UNIVERSITY OF CALIFORNIA

Los Angeles

Bayesian Selection Model with Shrinking Priors for Nonignorable Missingness

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy in Psychology

by

Juan Diego Vera

2023

ABSTRACT OF THE DISSERTATION

Bayesian Selection Model with Shrinking Priors for Nonignorable Missingness

by

Juan Diego Vera

Doctor in Philosophy in Psychology

University of California, Los Angeles, 2023

Professor Craig K. Enders, Chair

This study investigates the effectiveness of Bayesian variable selection (BVS) procedures in dealing with missing not at random (MNAR) data for identification in selection models. Three BVS-adapted selection models, namely Bayesian LASSO, horseshoe prior, and spike-and-slab prior, were compared, along with established missing data methods such as a model that assumes a missing at random (MAR) process and full-selection model. The results indicate that the spike-and-slab prior consistently outperformed other BVS methods in terms of accuracy and bias for various parameters, including slope estimates, residual variance, and intercept. When compared with the full-selection model, the spike-and-slab model exhibited superior performance across all parameters based on mean squared error (MSE) results.

Although the MAR and spike-and-slab models showed comparable performance for slope estimates, the spike-and-slab model consistently outperformed the MAR model in estimating residual variance and intercept. This comparable performance is attributed to the bias-variance tradeoff. The MAR model, while biased, demonstrated efficiency by estimating fewer parameters than selection models and obtaining robust support from the observed data. On the other hand, the spike-and-slab model outperformed the full-selection model, even when the full-selection model aligned with the true data-generating model. The adaptation of BVS to selection models, particularly through the spike-and-slab method, yielded promising results with unbiased estimates under various conditions. However, it is important to acknowledge that this study represents an initial exploration of this subject, and its scope was inherently limited. Finally, the BVS adaptations to the selection model was illustrated with data from a clinical-trial study.

The dissertation of Juan Diego Vera is approved.

Han Du

Amanda K. Montoya

Lara A. Ray

Craig K. Enders, Committee Chair

University of California, Los Angeles

2023

I dedicate this dissertation to my beloved wife, Kathy, and our wonderful daughter Martina Leila. Their patience and motivation have been my guiding light.

TABLE OF CONTENTS

List of Figures

List of Tables

Acknowledgements

I extend my heartfelt gratitude to all my mentors, collaborators, colleagues, and friends whose support, valuable feedback, and guidance have played an indispensable role in shaping this work. I am deeply indebted to my graduate advisor and committee chair, Professor Craig Enders, for his warm welcome to UCLA and for granting me the intellectual freedom to explore my research interests. His invaluable counsel, inspiration, encouragement, and continuous support have been instrumental throughout my graduate journey.

Special acknowledgments are due to Drs. Abbey Eisenhower and Alice Carter whose mentorship during my time at the ABCD Lab significantly contributed to my growth in the field of science. I am equally grateful to my friends and collaborators, Elizabeth Frenette, Luisa Buitrago, and Noah Hoch, who taught me the importance of not taking myself too seriously while still striving for long-term success. I would like to express my gratitude to Dr. Garrett Brady and Evan Marcus for setting the highest standards of academic excellence and for providing a source of both inspiration and amusement during my undergraduate years. Lastly, my thanks go out to my collaborators at the Semel Institute, including René Freichel, Drs. Sandra Loo, Agatha Lenartowicz, and Giorgia Michelini, who sparked my interest in attention science.

I owe a profound debt of gratitude to my incredible family: my parents, Jaime and Heidi, who worked tirelessly to provide me with the life they had dreamed for me. To my siblings, Nicole, Micaela, Eric, and Jaime, whose presence has been a constant source of comfort, and to my parents-in-law, Kathy and Rafael, for their unwavering support. I also want to give a special mention to my loyal companion, Paco, whose silent but steadfast presence has been a source of solace throughout my writing process. Finally, my sincere appreciation goes to my wife Kathy, for ensuring I maintained a balanced life filled with fun, love, and surf.

Vita /Biographical Sketch

EDUCATION

University of California Los Angeles (UCLA) **2018**
Masters, Psychology, Quantitative Methods
Advisor: Dr. Craig Enders

Florida State University (FSU) **2015**
Bachelor of Science in Statistics, *Cum Laude*
Bachelor of Science in Psychology, *Cum Laude*

PEER-REVIEWED PUBLICATIONS

- Sheldrick, R.C.; Frenette, E.C.; Vera, J.D.; Mackie, T.I.; Martinez-Pedraza, F.L.; Hoch, N.; Eisenhower, A.S.; Fettig, A.; Carter, A.S. (2019) What drives detection and diagnosis of autism spectrum disorder? Looking under the hood of a multi-stage screening process in Early Intervention. Journal of Autism and Developmental Disorders.
- Vera JD & Enders CK (2021) Is Item Imputation Always Better? An Investigation of Wave-Missing Data in Growth Models, Structural Equation Modeling: A Multidisciplinary Journal.
- Chorpita, B. F., Daleiden, E. L., Vera, J. D., & Guan, K. (2021). Creating a prepared mental health workforce: comparative illustrations of implementation strategies. Evidence-Based Mental Health
- Michelini, G., Salmastyan, G., Vera, J. D., & Lenartowicz, A. (2022). Event-related brain oscillations in attention-deficit/hyperactivity disorder (ADHD): a systematic review and meta-analysis. International Journal of Psychophysiology.
- Michelini, G., Lenartowicz, A., Vera, J. D., Bilder, R. M., McGough, J. J., McCracken, J. T., & Loo, S. K. (2022). Electrophysiological and Clinical Predictors of Methylphenidate, Guanfacine, and Combined Treatment Outcomes in Children With Attention-Deficit/Hyperactivity Disorder. Journal of the American Academy of Child & Adolescent Psychiatry.
- Vera, J.D., Freichel R.M, Lenartowicz A, Loo S.K., A Network Approach to Understanding The Role of Executive Functioning and Alpha Oscillations in Inattention and Hyperactivity-Impulsivity Symptom of ADHD. Journal of Attention Disorders (in press)

INTRODUCTION

The problem of missing data is relatively common in behavioral science research and, if not adequately handled, it can present various problems such as a reduction in statistical power, bias in the parameter estimates, and lack of sample representativeness (Enders, 2022; Little & Rubin, 2019). Missing data often arise for many different reasons; for example, some participants might refuse to respond to sensitive questions in a survey, participants in a clinical trial might drop out from the study because they are experiencing relief (or lack of), or participants with low income can miss visits to the lab because of lack of childcare or transportation. Each of these situations has a different reason for its missing observations, and as a result, researchers must consider the potential causes of missing values (Graham, 2009).

Thanks to research by Rubin and colleagues, missing data problems have been categorized into different mechanisms according to potential "reasons" for missingness (Little & Rubin, 2019; Mealli & Rubin, 2016; Rubin, 1976). These mechanisms are mainly interested in the relationship between missing data and the observed and unobserved responses in the dataset. Missing data are classified as: missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR). When the probability of being missing is the same for all observations, the missing data mechanism is considered MCAR. When the likelihood of data being missing does not depend on the unobserved data after conditioning on the observed data, then the missingness mechanism is considered MAR (or conditionally MAR; Graham, 2009). Finally, when the probability of data being missing depends on unobserved scores, even after conditioning on the observed data, it is then considered MNAR.

When the missing data mechanism is MNAR, the researcher must jointly model the missingness along with the substantive equation to avoid biases in parameter estimates and standard errors (Little, 2008; Rubin, 1976). Selection models and pattern mixture models are two of the most prominent MNAR-based methods that can jointly model both equations (Hedeker & Gibbons, 1997; Little, 2008; Michiels et al., 1999; Puhani, 2000; Sartori, 2003). These two modeling frameworks can prevent bias when estimating the parameters of the substantive model (of vital interest to investigators), as long as the missingness model is approximately correct. However, specifying and estimating the missingness equation to achieve unbiased estimates is extraordinarily difficult. Ibrahim, Chen, Lipsitz, and Herring (2005) summarize two different approaches to building a missingness model: (a) The model can be determined empirically from the observed data using traditional model selection approaches such as Akaike information criterion (AIC) and the Bayesian information criterion (BIC), and (b) sensitivity analyses can be performed that consider a range of different models. A main criticism of both options is that it requires the researcher to come up with candidate models for comparison (Ibrahim et al., 2005; Sterba & Gottfredson, 2015). This is particularly an issue when researchers do not have a substantive basis for constructing a missingness model. Under this circumstances, it is common for researchers to fit a large number of candidate missingness models, which increases the risk of overfitting the data (Ibrahim et al., 2005).

This dissertation is primarily interested in developing a flexible alternative to diagnostic indices or sensitivity analysis; however, I am not interested in recovering the true underlying missingness model, instead, I aim to apply variable selection methods to choose a set of adequate variables for the missingness model. It is possible, for example, that the true missingness model is never selected, but the substantive model parameters are estimated accurately because the missingness model effectively characterizes determinants of missing data.

Instead of using traditional criteria for model selection, I intend to investigate Bayesian variable selection (BVS) procedures that select the “best-fitting” variables for the missingness model. Although, there are numerous traditional variable selection procedures one could adopt (e.g., forward and backward selection, stepwise regression) (Borboudakis & Tsamardinos, 2019), I investigate fully-Bayesian variable selection procedures because they can provide the flexibility of simultaneously performing variable selection in the missingness model and treating missing data in the substantive model in one coherent analysis, in other words, the variable selection problem becomes part of the estimation. In this dissertation, I propose to use fully-Bayesian variable selection to eliminate unnecessary components of the missingness model, thereby improving estimation and reducing bias.

The three BVS procedures that I will investigate are (1) the Bayesian LASSO, (2) the spike-and-slab prior, (3) the horseshoe prior. The Bayesian LASSO is the Bayesian counterpart of a popular maximum likelihood approach called the Least Absolute Shrinkage and Selection Operator (LASSO), but instead of adding a penalty to the residual sum of squares, the Bayesian LASSO specifies a prior directly on the regression coefficients (Bhadra et al., 2019; Park & Casella, 2008). The spike-and-slab prior is an intuitive method for BVS because it uses a binary indicator variable for each regression coefficient, the purpose of which is to select which predictors are included or excluded (i.e., turned “on” or “off”) from the model (Bai et al., 2021; Bainter et al., 2020; Ishwaran & Rao, 2005). Finally, the horseshoe prior combines a global shrinkage parameter and a local shrinkage parameter, to allow a subset of predictors to have large and non-zero coefficients while shrinking irrelevant predictors towards zero (Carvalho et al., 2010). The goal of this study is to investigate if BVS methods can reduce nonresponse bias in the substantive model parameters when estimating selection models.

The remainder of the document is organized as follows. First, a brief introduction to the Bayesian framework is described. Second a short overview of missing data mechanism is presented. Third, a fully-Bayesian factored regression approach is explained and the MAR and MNAR missing data mechanisms are defined within this context. Then, I introduce in detail the three methods for Bayesian variable selection and discuss its relative strengths and weaknesses. Next, a fully-Bayesian factored regression approach with shrinking priors, a flexible model for handling missing data in a high dimensionality setting, will be described. Finally, a set of conditions for the simulation study, data generation procedures, and outcome variables will be stated.

Bayesian Framework

Given that the dissertation's focus is within the Bayesian framework, I briefly introduce the most relevant concepts to understand this thesis. The main idea of Bayesian inference is to combine information from the data with a prior distribution that represents a priori expectations about the parameters before looking at the data. The priors can be informed by previous research and accumulated knowledge, or it can be an “off-the-shelf” distribution that imposes as little information as possible on the data (Lynch, 2020). In other words, we specify a prior distribution for the parameter of interest, next, we combine this prior with a likelihood function that tells us how probable the data are under different parameter values. The result is a posterior distribution that describes the probability of different parameter values given the data (Chow & Hoijtink, 2017; Gelman et al., 1995; Kruschke, 2011; Lynch, 2020).

The posterior distribution of the parameter θ conditional on the data $\mathbf{Y}$ can be expressed using Bayes’ theorem.

$$f(\theta|\mathbf{Y}) = \frac{f(\mathbf{Y}|\theta)p(\theta)}{f(\mathbf{Y})} \propto f(\mathbf{Y}|\theta)f(\theta) \qquad (1)$$

Bayes' theorem combines the prior distribution of the parameter $f(\theta)$, the assumed distribution of the parameter before the data is collected, and the likelihood function $f(\mathbf{Y}|\theta)$, which is the relative probability of the data given different parameter values, to create the posterior distribution $f(\theta|\mathbf{Y})$. The denominator of the Bayes theorem, also called the marginal distribution of the data, functions as a normalizing constant and is often ignored because it does not depend on the parameters in θ**.** Therefore, the posterior distribution can also be described as proportional to the product of the prior and the likelihood function.

In general, deriving the posterior distribution directly is often challenging and difficult. When the focus is on one parameter, the analytical calculation of the posterior distribution is often tractable. However, when we fit a model with many parameters, this calculation can become very complex or impossible. Fortunately, iteration estimation methods such as Markov chain Monte Carlo (MCMC) can be used to estimate the posterior distribution of multiple parameters (Casella & George, 1992; Geman & Geman, 1984; Jackman, 2000; Metropolis et al., 1953).

Missing Data Mechanisms

In this section, I will further describe the three missing data mechanism: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). As described earlier in the introduction, missingness mechanisms are mainly concern with the relationship between missing data and the observed value of the variables in the dataset. Using notation from Rubin (1976), the data **Y** are portioned into observed scores and unseen scores, usually denoted $\mathbf{Y}_{\text{obs}}$ and $\mathbf{Y}_{\text{miss}}$, respectively; and missingness is represented by a binary indicator $\boldsymbol{M}$. Formally, missing data mechanisms are represented as the conditional distribution of $\boldsymbol{M}$ given the data $\mathbf{Y}$. To illustrate an example, I will use a hypothetical depression study, where the depression score is missing $\mathbf{Y}_{\text{miss}}$ and the variable age is observed $\mathbf{Y}_{\text{obs}}$**.**

MCAR is what researchers think of as unsystematic missingness, in the sense that missingness does not depend on the values of variables in the data and the distribution of the missing values are identical to those observed (Little & Rubin, 2019; Rubin, 1976). For example, some participants in a mental health study could have missing values on the depression score because the researcher lost their test scores by mistake.

$$f(\boldsymbol{M} = 1|\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{miss}}\ , \phi) = f(\boldsymbol{M} = 1|\phi) \tag{2}$$

The equation above the missingness indicator $\boldsymbol{M}$, if $\boldsymbol{M} = 1$, then the outcome is missing, and if $\boldsymbol{M} = 0$, then the outcome in observed. The term ϕ is a set of missingness model parameters that connect the data to the missingness indicator. In words, Equation 2 says that the cause of missingness is unrelated to any of the variables from dataset and is only related to ϕ which can be described as negligence from the researcher, where each participant had the same chance of having their score deleted.

Missing at random (MAR) is the most commonly assumed mechanism for missing data analyses. When the mechanism is MAR, it is assumed that the probability of missingness is only due to the observed scores (Little & Rubin, 2019). For example, participants in a depression study might miss a visit to the lab for reasons related to background variables such as age or income-level, but not to levels of depression itself.

$$f(\boldsymbol{M} = 1|\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{miss}}\,, \phi) = f(\boldsymbol{M} = 1|\mathbf{Y}_{\text{obs}}, \phi) \quad (3)$$

In words, the equation above says that after controlling for the observed scores (e.g., age in the bivariate example), all participants now have the same probability for missing data. That is, missingness is random after conditioning on observed data. The MAR mechanism is advantageous because the researcher can make statistical inferences about the model of interest by ignoring the specific causes of the missing data; that is, a researcher can perform the desired statistical analysis without also fitting an additional model that describes why data are missing (Little & Rubin, 2019; Rubin, 1976).

When missing data are related to specific missing values, then missing data is said to be MNAR. For example, participants in a depression study might miss a visit to the lab because they already feel better and no longer feel the need to stay in the study, in this case, the probability of missingness is related to the unobserved outcome variable and therefore interrelated with the substantive analysis of depression.

$$f(\boldsymbol{M} = 1|\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{miss}}\,, \phi) = f(\boldsymbol{M} = 1|\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{miss}}\,, \phi\,) \quad (4)$$

In this scenario, the researcher is recommended to jointly model both the underlying missingness process and the substantive analysis in order to mitigate or eliminate nonresponse bias (Little & Rubin, 2019).

The underlying missingness process can further be categorized by two distinct systems of missingness, focused and diffuse (Gomer & Yuan, 2021). Equation 4 is an example of a diffuse system, where missingness can relate to both the unseen and observed data. Returning to the depression example, Equation 4 shows a diffuse process that involves $\mathbf{Y}_{\text{obs}}$ as the age variable and $\mathbf{Y}_{\text{miss}}$ the unobserved depression score. In this scenario, age is also a predictor of the underlying missingness, where participants that come from an older generation are less likely to attend the study. Although the example above only involves one observed variable, a diffuse missingness process can involve any combination of covariates or auxiliary variables in addition to the unseen data (Gomer & Yuan, 2021).

Focused MNAR as described by Gomer and Yuan (2021) is when missingness depends only on the unobserved values $\mathbf{Y}_{\text{miss}}$ and not on the observed values in the data. For example, a focused MNAR process may occur when a responder drops out from a depression study because she already feels better, irrespective of any other variable in the study.

$$f(\boldsymbol{M} = 1 | \mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{miss}}\,, \phi) = f(\boldsymbol{M} = 1 | \mathbf{Y}_{\text{miss}}\,) \tag{5}$$

This is represented in Equation 5 where $\mathbf{Y}$ is a depression score and $\boldsymbol{M}$ is a missingness indicator. Equation 5 is simply saying that missingness is only conditional on one’s unseen depression score. Like diffuse MNAR, focused MNAR considers the interdependency of the unobserved or unmeasured outcome values by explicitly modeling the missingness process (Gomer & Yuan, 2021).

While it may be unlikely that a focused MNAR process accurately represents reality in a broad sense, the categorization of MNAR into distinct systems of missingness can still provide valuable insights for researchers. Firstly, it helps conceptualize the complexity of the missingness model as a spectrum, ranging from a simple model with just one predictor to a more intricate model with numerous predictors and potentially higher-order effects. This conceptualization proves particularly useful when conducting sensitivity analyses, where the examination of estimates across different missingness models is crucial, and when approaching MNAR analyses as a process of model development. In both scenarios, a focused MNAR serves as the initial step. Secondly, categorizing MNAR is beneficial due to the inherent challenges involved in implementing and estimating these models, which only intensify as more variables are incorporated into the missingness model. Even if a focused MNAR subtype may not accurately capture the true underlying missingness process, it may be the only estimable model in practical terms. Finally, MNAR subtypes play a pivotal role in the proposed simulation study, as this dissertation aims to investigate the behavior of variable selection methods under increasing levels of complexity in the true missingness model. The goal is to create scenarios that range from straightforward estimation of the model to scenarios where estimation becomes difficult or even impossible.

Fully-Bayesian Factored Regression Approach for MAR

This study will use a factored regression estimation procedure that tailors the distributions of missing values around the substantive analysis model. Missing data analysis using factored regression estimation has gain popularity recently because it allows for mixtures of categorical variables, continuous variables, interactions, non-linear terms, random coefficients, and skewed continuous variables. (Bartlett et al., 2015; Du et al., 2021; Enders et al., 2020; Erler et al., 2016; Lüdtke et al., 2020; Zhang & Wang, 2017). Although I could potentially deal with missing data by implementing full information maximum likelihood (FIML) or multiple imputation (MI), I will instead use a fully Bayesian factored regression approach for missing data handling. As noted previously, a fully Bayesian approach is appealing because it can implement two important tasks jointly in one estimation procedure; it can obtain inferences about the substantive parameters of interest, and it can perform variable selection by identifying relevant subsets of features that can predict missingness. This section will introduce the estimation of fully-Bayesian factored regression model using the MAR assumption that omits the missingness model. A subsequent section expands the approach to accommodate MNAR missingness.

To illustrate a factored regression approach, I will use an example from Enders (2022) where the substantive model has three predictors,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i = E(Y_i|\boldsymbol{X_i}) + \varepsilon_i \tag{6}$$

$\boldsymbol{X_i} = (X_{1i}, X_{2i}, X_{3i})$, all three covariates have missing observations. The term $E(Y_i|\boldsymbol{X_i})$ in Equation 6 is a predicted value and ε_i is the residual. I can express the multivariate distribution in the above equation as the product of four univariate conditional distributions by using the probability chain rule, where each univariate distribution corresponds to a regression model. As a reminder, the term $f(\cdot)$ just means that I am referencing a conditional distribution implied by regression model. The factored regression specification for the three-predictor model is as follows.

$$\begin{gathered} f(\boldsymbol{Y}, \boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3}) = f(\boldsymbol{Y}|\boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3}) \times f(\boldsymbol{X_1}|\boldsymbol{X_2}, \boldsymbol{X_3}) \\ \times f(\boldsymbol{X_2}|\boldsymbol{X_3}) \times f(\boldsymbol{X_3}) \end{gathered} \tag{7}$$

The joint distribution can be factorized using several different variable orders, and the key is to sequence the variables such that one of the conditional distributions aligns with the substantive analysis model. Accordingly, the first term after the equal sign in Equation 7 is the conditional distribution for the outcome $\boldsymbol{Y}$ given the predictors, and the remaining terms are the conditional distributions of the predictors. Assuming that all variables are continuous, below are the regression models that correspond to the conditional distributions of the predictors.

$$\begin{gathered} X_{1i} = \gamma_{01} + \gamma_{11}(X_{2i}) + \gamma_{21}(X_{3i}) + r_{1i} \\ X_{2i} = \gamma_{02} + \gamma_{12}(X_{3i}) + r_{2i} \\ X_{3i} = \gamma_{03} + r_{3i} \end{gathered} \tag{8}$$

The notation $\boldsymbol{\gamma}$ in the predictor's linear model are regression coefficient and $\boldsymbol{r}$ are residuals. Even though all models in our examples so far are linear, they can also contain interaction and non-linear effects, and the variables need not be normal or continuous.

