[[i]]$max_y <- max_y
  }
  return(df_vec)
}

process_trt <- function (df_comp, y) {

  dat <- df_comp[[2]]
  dat2 <- df_comp[[4]]
  xs <- unique(dat$X)
  xs2 <- unique(dat2$X)
  num_x <- dim(dat)[1]/length(xs)
  num_vars <- dim(dat)[2]

  if(length(xs) < 5){
    xs <- c(xs, NA)
    m <- data.frame(rep(paste0("X",
                               which(is.na(xs2))), num_x),
                    matrix(NA, ncol = num_vars-1, nrow = num_x))
    names(m) <- names(dat2)
    dat <- rbind(dat,m)
  }
  if(length(xs2) < 5){
    xs2 <- c(xs2, NA)
    m <- data.frame(rep(paste0("X",
                               which(is.na(xs2))), num_x),
                    matrix(NA, ncol = num_vars-1, nrow = num_x))
    names(m) <- names(dat2)
    dat2 <- rbind(dat2,m)
  }
  df_comp <- list(dat, dat2)
  nsims <- dim(df_comp[[2]])[1]/length(xs)
  colors1 = rep(c("#d0d1e6", "#d0d1e6",
                  "#d0d1e6", "#a6bddb",
                  "#a6bddb"),
                each = nsims)
  colors2 = rep(c("#0570b0", "#0570b0",
                  "#0570b0", "#045a8d",
                  "#045a8d"),
                each = nsims)
  max_y <- sapply(df_comp, function(x){
    return(max(x[,y], na.rm = T))
  })

  max_y <- max(max_y) + sd(max_y)/2

  for(i in 1:2){
    pep <- 1 - df_comp[[i]][,"PIP"]
    df_comp[[i]]$pep <- pep
  }
}

```

```

        if(i == 1){
          df_comp[[i]]$colors <- colors1
        } else {
          df_comp[[i]]$colors <- colors2
        }
        df_comp[[i]]$max_y <- max_y
      }
    }
    return(df_comp)
  }
}

process_y <- function (df, y_vec) {

  dat <- df[[2]]
  xs <- unique(dat$X)
  nsims <- dim(dat)[1]/length(xs)
  colors1 = rep(c("#d0d1e6", "#d0d1e6",
                  "#d0d1e6", "#a6bddb",
                  "#a6bddb"), each = nsims)
  colors2 = rep(c("#0570b0", "#0570b0",
                  "#0570b0", "#045a8d",
                  "#045a8d"), each = nsims)
  max_y <- apply(dat[,y_vec], 2, max, na.rm = T)
  max_y <- max(max_y) + sd(max_y)/2
  for(i in 1:2){
    pep <- 1 - dat[, "PIP"]
    dat$pep <- pep
    dat$colors1 <- colors1
    dat$colors2 <- colors2
    dat$max_y <- max_y
  }
  return(dat)
}

pres_d_plot <- function(df_vec, y = "d",
  mar_list = list(c(.25,4,4,0), c(.25,2,4, 2),
                  c(.25,0,4,4), c(4,4,.25,0),
                  c(4,2,.25, 2), c(4,0,.25,4)),
  m = rbind(c(1,2,3), c(4,5,6)),
  cor_vec = c(0.0,0.3,0.6,0.7,0.8,0.9),
  n_var = 5,
  plot_order = "forward",
  lab = "top", y_max = NULL){
  df_vec <- process_fun(df_vec, y = y)
  idx_d <- which(attributes(df_vec)$names == "summary")
  ylabel <- ifelse(y == "d", latex2exp::TeX("$Var_{MA}$ / $Var_{cts}$"),
    ifelse(y == "d_topCts",
      latex2exp::TeX("$Var_{cts}$ / $Var_{top}$"),
      ifelse(y == "d_topMA",
        latex2exp::TeX("$Var_{MA}$ / $Var_{top}$"),
        ifelse(y == "diff_ct",
          latex2exp::TeX("$SD_{cts}$ - $SD_{top}$"),

```

```

                                latex2exp::TeX("$SD_{MA} - SD_{top}$"))
                                )))

layout(m)
xs <- unique(df_vec[[2]][,1])
j <- 1
# cex_x <- c(5,4.5,4,3.5,3)
alpha_x <- 255*rep(0.4,5)
for(i in idx_d){

  # par(mar = mar_list[[j]], bg = "#f0f0f0")
  par(mar = mar_list[[j]], bg = "#ffffff")
  dat <- df_vec[[i]]
  y_max <- ifelse(is.null(y_max), dat$max_y[1], y_max)
  plot(dat[,y] ~ dat$pep,
        pch = 20,
        type = "n", ylim = c(0,y_max),
        xlim = c(0,1),
        xaxt = "n",
        yaxt = "n", xlab = "", ylab = "")
  abline(h = 1, lwd = 1.5, lty = 2, col = "#4d4d4d")
  abline(v = .5, lwd = 1.5, lty = 2, col = "#4d4d4d")
  if (lab == "top") {
    text(x = 0.2, y = y_max,
         paste("Cor = ", cor_vec[j]), pos = 1, cex = 2)
  } else {
    text(x = 0.2, y = 0.2,
         paste("Cor = ", cor_vec[j]), pos = 1, cex = 2)
  }
  # add axes
  if ( j %in% c(1,4)) {
    axis(side = 2,
         labels = as.character(round(seq(0, y_max, length = 11), 2)),
         at = seq(0, y_max, length = 11), tick = T, cex.axis = 1.75)
    mtext(ylabel, 2, line = 2.5)
  }
  if (j %in% c(4,5,6)){
    axis(side = 1, at = seq(0, 1, length = 11), tick = T,
         cex.axis = 1.75)
    mtext("PEP", 1, line = 2.5)
  }
  if( plot_order == "forward") {
    for(v in 1:n_var) {
      dp <- subset(dat, dat$X == xs[v])
      p_col <- col2rgb(unique(dp$colors))
      p_col <- rgb(red = p_col[1,], green = p_col[2,],
                   blue = p_col[3,], alpha = alpha_x[v],
                   maxColorValue = 255)
      points(dp[,y] ~ dp$pep,
             col = p_col,
             pch = 20, cex = 3.5)
    }
  }
}

```