When the predictors ($\boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3}$) are normally distributed, it is possible to simplify the previous full factorization by setting some of the coefficients to zero in the model. The factorization can be expressed as:

$$f(\boldsymbol{Y}, \boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3}) = f(\boldsymbol{Y}|\boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3}) \times f(\boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3}) \tag{9}$$

In this factorization, the terms $f(\boldsymbol{X_1}|\boldsymbol{X_2}, \boldsymbol{X_3}) \times f(\boldsymbol{X_2}|\boldsymbol{X_3}) \times f(\boldsymbol{X_3})$ have been dropped because it is assumed that $\boldsymbol{X_1}, \boldsymbol{X_2}$, and $\boldsymbol{X_3}$ are complete variables without any missing data. Therefore, they do not contribute to the missingness and can be excluded from the factorization. For the purpose of this dissertation, the factorization used will be the simplified version mentioned above, as only the outcome variable $\boldsymbol{Y}$ will have missing values.

Distribution of the Missing Outcome and Regressor Under MAR

Next, I will illustrate the distribution of the missing outcome Y_i and the distribution of the covariates X_{1i}, X_{2i}, X_{3i} by using factored regressions, which just means that I will break up the joint distribution into a product of conditional regression models. I denote the person-specific parameters by using the subscript i, where $i = (1, \ldots, n)$ and n is the sample size. Again, I will assume that outcome Y_i and all covariates in $\boldsymbol{X_i} = (X_{1i}, X_{2i}, X_{3i})$ have missing values to illustrate this example.

First, the distribution of the missing outcome Y_i is

$$f(Y_i|\boldsymbol{\beta}, \sigma_\varepsilon^2, \boldsymbol{X_i}) = N(E(Y_i|\boldsymbol{X_i}), \sigma_\varepsilon^2) \tag{10}$$

where this distribution is a normal distribution centered at the predicted value $E(Y_i|\boldsymbol{X_i})$ and the residual variance σ_ε^2 defines its spread. Drawing an observation from this distribution is equivalent to adding the predicted value and a random noise term from a normal distribution (Enders, 2022).

In contrast, the distribution of an incomplete regressor (X_{1i}, X_{2i}, X_{3i}) is more complicated because it must account for every model in which it appears. For example, $\boldsymbol{X_1}$ appears on the right side of the substantive regression in Equation 6 and on the left side of its own model in Equation 8. Therefore, the distribution of X_{1i} conditions on the other analysis variables via two sets of model parameters. Applying Bayes' theorem gives the following expression for the distribution of missing $\boldsymbol{X_1}$ scores.

$$f(X_{1i}|Y_i, X_{2i}, X_{3i}) \propto f(Y_i|X_{1i}, X_{2i}, X_{3i}) \times f(X_{1i}|X_{2i}, X_{3i}) \tag{11}$$

In words, Equation 11 says that the model-implied distribution of $\boldsymbol{X_1}$ is found by multiplying two univariate distributions, each of which aligns with a regression model.

$$\begin{gathered} f(Y_i|X_{1i},X_{2i},X_{3i}) \times f(X_{1i}|X_{2i},X_{3i}) = \\ N(E(Y_i|\boldsymbol{X_i}),\sigma_\varepsilon^2) \times N\left(E(X_{1i}|X_{2i},X_{3i}),\sigma_{r_1}^2\right) \end{gathered} \tag{12}$$

Equation 12 above specifies that the two univariate distributions are normal densities. The first distribution to the right of the equal sign is the exact same outcome distribution described in Equation 10, centered at the predicted value and σ_ε^2 defining its spread. The second distribution is centered at the predicted score for X_{1i} and its spread is defined by the residual variance $\sigma_{r_1}^2$.

If I drop any unnecessary scaling terms from the normal distribution functions, the conditional distribution of the X_{1i} regression is proportional to:

$$\begin{aligned} &\exp\left(-\frac{1}{2}\,\frac{(Y_i-(\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\beta_3X_{3i}))^2}{\sigma_\varepsilon^2}\right) \\ &\qquad\qquad \times \exp\left(-\frac{1}{2}\,\frac{(X_{1i}-(\gamma_{01}+\gamma_{11}X_{2i}+\gamma_{21}X_{3i}))^2}{\sigma_{r_1}^2}\right) \end{aligned} \tag{13}$$

Equation 13 shows the kernels of two normal curve functions, the first kernel depends on the substantive model parameters and the second function depends on the regressor model parameters.

Next, I can compute the product of the two normal curve functions and analytically combine them into a single distribution for $\boldsymbol{X_1}$.

$$f(X_{1i}|Y_i, X_{2i}, X_{3i}) = N\big(E(X_{1i}|Y_i, X_{2i}, X_{3i}), var(X_{1i}|Y_i, X_{2i}, X_{3i})\big) \tag{14}$$

$$E(X_{1i}|Y_i, X_{2i}, X_{3i}) =$$

$$var(X_{1i}|Y_i, X_{2i}, X_{3i}) \times \left(\frac{\gamma_{01} + \gamma_{11}X_{2i} + \gamma_{21}X_{3i}}{\sigma^2_{r_1}} + \frac{\beta_1(\, Y_i - \beta_0 - \beta_2 X_{2i} + \beta_3 X_{3i})}{\sigma^2_{\varepsilon}}\right)$$

$$var(X_{1i}|Y_i, X_{2i}, X_{3i}) = \left(\frac{1}{\sigma^2_{r_1}} + \frac{\beta_1^2}{\sigma^2_{\varepsilon}}\right)^{-1}$$

Now, the conditional distribution of $\boldsymbol{X_1}$ is represented as a normal distribution with two mean and variance expressions that are directly linked to the substantive and regressor model parameters. Although complicated, it is still manageable to analytically derive the predictive distribution of the regressor when it only appears in two models. Levy and Enders (2021) give general expressions for the missing data distributions, and specialized MCMC methods that approximate sampling from a complex multi-part distributions are also available (e.g., Metropolis–Hastings algorithm; Gilks et al., 1995; Hastings, 1970).

Auxiliary Variables

A standard recommendation is to include auxiliary variables that are not in the analysis model because it can reduce non-response bias, improve precision, or both (Collins et al., 2001; Graham, 2003, 2009). Fortunately, adding auxiliary variables using a factored regression specification is straightforward. To illustrate an example, I will add two auxiliary variables $\boldsymbol{A_1}$ and $\boldsymbol{A_2}$ to the three-predictor example. The factorization with auxiliary variables is as follows.

$$\begin{aligned} f(\boldsymbol{Y}, \boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3}, \boldsymbol{A_1}, \boldsymbol{A_2}) \\ &= f(\boldsymbol{A_1}|\boldsymbol{Y}, \boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3}, \boldsymbol{A_2}) \times f(\boldsymbol{A_2}|\boldsymbol{Y}, \boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3}) \\ &\times (\boldsymbol{Y}|\boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3}) \times f(\boldsymbol{X_1}|\boldsymbol{X_2}, \boldsymbol{X_3}) \times f(\boldsymbol{X_2}|\boldsymbol{X_3}) \times f(\boldsymbol{X_3}) \end{aligned} \tag{15}$$

The first two terms after the equal sign are auxiliary variable distributions that derive from regression models, the third term corresponds to the substantive analysis, and the final three terms are regressions for the incomplete predictors. Equation 15 specifies that the analysis variables predict the auxiliary variable, with the purpose of maintaining the desired interpretation of the substantive model parameters (i.e., placing auxiliary variables after $\boldsymbol{Y}$ would specify the additional variables as predictors, thus changing the meaning of the model parameters). The auxiliary variable regressions are shown below

$$\begin{aligned} A_{1i} &= \gamma_{04} + \gamma_{14}(Y_i) + \gamma_{24}(X_{1i}) + \gamma_{34}(X_{2i}) + \gamma_{44}(X_{3i}) + \gamma_{54}(A_{2i}) + r_{4i} \\ A_{2i} &= \gamma_{05} + \gamma_{15}(Y_i) + \gamma_{25}(X_{1i}) + \gamma_{35}(X_{2i}) + \gamma_{45}(X_{3i}) + r_{5i} \end{aligned} \tag{16}$$

MCMC Estimation

Finally, I illustrate a generic MCMC algorithm to estimate all parameters of interest. MCMC sampling is an iterative process in which a sequence of random variables is estimated by sampling values at random from a posterior distribution that depends on the previous drawn value. The Gibbs sampler, introduced by Geman and Geman (1984), is an MCMC method that breakdowns a complex multivariate distribution into a series of univariate problems. The Gibbs sampler iteratively estimates one parameter at a time while holding all other parameters constant.

The estimation recipe below shows the MCMC algorithmic steps where all analysis variables could be missing

0) Assign starting values to all parameters and missing values.

    a. Do for $t = 1$ to T iterations.

1) Estimate the regression coefficients of every regression model conditional on its residual variance and all imputations.

2) Estimate every regression model's residual variance conditional on its coefficients and all imputations.

3) Estimate each incomplete variable from a distribution, the shape of which depends on every model in which that variable appears

4) Repeat

The MCMC steps above estimate unknown parameters by drawing values at random from a probability distribution. First, it estimates the regression model coefficients from a multivariate normal distribution, conditioned on the current value of the residual variance at the tth iteration. Step 2 in the MCMC process then updates the value of the residual variance from a positively skewed inverse gamma distribution, given the updated regression coefficients and the filled-in data. Step 3 of the MCMC step estimates the incomplete variables based on the updated parameter values from step 1 and 2. The probability distribution for every incomplete variable is composed of every model in which that variable appears as shown in Equations 11-14. The conditional posterior distributions of the regression model parameters have a standard form and are widely available in the literature (Gelman et al., 1995; Levy & Enders, 2021; Lynch, 2020). I also describe these in more detail later in the paper.

Modeling frameworks for MNAR

The two major modeling frameworks for MNAR processes are selection models and pattern mixture models (Hedeker & Gibbons, 1997; Kenward, 1998; Little, 2008; Little & Wang, 1996; Michiels et al., 1999, 2002; Ratitch et al., 2013). These two frameworks decrease bias by introducing a model that describes the missingness process, but they do it in different manners. The pattern mixture model uses the binary missingness indicator as a predictor/moderator to form qualitatively distinct subgroups based on different missing data patterns, where every subgroup has its own parameter values (Fitzmaurice et al., 2008, pp. 409-431; Hedeker & Gibbons, 1997; Little, 1993). In contrast, the selection model uses the binary indicator as the outcome of the missingness model and directly models the relationship between the probability for missingness and its unobserved score (Galimard et al., 2016; Heckman, 1976, 1979; Ratitch et al., 2013).

To illustrate a pattern mixture model and a selection model, I will present an example of both models where outcome $\boldsymbol{Y}$ has missing observations and covariates $\mathbf{X}$ are all complete. Using factored regression specification, the pattern-mixture model factors the joint distribution of the outcome being measured and missingness mechanisms into the following set of univariate functions.

$$f(\boldsymbol{Y}, \boldsymbol{M}, \mathbf{X}) = f(\boldsymbol{Y}|\boldsymbol{M}, \mathbf{X}) \times f(\mathbf{X}|\boldsymbol{M}) \times f(\boldsymbol{M}) \tag{17}$$

The first term shows the missing data indicator as a predictor in the substantive model, where the distribution of the outcome is dependent on the missing data pattern. The third term $f(\boldsymbol{M})$ is a model that describes the pattern proportions.

The selection model simultaneously estimates two models, (1) the probability of a variable being missing, which includes the outcome variables and a set of covariates as predictors, and (2) the substantive model. Using a factored regression specification, I can factorize the joint distribution of the outcome $\boldsymbol{Y}$, missingness indicator $\boldsymbol{M}$, and covariates $\mathbf{X}$ into a sequence of univariate functions.

$$f(\boldsymbol{Y}, \boldsymbol{M}, \mathbf{X}) = f(\boldsymbol{M} | \boldsymbol{Y}, \mathbf{X}) \times f(\boldsymbol{Y} | \mathbf{X}) \times f(\mathbf{X}) \tag{18}$$

The first term in Equation 18 corresponds to a probit or logistic regression with the variables from the substantive model predicting the binary missing data indicator $\boldsymbol{M}$. The second term $f(\boldsymbol{Y}|\mathbf{X})$ is the substantive model, and the third term is the marginal distribution of $\mathbf{X}$. Even though differences in frameworks are evident (e.g., the missing data indicator functions as an outcome in one framework and a predictor in the other), they both require strict and unverifiable assumptions about data. Given that only some of the outcome values are observed, the parameters capturing the dependence between missingness and the outcomes cannot be fully estimated from the observed data and must be informed, at least partially, by information not specified by the substantive model. MNAR restrictions are untestable and if misspecified, it could produce estimates that contain more bias than those from a MAR analysis (Enders, 2022; Galimard et al., 2016, 2018; Molenberghs et al., 2004).

This dissertation is only interested in studying BVS under selection models and I will not investigate pattern-mixture models. As described earlier in this section, selection models estimate a missingness model with a missing data indicator $\boldsymbol{M}$ as its outcome. The main challenge of estimating a selection model is to find which variables to use in the missingness model that best explain missingness while avoiding too much overlap among the predictor sets in the two models (Ogundimu & Collins, 2019; Sartori, 2003). The literature refers to an *exclusion restriction* when a variable appears in the substantive model but not the missingness regression (Marra et al., 2017; Toomet & Henningsen, 2008). In general, the search for exclusion restrictions is important because it facilitates the estimation of selection models, which are prone to convergence failures in the absence of such restrictions.

BVS has a clear application for selection models because it can shrink unnecessary variables close to zero while simultaneously selecting variables that are associated with the missingness model. However, the application of BVS to pattern-mixture models is more challenging. Pattern-mixture models explicitly incorporate different missingness patterns, each with its own set of parameters. This increased complexity poses difficulties for directly applying BVS techniques due to the large number of parameters and potential model combinations that need to be considered. As a result, BVS for pattern-mixture models requires careful consideration and potentially modified approaches to address the increased complexity.

Fully-Bayesian Selection Model

Selection models estimate the relationship between the probability of missingness and the data by pairing the substantive analysis with an additional regression that models the underlying cause of missingness (Du et al., 2021; Heckman, 1979; Michiels et al., 1999; Ogundimu & Collins, 2019). Logistic and probit regression are both options to analyze the binary indicator for missingness. The current study will use a probit regression model because it is computationally simpler. The probit model represents the binary response as an latent and continuous normal distribution which denotes the propensity for missingness (Agresti, 2003; Albert & Chib, 1993; Johnson & Albert, 2006). This latent distribution has a threshold parameter φ that separates the normal distribution of latent scores into two non-overlapping sections, one for missing and the other section for observed data.

$$M_i = \begin{cases} 0 \ if \ M_i^* \leq \varphi \\ 1 \ if \ M_i^* > \varphi \end{cases} \tag{19}$$

Equation 19 shows the link between M_i, the observed missing data indicator for individual i, and M_i^*, the latent missing data variable for individual i. The threshold parameter φ is what divides the latent response distribution of $\boldsymbol{M}^*$ into sections for $\boldsymbol{M} = 1$ (missing observations) and $\boldsymbol{M} = 0$ (present observations).

As described earlier, the selection model uses two regressions, one for the substantive model and one for the missingness model. The description of the substantive and a missingness model for individual i are below:

$$\begin{gathered} Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i = E\left(Y_i \middle| X_{1i,} X_{2i,} X_{3i}\right) + \varepsilon_i \\ Y_i \sim N(E(Y_i|\boldsymbol{X_i}), \sigma_\varepsilon^2) \end{gathered} \tag{20}$$

$$\begin{gathered} M_i^\star = \gamma_0 + \gamma_1 Y_i + \gamma_2 X_{1i} + \gamma_3 X_{2i} + \gamma_4 X_{3i} + r_i \\ M_i^* \sim N(E(M_i^*|Y_i, \boldsymbol{X_i}), \sigma_r^2) \ I(Q_i) \\ \sigma_r^2 = 1 \end{gathered} \tag{21}$$

where $\boldsymbol{\beta}$ = ($\beta_0, \beta_1, \beta_2, \beta_3$) in Equation 20 are the estimable coefficients of the substantive model. The term σ_ε^2 is the variance and ε_i is the error term of the substantive regression. Equation 21 shows the missingness model, where the outcome $M_i^\star$ is a value from the normally distributed latent variable for individual i. Note that the residual variance σ_r^2 is set to 1 because we only observe its sign (whether the latent score is above or below the threshold of 0). The identifiability of $M_i^*$is not possible in Equation 21 (Albert & Chib, 1993; Du et al., 2021). The term $I(\cdot)$ in Equation 21 is an indicator function, Q_i is either equal to $\{M_i^* > \varphi\}$ or $\{M_i^* \leq \varphi\}$ corresponding to $M_i = 1$ or $M_i = 0$. The missingness equation uses $\boldsymbol{\gamma}$ to represent the probit regression coefficients, where $\boldsymbol{\gamma}$ = $(\gamma_0, \gamma_1, \dots, \gamma_4)$. In this example, the predictors in the missingness model are composed of the outcome variable Y_i and the same covariates from the substantive model. However, the missingness model can include all, some, or none of the covariates from the substantive model and it can also include variables that do not appear in the substantive model.

When constructing a missingness model, an important consideration is whether to include overlapping predictors from the substantive analysis (Ibrahim et al., 2005). Equation 21 is an example of a missingness model that includes all variables from the substantive model, where both regressions share the same predictors apart from $\boldsymbol{Y}$. If I were to try to estimate this selection model, I could encounter difficulties with convergence because the normality assumption alone identifies the model (Leung & Yu, 2000; Little & Rubin, 2019; Puhani, 2000). A common recommendation is then to eliminate shared redundancies between the two models (Ibrahim et al., 2005; Sartori, 2003); this could be achieved by using a subset of predictors from the substantive model or by adding an auxiliary variable to the missingness model that is unrelated to the substantive outcome. However, as Sartori (2003) noted, this practice can often turn into a "mad" search for an exclusion restriction which can lead to including unnecessary variables in the missingness model that do not satisfy exclusion restriction criteria (Bärnighausen et al., 2011). In addition, it is of practical importance to avoid including too many variables in the missingness model, as this could also translate into convergence issues (Ibrahim et al., 2005). Again, our focus is to improve estimation and reduce biases in the substantive model by investigating BVS as a data-driven option to simultaneously select variables that predict missingness, thereby decreasing the complexity of the missingness model.

Distribution of the Missing Outcome and Regressor Under MNAR

Next, I will illustrate the joint distribution of the substantive and missingness model variables by using factored regressions. To illustrate this example, I will assume that the outcome and covariates all have missing values. Using a factored specification, I can represent the joint distribution of the variables from both the substantive and missingness models into the following sequence of univariate functions:

$$\begin{aligned} f(\boldsymbol{Y},\boldsymbol{M},\mathbf{X}) = (\boldsymbol{M}|\boldsymbol{Y},\boldsymbol{X_1},\boldsymbol{X_2},\boldsymbol{X_3}) \times\ f(\boldsymbol{Y}|\boldsymbol{X_1},\boldsymbol{X_2},\boldsymbol{X_3}) \times f(\boldsymbol{X_1}|\boldsymbol{X_2},\boldsymbol{X_3}) \\ \times f(\boldsymbol{X_2}|\boldsymbol{X_3})\ \times\ f(\boldsymbol{X_3}) \end{aligned} \tag{22}$$

The first term to the right of the equals sign corresponds to the missingness model, the second term is the substantive model, the third and fourth terms are the conditional distribution for the first and second covariate, and the last term is the marginal distribution of the third covariate. The distributions of the incomplete covariates are composed by the product of all the univariate distributions in which it appears. In the case where a substantive covariate is also part of the missingness model, then the univariate distribution from the missingness model is also included in the distribution of its missing values (Du et al., 2021; Enders, 2022) For example, the distribution of $\boldsymbol{X_1}$ depends on three univariate distributions, each of which aligns with the first three models from Equation 22: the missingness model, the substantive model, and its own conditional distribution.

In contrast to the Equation 10 in the MAR section, the distribution of the outcome variable ***Y*** is more complex, as it must account for every model in which it appears (Du et al., 2021; Enders, 2022). Equation 22 shows ***Y*** as a variable in both the missingness and substantive model, therefore the distribution of the outcome's unobserved values is found by multiplying two univariate distributions, each of which aligns with the regression from Equation 20 and 21. If I drop any unnecessary scaling terms from the normal distribution functions, the model-implied distribution of Y_i is proportional to:

$$\begin{gathered} f(M_i^*|Y_i, \boldsymbol{X_i}) \times f(Y_i|\boldsymbol{X_i}) \propto \\ \exp\left(-\frac{1}{2}\,\frac{\left(M_i^* - (\gamma_0 +\ \gamma_1 Y_i +\ \gamma_2 X_{1i} + \gamma_3 X_{2i} + \gamma_4 X_{3i})\right)^2}{\sigma_{r_1}^2}\right) \\ \times \exp\left(-\frac{1}{2}\,\frac{(Y_i - (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i}))^2}{\sigma_\varepsilon^2}\right) \end{gathered} \tag{23}$$

where each of the two normal curve functions in Equation 23 depends on the substantive and missingness model parameters. As a side note, the selection model's residual variance $\sigma_{r_1}^2$ is fixed to 1, however I include $\sigma_{r_1}^2$ as a term in Equation 23 to keep the same structure as the regressor distribution in Equation 13 in the MAR section.

I can now multiply the two normal curve functions from Equation 23 and analytically combine them into a single distribution for Y_i. This results in a normal distribution with a mean and variance that depend on the substantive and missingness model parameters.

$$f(Y_i|M^*, \boldsymbol{X_i}) = N\big(E(Y_i|M_i^*, \boldsymbol{X_i}), var(Y_i|M_i^*, \boldsymbol{X_i})\big)$$

$$E(Y_i|M_i^*, \boldsymbol{X_i}) = var(Y_i|M_i^*, \boldsymbol{X_i}) \times$$

$$\left(\frac{\gamma_1(M_i^* + \gamma_0 + \gamma_2 X_{1i} + \gamma_3 X_{2i} + \gamma_4 X_{3i})}{\sigma_{r_1}^2} + \frac{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i}}{\sigma_\varepsilon^2}\right) \tag{24}$$

$$var(Y_i|M_i^*, \boldsymbol{X_i}) = \left(\frac{\gamma_1^2}{\sigma_{r_1}^2} + \frac{1}{\sigma_\varepsilon^2}\right)^{-1}$$

Auxiliary Variables

Selection models can easily extend to include auxiliary variables as predictors by using factored regression.

$$\begin{aligned} f(\boldsymbol{Y}, \boldsymbol{M}, \mathbf{X}, \mathbf{A}) = (\boldsymbol{M}|\boldsymbol{Y}, \boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3}, \boldsymbol{A_1}, \boldsymbol{A_2}) \times f(\boldsymbol{A_1}|\boldsymbol{Y}, \boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3}, \boldsymbol{A_2}) \\ \times f(\boldsymbol{A_2}|\boldsymbol{Y}, \boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3}) \times f(\boldsymbol{Y}|\boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3}) \times f(\boldsymbol{X_1}|\boldsymbol{X_2}, \boldsymbol{X_3}) \\ \times f(\boldsymbol{X_2}|\boldsymbol{X_3}) \times f(\boldsymbol{X_3}) \end{aligned} \tag{25}$$

The first term in Equation 25 is the missingness model, where the binary indicator $\boldsymbol{M}$ is conditional on the substantive model predictors $\mathbf{X}$ and two auxiliary variables $\boldsymbol{A_1}$ and $\boldsymbol{A_2}$. The second and third terms are the conditional distribution of the two auxiliary variables. Notice that like the MAR section, I specified a sequence where the analysis variables predict the auxiliary variables to maintain the desired interpretation of the substantive model parameters.

After including auxiliary variables in Equation 25, the distribution of the $\boldsymbol{Y}$ variable now appears in four models, as the outcome in the substantive model and as a predictor of the missingness indicator and the auxiliary variables. In contrast to Equation 23, where the distribution of the dependent variable is a product of two distributions, now the distribution of $\boldsymbol{Y}$

is a multi-part function that depends on the product of four univariate distributions. In general, when an auxiliary variable is included in the missingness model, the distribution of the outcome variable is far more complex, however, MCMC methods can also be used to approximate a complex multi-part distribution of missing values.

Collins et al. (2001) define the auxiliary variables in Equation 25 as Type A auxiliary variables because they both predict missingness and correlate with the analysis variables. In contrast, Type B auxiliary variables correlates with the analysis and not with the missingness indicator. An example of Type B auxiliary variables is found below,

$$\begin{aligned} f(\boldsymbol{Y},\boldsymbol{M},\mathbf{X},\mathbf{A}) = (\boldsymbol{M}|\boldsymbol{Y},\boldsymbol{X_1},\boldsymbol{X_2},\boldsymbol{X_3}) \times f(\boldsymbol{A_1}|\boldsymbol{Y},\boldsymbol{X_1},\boldsymbol{X_2},\boldsymbol{X_3},\boldsymbol{A_2}) \\ \times f(\boldsymbol{A_2}|\boldsymbol{Y},\boldsymbol{X_1},\boldsymbol{X_2},\boldsymbol{X_3}) \times f(\boldsymbol{Y}|\boldsymbol{X_1},\boldsymbol{X_2},\boldsymbol{X_3}) \times f(\boldsymbol{X_1}|\boldsymbol{X_2},\boldsymbol{X_3}) \\ \times f(\boldsymbol{X_2}|\boldsymbol{X_3}) \times f(\boldsymbol{X_3}) \end{aligned} \tag{26}$$

where the auxiliary variables $\boldsymbol{A_1}$ and $\boldsymbol{A_2}$ in the equation above are not correlated with the missingness indicator $\boldsymbol{M}$. Finally, Type C auxiliary variables correlate with the missing data indicator but not with substantive model variables.

$$\begin{aligned} f(\boldsymbol{Y},\boldsymbol{M},\mathbf{X},\mathbf{A}) = (\boldsymbol{M}|\boldsymbol{Y},\boldsymbol{X_1},\boldsymbol{X_2},\boldsymbol{X_3},\boldsymbol{A_1},\boldsymbol{A_2}) \times f(\boldsymbol{Y}|\boldsymbol{X_1},\boldsymbol{X_2},\boldsymbol{X_3}) \\ \times f(\boldsymbol{X_1}|\boldsymbol{X_2},\boldsymbol{X_3}) \times f(\boldsymbol{X_2}|\boldsymbol{X_3}) \times f(\boldsymbol{X_3}) \end{aligned} \tag{27}$$

As noted previously, past literature suggests that estimation problems may occur when the missingness model includes the same predictors from the substantive model or any other variables that are highly correlated with $\boldsymbol{Y}$ (Puhani, 2000; Stolzenberg & Relles, 1990, 1997). To avoid this concern, a sensible recommendation is to include auxiliary variables in the missingness model that correlate with the missing data indicator but not the substantive model outcome (Puhani, 2000; Toomet & Henningsen, 2008; Vella, 1998). Auxiliary variables of Type A and B would not be considered true exclusion restrictions and will not alleviate collinearity and can lead to specification error (Ogundimu, 2022). In contrast, adding Type C auxiliary variables in the selection model can reduce non-response bias and improve precision (Galimard et al., 2018; Marra et al., 2017).

MCMC Estimation

Now I will describe the MCMC estimation steps for the case where the outcome in the substantive model is MNAR and the missingness model also contain both covariates and auxiliary variables as predictors. The MCMC steps and posterior distributions come from Du et al. (2021).

MCMC steps:

0) Initialization step: set initial values for $\boldsymbol{\beta}^{(0)}, \sigma_{\varepsilon}^{2^{(0)}}, \boldsymbol{\gamma}^{(0)}, and\ M_i^{*(0)}$, and $Y_{i(miss)}^{(0)}$for the individuals who have missing outcome, and for the individuals who have missing covariates and auxiliary variables, set initial values

1) In the t^{th} iteration, considering the covariates in the substantive model, the imputed outcomes from the previous iteration (Y_i^{t-1}), and the residual variance of the substantive model in the previous iteration $(\sigma_{\varepsilon}^{2^{(t-1)}})$, I sample $\boldsymbol{\beta^t}$ from a multivariate normal distribution (MN). In this distribution, the variables and parameters are indicated by the symbol "·" and are conditioned on. The mean and covariance matrix of this distribution correspond to the ordinary least squares estimates of the coefficients and their parameter covariance, respectively.

$$f(\boldsymbol{\beta}|\cdot) = MN\left(\widehat{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}\right)$$

$$where\ \widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{Y} \tag{28}$$

$$and\ \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}} = \sigma_{\varepsilon}^{2}(\mathbf{X}'\mathbf{X})^{-1}$$

2) Given $\mathbf{X}$, Y_i^{t-1}, and $\boldsymbol{\beta}^{\boldsymbol{t}}$, we define $1/\sigma_{\varepsilon}^{2}$ as a gamma random variable and draw the reciprocal of residual variance (i.e., the precision) from a right-skewed gamma distribution. The terms df and S are hyperparameters from the prior. The shape parameter $\frac{N+df}{2}$ determines the height of the distribution, which in turn affects its skewness and heavy-tailed properties. The spread of the distribution is determined by the sum of squared residuals from the previous iteration, adjusted by the hyperparameter of the prior distribution S.

$$f(1/\sigma_{\varepsilon}^{2}|\cdot) = Gamma\left(\frac{N+df}{2}, \frac{(\boldsymbol{Y}-\boldsymbol{X\beta})'(\boldsymbol{Y}-\boldsymbol{X\beta})+S}{2}\right) \tag{29}$$

3) Using the imputed outcomes from the previous iteration Y_i^{t-1}, along with the other predictors in the missingness model (excluding the outcome variable), and the latent data $M_i^{*(t-1)}$, a sample of $\boldsymbol{\gamma}^{\boldsymbol{t}}$ is generated from a multivariate normal distribution. The parameters of this distribution, including the mean and variance, are determined by the latent data $\boldsymbol{M}^{*}$ and the current imputed data. The term $\boldsymbol{I}$ is an identity matrix and b is the variance of the prior distribution.

$$f(\boldsymbol{\gamma}|\cdot) = MN(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\gamma}}})$$

$$where\ \hat{\gamma} = \ \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\gamma}}}{}^{-1}\mathbf{Z}'\boldsymbol{M}^{*}$$

$$\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\gamma}}} = \left(\frac{1}{b} \times I + \mathbf{Z}'\mathbf{Z}\right)^{-1} \tag{30}$$

$$\mathbf{Z} = (\boldsymbol{Y}, \mathbf{X}, \mathbf{A})$$

4) The distribution of the latent variable M_i^* can now be defined. For individual i, considering their observed M_i, the imputed outcome $Y_i^{(t-1)}$, and the sampled values of $\boldsymbol{\gamma}^{(\boldsymbol{t})}$, I sample $M_i^{*(t)}$ from the equation below. In this equation, the latent variable scores are modeled using a truncated normal distribution. The indicator function $I()$ is used, and $\boldsymbol{Z_i}$ represents the *ith* row of $\mathbf{Z}$.

$$f(M_i^*|\cdot) = \begin{cases} N(\boldsymbol{Z_i}\boldsymbol{\gamma}, \sigma_r^2)\ I(M_i^* > \varphi)\ M_i = 1 \\ N(\boldsymbol{Z_i}\boldsymbol{\gamma}, \sigma_r^2)\ I(M_i^* \le \varphi)\ M_i = 0 \end{cases} \tag{31}$$

5) For every individual i who has missing outcome (i.e., $M_i = 1$), given the covariates and/or auxiliary variables, $\boldsymbol{\beta}^{(\boldsymbol{t})}$, $\sigma_\varepsilon^{2(t)}$, $\boldsymbol{\gamma}^{(\boldsymbol{t})}$, and $M_i^{*(t)}$, sample $Y_{i(miss)}^{(t)}$ from

$$\begin{aligned} f\left(\boldsymbol{Y_{(miss)}}|\cdot\right) \propto f(\boldsymbol{M}|\boldsymbol{Y}, \mathbf{X}, \mathbf{A}) \times f(\boldsymbol{A_1}|\boldsymbol{Y}, \mathbf{X}, \boldsymbol{A_2}) \times\ f(\boldsymbol{A_2}|\boldsymbol{Y}, \mathbf{X}) \\ \times\ f(\boldsymbol{Y}|\mathbf{X}) \end{aligned} \tag{32}$$

6) For the individual i who has a missing covariate or auxiliary variable, given $\boldsymbol{\beta}^{(\boldsymbol{t})}$, $\sigma_\varepsilon^{2(t)}$, $\boldsymbol{\gamma}^{(\boldsymbol{t})}$, $M_i^{*(t)}$, and $Y_{i(miss)}^{(t)}$ sample from

$$
\begin{aligned}
f\left(\boldsymbol{X}_{\mathbf{3}(\boldsymbol{miss})} \mid \cdot\right) &\propto f(\boldsymbol{M}|\boldsymbol{Y}, \mathbf{X}, \mathbf{A}) \times f(\boldsymbol{A}_{\mathbf{1}}|\boldsymbol{Y}, \mathbf{X}, \boldsymbol{A}_{\mathbf{2}}) \times \ldots \times f(\boldsymbol{Y}|\mathbf{X}) \\
&\quad \times f(\boldsymbol{X}_{\mathbf{1}}|\boldsymbol{X}_{\mathbf{2}}, \boldsymbol{X}_{\mathbf{3}}) \times \ \ldots \times f(\boldsymbol{X}_{\mathbf{3}}) \\
f\left(\boldsymbol{X}_{\mathbf{2}(\boldsymbol{miss})} \mid \cdot\right) &\propto f(\boldsymbol{M}|\boldsymbol{Y}, \mathbf{X}, \mathbf{A}) \ \times \ f(\boldsymbol{A}_{\mathbf{1}}|\boldsymbol{Y}, \mathbf{X}, \boldsymbol{A}_{\mathbf{2}}) \times \ldots \times f(\boldsymbol{Y}|\mathbf{X}) \\
&\quad \times \ f(\boldsymbol{X}_{\mathbf{1}}|\boldsymbol{X}_{\mathbf{2}}, \boldsymbol{X}_{\mathbf{3}}) \ \times \ f(\boldsymbol{X}_{\mathbf{2}}|\boldsymbol{X}_{\mathbf{3}}) \\
f\left(\boldsymbol{X}_{\mathbf{1}(\boldsymbol{miss})} \mid \cdot\right) &\propto f(\boldsymbol{M}|\boldsymbol{Y}, \mathbf{X}, \mathbf{A}) \ \times \ f(\boldsymbol{A}_{\mathbf{1}}|\boldsymbol{Y}, \mathbf{X}, \boldsymbol{A}_{\mathbf{2}}) \times \ldots \times f(\boldsymbol{Y}|\mathbf{X}) \\
&\quad \times f(\boldsymbol{X}_{\mathbf{1}}|\boldsymbol{X}_{\mathbf{2}}, \boldsymbol{X}_{\mathbf{3}}) \\
f\left(\boldsymbol{A}_{\mathbf{2}(\boldsymbol{miss})} \mid \cdot\right) &\propto f(\boldsymbol{M}|\boldsymbol{Y}, \mathbf{X}, \mathbf{A}) \times \ f(\boldsymbol{A}_{\mathbf{1}}|\boldsymbol{Y}, \mathbf{X}, \boldsymbol{A}_{\mathbf{2}}) \ \times \ f(\boldsymbol{A}_{\mathbf{2}}|\boldsymbol{Y}, \mathbf{X}) \\
f\left(\boldsymbol{A}_{\mathbf{1}(\boldsymbol{miss})} \mid \cdot\right) &\propto f(\boldsymbol{M}|\boldsymbol{Y}, \mathbf{X}, \mathbf{A}) \times \ f(\boldsymbol{A}_{\mathbf{1}}|\boldsymbol{Y}, \mathbf{X}, \boldsymbol{A}_{\mathbf{2}})
\end{aligned}
\tag{33}
$$

7) Repeat steps 1 to 6 until MCMC chains reach convergence and provide sufficient posterior samples.

Bayesian Variable Selection

Advances in machine-learning have popularized potential methods that can ameliorate problems with high dimensionality; for example, penalized regression is a statistical technique widely used to guard against overfitting because it can select variables out of a large set of variables that are relevant for predicting some outcome (Casella et al., 2010; Hesterberg et al., 2008; Van Erp et al., 2019; Wu & Lange, 2008). The central idea of penalized regression approaches is to add a penalty term to the minimization of the sum of squared residuals, with the goal of shrinking small coefficients towards zero while leaving large coefficients relatively intact (Hesterberg et al., 2008; Van Erp et al., 2019). The most common penalized regression is the least absolute shrinkage and selection operator (LASSO) introduced by Tibshirani (1996).

Below is the LASSO estimator

$$\widehat{\boldsymbol{\beta}}(\lambda) = argmin\left\{\frac{1}{n}\ \|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\| + \lambda\|\boldsymbol{\beta}\|_1\right\} \tag{34}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ is an n-dimensional vector containing the observations on the outcome variable, $\mathbf{X}$ is an $(n \times p)$ matrix of the observed scores on the p predictor variables, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ is a p-dimensional parameter vector of regression coefficients. The term λ is the tuning parameter that controls the level of shrinkage in the regression coefficient, where $\lambda\ =\ 0$ leads to the ordinary least squares solution (Tibshirani, 1996). The LASSO is the gold-standard of frequentist variable selection (Bai et al., 2021; Bhadra et al., 2019).

Recent advancements in BVS have led to the development of methods for achieving penalized regression within the Bayesian framework by leveraging prior distributions for unknown parameters (Chipman, 1996; George & McCulloch, 1997; O'Hara & Sillanpää, 2009). This approach of Bayesian penalization is gaining popularity in the social sciences (e.g. Bainter et al., 2020; Chen et al., 2022) and it has shown to be as effective or superior to its frequentist counterpart (Casella et al., 2010; Hans, 2009; Li & Lin, 2010).

Using the Bayesian framework for variable selection offers significant benefits. Firstly, in classical LASSO, the user must manually select the tuning parameter (λ), which can be challenging. In contrast, BVS allows for automatic tuning of the regularization parameter using hyperpriors. This eliminates the need for manual selection and provides a more data-driven approach to regularization. Similar to the LASSO, BVS often includes a parameter that controls shrinkage, but BVS methods like the Bayesian LASSO can estimate penalty parameters simultaneously with the model parameters within the same MCMC iteration. Another advantage of BVS over its frequentist counterpart is that Bayesian methods, in general, allow for the incorporation of complex models with hierarchical structures and dependencies. BVS methods can easily handle intricate model setups, such as incorporating group structures or additional random effects (Griffin & Brown, 2017; Van Erp et al., 2019). Additionally, a key advantage of the BVS over the classical LASSO is its natural handling of missing data. The Bayesian framework imputes missing values through the posterior distribution. In contrast, the classical LASSO would require a separate process to handle missing data prior to variable selection (Casella et al., 2010).

Bayesian penalization and variable selection is a very active area of research, and there is an immense variety of regularization methods, each with its advantages in terms of variable selection and prediction (Bai et al., 2021; O'Hara & Sillanpää, 2009; Van Erp et al., 2019). Broadly speaking, Bayesian procedures for penalization can be categorized into two categories of priors: (1) two-group model and (2) global-local shrinkage priors (Bhattacharya et al., 2015; Polson & Scott, 2012). The first category of prior places a discrete mixture of a point mass at zero (the spike) and a continuous density (the slab) on each parameter. The second category of priors places continuous shrinkage priors on the regression coefficients that selectively shrink coefficients to different degrees. For my thesis, I intend to employ three priors for penalizing the coefficients in the missingness model. These priors include one from the two-group models, namely the spike-and-slab prior (Ishwaran & Rao, 2005; Mitchell & Beauchamp, 1988). The remaining two priors fall under the global-local shrinkage priors category. The first of these is the Bayesian LASSO (Park and Casella 2008, Bhadra et al., 2017), which employs a Laplace prior. The second is the horseshoe prior (Carvalho et al., 2009). By utilizing these three Bayesian BVS methods, my aim is to select the most influential variables and eliminate unnecessary components from the missingness model.

Purpose of the Study

The current dissertation is predicated on the assumption that the researcher believes there is a diffuse missingness model that includes various potential predictors. As mentioned earlier, modeling diffuse processes poses a significant challenge due to the tendency of the missingness model to incorporate an excessive number of variables. Consequently, including all covariates and auxiliary variables is likely result in issues of non-identification or in overfitting, especially as the ratio of predictor variables to observation is high (Du et al., 2021; Ibrahim et al., 2005). As mentioned earlier, the absence of a variable in the missingness model despite its presence in the substantive model is referred to as an exclusion restriction in the literature. Identifying exclusion restrictions is crucial as they allow for the estimation of selection models, which may encounter convergence issues in their absence. Therefore, it is essential to carefully consider the inclusion of covariates and auxiliary variables to avoid non-identification or overfitting in the context of the diffuse missingness model.

My research focuses on exploring the application of BVS techniques as a means to introduce exclusion restrictions in the context of MNAR data. Specifically, I investigated the effectiveness of three approaches: the Bayesian LASSO, the horseshoe prior, and the spike-and-slab prior. The goal is to assess how well these methods can eliminate unnecessary components from the missingness model, thereby reducing nonresponse bias in the parameters of the substantive model.

There has been limited previous work on applying BVS methods to handle missing data. To the best of our knowledge, only two studies have investigated BVS methods to improve missing data handling. The first study by Zhao and Long (2016) introduced a Bayesian lasso imputation model, which demonstrated superior performance compared to other regularized regression methods such as LASSO, elastic net, and adaptive lasso when applied to multiple imputation (MI). The second study by Yamaguchi (2022) developed a built-in function for the Bayesian lasso imputation model, implemented within the framework of multiple imputation using chained equations. Their findings, in the context of longitudinal data analysis, indicated that incorporating the Bayesian LASSO into the imputation model increased statistical power compared to the regular imputation model, particularly with small sample sizes and high dimensionality.

There are several important distinctions between these two previous studies and my dissertation research. Firstly, while the previous studies applied the Bayesian LASSO to multiple linear regression, my study focuses on a probit model. Additionally, they employed a multiple imputation algorithm using chained equations, whereas I employed a fully Bayesian factored regression. Furthermore, in addition to the Bayesian LASSO, I am also investigating the spike-and-slab and the horseshoe prior as potential shrinking priors for the missingness model. Finally, my research is specifically targeted towards MNAR missingness, while the other studies investigated missingness mechanisms within the context of MAR.

In the following section, I will provide a technical overview of the Bayesian LASSO, spike-and-slab, and horseshoe prior. I will describe how these priors will be applied to model the underlying cause of missingness in a selection model. This will involve presenting the hierarchical representation of the prior, explaining the conditional posterior distributions, and outlining the sampling steps for the Gibbs sampler.

The Bayesian LASSO

As described before, Tibshirani (1996) proposed the frequentist LASSO as an alternative to the OLS estimator for multiple regression models. While OLS aims to minimize the sum of squared residuals, the LASSO introduces a penalty term that encourages sparse solutions by simultaneously shrinking coefficients towards zero and performing variable selection. Park and Casella (2008) proposed a Bayesian version of the LASSO by imposing independent and identical double-exponential priors on the regression parameters (Laplace prior; Park & Casella, 2008). The Laplace prior is illustrated below:

$$f(\boldsymbol{\beta}|\sigma_{\varepsilon}^{2}\ ) \ = \ \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma_{\varepsilon}^{2}}} exp\left\{\frac{-\lambda\left|\beta_{j}\right|}{\sqrt{\sigma_{\varepsilon}^{2}}}\right\} \tag{35}$$

This prior distribution is characterized by being centered around zero, having wide tails, and a shape that is unimodal and symmetrical. In Equation 35, the term β_j represents the regression coefficient for predictor j, λ serves as the shrinkage parameter in the prior, and σ_{ε}^{2} is the residual variance from the regression model. The role of the parameter λ in Equation 35 is analogous to the penalty shrinkage parameter in classical penalized LASSO regression. In Equation 35, λ controls the amount of shrinkage, similar to how the parameter λ in penalized LASSO regression determines the amount of shrinkage. When λ is set to zero, no shrinkage is applied, as any number raised to power of zero equals one. As λ increases, the amount of shrinkage imposed on the regression coefficients also increases.