```

    } else {
    for(v in n_var:1) {
      dp <- subset(dat, dat$X == xs[v])
      p_col <- col2rgb(unique(dp$colors))
      p_col <- rgb(red = p_col[1,], green = p_col[2,],
                  blue = p_col[3,], alpha = alpha_x[v],
                  maxColorValue = 255)
      points(dp[,y] ~ dp$pep,
            col = p_col,
            pch = 20, cex = 3.5)
    }
  }
  j <- j + 1
}
leg <- c(expression(X[1]),expression(X[2]),
         expression(X[3]), expression(X[4]), expression(X[5]))
fill <- c('#c7e9b4', '#7fcdbb', '#41b6c4', '#2c7fb8', '#253494')
legend("topright", legend = leg[1:n_var], pch = rep("+", n_var),
      fill = fill[1:n_var],
      cex = 1.75, bty = "n")
}

pres_trt_comp <- function(df_comp, y, cor_x,
  mar_list = list(c(.25,4,4,0), c(.25,2,4, 2),
    c(.25,0,4,4), c(4,4,.25,0), c(4,2,.25, 2),
    c(4,0,.25,4)),
  m = rbind(c(1,2,3), c(4,5,6)),
  n_var = 5,
  p = 5, labs = NULL, y_max = NULL){

  ylabel <- ifelse(y == "d", latex2exp::TeX("$Var_{MA}$ / $Var_{cts}$"),
    ifelse(y == "d_topCts",
      latex2exp::TeX("$Var_{cts}$ / $Var_{top}$"),
      ifelse(y == "d_topMA",
        latex2exp::TeX("$Var_{MA}$ / $Var_{top}$"),
        ifelse(y == "diff_ct",
          latex2exp::TeX("$SD_{cts}$ - $SD_{top}$"),
          latex2exp::TeX("$SD_{MA}$ - $SD_{top}$"))
        )))

  layout(m)
  df_comp <- process_trt(df_comp, y = y)
  dat <- df_comp[[1]]
  dat2 <- df_comp[[2]]
  xs <- unique(df_comp[[2]][,1])
  alpha_x <- 255*rep(0.4,5)

  j <- 1
  leg <- c(expression(X[1]),expression(X[2]),

```

```

        expression(X[3]), expression(X[4]), expression(X[5]))
for(i in 1:p){
  par(mar = mar_list[[j]], bg = "#ffffff")
  pdat <- subset(dat, X == xs[i])
  pdat2 <- subset(dat2, X == xs[i])
  y_max <- ifelse(is.null(y_max), pdat$max_y[1], y_max)
  plot(pdat[,y] ~ pdat$pep,
        pch = 20,
        type = "n", ylim = c(0,y_max),
        xlim = c(0,1),
        xaxt = "n",
        yaxt = "n", xlab = "", ylab = "")
  abline(h = 1, lwd = 1.5, lty = 2, col = "#4d4d4d")
  abline(v = .5, lwd = 1.5, lty = 2, col = "#4d4d4d")
  text(x = 0.8, y = pdat$max_y[1], leg[i], pos = 1, cex = 2)
  # add axes
  if ( j %in% c(1,4)) {
    axis(side = 2,
          labels = as.character(round(seq(0, y_max, length = 11), 2)),
          at = seq(0, y_max, length = 11), tick = T, cex.axis = 1.75)
    mtext(ylabel, 2, line = 2.5)
  }
  if (j %in% c(4,5,6)){
    axis(side = 1, at = seq(0, 1, length = 11), tick = T, cex.axis = 1.75)
    mtext("PEP", 1, line = 2.5)
  }

  p_col <- col2rgb(unique(pdat$colors))
  p_col <- rgb(red = p_col[1,], green = p_col[2,],
               blue = p_col[3,], alpha = alpha_x[i],
               maxColorValue = 255)
  p_col2 <- col2rgb(unique(pdat2$colors))
  p_col2 <- rgb(red = p_col2[1,], green = p_col2[2,],
               blue = p_col2[3,], alpha = alpha_x[i],
               maxColorValue = 255)
  points(pdat[,y] ~ pdat$pep,
         col = p_col,
         pch = 20, cex = 3.5)
  points(pdat2[,y] ~ pdat2$pep,
         col = p_col2,
         pch = 18, cex = 3.5)

  j <- j + 1
}
par(mar = mar_list[[j]], bg = "#ffffff")
plot(c(1,1), xaxt = "n", yaxt = "n", xlim = c(0,1), ylim = c(0,3.2),
     type = "n", xlab = "", ylab = "")
text(x = 0.5, y = 1.5,
     bquote("cor(~X[3]~", "~X[5]~") = "~.(cor_x)", cex = 2)
col1 <- unique(dat$colors)[2]
col2 <- unique(dat2$colors)[2]
col1 <- rgb(red = col2rgb(col1)[1,],

```

```

        green = col2rgb(col1)[2,],
        blue = col2rgb(col1)[3,],
        alpha = alpha_x[5], maxColorValue = 255)
col2 <- rgb(red = col2rgb(col2)[1,],
           green = col2rgb(col2)[2,],
           blue = col2rgb(col2)[3,],
           alpha = alpha_x[5], maxColorValue = 255)
if (!is.null(labs)) {
  labels <- c(labs[1], labs[2])
  legend("top",
        legend = labels,
        pch = c(20, 18),
        col = c(col1,col2), bty = "n", cex = 2)
}
}

pres_y_comp <- function(df, y_vec, cor_x,
  mar_list = list(c(.25,4,4,0), c(.25,2,4, 2),
                  c(.25,0,4,4), c(4,4,.25,0), c(4,2,.25, 2),
                  c(4,0,.25,4)),
  m = rbind(c(1,2,3), c(4,5,6)),
  n_var = 5,
  p = 5, labs = NULL,
  y_max = NULL){

  layout(m)
  df <- process_y(df, y_vec = y_vec)
  dat <- df
  xs <- unique(df$X)
  alpha_x <- 255*rep(0.4,5)
  leg <- c(expression(X[1]),expression(X[2]),
           expression(X[3]), expression(X[4]), expression(X[5]))
  j <- 1

  for(i in 1:p){
    par(mar = mar_list[[j]], bg = "#ffffff")
    pdat <- subset(dat, X == xs[i])
    y_max <- ifelse(is.null(y_max), pdat$max_y[1], y_max)
    plot(pdat[,y_vec[1]] ~ pdat$pep,
         pch = 20,
         type = "n", ylim = c(0,y_max),
         xlim = c(0,1),
         xaxt = "n",
         yaxt = "n", xlab = "", ylab = "")
    abline(h = 1, lwd = 1.5, lty = 2, col = "#4d4d4d")
    abline(v = .5, lwd = 1.5, lty = 2, col = "#4d4d4d")
    text(x = 0.8, y = y_max, leg[i], pos = 1, cex = 2)
    # add axes
    if ( j %in% c(1,4)) {
      axis(side = 2,

```