In Bayesian statistics, utilizing a conjugate prior, which is a prior from the same distribution family as the likelihood function, results in a posterior distribution that also falls within the same distribution family as the prior and likelihood. This implementation of a conjugate prior offers computational convenience by providing a closed-form expression for the conditional posterior distribution. As a results, model parameters can be efficiently updated using a Gibbs sampling procedure (Gelman et al., 1995; Lynch, 2020). Park and Casella (2008) exploit the fact that the Laplace distribution can be represented as a normal distribution with heterogeneous variances; this derivation is also called a scale mixture of normals (Andrews & Mallows, 1974). An advantage of using the scale mixture parameterization is that, by representing the Laplace distribution as normal densities, the prior can be considered conjugate to the likelihood function of the regression coefficients, and the posterior distribution can be estimated using standard Gibbs samplers.

When the Laplace prior is represented as a scale mixture of normal distributions, each regression coefficient j can be estimated using a hierarchical process (Andrews & Mallows, 1974; Eltoft et al., 2006). In this framework, each coefficient has a unique variance, and these variances themselves follow a distribution with a unique prior. The Bayesian LASSO falls under the category of global-local shrinkage priors, where the first step of the hierarchical process is to generate a parameter that determines the variance or spread of the regression coefficient distributions (global shinkage). The second step of the process, samples the coefficients themselves from the distribution that incorporate the coefficient-specific variance terms (local-shrinkage).

Given the scale mixture of normals representation, the full model can be expressed by using multiple hierarchical levels. The following hierarchical representation of the linear regression with Bayesian LASSO prior can be found in Park and Casella (2008):

$$\begin{gathered} \lambda^2 \sim Gamma(r, \delta), where\ \lambda^2 > 0\ (r > 0, \delta > 0) \\ \sigma_\varepsilon^2, \tau_1^2, \dots, \tau_p^2 \sim f(\sigma_\varepsilon^2) d\sigma_\varepsilon^2\ Exp\left(\frac{\lambda^2}{2}\right) \end{gathered} \tag{36}$$

$$\sigma_\varepsilon^2, \tau_1^2, \dots, \tau_p^2 > 0$$

$$\boldsymbol{\beta}|\sigma_\varepsilon^2, \tau_1^2, \dots, \tau_p^2 \sim MN(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{D}_\tau), \;\; where\ \mathbf{D}_\tau = diag(\tau_1^2, \dots, \tau_p^2)$$

$$y_i|\alpha, \boldsymbol{X_i}, \boldsymbol{\beta}, \sigma_\varepsilon^2 \sim N(\alpha + \boldsymbol{X}_\mathbf{i}\boldsymbol{\beta}, \sigma_\varepsilon^2)$$

$$f(\alpha) \propto 1$$

In Equation 36, the subscript j ranges from 1 to p, where j represents the index for the predictors. Additionally, the subscript p represents the total number of predictors in the model. The first line in Equation 36 is the prior for the shrinkage parameter λ^2. One option proposed by Park and Casella (2008) is to assign a diffuse hyperprior to λ^2, specifically a gamma prior with shape r and rate δ. This choice is advantageous as it enables a straightforward extension of the Gibbs sampler (Park & Casella, 2008).

The second line of Equation 36 represents the joint prior distribution for the residual variance σ_ε^2 and the predictor-specific $\tau_1^2, \dots, \tau_p^2$. This joint prior distribution incorporates an improper prior density for σ_ε^2, specifically $f(\sigma_\varepsilon^2) = 1/\sigma_\varepsilon^2$, as well as an exponential distribution that incorporates the global shrinkage parameter λ^2. The shrinkage parameter λ^2 determines the amount of overall shrinkage of the regression coefficients, with extreme values resulting in smaller variation among the τ_j^2. When the parameter λ^2 is near 0 there is virtually no shrinkage in any of the coefficients, and when λ is relatively large then all coefficients are shrunk towards zero.

The fourth line of Equation 36 says that the priors of the regression coefficients $\boldsymbol{\beta}$ are normal distributions centered at zero and a covariance matrix that includes a diagonal matrix $\mathbf{D}_\tau$ with a vector $(\tau_1^2, \dots, \tau_p^2)$. More specifically, each regression coefficient β_j has a prior with its center at 0 and a specific term τ_j^2. The variance of the prior determines the extent of the predictor-specific shrinkage: if a predictor has a salient influence on the outcome, the variance of the prior will be wide, resulting in less shrinkage. Conversely, if the predictor has a minor influence in the outcome, then the variance will be narrow, and it will shrink the regression coefficient relatively close to zero. In the fifth line of Equation 36, the distribution of the outcome variable y_i is described. It follows a normal distribution centered around $\alpha + \boldsymbol{X}_{\mathbf{i}}\boldsymbol{\beta}$**,** where $\boldsymbol{X_i}$ represents the predictor variables for observation i, and $\boldsymbol{\beta}$ represents the regression coefficients. The residual variance of the distribution is represented by σ_ε^2. Finally, the last line of Equation 36 is a non-informative prior for the intercept. No shrinking prior is applied to the intercept since the objective is to shrink the slopes while leaving the intercept unaffected.

Bayesian LASSO for Probit Model

Thus far, my focus has mainly revolved around discussing the Bayesian Lasso in the context of multiple regression scenarios. However, in this dissertation, I will be employing the BVS techniques to model the underlying cause of missingness in a selection model. The missingness model can be represented as follows:

$$\begin{gathered} M_i^\star = \alpha + \ \gamma_1 Y_i + \ \gamma_2 X_{1i} + \gamma_3 X_{2i} + \gamma_4 X_{3i} + \gamma_5 A_{1i} + r_i \\ M_i^* | \boldsymbol{\gamma} \sim N(\alpha + \boldsymbol{Z_i^T}\boldsymbol{\gamma}, \sigma_r^2)\, I(Q_i) \\ \sigma_r^2 = 1 \end{gathered} \tag{37}$$

In Equation 37, the missingness model is similar to Equation 21. However, there is a new addition of an auxiliary variable A_1 and the intercept is represented as α. This missingness model specifically follows a probit model form.

Bae & Mallick (2004) introduced the application of the Bayesian Lasso in a probit model, specifically for gene expression classification. Similar to the Bayesian Lasso in linear regression, the probit model was assigned a Laplace prior for the regression coefficients to encourage sparsity. This approach ensured that parameters with minimal impact on the outcome were effectively shrunk towards zero. Since then, the application of the Bayesian Lasso in the probit model has been adopted in various studies, such as economic forecasting (Yang et al., 2019), and sports data analysis (Gao, 2018). However, to the best of my knowledge, this will be the first instance of employing the Bayesian Lasso within a selection model for MNAR data.

In order to illustrate the distinctions between the linear regression and probit versions of the Bayesian LASSO, I will provide the hierarchical representation of the full probit model incorporating a Bayesian LASSO prior:

$$\begin{gathered}
\lambda^2 \sim Gamma(r, \delta), where\ \lambda^2 > 0\ (r > 0, \delta > 0) \\
\tau_1^2, \ldots, \tau_p^2 \sim Exp\left(\frac{\lambda^2}{2}\right) \\
\tau_1^2, \ldots, \tau_p^2 > 0 \\
\boldsymbol{\gamma} | \tau_1^2, \ldots, \tau_p^2 \sim MN(\mathbf{0}, \mathbf{D}_{\boldsymbol{\tau}}),\ \ where\ \mathbf{D}_{\boldsymbol{\tau}} = diag(\tau_1^2, \ldots, \tau_p^2) \\
M_i^* | \boldsymbol{\gamma} \sim N(\alpha + \boldsymbol{Z_i^T} \boldsymbol{\gamma}, \sigma_r^2)\ I(Q_i), \\
f(\alpha)\ \propto\ 1
\end{gathered} \tag{38}$$

Equation 38 introduces the hierarchical representation of the full probit model with a Bayesian LASSO prior. In this equation, the subscript j ranges from 1 to p, where j represents the index for the predictors. Additionally, the subscript p represents the total number of predictors in the model. I will now outline the differences between the hierarchical representation of the full linear model in Equation 36 and the probit model presented in Equation 38.

The first line in Equation 38, which describes the prior of λ^2, remains unchanged between the linear regression (Equation 36) and probit model. Further details regarding this line can be found in the description of Equation 36. The second line in Equation 38 describes the prior joint distribution of $\tau_1^2, \dots, \tau_p^2$. Unlike in the linear regression case, where the residual variance is explicitly modeled, the probit model fixes the residual variance σ_r^2 at 1. Consequently, there is no need for a prior distribution of the residual variance. Instead, the prior distribution of $\tau_1^2, \dots, \tau_p^2$ is represented by an exponential distribution with rate parameter set at $\lambda/2$. The fourth line in Equation 38 describes the priors of the regression coefficients $\boldsymbol{\gamma}$. The only distinction from the priors of the regression coefficients $\boldsymbol{\beta}$ in the linear regression (Equation 36) is that the residual variance σ_r^2 is fixed at one. Similar to Equation 36, the regression coefficients $\boldsymbol{\gamma}$ are normal distributions centered at zero and a covariance matrix that includes a diagonal matrix $\mathbf{D}_{\boldsymbol{\tau}}$ with a vector $\left(\tau_1^2, \dots, \tau_p^2\right)$.

The fifth line in Equation 38 addresses the requirements of the probit model, where a latent variable M_i^* is introduced to establish a connection between the binary response M_i and the regression parameters $\boldsymbol{\gamma}$. The distribution of M_i^* follows a normal distribution with a fixed residual variance σ_r^2 of one, centered around $\alpha + \boldsymbol{Z}_{\boldsymbol{i}}^{\boldsymbol{T}}\boldsymbol{\gamma}$. The term $I(\cdot)$ is an indicator function, Q_i is either equal to $\{M_i^* > \varphi\}$ or $\{M_i^* \leq \varphi\}$, the threshold parameter φ is used to divide the latent response distribution of M_i^* into distinct sections corresponding to $M_i = 1$ (missing observations) and $M_i = 0$ (present observations). Furthermore, the last line in Equation 38 corresponds to a non-informative prior for the intercept. Again, no shrinking prior is applied to the intercept since the objective is to shrink the slopes while leaving the intercept unaffected.

Bayesian LASSO for Selection Model Computation

In this section, I will provide the conditional posterior distributions for MCMC estimation of the selection model, encompassing parameters from both the substantive model and the missingness model. I will also present the sampling steps from the posterior densities using the Gibbs sampler. This approach allows for the estimation of the selection model through iterative sampling.

Step 1: Considering the covariates in the substantive model $\mathbf{X}$, the imputed outcomes from the previous iteration $Y_i^{(t-1)}$, and the residual variance of the substantive model in the previous iteration $\sigma_\varepsilon^{2(t-1)}$, draw regression coefficients $\boldsymbol{\beta}^{(\boldsymbol{t})}$ from a multivariate normal distribution (MN). Assuming a uniform prior for the regression coefficients, $f(\boldsymbol{\beta}) \propto 1$.

$$\begin{gathered} f(\boldsymbol{\beta}|\boldsymbol{Y},\mathbf{X},\sigma_\varepsilon^2) = MN\left(\widehat{\boldsymbol{\beta}},\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}\right) \\ where\ \widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{Y} \\ and\ \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}} = \sigma_\varepsilon^2(\mathbf{X}'\mathbf{X})^{-1} \end{gathered} \tag{39}$$

Step 2: Given the covariates in the substantive model $\mathbf{X}$, the imputed outcomes from the previous iteration $Y_i^{(t-1)}$, and $\boldsymbol{\beta}^{(t)}$, drawn the reciprocal of residual variance $\sigma_\varepsilon^{2(t)}$ (i.e., the precision) from a right-skewed gamma distribution

$$f(1/\sigma_\varepsilon^2|\boldsymbol{\beta},\boldsymbol{Y},\mathbf{X}) = Gamma\left(\frac{N+df}{2},\frac{(\boldsymbol{Y}-\boldsymbol{X\beta})'(\boldsymbol{Y}-\boldsymbol{X\beta})+S}{2}\right) \tag{40}$$

where the terms df and S are hyperparameters from the prior. Both hyperparameters df and S were specified to zero, which corresponds to a Jeffreys prior. The shape parameter $\frac{N+df}{2}$ determines the height of the distribution, which in turn affects its skewness and heavy-tailed properties. The spread of the distribution is determined by the sum of squared residuals from the previous iteration, adjusted by the hyperparameter of the prior distribution S. The term N represents the total number of observations.

Step 3: By utilizing the imputed outcomes $Y_i^{(t-1)}$ from the previous iteration, along with the other predictors $\mathbf{X}$ and $\boldsymbol{A_1}$ in the missingness model, as well as the latent data $M_i^{*(t-1)}$ from the previous iteration, draw missingness model regression coefficients $\boldsymbol{\gamma}^{(\boldsymbol{t})}$ from a multivariate normal distribution

$$\begin{gathered}
f(\boldsymbol{\gamma} \middle| \boldsymbol{M}^{*}, \boldsymbol{\tau}^{\mathbf{2}}, \boldsymbol{M}, \mathbf{Z}) = \ MN\left(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\gamma}}}\right) \\
where\ \widehat{\boldsymbol{\gamma}} = \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\gamma}}}\mathbf{Z}^{\mathbf{T}}\boldsymbol{M}^{*} \\
and\ \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\gamma}}} = (\mathbf{Z}^{\mathbf{T}}\mathbf{Z})^{-1} + (\mathbf{D}_{\boldsymbol{\tau}})^{-1} \qquad (41) \\
\mathbf{D}_{\boldsymbol{\tau}} = diag(\tau_1^2, \dots, \tau_p^2) \\
\mathbf{Z} = (\boldsymbol{Y}, \mathbf{X}, \boldsymbol{A_1})
\end{gathered}$$

where the center of the distribution is determined by $\mathbf{Z}^{\mathbf{T}}\boldsymbol{M}^{*}$and the covariance matrix $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\gamma}}}$**.** The covariance matrix includes a diagonal matrix $\mathbf{D}_{\boldsymbol{\tau}}$ with a vector $\left(\tau_1^2, \dots, \tau_j^2\right)^{t-1}$ from the previous iteration, populating its diagonal elements. The purpose of the diagonal matrix $\mathbf{D}_{\boldsymbol{\tau}}$ is to shrink the coefficients towards zero, where only some of the coefficients will be shrunk effectively to zero.

Step 4: By using the missingness model predictors $\mathbf{Z}$, the latent data $M_i^{*(t-1)}$ from the previous iteration, and the regression coefficients $\boldsymbol{\gamma}^{(t)}$, draw the missingness model intercept $\alpha^{(t)}$ from a normal distribution.

$$
\begin{gathered}
f(\alpha) = N(m_{prior}, v_{prior}) \\
\overline{M^*} = \frac{\sum_{i=1}^{N} M_i^* - (\boldsymbol{Z_i\gamma})}{N} \\
\hat{\alpha} = \frac{\sigma_r^2 m_{prior} + \ N\overline{M^*}v_{prior}}{Nv_{prior} + \sigma_r^2} \\
\sigma_{\hat{\alpha}}^2 = \frac{\sigma_r^2 \ v_{prior}}{Nv_{prior} + \sigma_r^2} \\
f(\alpha|\boldsymbol{M}^*, \boldsymbol{\gamma}, \boldsymbol{M}, \mathbf{Z}) = \ N(\hat{\alpha}, \sigma_{\hat{\alpha}}^2)
\end{gathered}
\tag{42}
$$

The prior distribution for the intercept α is specified as a normal distribution with a mean m_{prior} and a variance v_{prior}. In order to ensure stable estimation, I have set the prior mean m_{prior} to 0.01 and the prior variance v_{prior} to 5. The term $\overline{M^*}$ represents the sample mean of the intercept, while $\hat{\alpha}$ and $\sigma_{\hat{\alpha}}^2$ represent the conditional distribution mean and variance of the intercept, respectively.

Step 5: By utilizing the previously sampled shrinkage parameter $\lambda^{2(t-1)}$, and the sampled missingness model regression slopes $\boldsymbol{\gamma}^{(t)}$**,** a sample for the vector $\tau_j^{2(t)}$ is generated from the full conditional posterior which follow an inverse Gaussian.

$$f\left({\tau^2}_j^{-1} \middle| \boldsymbol{\tau^2_{-j}}, \boldsymbol{\gamma}, \lambda\right) = IG\left(\frac{\lambda}{|\gamma_j|}, \lambda^2\right) \tag{43}$$

The mean of this inverse Gaussian distribution is determined by the ratio $\lambda/|\gamma_j|$, where the data informs the mean of the distribution through the absolute value of the regression slope γ_j. If the regression slope is small, indicating a weaker relationship, the mean of the distribution will be larger. Consequently, drawing a value of τ_j^2 from this distribution will result in shrinking the regression coefficient γ_j closer to zero.

Step 6: Using the sampled $\tau_j^{2(t)}$ from the current iteration, Generate a sample for the parameter $\lambda^{2(t)}$ from its full conditional posterior

$$f\left(\lambda^2 \middle| \tau_1^2, \dots, \tau_p^2\right) = Gamma\left(p + r, \frac{1}{2}\sum_{j=1}^{p} \tau_j^2 + \delta\right) \tag{44}$$

Where the subscript j ranges from 1 to p, where j represents the index for the predictors and subscript p is the total number of predictors in the missingness model. The terms r and δ are the hyperparameters from the prior of λ^2 described in Equation 36. Both hyperparameters were set to one, which sets the prior distribution to be a right-skewed distribution that starts at zero, peaks at zero, and decreases exponentially as the values increase.

Step 7: I can now proceed to compute the latent variable $M_i^{*(t)}$. For each individual i, considering their observed value of M_i, the imputed outcome $Y_i^{(t-1)}$, and the sampled values of $\boldsymbol{\gamma}^{(t)}$, I can sample $M_i^{*(t)}$ from the distribution below.

$$f(M_i^*|\boldsymbol{\gamma}, \alpha, \mathbf{Z}, \boldsymbol{M}) = N\big(\alpha + \boldsymbol{Z_i^T\gamma}, \sigma_r^2\big) I(Q_i) \tag{45}$$

where $I(\cdot)$ is an indicator function, Q_i is either equal to $\{M_i^* > \varphi\}$ or $\{M_i^* \leq \varphi\}$ corresponding to $M_i = 1$ or $M_i = 0$. The conditional distribution of M_i^* is a normal distribution with a residual variance σ_r^2 fixed at one and a center determined by the $\alpha + \boldsymbol{Z_i^T\gamma}$**.**

Step 8: Given $\boldsymbol{\beta^{(t)}}$, $\sigma_\varepsilon^{2(t)}$, $\boldsymbol{\gamma^{(t)}}$, and $M_i^{*(t)}$, for every individual i who has missing outcome (i.e., $M_i = 1$), draw $Y_{i(miss)}^{(t)}$ from

$$f\left(Y_{i_{(miss)}}|M_i, \boldsymbol{X_i}, A_{1(i)}\right) \propto f\big(M_i\big|Y_i, \boldsymbol{X_i}, A_{1(i)}\big) \times f\big(A_{1(i)}\big|Y_i, \boldsymbol{X_i}\big) \times f(Y_i|\boldsymbol{X_i}) \tag{46}$$

The conditional distribution of $Y_{i_{(miss)}}^{(t)}$ must account for every model in which it appears. It is obtained by multiplying three separate distributions. The first distribution represents the model for missingness, the second distribution is the conditional distribution for the auxiliary variable, and the third distribution corresponds to the substantive model. The conditional distribution of $Y_{i_{(miss)}}^{(t)}$ is complex and cannot be derived analytically. Therefore I used the Metropolis–Hastings algorithm, a specialized MCMC method, to approximate sampling from this complex multi-part distribution (Gilks et al., 1995; Hastings, 1970)

The Horseshoe Prior

The horseshoe prior, initially proposed by Carvalho et al. (2009), is a special type of shrinkage prior. It exhibits a symmetric distribution around zero, with fat tails and an infinitely large spike at zero. The horseshoe reduces variance by eliminating noise variables completely and introduces less bias by preserving informative variables intact (Polson & Scott, 2010). Similar to the Bayesian LASSO, the horseshoe prior falls under the category of global-local shrinkage priors. Its main concept involves employing a global shrinkage parameter that can potentially shrink all regression coefficients, while also incorporating a local shrinkage parameter that allows certain informative coefficients to escape shrinkage (Bhadra et al., 2019). The visual representation of the horseshoe prior is provided below:

$$\begin{gathered}\beta_j \,|\lambda_j \sim N\left(0, \lambda_j^2 \tau^2\right) \\ \lambda_j \sim C^+(0{,}1) \\ \tau|\sigma_\varepsilon \sim C^+(0, \sigma_\varepsilon)\end{gathered} \tag{47}$$

The prior distribution for each coefficient β_j is modeled as a normal distribution centered at zero and a variance of $\lambda_j^2 \tau^2$. The local shrinkage parameter λ_j follows a standard half-Cauchy distribution, and λ_j controls the specific shrinkage strength of the β_j regression coefficient. The global shrinkage parameter τ also follows a standard half-Cauchy distribution, where τ controls the overall shrinkage of all regression coefficients (Carvalho et al., 2010; Makalic & Schmidt, 2016b).

The horseshoe prior offers several advantages. Firstly, the half-Cauchy distribution has a mode near zero and slow decaying tails. By concentrating an infinite density near zero, the horseshoe prior places a significant amount of prior mass near the true parameter when $\beta_j = 0$. This characteristic leads to fast convergence to the correct estimate of the sampling density, surpassing other global-local shrinkage priors such as the Bayesian Lasso (Carvalho et al., 2010). Secondly, the horseshoe estimator is asymptotically unbiased, which means that as the sample size (i.e., the number of observations) tends towards infinity, the parameter estimates obtained using this prior converge to the true values without any systematic bias. This distinguishes the horseshoe prior from the Bayesian Lasso, as the Bayesian Lasso does not achieve asymptotic unbiasedness (Carvalho et al., 2010).