```

        labels = as.character(round(seq(0, y_max, length = 11), 2)),
        at = seq(0, y_max, length = 11), tick = T, cex.axis = 1.75)
    }
    if (j %in% c(4,5,6)){
        axis(side = 1, at = seq(0, 1, length = 11), tick = T, cex.axis = 1.75)
        mtext("PEP", 1, line = 2.5)
    }

    p_col <- col2rgb(unique(pdat$colors1))
    p_col <- rgb(red = p_col[1,], green = p_col[2,],
                blue = p_col[3,], alpha = alpha_x[i],
                maxColorValue = 255)
    p_col2 <- col2rgb(unique(pdat$colors2))
    p_col2 <- rgb(red = p_col2[1,], green = p_col2[2,],
                blue = p_col2[3,], alpha = alpha_x[i],
                maxColorValue = 255)
    points(pdat[,y_vec[1]] ~ pdat$pep,
           col = p_col,
           pch = 20, cex = 3.5)
    points(pdat[,y_vec[2]] ~ pdat$pep,
           col = p_col2,
           pch = 18, cex = 3.5)

    j <- j + 1
}
par(mar = mar_list[[j]], bg = "#ffffff")
plot(c(1,1), xaxt = "n", yaxt = "n", xlim = c(0,1), ylim = c(0,3.2),
     type = "n", xlab = "", ylab = "")
text(x = 0.5, y = 1.5, bquote("cor(~X[3]~", "~X[5]~") = "~.(cor_x)",
    cex = 2.5, font = 2, col = "black")
col1 <- unique(dat$colors1)[2]
col2 <- unique(dat$colors2)[2]
col1 <- rgb(red = col2rgb(col1)[1,],
            green = col2rgb(col1)[2,],
            blue = col2rgb(col1)[3,],
            alpha = alpha_x[5], maxColorValue = 255)
col2 <- rgb(red = col2rgb(col2)[1,],
            green = col2rgb(col2)[2,],
            blue = col2rgb(col2)[3,],
            alpha = alpha_x[5], maxColorValue = 255)
if (!is.null(labs)) {
    labels <- c(labs[1], labs[2])
    legend("top",
           legend = labels,
           pch = c(20, 18),
           col = c(col1,col2), bty = "n", cex = 2)
}
}

```

```
process_y_indiv <- function (df, y) {
```

```

dat <- df[[2]]
xs <- unique(dat$X)
nsims <- dim(dat)[1]/length(xs)
colors2 = rep(c("#0570b0", "#0570b0",
                "#0570b0", "#045a8d",
                "#045a8d"), each = nsims)

max_y <- max(dat[,y], na.rm = T)
max_y <- max(max_y) + 0.25
for(i in 1:2){
  pep <- 1 - dat[, "PIP"]
  dat$pep <- pep
  dat$colors <- colors2
  dat$max_y <- max_y
}
return(dat)
}

pres_y <- function(df, y, cor_x,
                  mar_list = list(c(.25,4,4,0), c(.25,2,4, 2),
                                c(.25,0,4,4), c(4,4,.25,0),
                                c(4,2,.25, 2), c(4,0,.25,4)),
                  m = rbind(c(1,2,3), c(4,5,6)),
                  n_var = 5,
                  p = 5, y_max = NULL){

  layout(m)
  df <- process_y_indiv(df, y = y)
  dat <- df
  xs <- unique(df$X)
  alpha_x <- 255*rep(0.4,5)
  leg <- c(expression(X[1]),expression(X[2]),
            expression(X[3]), expression(X[4]), expression(X[5]))
  ylabel <- ifelse(y == "d", latex2exp::TeX("$Var_{MA}$ / $Var_{cts}$"),
                  ifelse(y == "d_topCts",
                          latex2exp::TeX("$Var_{cts}$ / $Var_{top}$"),
                          ifelse(y == "d_topMA",
                                  latex2exp::TeX("$Var_{MA}$ / $Var_{top}$"),
                                  ifelse(y == "diff_ct",
                                          latex2exp::TeX("$SD_{cts}$ - $SD_{top}$"),
                                          latex2exp::TeX("$SD_{MA}$ - $SD_{top}$"))
                                )))

  j <- 1

  for(i in 1:p){
    par(mar = mar_list[[j]], bg = "#ffffff")
    pdat <- subset(dat, X == xs[i])
    y_max <- ifelse(is.null(y_max), pdat$max_y[1], y_max)
    plot(pdat[,y] ~ pdat$pep,
         pch = 20,
         type = "n", ylim = c(0,y_max),

```

```

        xlim = c(0,1),
        xaxt = "n",
        yaxt = "n", xlab = "", ylab = "")
abline(h = 1, lwd = 1.5, lty = 2, col = "#4d4d4d")
abline(v = .5, lwd = 1.5, lty = 2, col = "#4d4d4d")
text(x = 0.8, y = y_max, leg[i], pos = 1, cex = 2)
# add axes
if ( j %in% c(1,4)) {
  axis(side = 2,
        labels = as.character(round(seq(0, y_max, length = 11), 2)),
        at = seq(0, y_max, length = 11), tick = T, cex.axis = 1.75)
  mtext(ylab, 2, line = 2.5)
}
if (j %in% c(4,5,6)){
  axis(side = 1, at = seq(0, 1, length = 11), tick = T, cex.axis = 1.75)
  mtext("PEP", 1, line = 2.5)
}

p_col <- col2rgb(unique(pdat$colors))
p_col <- rgb(red = p_col[1,], green = p_col[2,],
             blue = p_col[3,], alpha = alpha_x[i],
             maxColorValue = 255)
points(pdat[,y] ~ pdat$pep,
       col = p_col,
       pch = 20, cex = 3.5)
j <- j + 1
}
par(mar = mar_list[[j]], bg = "#ffffff")
plot(c(1,1), xaxt = "n", yaxt = "n", xlim = c(0,1), ylim = c(0,3.2),
     type = "n", xlab = "", ylab = "")
text(x = 0.5, y = 1.5, bquote("cor(\"~X[3]~\", \"~X[5]~\") = \"~.(cor_x)\",
                              cex = 2.5, font = 2, col = "black")
}

final_bean_d <- function(df_vec, y,
                        cor_vec = c(0, 0.3, 0.6, 0.7, 0.8, 0.9),
                        mar_list = list(c(.25,4,4,0), c(.25,2,4, 2),
                                       c(.25,0,4,4), c(4,4,.25,0),
                                       c(4,2,.25, 2), c(4,0,.25,4)),
                        m = rbind(c(1,2,3), c(4,5,6))) {
  layout(m)
  idx_d <- which(attributes(df_vec)$names == "summary")
  max_y <- sapply(df_vec[idx_d], function(x) {
    return(max(x[,y], na.rm = T))})
  min_y <- sapply(df_vec[idx_d], function(x) {
    return(min(x[,y], na.rm = T))})
  num_obs <- lapply(df_vec[idx_d], function(x) {
    xs <- unique(x$X)
    obs <- rep(NA, 5)
    nas <- rep(NA, 5)
    for(i in 1:5){

```