Similar to the Laplace prior used in the Bayesian LASSO approach, the half-Cauchy distribution in the horseshoe prior can also be represented hierarchically (Carvalho et al., 2010; Piironen & Vehtari, 2017). The full model can be expressed by incorporating multiple hierarchical levels. Below is the hierarchical representation of linear regression using the horseshoe prior, also found in Makalic & Schmidt (2016a):

$$
\begin{gathered}
\nu_1, \dots, \nu_p, \xi \sim InvGam\left(\frac{1}{2}, 1\right) \\
\tau^2 | \xi \sim InvGam\left(\frac{1}{2}, \frac{1}{\xi}\right) \\
\lambda_j^2 | \nu_j \sim InvGam\left(\frac{1}{2}, \frac{1}{\nu_j}\right) \\
\frac{1}{\sigma_\varepsilon^2} \sim Gamma(df, S), where\ df, S > 0 \\
\beta_j | \lambda_j^2, \tau^2, \sigma_\varepsilon^2 \sim N\left(0, \lambda_j^2 \tau^2 \sigma_\varepsilon^2\right) \\
y_i | \alpha, \boldsymbol{X_i}, \boldsymbol{\beta}, \sigma_\varepsilon^2 \sim N(\alpha + \boldsymbol{X}_{\mathbf{i}} \boldsymbol{\beta}, \sigma_\varepsilon^2)
\end{gathered}
\tag{48}
$$

$$f(\alpha) \propto 1$$

The half-Cauchy distribution can be represented as a scale mixture, which involves using two latent variables ν_j and ξ. By employing this scale mixture representation, it becomes possible to establish conjugate conditional posterior distributions for all parameters. This facilitates the use of Gibbs sampling, as it simplifies the sampling process (Makalic & Schmidt, 2016b).

In the first line of Equation 48, each of the latent variables ν_j and ξ are independent, such as they each have their own inverse-gamma prior distribution. The global and local shrinkage terms, τ^2 and λ_j^2, also have an inverse gamma prior distribution, which are conditional to their respective latent variable. The fourth line in Equation 48 is the prior distribution for the reciprocal of the residual variance σ_ε^2 which follows a gamma distribution with hyperparameters df and S. The fifth line in Equation 48 represents the prior distribution of the regression coefficient β_j, where the regression coefficient β_j follow a normal distribution centered at zero and a variance that is influenced by the global-local shrinkage parameters $\lambda_j^2\tau^2$.

In the sixth line of Equation 48, the distribution of the outcome variable y_i is described. It follows a normal distribution centered around $\alpha + \boldsymbol{X}_\mathbf{i}\boldsymbol{\beta}$**,** where $\boldsymbol{X_i}$ represents the predictor variables for observation i, and $\boldsymbol{\beta}$ represents the regression coefficients. The residual variance of the distribution is represented by σ_ε^2. Finally, the last line of Equation 48 is a non-informative prior for the intercept. No shrinking prior is applied to the intercept since the objective is to shrink the slopes while leaving the intercept unaffected.

Horseshoe for Probit Model

In this dissertation, I will be employing the BVS techniques to model the underlying cause of missingness in a selection model, which takes the form of a probit model. There are few studies that have adapted the horseshoe prior to a probit model. Maity, Carroll, and Mallik (2019) developed a Bayesian hierarchical model to jointly model the survival time and the classification of the cancer stages. To deal with the high dimensionality, they used horseshoe prior on a probit model to identify significant predictors of cancer. Another study, that adapted the horseshoe prior to a probit regression was Terenin, Dong, and Draper (2019). They developed a horseshoe probit regression algorithm based on the probit model described in Albert and Chib (1993), and combined it with the hierarchical representation of the horseshoe in Makalic and Schmidt (2016).

Below, I will provide the hierarchical representation of the full probit model incorporating a horseshoe prior:

$$
\begin{gathered}
\nu_1, \ldots, \nu_p, \xi \sim InvGam\left(\frac{1}{2}, 1\right) \\
\tau^2 | \xi \sim InvGam\left(\frac{1}{2}, \frac{1}{\xi}\right) \\
\lambda_j^2 | \nu_j \sim InvGam\left(\frac{1}{2}, \frac{1}{\nu_j}\right) \\
\beta_j | \lambda_j^2, \tau^2 \sim N\left(0, \lambda_j^2 \tau^2\right) \\
M_i^* | \boldsymbol{\gamma} \sim N(\alpha + \boldsymbol{Z_i^T} \boldsymbol{\gamma}, \sigma_r^2)\ I(Q_i) \\
f(\alpha) \propto 1
\end{gathered}
\tag{49}
$$

Equation 49 introduces the hierarchical representation of the full probit model with a horseshoe prior. In this equation, the latent variables ν_j and ξ retain the same prior as shown in Equation 48. The prior for the global and local shrinkage parameters τ^2 and λ_j^2 also remains unchanged from the linear regression equation. It is important to note that unlike the linear regression case, where the residual variance σ_ε^2 is explicitly modeled with a prior distribution, in the probit model, the residual variance σ_r^2 is fixed at 1. The fourth line in Equation 49 describes the prior of the regression coefficient γ_j. The only distinction from the prior of the regression coefficients β_j in the linear regression (Equation 48) is that the residual variance σ_r^2 is fixed at one. Similar to Equation 48, the regression coefficient γ_j follow a normal distribution centered at zero and a variance that depends on the global-local shrinkage parameters $\lambda_j^2\tau^2$.

The fifth line of Equation 49 represents the conditional distribution of M_i^*. It follows a normal distribution with a fixed residual variance σ_r^2 of one, centered around $\alpha + \boldsymbol{Z_i^T\gamma}$. Here, the term $I(\cdot)$ represents an indicator function, Q_i is either equal to $\{M_i^* > \varphi\}$ or $\{M_i^* \leq \varphi\}$, the threshold parameter φ is used to divide the latent response distribution of M_i^* into distinct sections corresponding to $M_i = 1$ (missing observations) and $M_i = 0$ (present observations). The last line in Equation 49 represents a non-informative prior for the intercept. Similar to the previous case, no shrinking prior is applied to the intercept because the objective is to shrink the slopes while keeping the intercept unaffected.

Horseshoe Prior for Selection Model Computation

In this section, I will provide the conditional posterior distributions for MCMC estimation of the selection model, encompassing parameters from both the substantive model and the missingness model. I will also present the sampling steps from the posterior densities using the Gibbs sampler. This approach allows for the estimation of the selection model through iterative sampling.

Step 1: Considering the covariates in the substantive model **X**, the imputed outcomes from the previous iteration $Y_i^{(t-1)}$, and the residual variance of the substantive model in the previous iteration $\sigma_\varepsilon^{2(t-1)}$, draw regression coefficients $\boldsymbol{\beta}^{(\boldsymbol{t})}$ from a multivariate normal distribution (MN). Assuming a uniform prior for the regression coefficients, $f(\boldsymbol{\beta}) \propto 1$.

$$
\begin{gathered}
f(\boldsymbol{\beta}|\boldsymbol{Y},\mathbf{X},\sigma_\varepsilon^2) = MN\left(\widehat{\boldsymbol{\beta}},\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}\right) \\
where\ \widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{Y} \qquad (50) \\
and\ \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}} = \sigma_\varepsilon^2(\mathbf{X}'\mathbf{X})^{-1}
\end{gathered}
$$

Step 2: Given the covariates in the substantive model $\mathbf{X}$, the imputed outcomes from the previous iteration $Y_i^{(t-1)}$, and $\boldsymbol{\beta}^{(\boldsymbol{t})}$, drawn the reciprocal of residual variance $\sigma_\varepsilon^{2(t)}$ (i.e., the precision) from a right-skewed gamma distribution

$$
f(1/\sigma_\varepsilon^2|\boldsymbol{\beta},\boldsymbol{Y},\mathbf{X}) = Gamma\left(\frac{N+df}{2},\frac{(\boldsymbol{Y}-\boldsymbol{X\beta})'(\boldsymbol{Y}-\boldsymbol{X\beta})+S}{2}\right) \qquad (51)
$$

where the terms df and S are hyperparameters from the prior. Both hyperparameters df and S were specified to zero, which corresponds to a Jeffreys prior. The shape parameter $\frac{N+df}{2}$ determines the height of the distribution, which in turn affects its skewness and heavy-tailed properties. The spread of the distribution is determined by the sum of squared residuals from the previous iteration, adjusted by the hyperparameter of the prior distribution S. The term N represents the total number of observations.

Step 3: By utilizing the imputed outcomes $Y_i^{(t-1)}$ from the previous iteration, along with the other predictors $\mathbf{X}$ and $\boldsymbol{A_1}$ in the missingness model, as well as the latent data $M_i^{*(t-1)}$ from the previous iteration, draw missingness model regression coefficients $\boldsymbol{\gamma}^{(\boldsymbol{t})}$ from a multivariate normal distribution

$$
\begin{gathered}
f(\boldsymbol{\gamma}|\boldsymbol{M}^{*},\boldsymbol{\tau}^{\mathbf{2}},\boldsymbol{M},\mathbf{Z}) = \ MN\left(\widehat{\boldsymbol{\gamma}},\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\gamma}}}\right) \\
where\ \widehat{\boldsymbol{\gamma}} = \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\gamma}}}\mathbf{Z}^{\mathbf{T}}\boldsymbol{M}^{*} \\
and\ \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\gamma}}} = (\mathbf{Z}^{\mathbf{T}}\mathbf{Z})^{-1} + (\mathbf{D}_{\boldsymbol{\tau}})^{-1} \\
\mathbf{D}_{\boldsymbol{\tau}} = diag(\lambda_1^{-2},\dots,\lambda_p^{-2})/\tau^2 \\
\mathbf{Z} = (\boldsymbol{Y},\mathbf{X},\boldsymbol{A}_{\mathbf{1}})
\end{gathered}
\tag{52}
$$

where the conditional posterior distribution of $\boldsymbol{\gamma}$ is a normal distribution with its center determined by $\mathbf{Z}^{\mathbf{T}}\boldsymbol{M}^{*}$ and the covariance matrix $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\gamma}}}$. The covariance matrix includes a diagonal matrix $\mathbf{D}_{\boldsymbol{\tau}}$ with a vector $\left(\lambda_1^{-2},\dots,\lambda_p^{-2}/\tau^2\right)^{(t-1)}$ from the previous iteration, populating its diagonal elements. The purpose of the diagonal matrix $\mathbf{D}_{\boldsymbol{\tau}}$ is to shrink the coefficients towards zero, where only some of the coefficients will be shrunk effectively to zero.

Step 4: By using the missingness model predictors $\mathbf{Z}$, the latent data $M_i^{*(t-1)}$ from the previous iteration, and the regression coefficients $\boldsymbol{\gamma}^{(\boldsymbol{t})}$, draw the missingness model intercept $\alpha^{(t)}$ from a normal distribution.

$$
\begin{gathered}
f(\alpha) = N(m_{prior},v_{prior}) \\
\overline{M^{*}} = \frac{\sum_{i=1}^{N} M_i^{*} - (\boldsymbol{Z}_{\boldsymbol{i}}\boldsymbol{\gamma})}{N} \\
\hat{\alpha} = \frac{\sigma_r^2 m_{prior} + \ N\overline{M^{*}}v_{prior}}{Nv_{prior} + \sigma_r^2}
\end{gathered}
\tag{53}
$$

$$\sigma_{\hat{\alpha}}^2 = \frac{\sigma_r^2 \, v_{prior}}{N v_{prior} + \sigma_r^2}$$

$$f(\alpha | \boldsymbol{M}^*, \boldsymbol{\gamma}, \boldsymbol{M}, \mathbf{Z}) = \; N(\hat{\alpha}, \sigma_{\hat{\alpha}}^2)$$

The prior distribution for the intercept α is specified as a normal distribution with a mean m_{prior} and a variance v_{prior}. In order to ensure stable estimation, I have set the prior mean m_{prior} to 0.01 and the prior variance v_{prior} to 5. The term $\overline{M^*}$ represents the sample mean of the intercept, while $\hat{\alpha}$ and $\sigma_{\hat{\alpha}}^2$ represent the conditional distribution mean and variance of the intercept, respectively.

Step 5: By utilizing the previously sampled shrinkage parameter $\tau^{2(t-1)}$, a sample for $\xi^{-1(t)}$ is generated from a full conditional posterior which follow an exponential distribution.

$$f(\xi^{-1} | \tau^{-2}) = Exp\left(1 + \frac{1}{\tau^2} \right) \tag{54}$$

Step 6: By utilizing the sampled $\xi^{-1(t)}$, the sampled missingness model regression slopes $\boldsymbol{\gamma}^{(t)}$, and the previously sampled local shrinkage parameter $\lambda_j^{-2(t-1)}$ a sample for $\tau^{2(t)}$ is generated from the full conditional posterior which follow a gamma distribution.

$$f(\tau^2 | \boldsymbol{\gamma}, \boldsymbol{\lambda^2}, \xi^{-1}) = Gamma\left(\frac{p+1}{2}, \frac{1}{\xi} + \frac{1}{2} \sum_{j=1}^{p} \frac{\gamma_j^2}{\lambda_j^2} \right) \tag{55}$$

The term p is the total number of predictors in the missingness model. The term ξ is the latent variables generated from the hierarchical expression of the half-Cauchy prior, λ_j^2 is the local shrinkage parameter, and γ_j^2 is the squared transformation of the regression coefficients from the missingness model.

Step 7: By utilizing the previous sampled local shrinkage parameter $\lambda_j^{-2(t-1)}$, a sample for ν_j^{-1} is generated from the full conditional posterior which follow an exponential distribution.

$$f\left(\nu_j^{-1}\,|\lambda_j^{-2}\right) = Exp\left(1+\frac{1}{\lambda_j^2}\right) \tag{56}$$

Step 8: By using the missingness model regression coefficients $\boldsymbol{\gamma}^{(t)}$, the global shrinkage parameter $\tau^{2(t)}$, and a vector of latent variable $\nu_j^{(t)}$, draw a vector of local shrinkage parameters $\lambda_j^{2(t)}$ from an exponential distribution.

$$f\left(\lambda_j^2\middle|\gamma_j,\tau^2,\nu_j\right) = Exp\left(\frac{1}{\nu_j}+\frac{\gamma_j^2}{2\tau^2}\right) \tag{57}$$

The rate parameter for this distribution is informed by the data via γ_j^2 and by the global shrinkage parameter τ^2.

Step 9: I will now proceed with computing the latent variable $M_i^{*(t)}$. For each individual i, taking into account their observed value of M_i, the imputed outcome $Y_i^{(t-1)}$, and the sampled values of $\boldsymbol{\gamma}^{(t)}$, I can sample $M_i^{*(t)}$ from the distribution below.

$$f(M_i^*|\boldsymbol{\gamma}, \alpha, \mathbf{Z}, \boldsymbol{M}) = N\big(\alpha + \boldsymbol{Z_i^T}\boldsymbol{\gamma}, \sigma_r^2\big)I(Q_i) \tag{58}$$

where $I(\cdot)$ is an indicator function, Q_i is either equal to $\{M_i^* > \varphi\}$ or $\{M_i^* \leq \varphi\}$ corresponding to $M_i = 1$ or $M_i = 0$. The conditional distribution of M_i^* is a normal distribution with a residual variance σ_r^2 fixed at one and a center determined by the $\alpha + \boldsymbol{Z_i^T\gamma.}$

Step 10: Given $\boldsymbol{\beta}^{(\boldsymbol{t})}$, $\sigma_\varepsilon^{2(t)}$, $\boldsymbol{\gamma}^{(\boldsymbol{t})}$, and $M_i^{*(t)}$, for every individual i who has missing outcome (i.e., $M_i = 1$), draw $Y_{i(miss)}^{(t)}$ from

$$f\left(Y_{i_{(miss)}}|M_i, \boldsymbol{X_i}, A_{1(i)}\right) \propto f\big(M_i\big|Y_i, \boldsymbol{X_i}, A_{1(i)}\big) \times f\big(A_{1(i)}\big|Y_i, \boldsymbol{X_i}\big) \times f(Y_i|\boldsymbol{X_i}) \tag{59}$$

The conditional distribution of $Y_{i_{(miss)}}^{(t)}$ must account for every model in which it appears. It is obtained by multiplying three separate distributions. The first distribution represents the model for missingness, the second distribution is the conditional distribution for the auxiliary variable, and the third distribution corresponds to the substantive model. The conditional distribution of $Y_{i_{(miss)}}^{(t)}$ is complex and cannot be derived analytically. Therefore I used the Metropolis–Hastings algorithm, a specialized MCMC method, to approximate sampling from this complex multi-part distribution (Gilks et al., 1995; Hastings, 1970)

The Spike-and-Slab Prior

Spike-and-slab prior was initially proposed by Mitchell and Beauchamp (1988) for BVS in the context of linear regression models, and it has been used extensively for variable selection (George and McCulloch, 1993; Ishwaran and Rao, 2005). These prior combines two components: a spike component and a slab component. The spike component assigns high prior probability, such as a point mass at zero or a normal distribution with a very narrow variance, to exclude regression coefficients from the model. The slab component assigns non-zero prior probability, such as uniform or normal with a wide variance, to allow the inclusion of regression coefficient in the model.

The spike-and-slab distribution, introduced by Mitchell and Beauchamp (1988), combines a point mass at zero (known as a Dirac delta spike) with a slab represented by a truncated uniform distribution. While this formulation is conceptually straightforward, it presents substantial computational challenges due to the need to calculate marginal likelihoods (Bai et al., 2021; Ishwaran & Rao, 2005; Rockova, 2013). A computationally superior variation of the spike-and-slab prior was introduced by George and McCulloch (1993). This variation replaces the point mass at zero with a continuous normal prior distribution with a very small variance. Similarly, the truncated uniform distribution is substituted with a normal prior distribution with a very large variance. This modification enhances computational feasibility in variable selection, as it enables the spike-and-slab to be represented hierarchically, similar to the Bayesian LASSO in the previous section, by formulating it as a scale mixture of normal distributions (Ishwaran & Rao, 2005).

Now let's consider how the spike and slab operate in the classical linear regression model with regression coefficients $\boldsymbol{\beta}$ and residual variance σ_ε^2. The hierarchical representation of the full linear model with spike-and-slab priors is presented below:

$$
\begin{gathered}
\frac{1}{\sigma_\varepsilon^2} \sim Gamma(df, S), where\ df, S > 0 \\
\delta_j \sim Bernoulli(w) \\
\beta_j | \delta_j, \sigma_\varepsilon^2 \sim \left(1 - \delta_j\right) N(0, \sigma_\varepsilon^2 \tau_0^2)\ + \delta_j N(0, \sigma_\varepsilon^2 \tau_1^2) \\
y_i | \alpha, \boldsymbol{X_i}, \boldsymbol{\beta}, \sigma_\varepsilon^2 \sim N(\alpha +\ \mathbf{X}_{\mathbf{i}} \boldsymbol{\beta}, \sigma_\varepsilon^2) \\
f(\alpha)\ \propto\ 1
\end{gathered}
\qquad (60)
$$

The first line in Equation 60 is the prior distribution for the reciprocal of the residual variance σ_ε^2 which follows a gamma distribution with hyperparameters df and S. The second line in Equation 59 is the prior distribution of δ_j. Here, δ_j represents an auxiliary indicator that determines the presence ($\delta_j = 1$) or absence ($\delta_j = 0$) of predictor *j* in the model. The purpose of these indicator variables is to denote whether a variable belongs to the “slab” or “spike” part of the prior.

The spike-and-slab prior treats the complete set of indicator variables $\boldsymbol{\delta}$ as unknown parameters to be estimated; this estimation then combines variable selection with the estimation of the regression parameters (Bai et al., 2021). Since the indicator variables are unknown parameters, we need to specify a prior for each δ_j. We do so by utilizing a Bernoulli distribution with parameter w. This parameter w sets the prior probability of including or excluding a variable (George & McCulloch, 1997; Lee et al., 2003). In some cases, a researcher might believe that every predictor has an equal chance to enter the model, and a value of 0.5 for w is recommended (George & McCulloch, 1997). Choosing a standard weight of 0.5 implies that each predictor has a 50/50 prior probability of being included in the model. The value for w is then up to the investigator; in some cases, a data-based approach may be a good way of specifying this prior (O'Hara & Sillanpää, 2009).

The third line in Equation 60 represents the prior distribution of the regression coefficient β_j. This prior is a combination of two components: a peaked prior around zero (referred to as the spike) and a high variance prior (referred to as the slab). The values of the regression coefficients depend on the estimated indicator variable $\delta_{\boldsymbol{j}}$, which can change from one iteration to another. When δ_j = 0 the coefficient uses the spike as the prior. In the given equation, the spike is represented by a normal distribution with a hyperparameter τ_0^2 in the variance. This spike component effectively shrinks the regression coefficients towards zero. On the other hand, when δ_j = 1, the coefficient uses the slab as the prior. The slab is specified as a normal distribution $N(0, \tau_1^2)$ with a wide variance, determined by the hyperparameter τ_1^2. This slab component allows for more flexibility, allowing the regression coefficients to take on a wider range of values.

In the fourth line of Equation 60, the distribution of the outcome variable y_i is described. It follows a normal distribution centered around $\alpha + \boldsymbol{X}_\mathbf{i}\boldsymbol{\beta}$, where $\boldsymbol{X_i}$ represents the predictor variables for observation i, and $\boldsymbol{\beta}$ represents the regression coefficients. The residual variance of the distribution is represented by σ_ε^2. Finally, the last line of Equation 59 is a non-informative prior for the intercept. No shrinking prior is applied to the intercept since the objective is to shrink the slopes while leaving the intercept unaffected.

Spike-and-Slab Prior for Probit Model

Most developments in the spike-and-slab prior for BVS has occurred in the context of the classical linear regression model. However, in this dissertation, I will be employing the BVS techniques to the missingness model, which has the form of a probit model. Publication in biomedical research have adapted the spike-and-slab for probit model for classification or prediction of binary outcomes (Lee et al., 2003; Russu et al., 2012; Yang et al., 2019). The scalable spike-and-slab R package also provides an algorithm for probit regression with a spike-and-slab prior (Biswas et al., 2022). Nevertheless, as far as my knowledge extends, this dissertation represents the first instance of utilizing the spike-and-slab prior with missing data. By employing the spike-and-slab prior in this context, the study expands the scope of the spike-and-slab prior beyond its conventional applications.