```

      nas[i] <- sum(is.na(subset(x, X == xs[i]))[,y])
      obs[i] <- 300-nas[i]
    }
    return(c(obs))
  })
min_y <- min(min_y) - 30*sd(min_y)
xs <- unique(df_vec[[2]][,1])
nsims <- dim(df_vec[[2]])[1]/length(xs)
yt <- ifelse(y == "d", latex2exp::TeX("$Var_{MA}$ / $Var_{cts}$"),
  ifelse(y == "d_topCts",
    latex2exp::TeX("$Var_{cts}$ / $Var_{top}$"),
    ifelse(y == "d_topMA",
      latex2exp::TeX("$Var_{MA}$ / $Var_{top}$"),
      ifelse(y == "diff_ct",
        latex2exp::TeX("$SD_{cts}$ - $SD_{top}$"),
        ifelse(y == "diff_MAt",
          latex2exp::TeX("$SD_{MA}$ - $SD_{top}$"),
          "PIP")
        )))
  )))
yl <- ifelse(y == "PIP", 1,
  ifelse(y == "d_topCts", 1.25, 3.25))
hline <- ifelse(y == "PIP", NA,
  ifelse(y == "diff_ct" | y == "diff_MAt", 0, 1))
j <- 1

df_vec <- process_fun(df_vec, y = y)

for(i in idx_d){
  par(mar = mar_list[[j]], bg = "#ffffff")
  if(j %in% c(1,2,3)){
    beanplot::beanplot(df_vec[[i]][,y] ~ df_vec[[i]][,1],
      yaxt = "n", log = "", xaxt = "n",
      ylim = c(min_y, max_y[3] + sd(max_y[1:3])/4))
    text(1:5, min_y, as.character(unlist(num_obs[j])),
      cex = 1.5, font = 2)
    text(1.75, max_y[3],
      bquote("Cor(~X[3]~, ~X[5]~) = ~.(cor_vec[j])"),
      cex = 1.5, font = 2)
  } else {
    beanplot::beanplot(df_vec[[i]][,y] ~ df_vec[[i]][,1],
      yaxt = "n", log = "", xaxt = "n",
      ylim = c(min_y, max(max_y)+sd(max_y)/2))
    axis(1, at = 1:5, labels = c(latex2exp::TeX("$X_1$"),
      latex2exp::TeX("$X_2$"),
      latex2exp::TeX("$X_3$"),
      latex2exp::TeX("$X_4$"),
      latex2exp::TeX("$X_5$")), cex.axis = 1.75)
    text(1:5, min_y, as.character(unlist(num_obs[j])), cex = 1.5, font = 2)
    text(1.75, max(max_y),
      bquote("Cor(~X[3]~, ~X[5]~) = ~.(cor_vec[j])"),
      cex = 1.5, font = 2)
  }
}

```

```

}
if (j == 1) {
  axis(side = 2,
        labels = as.character(round(seq(0, max_y[3] + sd(max_y[1:3])/4,
                                      length = 11), 2)),
        at = seq(0, max_y[3] + sd(max_y[1:3])/4, length = 11),
        tick = T, cex.axis = 1.5)
  mtext(yt, 2, line = 2.5)
}
if (j == 4) {
  axis(side = 2,
        labels = as.character(round(seq(0, max(max_y) + sd(max_y)/4,
                                      length = 11), 2)),
        at = seq(0, max(max_y) + sd(max_y)/2, length = 11),
        tick = T, cex.axis = 1.5)
  mtext(yt, 2, line = 2.5)
}
abline(h = hline, lwd = 2, lty = 2, col = "#b2182b")
j <- j + 1
}
}

final_bean_diff <- function(df_vec, y,
                           cor_vec = c(0, 0.3, 0.6, 0.7, 0.8, 0.9),
                           mar_list = list(c(.25,4,4,0), c(.25,2,4, 2),
                                           c(.25,0,4,4), c(4,4,.25,0),
                                           c(4,2,.25, 2), c(4,0,.25,4)),
                           m = rbind(c(1,2,3), c(4,5,6))) {
  layout(m)
  idx_d <- which(attributes(df_vec)$names == "summary")
  max_y <- sapply(df_vec[idx_d], function(x) {
    return(max(x[,y], na.rm = T))})
  min_y <- sapply(df_vec[idx_d], function(x) {
    return(min(x[,y], na.rm = T))})
  num_obs <- lapply(df_vec[idx_d], function(x) {
    xs <- unique(x$X)
    obs <- rep(NA, 5)
    nas <- rep(NA, 5)
    for(i in 1:5) {
      nas[i] <- sum(is.na(subset(x, X == xs[i]))[,y])
      obs[i] <- 300-nas[i]
    }
    return(c(obs))
  })
  min_y <- min(min_y) - 30*sd(min_y)
  xs <- unique(df_vec[[2]][,1])
  nsims <- dim(df_vec[[2]])[1]/length(xs)
  yt <- ifelse(y == "d", latex2exp::TeX("$Var_{MA}$ / $Var_{cts}$"),
              ifelse(y == "d_topCts",
                    latex2exp::TeX("$Var_{cts}$ / $Var_{top}$"),
                    ifelse(y == "d_topMA",

```

```

        latex2exp::TeX("$Var_{MA}$ / $Var_{top}$"),
        ifelse(y == "diff_ct",
            latex2exp::TeX("$SD_{cts} - SD_{top}$"),
            ifelse(y == "diff_MAt",
                latex2exp::TeX("$SD_{MA} - SD_{top}$"),
                "PIP")
            ))))
yl <- ifelse(y == "PIP", 1,
            ifelse(y == "d_topCts", 1.25, 3.25))
hline <- ifelse(y == "PIP", NA,
               ifelse(y == "diff_ct" | y == "diff_MAt", 0, 1))
j <- 1

df_vec <- process_fun(df_vec, y = y)

for(i in idx_d){
  par(mar = mar_list[[j]], bg = "#ffffff")
  if(j %in% c(1,2,3)){
    beanplot::beanplot(df_vec[[i]][,y] ~ df_vec[[i]][,1],
                        yaxt = "n", log = "", xaxt = "n",
                        ylim = c(min_y, max(max_y)+sd(max_y)/2))
    text(1:5, min_y, as.character(unlist(num_obs[j])), cex = 1.5, font = 2)
    text(1.75, max(max_y),
         bquote("Cor(~X[3]~, ~X[5]~) = ~.(cor_vec[j])"),
         cex = 1.5, font = 2)
  } else {
    beanplot::beanplot(df_vec[[i]][,y] ~ df_vec[[i]][,1],
                        yaxt = "n", log = "", xaxt = "n",
                        ylim = c(min_y, max(max_y)+sd(max_y)/2))
    axis(1, at = 1:5, labels = c(latex2exp::TeX("$X_1$"),
                                  latex2exp::TeX("$X_2$"),
                                  latex2exp::TeX("$X_3$"),
                                  latex2exp::TeX("$X_4$"),
                                  latex2exp::TeX("$X_5$")), cex.axis = 1.75)
    text(1:5, min_y, as.character(unlist(num_obs[j])), cex = 1.5, font = 2)
    text(1.75, max(max_y),
         bquote("Cor(~X[3]~, ~X[5]~) = ~.(cor_vec[j])"),
         cex = 1.5, font = 2)
  }
  if (j %in% c(1,4)) {
    axis(side = 2,
         labels = as.character(round(seq(0, max(max_y) + sd(max_y)/4,
                                         length = 11), 2)),
         at = seq(0, max(max_y) + sd(max_y)/2, length = 11),
         tick = T, cex.axis = 1.5)
    mtext(yt, 2, line = 2.5)
  }
  abline(h = hline, lwd = 2, lty = 2, col = "#b2182b")
  j <- j + 1
}
}

```

### C.3 Recreate the Simulation

The following code generates the same results presented in Chapter 3.

```
# reference case
#~~~~~#
# 300 realizations #
#~~~~~#

set.seed(1)
cor_0 <- sim_d2(300, cor_x = 0.0, tol = 0.01)

set.seed(6)
cor_3 <- sim_d2(300, cor_x = 0.3, tol = 0.01)

set.seed(5)
cor_6 <- sim_d2(300, cor_x = 0.6, tol = 0.01)

set.seed(4)
cor_7 <- sim_d2(300, cor_x = 0.7, tol = 0.01)

set.seed(3)
cor_8 <- sim_d2(300, cor_x = 0.8, tol = 0.01)

set.seed(8)
cor_9 <- sim_d2(300, cor_x = 0.9, tol = 0.01)

df_vec <- c(cor_0, cor_3, cor_6, cor_7, cor_8, cor_9)

# comparing signal to noise Truth = YES

set.seed(1)
cor_0s1 <- sim_d2(300, cor_x = 0.0, tol = 0.01, sig_y = 1)

set.seed(6)
cor_3s1 <- sim_d2(300, cor_x = 0.3, tol = 0.01, sig_y = 1)

set.seed(5)
cor_6s1 <- sim_d2(300, cor_x = 0.6, tol = 0.01, sig_y = 1)

set.seed(4)
cor_7s1 <- sim_d2(300, cor_x = 0.7, tol = 0.01, sig_y = 1)

set.seed(3)
```

```

cor_8s1 <- sim_d2(300, cor_x = 0.8, tol = 0.01, sig_y = 1)

set.seed(8)
cor_9s1 <- sim_d2(300, cor_x = 0.9, tol = 0.01, sig_y = 1)

df_vec_s1 <- c(cor_0s1, cor_3s1, cor_6s1, cor_7s1, cor_8s1, cor_9s1)

# comparing truth in the model set sig-y = low

# NO5
set.seed(1)
cor_0n5 <- sim_d2(300, cor_x = 0.0, tol = 0.01, case = 2)

set.seed(6)
cor_3n5 <- sim_d2(300, cor_x = 0.3, tol = 0.01, case = 2)

set.seed(5)
cor_6n5 <- sim_d2(300, cor_x = 0.6, tol = 0.01, case = 2)

set.seed(4)
cor_7n5 <- sim_d2(300, cor_x = 0.7, tol = 0.01, case = 2)

set.seed(3)
cor_8n5 <- sim_d2(300, cor_x = 0.8, tol = 0.01, case = 2)

set.seed(8)
cor_9n5 <- sim_d2(300, cor_x = 0.9, tol = 0.01, case = 2)

df_vec_n5 <- c(cor_0n5, cor_3n5, cor_6n5, cor_7n5, cor_8n5, cor_9n5)

# NO 4
set.seed(1)
cor_0n4 <- sim_d2(300, cor_x = 0.0, tol = 0.01, truth4 = T)

set.seed(6)
cor_3n4 <- sim_d2(300, cor_x = 0.3, tol = 0.01, truth4 = T)

set.seed(5)
cor_6n4 <- sim_d2(300, cor_x = 0.6, tol = 0.01, truth4 = T)

set.seed(4)
cor_7n4 <- sim_d2(300, cor_x = 0.7, tol = 0.01, truth4 = T)

set.seed(3)
cor_8n4 <- sim_d2(300, cor_x = 0.8, tol = 0.01, truth4 = T)

set.seed(8)
cor_9n4 <- sim_d2(300, cor_x = 0.9, tol = 0.01, truth4 = T)

```

```

df_vec_n4 <- c(cor_0n4, cor_3n4, cor_6n4, cor_7n4, cor_8n4, cor_9n4)

# comparing truth in the model set for sig to noise = High

## truth = N04, sig = 1

set.seed(1)
cor_0n41 <- sim_d2(300, cor_x = 0.0, tol = 0.01, truth4 = T, sig_y = 1)

set.seed(6)
cor_3n41 <- sim_d2(300, cor_x = 0.3, tol = 0.01, truth4 = T, sig_y = 1)

set.seed(5)
cor_6n41 <- sim_d2(300, cor_x = 0.6, tol = 0.01, truth4 = T, sig_y = 1)

set.seed(4)
cor_7n41 <- sim_d2(300, cor_x = 0.7, tol = 0.01, truth4 = T, sig_y = 1)

set.seed(3)
cor_8n41 <- sim_d2(300, cor_x = 0.8, tol = 0.01, truth4 = T, sig_y = 1)

set.seed(8)
cor_9n41 <- sim_d2(300, cor_x = 0.9, tol = 0.01, truth4 = T, sig_y = 1)

df_vec_n41 <- c(cor_0n41, cor_3n41, cor_6n41, cor_7n41, cor_8n41, cor_9n41)

# truth = N05, sig = 1

set.seed(1)
cor_0n51 <- sim_d2(300, cor_x = 0.0, tol = 0.01, case = 2, sig_y = 1)

set.seed(6)
cor_3n51 <- sim_d2(300, cor_x = 0.3, tol = 0.01, case = 2, sig_y = 1)

set.seed(5)
cor_6n51 <- sim_d2(300, cor_x = 0.6, tol = 0.01, case = 2, sig_y = 1)

set.seed(4)
cor_7n51 <- sim_d2(300, cor_x = 0.7, tol = 0.01, case = 2, sig_y = 1)

set.seed(3)
cor_8n51 <- sim_d2(300, cor_x = 0.8, tol = 0.01, case = 2, sig_y = 1)

set.seed(8)
cor_9n51 <- sim_d2(300, cor_x = 0.9, tol = 0.01, case = 2, sig_y = 1)

df_vec_n51 <- c(cor_0n51, cor_3n51, cor_6n51, cor_7n51, cor_8n51, cor_9n51)

```

## C.4 Recreating Simulation Study Results

We discussed results from select treatments in Chapter 3. Here, we provide instructions for creating plots to summarize results from other treatments. For all of the plotting functions discussed below, the `y = "d"` tells the function to use  $Var_{MA}/Var_{cts}$  as the response variable. The argument `y` can take on the following characters:  $Var_{cts}/Var_{top} = \text{"d\_topCts"}$ ,  $Var_{MA}/Var_{top} = \text{"d\_topMA"}$ ,  $SD_{MA} - SD_{top} = \text{"diff\_MA"}$ , and  $SD_{cts} - SD_{top} = \text{"diff\_ct"}$ .

### C.4.1 Using `pres_d_plot` to Visualize Results

Use `pres_d_plot` to plot results from all levels of correlation for fixed levels of truth in the model set and signal to noise. Specify the number of variables with `n_var` and the upper limit for the  $y$ -axis with `y_max` (Figure C.1).