Now I will present the hierarchical representation of a probit regression with spike-and-slab prior for the missingness model:

$$\begin{gathered}\delta_j \sim Bernoulli(w) \\ \gamma_j|\delta_j \sim \left(1-\delta_j\right)N(0,\tau_0^2)\ +\delta_j N(0,\tau_1^2) \\ M_i^*|\boldsymbol{\gamma} \sim N(\alpha+\boldsymbol{Z_i^T}\boldsymbol{\gamma},\sigma_r^2)\ I(Q_i), \\ f(\alpha)\ \propto\ 1\end{gathered} \tag{61}$$

Equation 61 introduces the hierarchical representation of the full probit model with a spike-and-slab prior. In this equation, the subscript j ranges from 1 to p, where j represents the index for the predictors. Additionally, the subscript p represents the total number of predictors in the model. I will now outline the differences between the hierarchical representation of the full linear model in Equation 60 and the probit model presented in Equation 61.

The first line in Equation 61, which describes the prior of δ_j, remains unchanged between the linear regression (Equation 60) and probit model. Further details regarding the prior of δ_j can be found in the description of Equation 60. Notice that the residual variance σ_r^2 of the probit model does not have a prior distribution, unlike in the linear regression case, where the residual variance is explicitly modeled, the residual variance σ_r^2 is fixed at 1. The second line in Equation 61 describes the prior of the regression coefficient γ_j. The only distinction from the prior of the regression coefficients β_j in the linear regression (Equation 60) is that the residual variance σ_r^2 is fixed at one. Similar to Equation 60, the regression coefficient γ_j follow a normal distribution centered at zero and a variance that depends on the estimated indicator variables δ_j.

The fourth line of Equation 61 represents the conditional distribution of M_i^* which follows a normal distribution with a fixed residual variance σ_r^2 of one, centered around $\alpha + \boldsymbol{Z_i^T}\boldsymbol{\gamma}$. The term $I(\cdot)$is an indicator function, Q_i is either equal to $\{M_i^* > \varphi\}$ or $\{M_i^* \leq \varphi\}$, the threshold parameter φ is used to divide the latent response distribution of M_i^* into distinct sections corresponding to $M_i = 1$ (missing observations) and $M_i = 0$ (present observations). The last line in Equation 61 corresponds to a non-informative prior for the intercept. Again, no shrinking prior is applied to the intercept since the objective is to shrink the slopes while leaving the intercept unaffected.

Spike-and-Slab for Selection Model Computation

In this section, I will provide the conditional posterior distributions for MCMC estimation of the selection model. A spike-and-slab prior will be included on the missingness model. Additionally, I will outline the sampling steps involved in the Gibbs sampler, which facilitates the estimation of the selection model by iteratively sampling from the posterior densities.

Step 1: Considering the covariates in the substantive model **X**, the imputed outcomes from the previous iteration $Y_i^{(t-1)}$, and the residual variance of the substantive model in the previous iteration $\sigma_\varepsilon^{2(t-1)}$, draw regression coefficients $\boldsymbol{\beta}^{(\boldsymbol{t})}$ from a multivariate normal distribution (MN). Assuming a uniform prior for the regression coefficients, $f(\boldsymbol{\beta}) \propto 1$.

$$
\begin{gathered}
f(\boldsymbol{\beta}|\boldsymbol{Y},\mathbf{X},\sigma_\varepsilon^2) = MN\left(\widehat{\boldsymbol{\beta}},\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}\right) \\
where\ \widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{Y} \\
and\ \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}} = \sigma_\varepsilon^2(\mathbf{X}'\mathbf{X})^{-1}
\end{gathered}
\tag{62}
$$

Step 2: Given the covariates in the substantive model **X**, the imputed outcomes from the previous iteration $Y_i^{(t-1)}$, and $\boldsymbol{\beta}^{(t)}$, drawn the reciprocal of residual variance $\sigma_\varepsilon^{2(t)}$ (i.e., the precision) from a right-skewed gamma distribution

$$
f(1/\sigma_\varepsilon^2|\boldsymbol{\beta},\boldsymbol{Y},\mathbf{X}) = Gamma\left(\frac{N+df}{2},\frac{(\boldsymbol{Y}-\boldsymbol{X\beta})'(\boldsymbol{Y}-\boldsymbol{X\beta})+S}{2}\right) \tag{63}
$$

where the terms df and S are hyperparameters from the prior. Both hyperparameters df and S were specified to zero, which corresponds to a Jeffreys prior. The shape parameter $\frac{N+df}{2}$ determines the height of the distribution, which in turn affects its skewness and heavy-tailed properties. The spread of the distribution is determined by the sum of squared residuals from the previous iteration, adjusted by the hyperparameter of the prior distribution S. The term N represents the total number of observations.

Step 3: A sample of $\boldsymbol{\gamma}^{(\boldsymbol{t})}$ is generated by incorporating the imputed outcomes $Y_i^{(t-1)}$, the latent data $M_i^{*(t-1)}$, and the indicator variable $\boldsymbol{\delta}^{(\boldsymbol{t-1})}$ from the previous iteration, along with the other predictors $\mathbf{X}$ and $\boldsymbol{A_1}$ in the missingness model.

$$
\begin{gathered}
f(\boldsymbol{\gamma}|\boldsymbol{\delta}, \boldsymbol{M}^*, \boldsymbol{M}, \mathbf{Z}) = MN\left(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\gamma}}}\right) \\
where\ \widehat{\boldsymbol{\gamma}} = \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\gamma}}} \mathbf{Z}^{\mathbf{T}} \boldsymbol{M}^* \\
and\ \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\gamma}}} = (\mathbf{Z}^{\mathbf{T}}\mathbf{Z})^{-1} + (\mathbf{D}_{\boldsymbol{\tau}})^{-1} \qquad (64) \\
\mathbf{D}_{\boldsymbol{\tau}} = diag((\mathbf{1} - \boldsymbol{\delta})\tau_0^2 + \boldsymbol{\delta}\tau_1^2) \\
\mathbf{Z} = (\boldsymbol{Y}, \mathbf{X}, \boldsymbol{A_1})
\end{gathered}
$$

In words, Equation 64 says that the missingness model's regression coefficients are drawn from a multivariate normal distribution and the center and variance are determined by $\mathbf{Z}^{\mathbf{T}}\boldsymbol{M}^*$ and the covariance matrix $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\gamma}}}$. The covariance matrix includes a diagonal matrix $\mathbf{D}_{\boldsymbol{\tau}}$ with a vector $\left((\mathbf{1} - \boldsymbol{\delta})\tau_0^2 + \boldsymbol{\delta}\tau_1^2\right)^{t-1}$ from the previous iteration, populating its diagonal elements. The purpose of the diagonal matrix $\mathbf{D}_{\boldsymbol{\tau}}$ is to set coefficients to either the spike or the slab state.

Step 4: By using the missingness model predictors $\mathbf{Z}$, the latent data $M_i^{*(t-1)}$ from the previous iteration, and the regression coefficients $\boldsymbol{\gamma}^{(\boldsymbol{t})}$, draw the missingness model intercept $\alpha^{(t)}$ from a normal distribution.

$$f(\alpha) = N(m_{prior}, v_{prior})$$

$$\overline{M^*} = \frac{\sum_{i=1}^{N} M_i^* - (\boldsymbol{Z_i\gamma})}{N}$$

$$\hat{\alpha} = \frac{\sigma_r^2 m_{prior} + \ N\overline{M^*} v_{prior}}{Nv_{prior} + \sigma_r^2} \tag{65}$$

$$\sigma_{\hat{\alpha}}^2 = \frac{\sigma_r^2 \ v_{prior}}{Nv_{prior} + \sigma_r^2}$$

$$f(\alpha|\boldsymbol{M}^*, \boldsymbol{\gamma}, \boldsymbol{M}, \mathbf{Z}) = \ N(\hat{\alpha}, \sigma_{\hat{\alpha}}^2)$$

The prior distribution for the intercept α is specified as a normal distribution with a mean m_{prior} and a variance v_{prior}. To ensure stable estimation, I have set the prior mean m_{prior} to 0.01 and the prior variance v_{prior} to 5. The term $\overline{M^*}$ represents the sample mean of the intercept, while $\hat{\alpha}$ and $\sigma_{\hat{\alpha}}^2$ represent the conditional distribution mean and variance of the intercept, respectively.

Step 5: To sample $\delta_j^{(t)}$, the following procedure is followed:

$$w = 1/p$$

$$P_0 = \ \log(1-w) + \log\left(N\left(\gamma_j; 0, \tau_0^2\right)\right)$$

$$P_1 = \ \log(w) + \log\left(N\left(\gamma_j; 0, \tau_1^2\right)\right) \tag{66}$$

$$prob_j = \frac{1}{1 + \exp(P_0 - P_1)}$$

$$f\left(\delta_j\middle|\boldsymbol{\gamma}, \boldsymbol{M}^*, \boldsymbol{M}\right) = Bernoulli\left(prob_j\right)$$

In words, Equation 66 says that the indicator variable δ_j is drawn from a Bernoulli distribution. The probability that each variable in $\boldsymbol{Z}$ is selected to the missingness model is determined by

γ_j and the hyperparameters included by the prior distributions (w, τ_1^2, τ_0^2). The probability $prob_j$ represents the likelihood of δ_j being equal to one, which indicates if the predictor γ_j is included in the model. The hyperparameter w is initially set to be 1 divided by the total number of predictors in the initial model (p). For example, if there are five predictors, w would be set to 0.2. Finally, hyperparameters τ_0^2 and τ_1^2, are set to $1/\sqrt{N}$ and 1, respectively, where N represents the sample size.

Step 6: I will now proceed with computing the latent variable $M_i^{*(t)}$. For each individual i, taking into account their observed value of M_i, the imputed outcome $Y_i^{(t-1)}$, and the sampled values of $\boldsymbol{\gamma}^{(t)}$, I can sample $M_i^{*(t)}$ from the distribution below.

$$f(M_i^* | \boldsymbol{\gamma}, \alpha, \mathbf{Z}, \boldsymbol{M}) = N\left(\alpha + \boldsymbol{Z_i^T}\boldsymbol{\gamma}, \sigma_r^2\right) I(Q_i) \tag{67}$$

where $I(\cdot)$ is an indicator function, Q_i is either equal to $\{M_i^* > \varphi\}$ or $\{M_i^* \leq \varphi\}$ corresponding to $M_i = 1$ or $M_i = 0$. The conditional distribution of M_i^* is a normal distribution with a residual variance σ_r^2 fixed at one and a center determined by the $\alpha + \boldsymbol{Z_i^T}\boldsymbol{\gamma}$**.**

Step 7: Given $\boldsymbol{\beta}^{(t)}$, $\sigma_\varepsilon^{2(t)}$, $\boldsymbol{\gamma}^{(t)}$, and $M_i^{*(t)}$, for every individual i who has missing outcome (i.e., $M_i = 1$), draw $Y_{i(miss)}^{(t)}$ from

$$f\left(Y_{i_{(miss)}} | M_i, \boldsymbol{X_i}, A_{1(i)}\right) \propto f\left(M_i \middle| Y_i, \boldsymbol{X_i}, A_{1(i)}\right) \times f\left(A_{1(i)} \middle| Y_i, \boldsymbol{X_i}\right) \times f(Y_i | \boldsymbol{X_i}) \tag{68}$$

The conditional distribution of $Y_{i_{(miss)}}^{(t)}$ must account for every model in which it appears. It is obtained by multiplying three separate distributions. The first distribution represents the model for missingness, the second distribution is the conditional distribution for the auxiliary variable, and the third distribution corresponds to the substantive model. The conditional distribution of $Y_{i_{(miss)}}^{(t)}$ is complex and cannot be derived analytically. Therefore I used the Metropolis–Hastings algorithm to approximate sampling from this complex multi-part distribution (Gilks et al., 1995; Hastings, 1970)

METHODS

The methods section will present a summary of the population model utilized in the simulation. Following that, an outline of the simulation conditions will be provided, highlighting the different scenarios examined. Next, the data generation process for the simulation will be described, along with an explanation of the estimation procedure. Next, implementation details and outcome measures will be discussed, addressing how the simulation was evaluated. Finally, a description of the software employed will be provided.

Population Models

The equation provided (Equation 68) represents a linear regression equation for the substantive model, which includes **X** predictors. In this equation, $\boldsymbol{\beta}$ represents the regression coefficients, ε_i denotes the error term of the substantive regression, and σ_ε^2 represents the residual variance.

$$\begin{gathered} Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \beta_6 X_{6i} + \varepsilon_i \\ Y_i \sim N(E(Y_i|\boldsymbol{X_i}), \sigma_\varepsilon^2) \end{gathered} \tag{68}$$

In the simulation design, the substantive model was defined with a fixed set of six predictors, and this configuration remained consistent across all simulations. However, the population regression coefficients were manipulated to reflect different correlation level among predictors in **X**. In addition to contributing to the between-subject factor, the substantive model played a crucial role in evaluating the simulation's outcomes. The bias and precision of the coefficients in the substantive model were assessed. This evaluation allowed for comparisons between different conditions and estimation procedures. Further details regarding these outcome measures will be explained in the subsequent subsection.

Additionally, the population model will include one auxiliary variable and a missingness model that will vary in complexity. Using notation from the introduction section, the full factorization is expressed as:

$$\begin{aligned} f(\boldsymbol{Y},\boldsymbol{M},\mathbf{X},\boldsymbol{A_1}) = {} & f(\boldsymbol{M}|\boldsymbol{Y},\boldsymbol{X_1},\boldsymbol{X_2},\boldsymbol{X_3},\boldsymbol{X_4},\boldsymbol{X_5},\boldsymbol{X_6},\boldsymbol{A_1}) \\ & \times f(\boldsymbol{A_1}|\boldsymbol{Y},\boldsymbol{X_1},\boldsymbol{X_2},\boldsymbol{X_3},\boldsymbol{X_4},\boldsymbol{X_5},\boldsymbol{X_6}) \\ & \times f(\boldsymbol{Y}|\boldsymbol{X_1},\boldsymbol{X_2},\boldsymbol{X_3},\boldsymbol{X_4},\boldsymbol{X_5},\boldsymbol{X_6}) \\ & \times f(\boldsymbol{X_1},\boldsymbol{X_2},\boldsymbol{X_3},\boldsymbol{X_4},\boldsymbol{X_5},\boldsymbol{X_6}) \end{aligned} \tag{69}$$

Next, I propose to simplify the full factorization by setting some of the coefficients in the factorization to zero in the population model.

Auxiliary variable A_1 in this simulation represents Type C auxiliary variables, as defined by Collins et al. (2001). These variables are only correlated with the missingness indicator M and do not have any correlation with the outcome variable Y or predictors X_1-X_6. To achieve this, the slopes coefficient for predictors **X** and outcome $\boldsymbol{Y}$ in the second function in Equation 69 were set to zero. By simplifying the factorization of the auxiliary variable $\boldsymbol{A_1}$ in this way, it is assumed that $\boldsymbol{A_1}$ is solely correlated with the missingness indicator.

The rationale behind this choice is driven by the primary interest in auxiliary variables that provide information about the predictability of missingness in the outcome. The correlation of these variables with X is less important since the X variables will be observed. Another significant reason for excluding X from the auxiliary variable factorization is that selection models tend to be better estimated when a subset of predictor variables is not shared with the substantive model. Consequently, linking the auxiliary variables solely to the missingness indicator introduces exclusion restrictions that facilitate estimation.

Finally, a key feature of the simulation design is to manipulate or vary the complexity of the missingness model. In this case, variables Y and A_1 were set to always be predictors in the missingness model. The complexity of the missingness model is manipulated by varying the number of predictors in X that also appear in the population data-generating model. Specifically, 60% of the explained variance is allocated to Y, while the remaining variation is evenly distributed among the predictors.

For this simulation, there will be four levels of complexity, each differing in the number of substantive predictors. The models are as follows: a model with (a) only $\boldsymbol{Y}$ and $\boldsymbol{A_1}$ predicting $\boldsymbol{M}^*$, (b) $\boldsymbol{Y}$, $\boldsymbol{X_1}$, $\boldsymbol{X_2}$, and $\boldsymbol{A_1}$ predicting $\boldsymbol{M}^*$, (c) $\boldsymbol{Y}$, $\boldsymbol{X_1}$, $\boldsymbol{X_2}$, $\boldsymbol{X_3}$, $\boldsymbol{X_4}$, and $\boldsymbol{A_1}$ predicting $\boldsymbol{M}^*$, and (f) $\boldsymbol{Y}$, $\boldsymbol{X_1}$, $\boldsymbol{X_2}$, $\boldsymbol{X_3}$, $\boldsymbol{X_4}$, $\boldsymbol{X_5}$, $\boldsymbol{X_6}$, and $\boldsymbol{A_1}$, predicting $\boldsymbol{M}^*$. Below are the factorizations of the four true data-generating models, ordered from most complex to least complex.

$$\begin{aligned} f(\boldsymbol{Y},\boldsymbol{M},\mathbf{X},\boldsymbol{A_1}) = (\boldsymbol{M}^*|\boldsymbol{Y},\boldsymbol{A_1}) \times f(\boldsymbol{A_1}) \times f(\boldsymbol{Y}|\boldsymbol{X_1},\boldsymbol{X_2},\boldsymbol{X_3},\boldsymbol{X_4},\boldsymbol{X_5},\boldsymbol{X_6}) \\ \times f(\boldsymbol{X_1},\boldsymbol{X_2},\boldsymbol{X_3},\boldsymbol{X_4},\boldsymbol{X_5},\boldsymbol{X_6}) \end{aligned} \tag{70}$$

$$f(\boldsymbol{Y}, \boldsymbol{M}, \mathbf{X}, \boldsymbol{A_1}) = (\boldsymbol{M}^{*}|\boldsymbol{Y}, \boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{A_1}) \times f(\boldsymbol{A_1}) \\ \times f(\boldsymbol{Y}|\boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3}, \boldsymbol{X_4}, \boldsymbol{X_5}, \boldsymbol{X_6}) \times f(\boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3}, \boldsymbol{X_4}, \boldsymbol{X_5}, \boldsymbol{X_6}) \tag{71}$$

$$f(\boldsymbol{Y}, \boldsymbol{M}, \mathbf{X}, \boldsymbol{A_1}) = (\boldsymbol{M}^{*}|\boldsymbol{Y}, \boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3}, \boldsymbol{X_4}, \boldsymbol{A_1}) \times f(\boldsymbol{A_1}) \\ \times f(\boldsymbol{Y}|\boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3}, \boldsymbol{X_4}, \boldsymbol{X_5}, \boldsymbol{X_6}) \times f(\boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3}, \boldsymbol{X_4}, \boldsymbol{X_5}, \boldsymbol{X_6}) \tag{72}$$

$$f(\boldsymbol{Y}, \boldsymbol{M}, \mathbf{X}, \boldsymbol{A_1}) = (\boldsymbol{M}^{*}|\boldsymbol{Y}, \boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3}, \boldsymbol{X_4}, \boldsymbol{X_5}, \boldsymbol{X_6}, \boldsymbol{A_1}) \times f(\boldsymbol{A_1}) \\ \times f(\boldsymbol{Y}|\boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3}, \boldsymbol{X_4}, \boldsymbol{X_5}, \boldsymbol{X_6}) \times f(\boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3}, \boldsymbol{X_4}, \boldsymbol{X_5}, \boldsymbol{X_6}) \tag{73}$$

Simulation Conditions

As described before, the first between-subject condition is the complexity of the missingness model. The factorization in Equations 70-73 shows four conditions, where the level of complexity varies from including: (a) $\boldsymbol{Y}$ and the auxiliary variable $\boldsymbol{A_1}$, (b) $\boldsymbol{Y}$, auxiliary variable $\boldsymbol{A_1}$, and two predictors in $\mathbf{X}$, (c) $\boldsymbol{Y}$, auxiliary variable $\boldsymbol{A_1}$, and four predictors in $\mathbf{X}$, and (d) Y, auxiliary variable $\boldsymbol{A_1}$, and all six predictors in $\mathbf{X}$. The second condition manipulated the correlation between the covariates in $\mathbf{X}$, where I use $\rho =.10$ and $\rho =.40$ to cover both low and high collinearity among covariates. The third condition is sample size, and I simulated datasets of 100, 200, and 400 observations. I used these three sample sizes to evaluate simulation behavior with small and large sample sizes and at the same time keep a range that generalizes to typical social science data. In total, I used a combination of 144 between-subject design cells.

The dissertation proposal involved manipulating the rate of missing data for variable Y in order to simulate both a high (30%) and low (10%) rate of missing data. However, due to time constraints given the computational burden of the simulation, I made adjustments and decided to only simulate a high rate of missing data in order to illustrate worse-case scenario. Additionally, I made two other modifications to the original proposal regarding the simulation conditions. Initially, I had intended to include a condition with a sample size of 800. However, I decided against it for two reasons. Firstly, it proved to be computationally infeasible as it resulted in excessively long running times, even though it eventually converged. Secondly, including this condition was not essential for the study's objectives. Our primary focus lies on scenarios with low sample sizes and a considerable number of potential variables to be included in the missingness model. The last modification involved reducing the number of conditions for the complexity of the missingness model. Upon closer examination, it was unclear what unique information certain conditions would provide compared to others. I employed the four complexity conditions represented in Equations 70 to 73 as they serve as a solid reference point for the varying degrees of complexity in the missingness model

Within the simulation, there were six fitted models that served as the sole within-subject factor. The chosen models were as follows: (a) an analysis that omits the missingness model and assumes MAR, (b) a selection model with no BVS that corresponds to the true data-generating model, (c) a full-selection model with the most complex factorization without BVS, and a selection model with the most complex factorization with three types of BVS applied to the missingness model, (d) the Bayesian LASSO (Gao, 2018; Park & Casella, 2008), (e) the horseshoe prior (Makalic & Schmidt, 2016a; Terenin et al., 2019), and (f) the spike-and-slab prior (Biswas et al., 2022; Ishwaran & Rao, 2005).

Data Generation

Data generation followed the four factorizations outlined in Equations 70-73. In each factorization, the rightmost term corresponds to the multivariate normal distribution for the predictor matrix $\mathbf{X}$, where $\mathbf{X}$ represented the matrix of predictors in the substantive model. The term immediately to the left represents the substantive model itself. To generate data for both terms, I considered the correlation between $\mathbf{X}$ covariates, which is a manipulated condition in the simulations, as well as the R-squared value for the substantive model, indicating the total variation in $\boldsymbol{Y}$ that is explained by the predictors in $\mathbf{X}$. The R-squared of the substantive model remained constant at $R^2 = .13$, while correlation between $\mathbf{X}$ covariates varied between a low (.10) and high (.40) correlation. To achieve this, I generated multivariate standard normal data for $\boldsymbol{Y}$ and $\mathbf{X}$ using their correlation matrix. The $\boldsymbol{Y}$ variable was standardized with a mean of zero and a standard deviation of one, and each regression coefficient in the substantive model equally contributed to the explained variability. To generate the variables $\mathbf{X}$ and $\boldsymbol{Y}$ based on their correlation matrix, I employed the R package 'mvtnorm' (Genz et al., 2021), which provided the necessary functionality to generate multivariate standard normal data.

Subsequently, data can be generated for A_1. To achieve this, I propose utilizing a univariate normal function, since A_1 is uncorrelated with all X and Y variables. In order to simulate the normally distributed latent variable M^*, I needed to determine the regression coefficients of the missingness model, which corresponds to the first term after the equal sign in Equations 70-73. It was not necessary to solve for the residual variance as this is already set to 1.

For all four proposed missingness models, the proportion of variance in the missingness indicator that could be explained by the independent variables was set to $R^2 = .30$. This choice was made to ensure a relatively strong selection mechanism (McKelvey & Zavoina, 1975).

Furthermore, approximately 60% of the explained variance was allocated to Y, while the remaining explained variation was distributed equally across the predictors. By imposing this proportional constraint, I was able to solve for the regression coefficients of the missingness model, denoted as the $\boldsymbol{\gamma}$ coefficients in previous equations.

After solving for the missingness model parameters, I proceeded to simulate M^* by plugging in a set of variables into the regression equation, generating predicted values for M^*, and adding a standard normal residual to create a dataset of simulated M^* scores. Subsequently, I converted $\boldsymbol{M}^*$ to the binary $\boldsymbol{M}$ variable using Equation 19 from the introduction. Specifically, any M_i^* value exceeding the fixed threshold φ was set as $M_i = 1$, while values below or equal to the thresholds were designated as $M_i = 0$. In the dataset, the outcome variable Y_i was assigned a missing value whenever M_i =1, representing a 30% rate of missing observations to simulate a substantial amount of missing data.

Outcome Variables

Four outcomes measures were computed to assess the performance of the substantive model parameters: convergence rates, percent bias, standardized bias, and mean square error (MSE). I used two different approaches to measure bias. The first approach, which I call "percent bias," involves calculating the difference between an average estimate and the true value, dividing that difference by the true value. However, this method cannot be used to measure bias in the focal model's intercept, which has a true value (denominator) equal to zero. Therefore, I included a second method, which I refer to as standardized bias.

The first outcome measure is convergence rates. Convergence was evaluated using potential scale reduction factors (PSR; Gelman & Rubin, 1992). The PSR is a statistical method used to assess the convergence of multiple chains in a MCMC simulation. It determines if independent chains generate estimates with similar means and variation, thus indicating that the distributions are stationary, and their center and spread do not change with additional iterations. The PSR formula, shown below, compares the variation between chains to the variation within each chain.

$$PSR = \sqrt{\frac{Within + Between\ Chain\ Variance}{Within\ Chain\ Variance}} \tag{74}$$

It employs the between-group mean square obtained from an analysis of variance (ANOVA) to quantify the mean differences between chains, and the within-group mean square from ANOVA to quantify the pooled variance within each chain. The sum of the between and within variances (total variance) is divided by the within-chain variance and then square-rooted to define the PSR. If there is minimal discrepancy between chains, the total variance in the numerator will be comparable to the denominator, resulting in a PSR value close to 1. If the chains are still diverging and failing to reach a stable estimate, the PSR will exceed 1. A PSR value close to 1 indicates that the chains have converge and mix well. The most used cutoff values for PSR convergence are 1.10 (Gelman et al., 2013) and 1.05 (Asparouhov & Muthén, 2010), meaning that a PSR value less than or equal to 1.10 or 1.05 is usually considered indicative of convergence. (Gelman et al., 2004).

Replications that produce PSR values greater than 1.05 within the burn-in period were considered convergence failures.

$$convergence\ rate = \frac{(number\ of\ replications\ that\ converged)}{(total\ number\ of\ replications)} \qquad (75)$$

Convergence rate was then defined as the ratio of the number of replications that converged to the total number of replications.

The second outcome measure is percent bias. This measure was calculated as the difference between an average estimate and the true value divided by the true value, and then multiplied by 100 to create a percentage (i.e., bias as a percentage of the true value).

$$percent\ bias = \frac{(average\ estimate) - (true\ parameter)}{(true\ parameter)} \times 100 \qquad (74)$$

A commonly accepted criterion is that relative bias values should be less than 10% in absolute value. (Finch, West, & MacKinnon, 1997; Kaplan, 1988).

Standardized bias was computed by dividing the difference between the average estimate and the true value, and then dividing that difference by the theoretical standard error (*SE*).

$$\begin{gathered} standarized\ bias = \frac{(average\ estimate) - (true\ parameter)}{(theoretical\ SE)} \\ theoretical\ SE = sqrt\left(diag\left(cov(\boldsymbol{\beta})\right)\right) \\ cov(\boldsymbol{\beta}) = \sigma_{\varepsilon}^{2}(\boldsymbol{\Sigma}_{\mathbf{X}}N)^{-1} \end{gathered} \qquad (75)$$

The second line of Equation 75 provides the definition of the theoretical *SE*, which is calculated as the square root of the diagonal vector derived from the population covariance matrix of the regression coefficients $\boldsymbol{\beta}$. Moving on to the third line of Equation 75, we find the computation of the covariance matrix for the population regression coefficients $\boldsymbol{\beta}$. Within this equation, σ_{ε}^{2} represents the population residual variance, N signifies the sample size, and $\boldsymbol{\Sigma}_{\mathbf{X}}$ denotes the population covariance matric of predictors.

Theoretical *SE* defines the spread of the sampling distribution of complete-data estimates around the true value, and the resulting bias values indicate where the average estimate falls in this theoretical sampling distribution. The standardized bias outcome employs a threshold of 0.40, after which bias can hinder efficiency, convergence, and error rates (Collins et al., 2001). Each bias measure has its advantages and disadvantages. The advantage of percent bias is that it is easy to understand, but it is not always clear if large values in percent bias are practically important. Additionally, percent bias fails to account for sample size; a bias value of 10% would indicate a substantial distortion in large sample sizes because the sampling distribution is narrower and has less variability, whereas the same 10% bias value may not be as meaningful when sample size is small, and sampling variance is high. Both methods produce slightly different results, but generally show similar trends.

The fouth criteria I used in the study was MSE. The MSE is a composite measure that captures the accuracy and precision of an estimator, as it equals the squared bias plus the sampling variance of the parameter estimate.

$$\mathrm{MSE} = \frac{1}{1000}\sum(\hat{\theta} - \theta)^{2} \tag{76}$$

where $\widehat{\theta}$ is the parameter estimate from a particular replication within a given design cell, θ is the population parameter, and 1000 is the number of replications within a given design cell. To facilitate interpretation, I used MSE ratios to evaluate whether the any of the missing data methods, increases precision relative to a theoretical complete-data sample. I computed these MSE ratios by dividing the MSE from one of the missing data methods, by the squared theoretical standard error (*SE*). The squared theoretical *SE* represents the expected MSE of an unbiased complete-data estimator, as it only considers sampling variance without the bias component. Values closer to 1 indicate lower variance and better precision. These four outcome measures were calculated to compare the performance of the six within-subject models. The Bayesian LASSO, spike-and-slab, horseshoe prior, full model, MAR model, and the true data-generating model estimates were compared.

Software Implementation

For the purpose of data generation, I utilized the R programming language for statistical computing (R Development Core Team, 2017). The MCMC algorithm and analysis for the full-selection model, MAR model, and the true model were implemented using Blimp version 3.2 (Keller & Enders, 2021). Estimation and analysis using the Bayesian LASSO, horseshoe, and spike-and-slab selection models were performed through custom R functions developed specifically for this project. These functions were created as there is currently no existing software program that handles missing data in selection models with a BVS adaptation.

To ensure the correctness of the custom programs, I fitted complete data using R functions from existing R packages that computed BVS in probit models and compared the outputs with my custom programs. For comparing the spike-and-slab and horseshoe custom programs, I used the R packages scalable spike-and-slab (Biswas et al., 2022) and horseshoe nlm (Maity et al., 2019), respectively. It should be noted that there was no R package available to fit a Bayesian Lasso to a probit model, to the best of my knowledge. Additionally, the computation of outcome measures on the estimates of the substantive model was performed using the R programming language for statistical computing.

Before starting the MCMC simulation with the Gibbs sampler, it is crucial to establish a sufficiently long burn-in period. The burn-in period is the initial phase of the simulation where the chain explores the parameter space and moves towards the desired outcome. Its purpose is to discard initial samples that may introduce bias or be far from convergence (Gelman et al., 2013; Little & Rubin, 2019; Rubin, 1976). To determine the appropriate burn-in period, I employed different methods depending on the software used. For the full, MAR, and true models, I utilized the Blimp software package. Blimp divides the burn-in period into 20 equal intervals and computes the PSR at the end of each interval. By examining the PSR values in Blimp's convergence diagnostic output, I identified a suitable value for the burn-in period, applying a threshold of 1.05 to the PSR.

For the Bayesian LASSO, horseshoe, and spike-and-slab selection models, I employed the coda package (Plummer et al., 2006) in R to obtain PSR values. Prior to conducting the MCMC estimation of the BVS models, I performed exploratory MCMC estimations and monitored the PSR values every 10,000 iterations. If the PSR convergence rates remained below 1.05 for five consecutive checks, I considered the number of iterations used in the third check as the final burn-in period. This approach was necessary because occasionally the PSR would momentarily dip below 1.05 in one check but then exceed the threshold in subsequent checks.

RESULTS

The results section will first focus on the convergence rates, followed by an analysis of bias and accuracy in the Bayesian variable selection (BVS) methods, full selection model, missing at random (MAR) model, and the true data-generating model. To ensure clarity in presentation, the discussion will begin by comparing just the BVS methods, followed by a comparison of the best BVS method against the full selection, MAR, and true models.

Convergence Rates

Table 1 displays the convergence rates of six methods: Bayesian LASSO, horseshoe prior, MAR model, full selection model, spike-and-slab prior, and true model. The table is divided into four subsections based on the complexity conditions of the missingness model, with each subsection containing three columns for the sample size conditions of 100, 200, and 400. Moreover, the table specifies the finalized burn-in periods applied in each sample size condition. It is important to note that the burn-in period varies for each method. This decision was made considering the runtime of the custom scripts utilized for the BVS methods. To ensure efficient execution, I chose not to prolong the runtime of any method unnecessarily.

Across the missingness model complexity condition, the Bayesian LASSO and horseshoe prior methods exhibited convergence rates of 42%-57% and 40%-55%, respectively. This means that out of 100 iterations, only 420 to 570 iterations converged for the Bayesian LASSO and only 400 to 550 iterations converged for the horseshoe prior, depending on sample size condition. Even though rates improved with a larger sample size, it did not change at different levels of complexity of the missingness model. In contrast, the spike-and-slab method had its highest convergence rates (~60%) at the lower sample size condition, and these rates decreased as sample size increased. The convergence rates of the spike-and-slab method were found to significantly vary with changes in the complexity of the missingness model, where convergence improved at lower complexity levels. For instance, at a sample size of 400 and in the simplest missingness model scenario, the spike-and-slab method demonstrated a convergence rate of 71%, whereas in the most complex missingness model condition, it was only 35%.

The full selection model had consistently low convergence rates, but these did improve with increasing sample size. The full selection model convergence rate also decreased at the least complex missingness model condition in comparison to the other complexity conditions. The true model had lower convergence rates compared to the BVS methods, particularly under low sample sizes and more complex missingness models. Finally, the misspecified method with an MAR assumption had a convergence rate ranged from 97%-100% under all possible conditions. It is important to note that the results presented in the following sections only consider iterations that converged. Therefore, methods with higher convergence rates, such as the MAR model, will have a larger number of iterations included in the analysis compared to methods with lower convergence rates, such as the full-selection model.

Comparison of Bayesian Variable Selection Methods

I will use trellis plots to compare different BVS methods based on percent and standardized bias, and MSE ratio. These plots illustrate the impact of sample size, predictor intercorrelation, and complexity of the missingness model. Each figure consists of four row-panels with specific configurations. In the first row panel, a sample size of 100 and a predictor intercorrelation of .10 are presented. The second row panel displays the same predictor correlation of .10 and the percentage bias for a sample size of 400. Moving to the third and fourth row panels, they exhibit sample sizes of 100 and 400, respectively, with a predictor intercorrelation of .40.

Each row-panel focuses on percent bias, standardized bias, or MSE ratio for the substantive model parameters. However, some specific rules apply to the values displayed in the trellis plots. For percent bias, if the estimate exceeds 50% (the minimum and maximum values shown), it will be presented as 50%. When measuring standardized bias, any value below -1 will be displayed as -1, and any value greater than 1 will be shown as 1. In the case of MSE ratio, any value below 0.90 is presented as 0.90, and any value exceeding three is displayed as three. It's important to note that the actual percent bias, standardized bias, or MSE ratio value are provided in the corresponding description of each plot.

The following sections will be divided into four subparts, each corresponding to a specific complexity condition of the missingness model. Each subpart will include three trellis plots. The first two figures compare percent bias and standardized bias for the spike-and-slab, Bayesian LASSO, and the horseshoe prior; while the last figure compares the MSE ratio for the same three BVS methods. In the first subpart, Figures 1 to 3 represent a true missingness model involving the outcome variable Y, predictors X_1 through X_6 and auxiliary variable A_1, which falls under complexity 4. Moving on to the second subpart, Figures 4 to 6 showcase the results of BVS methods under a true missingness model where Y, X_1 through X_4, and A_1 act as predictors, which corresponds to complexity 3. Next, the third subpart, Figures 7 to 9, present a true missingness model where the outcome variable Y, predictors X_1 and X_2, and auxiliary variable A_1 are significant predictors, denoting complexity 2. Lastly, the fourth subsection will revolve around Figures 10 to 12. These figures pertain to a true missingness model where only Y and A_1 are significant predictors, representing complexity.

Complexity 4

Figures 1 to 3 illustrate the most complex true missingness model, which includes all variables from the focal regression and the auxiliary variable. All BVS models are always fitting the most complex missingness model, which means that in the complexity 4 condition, BVS model are fitting the true data generating model.

Percent and Standardize Bias

In the methods section, I described two different approaches to measure bias. The first approach, which I call “percent bias,” involves calculating the difference between an average estimate and the true value, dividing that difference by the true value, and then multiplying the result by 100 to get a percentage (Equation 74). Standardized bias involves calculating the difference between an average estimate and the true value, and then dividing that difference by the theoretical *SE* (Equation 75). In my presentation of results, I will provide both measures of bias, but will explain percent bias in more detail and give a brief overview of standardized bias, focusing on their differences. Please note that bias in the intercept estimate will only be measured using standardized bias, while *R*-squared bias will only be measured using the percent bias method.

Figure 1 displays percent bias, and Figure 2 shows standardized bias measured in theoretical complete-data *SE* units. As previously mentioned, the cut-off for percent bias is ±10%, and the standardized bias cut-off is ±0.40, which are marked in both figures using vertical dashed lines (Collins et al., 2001; L. K. Muthén & Muthén, 2002). The bias results for BVS methods revealed little difference between the Bayesian LASSO and the horseshoe prior. This similarity can be attributed to the fact that both priors stem from the same category of global-local shrinkage priors. Global-local shrinkage priors use a continuous shrinkage function to model the prior distribution of the regression coefficients. In contrast, the spike-and-slab is referred as part of the two-group model family, because it separates regression coefficients into "important" coefficients with non-zero values and "unimportant" coefficients with zero values (Polson & Scott, 2010). For concise writing, I will refer to the Bayesian LASSO and horseshoe prior as global-shrinkage priors in the following results.

In the first row panel of both figures, where the sample size is 100 and the intercorrelation of predictors is .10, the spike-and-slab prior produced less biased estimates than the global-shrinkage priors (horseshoe and Bayesian LASSO) across the regression slope estimates, which were the main focus of the substantive model. As the bias values were fairly similar across the regression coefficients, I will describe the average bias across slopes. On average, the spike-and-slab had -12% bias in the slope estimates for all six predictors in comparison to the -25% average bias for the global-shrinkage priors. The residual variance estimates had virtually no bias for the spike-and-slab, whereas the global-shrinkage priors showed a bias of nearly 39%. In Figure 2, it can be observed that the standardized bias estimates for the intercept parameter were -0.19 and -0.32 for the spike-and-slab and global-shrinkage priors, respectively. The regression coefficients and residual variance displayed a similar pattern in both Figure 1 and 2, with the spike-and-slab exhibiting less bias than the horseshoe prior and Bayesian LASSO for almost all parameters, except for the *R*-squared estimate.

The *R*-squared estimate is an important parameter to consider because it is a composite function of multiple estimates, including the squared slope coefficients, residual variance, and predictor variances. Essentially, the *R*-squared represents a composite parameter that represents the cumulative effect of the biases in the individual parameters. In Figure 1's first row panel, the spike-and-slab model had a highly biased *R*-squared estimate of 72%, while the global-shrinkage prior had a bias of 46% on average. Upon closer inspection of the density plots for *R*-squared and residual variance estimates in the supplementary materials, it became clear that the spike-and-slab model had unbiased estimates for residual variance, but the distribution was very narrow, resulting in low variability.

This lack of variability led to a consistent overestimation of *R*-squared because the sampling distribution did not include a sufficient number of estimates with high residual variation. On the other hand, Bayesian LASSO and horseshoe models had more asymmetric distributions with long right tails, indicating larger residual variance estimates. Replications with high residual variances naturally reduced the mean estimate of *R*-squared, resulting in a lower bias for the composite parameter. In the spike-and-slab model, consistently obtaining accurate but too small residual variance estimates resulted in an inflated *R*-squared estimate.

Moving on to the second row panel in Figure 1, when the sample size increased to 400, the percentage bias for all BVS methods decreased. The spike-and-slab once again exhibited the smallest bias with an average of -8%, while the global-local shrinkage priors had an average bias of -15% between the substantive model predictors. The residual variance estimates for the spike-and-slab prior again had almost no bias, while the global-local shrinkage priors showed a significant decrease in bias under the *N* = 400 condition, with a 10% bias. The *R*-squared estimate also showed a decrease in bias with a 10% biased estimate for the spike-and-slab prior, while the global-local shrinkage priors had an average bias of 4%.

In the second row panel of Figure 2, it is evident that the spike-and-slab prior had lower bias for all parameters compared to the global-local shrinkage priors. The regression coefficients and residual variance in the second row panel exhibited a similar trend to that seen in Figure 1. However, there is one notable difference between the two figures. Unlike percent bias (Figure 1), standardized bias (Figure 2) did not show a noticeable decrease in bias as sample size increased for all parameters (first row panel versus second row panel). This discrepancy can be attributed to the fact that both the raw bias and theoretical *SE* (which form the numerator and denominator of the standardized bias calculation, respectively) decrease as the sample size increases. As a result, the mean estimates from the two sample sizes fall at approximately the same position in their respective complete-data sampling distributions.

The third row panel of the trellis plot depicted a stronger predictor intercorrelation of .40 and a sample size of 100, allowing for a comparison of the simple effect of predictor correlation (ρ = .10 vs ρ = .40) with the first-row panel. As shown in Figure 1, bias across the slope estimates for all methods increased with highly correlated predictors. The spike-and-slab prior exhibited a bias of -18% (-12% in the first row panel), while the global-local shrinkage priors showed an average bias of -35% (-25% in the first row panel). The presence of highly correlated predictors did not influence the bias in the estimates of residual variance or the *R*-squared. For the spike-and-slab model, there was no bias observed in residual variance, while the global-local shrinkage priors exhibited a bias of approximately 38% (39% in the first row panel). The *R*-squared estimate had a biased estimate of 70% (72% in the first row panel) for the spike-and-slab and an average bias of 45% (46% in the first row panel) for the global-local shrinkage prior. Moving to the third row panel in Figure 2, there was no influence of highly correlated predictor in the bias estimate of the intercept parameter. The standardized bias estimates for the intercept parameter were -0.19 and -0.32 for the spike-and-slab and global-local shrinkage priors, respectively.

The fourth row panel in Figure 1 and Figure 2, displayed a predictor correlation of .40 and a higher sample size of 400, enabling us to investigate whether the effect of the predictor correlations varied as a function of sample size. If there was an interaction between predictor correlation and sample size, a difference between the first and third row panel would not be the same as the difference between the second and fourth row panel. Both Figures 1 and 2 demonstrated that the difference between intercorrelation conditions remained uniform across both sample size conditions. For instance, the disparities between the first and third row panel in Figure 1 indicated that all BVS methods were more biased with strongly intercorrelated predictors in the *N* = 100 condition. However, the spike-and-slab method remained the most unbiased. This pattern repeated itself in the *N* = 400 condition, where differences in bias between the second and fourth row panel also showed the same effect of predictor correlation.

MSE Ratio

A useful way to measure overall accuracy is by using MSE. This metric calculates the average of the squared differences between an estimate and the true value of a parameter. Figure 3 illustrate the MSE ratios for the complexity 4 conditions, which is the most complex true missingness model, including all variables from the focal regression and the auxiliary variable.