```
pres_d_plot(df_vec, y = "d", n_var = 5, y_max = 3.5)
```

Use `plot_order = "backward"` to plot results from  $X_5$  first and  $X_1$  last (Figure C.2).

```
pres_d_plot(df_vec, y = "d", n_var = 5, plot_order = "backward", y_max = 3.5)
```

To create plots from different treatments use a different list of results (e.g., to plot results from high signal to noise and truth in the model set, use `df_vec.s1` in place of `df_vec`).

### C.4.2 Using `pres_y` to Visualize Results

It can be useful to visualize results from one individual treatment combination for all variables considered (in separate plots), use `pres_y`. The correlation with `cor_x`

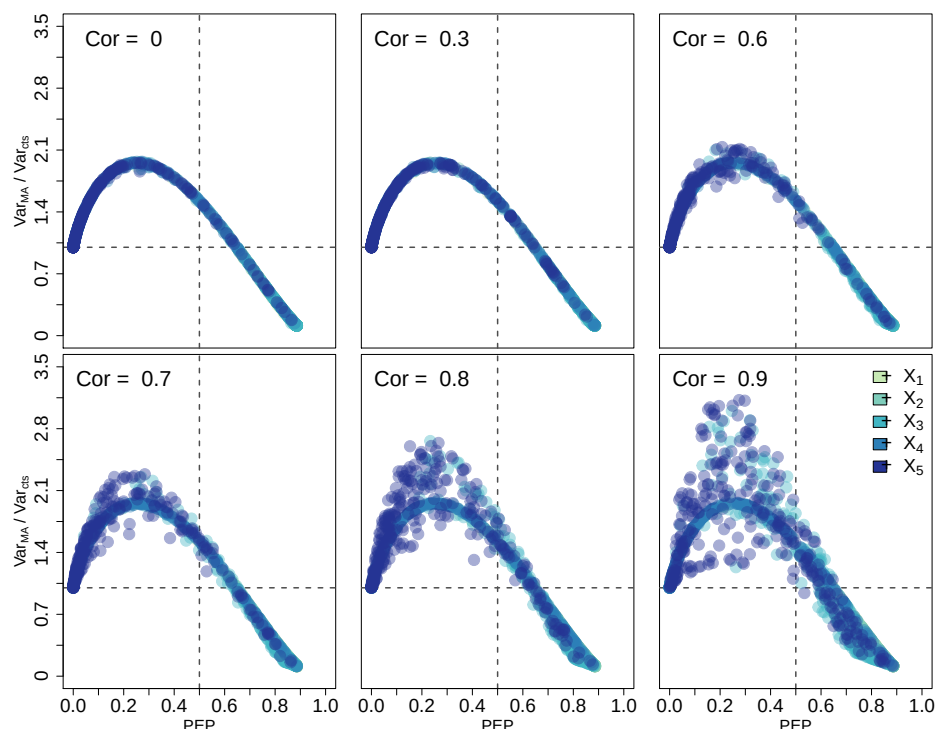


Figure C.1: Low signal to noise, truth = YES, all correlations, plotted X1:X5

must be specified, and there is an option to specify a maximum value for the  $y$ -axis with `y_max` (Figure C.3). Note, rather than using the list of results for all correlations and fixed levels of signal to noise and truth in the model set, the individual results from one treatment are used to create the plot (i.e., `cor_y` instead of `df_vec`).

```
pres_y(cor_9, y = "d", cor_x = 0.9, y_max = 3.5)
```

#### C.4.3 Using `pres_trt_compare` to Visualize Results

Use `pres_trt_compare` to compare results from two specific treatments. Here, I use this function to compare low signal to noise to high signal to noise for truth in the model set and correlation equal to 0.9. Specify the upper limit for the  $y$ -axis with

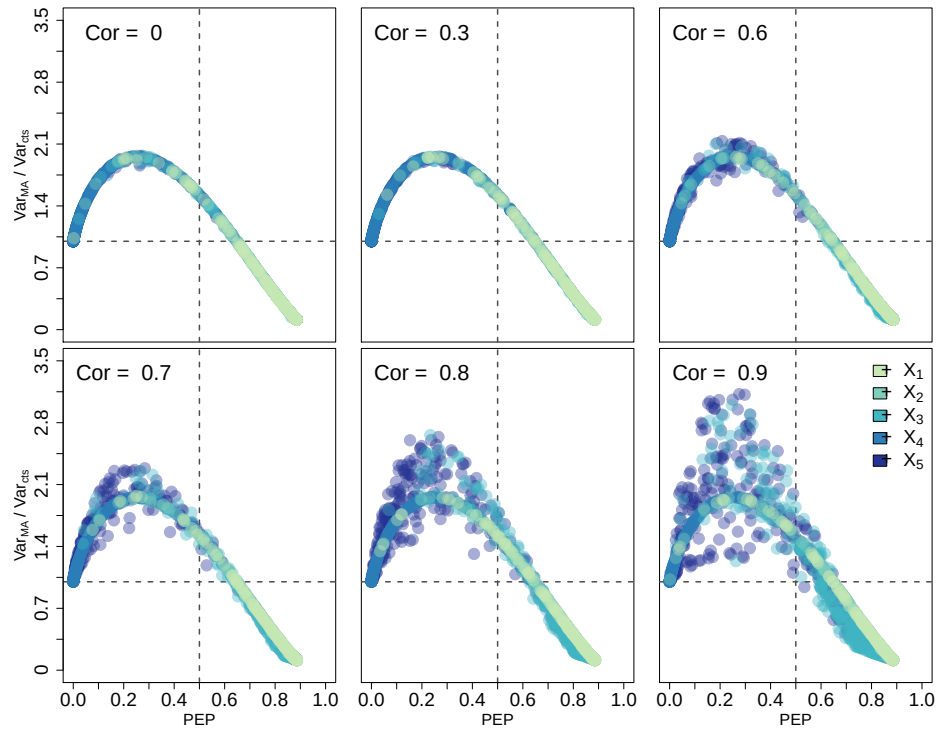


Figure C.2: Low signal to noise, truth = YES, all correlations, plotted X5:X1

`y_max`, the response variable with `y`, the correlation between  $X_3$  and  $X_5$  with `cor_x`, and labels for the treatments being compared with `labs` (Figure C.4).

```
pres_trt_comp(c(cor_9, cor_9s1), y = "d", cor_x = 0.9,
              labs = c("sig-y = 2.5, YES", "sig-y = 1, YES"), y_max = 3.5)
```

#### C.4.4 Using `pres_y_compare` to Visualize Results

Use `pres_y_compare` to compare results from two response variables for a particular treatment. Here, I use this function to compare  $Var_{MA}/Var_{cts}$  to  $Var_{MA}/Var_{top}$  for low signal to noise, truth in the model set, and correlation equal to 0.9. Specify the upper limit for the  $y$ -axis with `y_max`, the response variable with `y`, the correlation between  $X_3$  and  $X_5$  with `cor_x`, and labels for the ratios of variances being compared with `labs`.

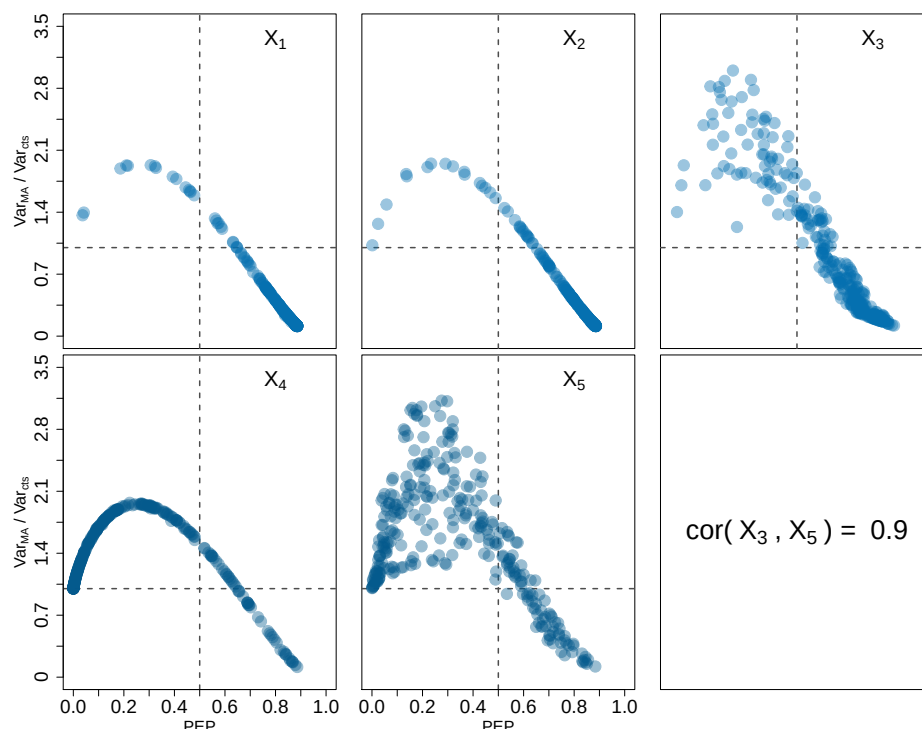


Figure C.3: Low signal to noise, truth = YES, cor = 0.9

```
pres_y_comp(cor_9, y_vec = c("d", "d_topMA"),
            labs = c("Var_MA/Var_cts", "Var_MA/Var_top"), cor_x = 0.9, y_max = 8)
```

#### C.4.5 Using final\_bean\_d to Visualize Results

Use `final_bean_d` to visualize results for  $Var_{cts}/Var_{top}$  across all correlations for fixed levels of signal to noise and truth in the model set (Figure C.6).

```
final_bean_d(df_vec, y = "d_topCts")
```

#### C.4.6 Using final\_bean\_diff to Visualize Results

Use `final_bean_diff` to visualize results for  $SD_{MA} - SD_{top}$  across all correlations for fixed levels of signal to noise and truth in the model set (Figure C.7).

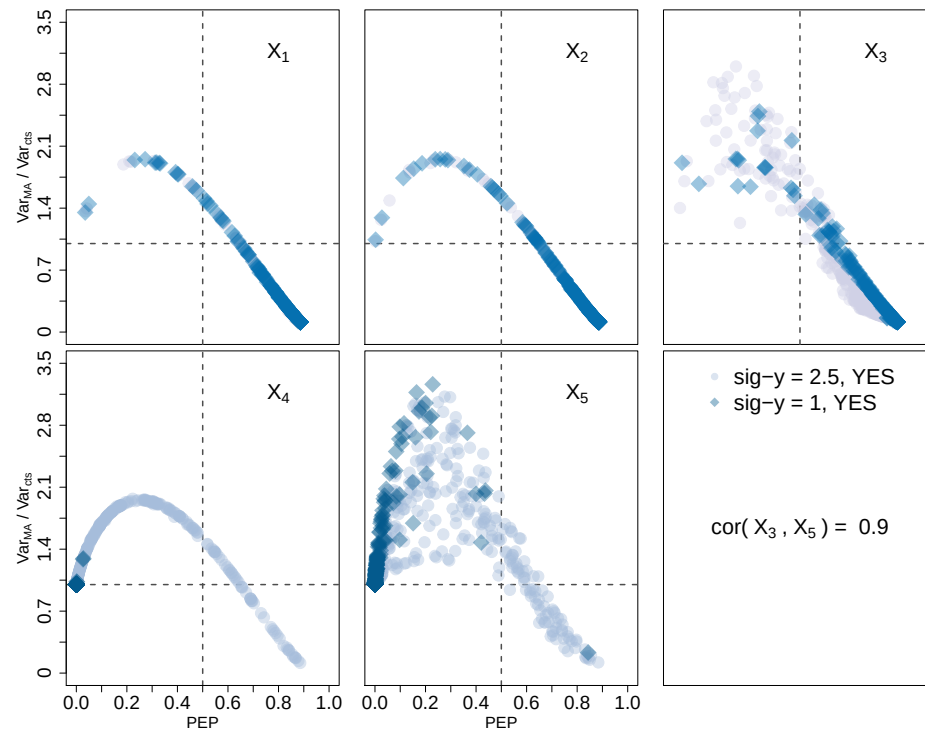


Figure C.4: Comparing signal to noise: Low vs. High signal to noise, truth = YES, all correlations

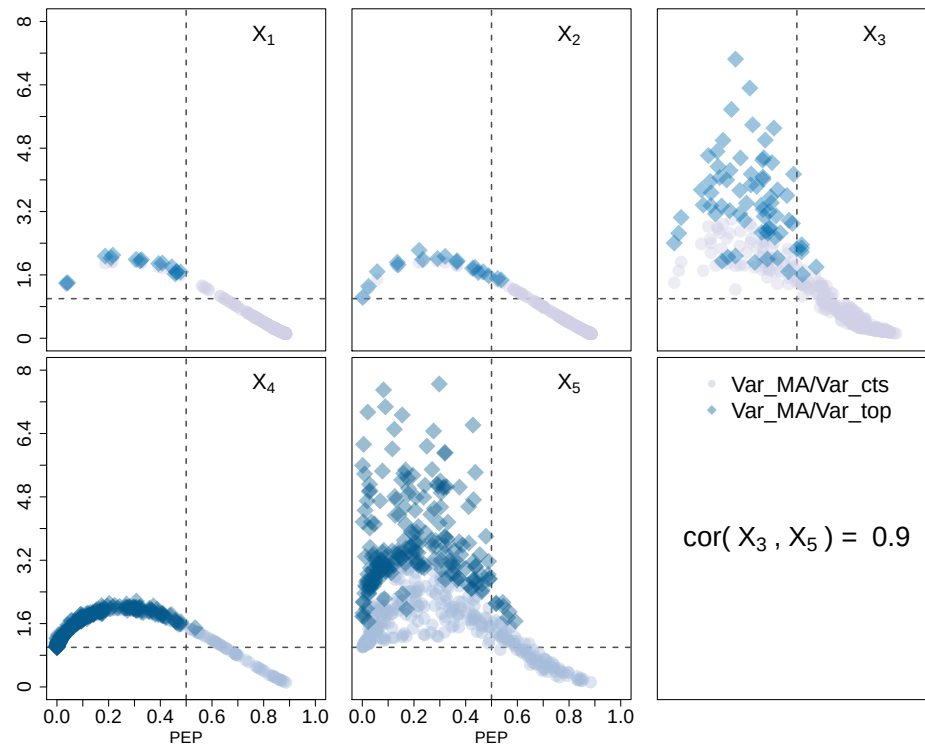


Figure C.5: Comparing two ratios of variances: Low signal to noise, truth = YES, all correlations

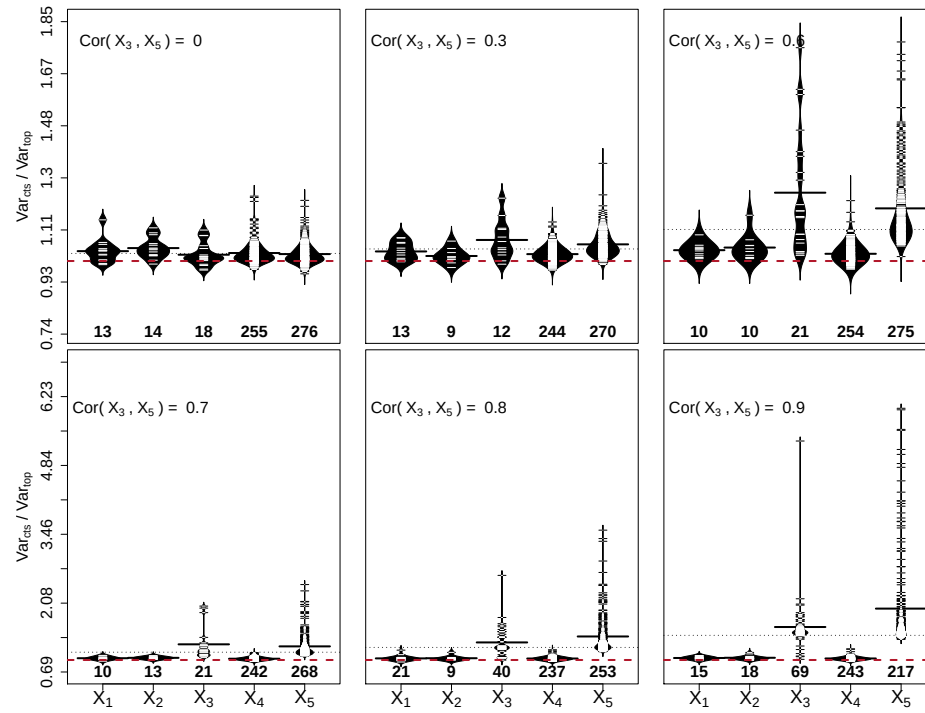


Figure C.6: Low signal to noise, truth = YES, cor = 0.9

```
final_beam_d(df_vec, y = "diff_MAt")
```

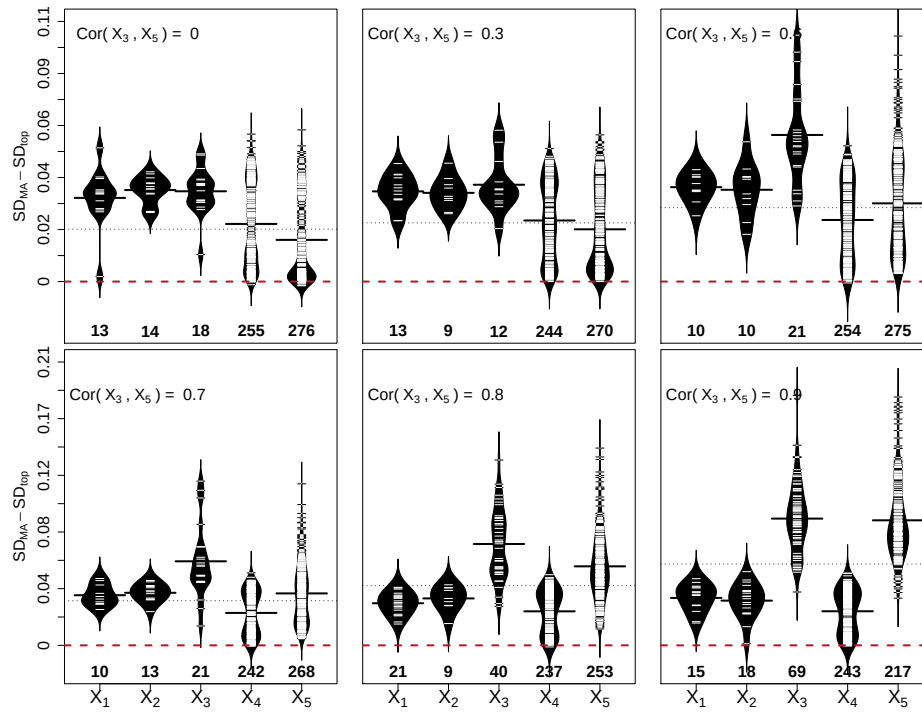


Figure C.7: Low signal to noise, truth = YES, cor = 0.9