In the first row panel Figure 3, where the sample size was 100 and the intercorrelation of predictors was .10, the spike-and-slab prior produced more accurate estimates than the global-shrinkage priors (horseshoe and Bayesian LASSO) across the regression slope estimates, residual variance, and intercept. However, differences in accuracy between methods was more pronounce for the intercept and residual variance parameters. Looking at the slope estimates, the spike-and-slab MSE was 1.58 times larger than the complete-data MSE for all six predictors, in contrast the MSE ratio for the global-shrinkage priors was 1.87 on average. The spike-and-slab estimates for the intercept and residual variance were substantially more accurate in comparison. The residual variance estimates from the spike-and-slab approach were, on average, 2.20 times larger than the MSE from complete-data analysis, while the corresponding ratio for the intercept parameter was 1.74. In contrast, the global-shrinkage priors exhibited much higher MSE ratios for the residual variance (13.15) and the intercept (5.86).

In the second row panel of Figure 3, a larger sample size of 400 is depicted, allowing for a comparison of the impact of sample size with the first row panel. The increased sample size had an impact on the accuracy of both the intercept parameter and the residual variance, but only for the global-local shrinkage priors. In the case of the spike-and-slab approach, the sample size affected only the accuracy of the intercept. With the global-local shrinkage priors, the accuracy of the residual variance estimates improved as the sample size increased. In the condition with N = 400, the MSE ratio was 5.38, whereas in the N = 100 condition, it was 13.15. This suggests that larger sample sizes led to more accurate estimates of the residual variance when using the global-local shrinkage prior methods.

The intercept parameter shows results in the opposite direction. Specifically, at N = 400, the MSE (accuracy) of the global-local shrinkage prior intercept was 10.40 times larger than the complete-data MSE. However, at N = 100, this MSE was only about 5.86 times larger. For the spike-slab, the MSE ratio was 1.74 and 2.05 for N = 100 and N = 400 conditions, respectively. Initially, it may seem counterintuitive to have lower accuracy with a larger sample size, however both the squared bias in the numerator and theoretical MSE in the denominator decrease with increasing sample size. For the intercept parameter, it appears that the quantity in the denominator shrinks at a proportionally faster rate than the numerator. For all other estimates/methods, increasing the sample size did not proportionally improve their accuracy relative to the true values. The accuracies of the missing data estimates maintained similar magnitudes compared to the squared theoretical *SE* (complete-data MSE).

In Figure 3, the third row panel shows a stronger intercorrelation of .40 among the predictors and a sample size of 100. This allowed for a comparison of the impact of predictor intercorrelation ($\rho = .10$ vs $\rho = .40$) with the first-row panel. The stronger predictor intercorrelation had a noticeable effect on the slope coefficients of the global-local shrinkage priors. At a predictor intercorrelation of .10, the slope estimates had an average MSE ratio of 1.87. However, with a higher predictor intercorrelation of .40, the accuracy improved, resulting in a MSE ratio of 1.67. Although the accuracy of the slope coefficient estimates was more comparable among the BVS methods, the spike-and-slab approach still yielded the most accurate results. In terms of the residual variance and intercept, increasing the predictor intercorrelation did not lead to a proportional improvement in accuracy compared to the true values. Similar to the first and second row panels, the spike-and-slab method produced substantially more accurate estimates for both the residual variance and intercept parameters.

The fourth row panel of the trellis-plot displays a predictor intercorrelation of .40 and a higher sample size of 400. This allows for an examination of whether the impact of predictor correlations varies depending on the sample size. If there were an interaction between predictor intercorrelation and sample size, the difference observed between the first and third row panels would not be the same as the difference between the second and fourth row panels. Or equivalently, the difference between the first and second row panels would be approximately equivalent to the difference between the third and fourth row panels. However, the findings in Figure 3 indicate that there is no interaction between predictor correlations and sample size conditions.

Summary of Results for Complexity 4

The findings of Complexity 4 are presented in Figures 1-3, showcasing the outcomes of the most complex missingness model condition. These figures display the bias, standardized bias, and MSE ratio. The study compared three BVS models across various conditions, including sample size, intercorrelation, and missing data complexity. The results, as indicated by the relative and standardized bias, revealed that the spike-and-slab prior yielded less biased estimates compared to the global-local shrinkage priors (horseshoe and Bayesian LASSO) for the intercept, slope coefficients, and residual variance estimates. Because bias only indicates deviations from the average estimate, it's important to note that an estimate with low bias may still be inaccurate due to wide variation in the estimates. To account for this, the MSE of the estimates was also examined. The MSE ratio results demonstrated that the spike-and-slab prior outperformed the global-shrinkage priors in terms of accuracy for slope estimates, residual variance, and intercept across all conditions. Generally, the accuracy of slope coefficient estimates did not differ significantly, although the spike-and-slab approach still yielded the most precise results. On the other hand, for residual variance and intercept, the spike-and-slab method consistently produced substantially more accurate estimates compared to the global-local prior methods.

Complexity 3

Figures 4 to 6 illustrate the condition where the true missingness model was comprised of the outcome variable, slope coefficients X_1 through X_4, and one auxiliary variable as predictors. This condition is compared to the complexity 4 condition, which includes not only these variables but also the slope coefficients X_5 and X_6 as predictors of the true missingness model. As a reminder, all BVS model fitted the most complex missingness model, which for the complexity 3 condition, means that these models are overfitted and are using variable selection to identify exclusion restrictions.

Percent and Standardize Bias

After examining Figures 4 and 5, significant disparities in bias were observed between the slope coefficients that acted as predictors of missingness in the data-generating model and those that were not involved in the missingness prediction. As a result, I will provide distinct explanations for the average bias regarding the slope coefficients that served as predictors (X_1 through X_4) of missingness in the data-generating model and those that were not predictors (X_5 and X_6).

In the first row panel of both figures, where the sample size is 100 and the intercorrelation of predictors is .10, the spike-and-slab prior produced less biased estimates than the global-shrinkage priors across all the regression slope estimates, residual variance, and intercept. However, it should be noted that the global-shrinkage priors exhibited less bias specifically for the *R*-squared parameter. The spike-and-slab model displayed a highly biased estimate of 71%, while the global-shrinkage prior had an average bias of 46%. Similar to the complexity 4 condition, the inflated *R*-squared estimate for the spike-and-slab model can be attributed to the excessively small spread of the distribution of the residual variance estimates.

Moving to the slope parameters in the first row panel of Figure 4, the spike-and-slab had -18% bias in the slope estimates for X_1 through X_4 slope estimates (predictors of missingness in the true data-generating model) in comparison to the -35% average bias for the global-shrinkage priors. For the slope estimates X_5 and X_6 (not predictors in the true data-generating model), the spike-and-slab showed a bias of 5%, on average, and the global-local priors showed a bias of 14% on average. The residual variance estimates had virtually no bias for the spike-and-slab, whereas the global-shrinkage priors showed a bias of nearly 37%. The standardized bias for the intercept estimates in Figure 5 were -0.18 and -0.33 for the spike-and-slab and global-shrinkage priors, respectively. In general, the two measures of bias in Figure 4 and 5 showed the same pattern, with the spike-and-slab exhibiting less bias than the horseshoe prior and Bayesian LASSO. The one exception was the *R*-squared estimate.

In the second row panel of Figure 4, as the sample size increased to 400, the percent bias decreased for all BVS methods, with the spike-and-slab method exhibiting the smallest bias. Specifically, in the *N* = 400 condition, the spike-and-slab method had a 10% bias for the X_1 through X_4 slope estimates, compared to an -18% bias in the *N* = 100 condition. The global-local shrinkage priors also experienced a decrease in bias, from -35% to -18% in the *N* = 400 condition. For the residual variance estimates, the spike-and-slab prior continued to show nearly no bias, while the global-local shrinkage priors demonstrated a substantial reduction in bias from 37% in the *N* = 100 condition to 10% in the *N* = 400 condition. The *R*-squared estimate also exhibited a decrease in bias, with the spike-and-slab prior having an 11% biased estimate, while the global-local shrinkage priors had an average bias of 5%.

The second row panel of Figure 5, the standardized bias is depicted, providing insight into the position of the average estimates within a theoretical sampling distribution. There were no noticeable differences between the first and second row panels for the spike-and-slab. This result suggests that the mean estimates of the spike-and-slab method from the two sample sizes fell approximately at the same position in their respective complete-data sampling distributions. On the other hand, the global-local prior showed a slight increase in standardized bias for the X_1 through X_4 slope estimates, indicating that in the *N* = 400 conditions, bias did not decrease as rapidly as in its respective complete-data sampling distribution.

The third row panel of the trellis plot in Figure 4 depicted estimates from the condition with a stronger predictor correlation of .40, along with a sample size of 100. This allowed for a comparison of the effect of predictor correlation ($\rho = .10$ vs $\rho = .40$) with the first row panel. As depicted in Figure 4, the average bias across the slope estimates X_1 through X_4 slightly increased for all methods. The spike-and-slab prior exhibited a bias of -23% (-18% in the first-row panel), while the global-local shrinkage priors demonstrated an average bias of -40% (-35% in the first-row panel). However, the presence of highly correlated predictors did not influence the bias in the slope estimates X_5 and X_6 , residual variance, or the R-squared. Moving to the third row panel in Figure 5, the standardized bias displayed a similar pattern. The presence of highly correlated predictors did not affect the slope estimates X_5 and X_6, residual variance, R-squared, and intercept, and only slightly reduced the bias in the slope estimates X_1 through X_4.

The fourth row panel in Figure 4 and Figure 5, displayed a predictor correlation of .40 and a higher sample size of 400, enabling us to investigate whether the effect of the predictor correlations varied as a function of sample size. If there was an interaction between predictor correlation and sample size, a difference between the first and third row panel would not be the same as the difference between the second and fourth row panel. Both Figure 4 and 5 demonstrated that the difference between predictor intercorrelation remained uniform across both sample size conditions. These results indicate that there is no interaction between predictor intercorrelation and sample size conditions.

Finally, the impact of missingness model complexity can be compared the between the complexity 4 and complexity 3 conditions. One notable difference between the two complexities is that in the complexity 4 condition, all slope coefficients within a row panel display similar bias values. However, in the complexity 3 condition, there is a distinct pattern where slope coefficients X_1 through X_4 generally exhibit higher bias compared to slope coefficients X_5 and X_6. This distinction between slope coefficients in complexity 3 arises because X_5 and X_6 are not predictors of the data generating model in the complexity 3 condition.

MSE Ratio

Figure 6 illustrate the MSE ratios for the complexity 3 conditions. Under this complexity condition, the true data generating missingness model was comprised of the outcome variable, slope coefficients $\boldsymbol{X_1}$ through $\boldsymbol{X_4}$, and one auxiliary variable as predictors. This figure displays the MSE ratio.

In the first row panel of Figure 6, with a sample size of 100 and a predictor correlation of .10, the spike-and-slab prior yielded more accurate estimates (i.e., MSE ratios closer to 1) compared to the global-shrinkage priors for the regression slope estimates, residual variance, and intercept. However, the differences in accuracy between the methods were more pronounced for the intercept and residual variance parameters. When examining the slope estimates X_1 through X_4, the spike-and-slab MSE was 1.57 times larger than the MSE from complete-data analysis, while the average MSE ratio for the global-shrinkage priors was 1.91. The bias values for the slope estimates X_5 and X_6 did not exhibit a substantial difference between the methods, although the spike-and-slab approach remained the more accurate method. For the residual variance estimates, the spike-and-slab method yielded, on average, a ratio of 2.12 between the MSE and the complete-data MSE and a MSE ratio of 1.72 for the intercept parameter. In comparison, the global-shrinkage priors displayed much higher MSE ratios for the residual variance (12.72) and the intercept (5.67).

In the second row panel of Figure 6, a larger sample size of 400 was used, enabling a comparison of the impact of sample size with the first row panel. The increased sample size affected the accuracy of the intercept parameter and the residual variance for the global-local shrinkage prior methods. On the other hand, for the spike-and-slab approach, the sample size only influenced the accuracy of the intercept. With the global-local shrinkage priors, the accuracy of the residual variance estimates improved as the sample size increased. In the condition with *N* = 400, the MSE ratio was 5.61, indicating greater accuracy compared to the *N* = 100 condition where the MSE ratio was 12.72. This suggests that larger sample sizes lead to more accurate estimates of the residual variance when using the global-local shrinkage prior methods.

However, the results for the intercept parameter showed the opposite trend. Specifically, at *N* = 400, the MSE of the global-local shrinkage prior intercept was 10.35 times larger than the complete-data MSE. In contrast, at *N* = 100, this MSE was only about 5.67 times larger. For the spike-and-slab approach, the MSE ratio was 1.72 for the *N* = 100 condition and 2.29 for the *N* = 400 condition. Regarding the intercept parameter, it seems that the denominator of the MSE ratio, the squared theoretical *SE*, decreases at a proportionally faster rate than the numerator. For all other estimates and methods, increasing the sample size did not lead to proportionally improved accuracy relative to the true values. The accuracies of the missing data estimates remained at similar magnitudes compared to the squared theoretical *SE*.

In Figure 6, the third row panel presents the results for a stronger predictor intercorrelation of .40 and a sample size of 100. This allows for a comparison of the impact of predictor intercorrelation ($\rho = .10$ vs $\rho = .40$) with the first-row panel. The stronger predictor intercorrelation had a moderate effect on the slope coefficients X_1 through X_4 of the global-local shrinkage priors. When the predictor correlation condition was .10, the slope estimates X_1 through X_4 had an average MSE ratio of 1.91. However, with a stronger predictor intercorrelation of .40, the accuracy improved, resulting in a MSE ratio of 1.73. In contrast, the slope coefficients X_5 and X_6 showed no difference in MSE ratio between the .10 and .40 intercorrelation conditions. Regarding the residual variance and intercept, increasing the predictor intercorrelation did not lead to a proportional improvement in accuracy compared to the true values. Similar to the first and second row panels, the spike-and-slab method produced more accurate estimates for all the slope coefficients and substantially more accurate estimates for both the residual variance and intercept parameters.

In the fourth row panel of the trellis plot, a predictor correlation of .40 and a higher sample size of 400 are depicted. This allows for an investigation of whether the effect of predictor correlations varies depending on the sample size. If there is an interaction between predictor correlation and sample size, the difference observed between the first and third row panels would not be the same as the difference between the second and fourth row panels. Alternatively, if there is no interaction, the difference between the first and second row panels would be approximately equivalent to the difference between the third and fourth row panels. However, the findings in Figure 6 indicate that there is no interaction between predictor correlations and sample size conditions, because the difference in ratios between predictor intercorrelation .10 and .40 were approximately equivalent in the 100 and 400 sample size conditions.

Lastly, the impact in accuracy of missingness model complexity between the complexity 4 and complexity 3 conditions can be compared. One notable difference between the two is that in the complexity 4 condition, all slope coefficients within a row panel display similar values for the MSE ratio. However, in the complexity 3 condition, there is a distinct pattern where slope coefficients X_1 through X_4 generally exhibit less accuracy than the slope coefficients X_5 and X_6. Again, this distinction between slope coefficients in complexity 3 arises because X_5 and X_6 are not predictors of missingness in the true data generating model.

Summary of Results for Complexity 3

The results of Complexity 3, depicted in Figures 4 to 6, provide insights into the scenario where the true missingness model only has outcome variable Y, A_1, and X_1 through X_4 as its predictors. The findings, based on relative and standardized bias, indicate that the spike-and-slab prior yields less biased estimates compared to the global-local shrinkage priors for the intercept, slope coefficients, and residual variance. To assess the variation in the estimates, the MSE is also examined. The MSE ratio results demonstrate that the spike-and-slab prior outperforms the global-local shrinkage priors in terms of accuracy for slope estimates, residual variance, and intercept across all conditions. Notably, the spike-and-slab method consistently produced substantially more accurate estimates for residual variance and intercept compared to the global-local prior methods. An important difference between complexity 4 and 3 is in the bias and accuracy of slope coefficient estimates. In Complexity 3 there is a distinction between the slope coefficients X_5 and X_6 (not included as predictors in the true data-generating model) and the slope coefficients X_1 through X_4. The former exhibit less bias and greater accuracy compared to the latter. Complexity 4 did not show this difference between slope coefficients.

Complexity 2

Figures 7 to 9 illustrate the percent bias, standardized bias, and MSE ratio, respectively, for the condition where the true data-generating model was comprised of the outcome variable, variables X_1 and X_2, and one auxiliary variable as predictors. The complexity 2 condition does not include variables X_3 through X_6 as predictors of missingness in the data-generating model. As a reminder, All BVS models fitted the most complex missingness model including outcome variable, variables X_1 through X_6, and one auxiliary variable as predictors, and will use variable selection to find exclusion restrictions.

Percent and Standardize Bias

Similar to the complexity 3 condition, Figures 7 and 8 showed notable differences in bias between the slope coefficients that serve as predictors and those that do not contribute to the data-generating model. As a result, I will provide separate descriptions of the average bias for slopes that are predictors (X_1 and X_2) and those that are not predictors (X_3 through X_6) of missingness in the data-generating model.

In the first row panel of both figures, where the sample size is 100 and the intercorrelation of predictors is .10, the spike-and-slab method yielded more accurate estimates compared to the global-local shrinkage priors across various regression slope estimates, residual variance, and intercept. The only exception was the *R*-squared parameter, where the global-local shrinkage priors demonstrated less bias at 47% compared to the 73% bias from the spike-and-slab approach. Similar to the complexity 4 and 3 conditions, the spike-and-slab model's inflated *R*-squared estimate could be attributed to the excessively narrow distribution of the residual variance estimates. For the slope estimates of variables X_1 and X_2 (predictors in the data-generating model), the spike-and-slab method exhibited a bias of -26%, whereas the global-shrinkage priors showed an average bias of -49%.

Regarding the slope estimates of variables X_3 through X_6 (not predictors of missingness in the data-generating model), the spike-and-slab method displayed an average bias of -7%, while the global-local shrinkage priors exhibited an average bias of -14%. The spike-and-slab method showed virtually no bias in the residual variance estimates, while the global-shrinkage priors had a bias of 39%. In Figure 8, the standardized bias for the intercept estimates was -0.19 for the spike-and-slab method and -0.32 for the global-shrinkage priors. In general, both Figures

7 and 8 exhibited a consistent pattern where the spike-and-slab method yielded less bias compared to the horseshoe prior and Bayesian LASSO for most parameters.

In the second row panel of Figure 7, as the sample size increased to 400, the percent bias decreased for all BVS methods, with the spike-and-slab method exhibiting the smallest bias. Specifically, for the X_1 and X_2 slope estimates in the *N* = 400 condition, the spike-and-slab and global-local shrinkage methods had -18% and -32% bias, respectively, compared to -26% and 49% bias in the *N* = 100 condition. However, looking at the standardized bias in the second row panel of Figure 8, it can be observed that those same X_1 and X_2 slope estimates had a slight increase in standardized bias. This shows that even though bias did decrease with the increase of sample size, bias did not decrease as rapidly as in its *N* = 400 expected complete-data sampling distribution. For the residual variance estimates, the spike-and-slab prior continued to show nearly no bias, while the global-local shrinkage priors demonstrated a substantial reduction in bias from 39% in the *N* = 100 condition to 11% in the *N* = 400 condition. The R-squared estimate also exhibited a decrease in bias, with the spike-and-slab prior having an 12% biased estimate, while the global-local shrinkage priors had an average bias of 4%.

In the trellis plot of Figure 7, the third row panel examined a stronger predictor intercorrelation of .40, with a sample size of 100, allowing for a comparison of the effect of predictor correlation ($\rho = .10$ vs. $\rho = .40$) with the first row panel. As depicted in Figure 7, all methods showed a moderate increase in average bias across the slope estimates X_1 and X_2 when the predictor correlation increased. The spike-and-slab prior exhibited a bias of -36% (compared to -26% in the first-row panel), while the global-local shrinkage priors demonstrated an average bias of -61% (compared to -49% in the first-row panel). However, when looking at the standardized bias in the third row panel of Figure 8, it can be observed that the same X_1 and

X_2 slope estimates did not increase in .40 predictor correlation condition. This suggests that although percent bias increased in the .40 correlation condition, the mean estimates from the two predictor correlation conditions fell approximately at the same position in their respective complete-data sampling distributions.

Moving on to other parameters in the third row panel, the presence of highly correlated predictors did not influence the bias in the slope estimates X_3 through X_6, the residual variance, or the *R*-squared. This same pattern was observed in Figure 8, where strongly intercorrelated predictors did not influence these parameters, including the intercept.

In the fourth row panel of Figure 7 and Figure 8, a predictor intercorrelation of .40 and a higher sample size of 400 were presented, allowing for an examination of whether the effect of predictor correlations varied with sample size. If there was an interaction between predictor correlation and sample size, the difference between the first and third row panel would not be the same as the difference between the second and fourth row panel. However, both Figure 7 and Figure 8 demonstrated that the difference between the intercorrelation conditions remained consistent across both sample size conditions. These results indicate that there is no interaction between predictor correlations and sample size conditions.

Finally, we can examine the effect of varying complexity levels in the missingness model on bias by comparing three conditions: complexity 4, complexity 3, and complexity 2. Complexity 2 and 3 show the same pattern, where the variables that were excluded from the data generating model generally exhibit lower bias compared to the variables that were included. When comparing complexity 3 and 2, a noticeable difference in bias is observed for the predictors of missingness included in the data-generating model (X_1 through X_4, in complexity 3 and X_1 through X_2 in complexity 2). For example, when comparing the first row panel of Figure

4 and Figure 7, the average bias for slope coefficients X_1 through X_4 in complexity 3 was -18% for the spike-and-slab and -35% for the global-local shrinkage priors. In contrast, the average bias for slope coefficients X_1 and X_2 was larger, with a spike-and-slab bias of -26% and a global-local shrinkage prior bias of -49% (Figure 7). This trend is consistent across different sample sizes and intercorrelations.

MSE Ratio

Figure 9 illustrate the condition where the true missingness model was comprised of the outcome variable, slope coefficients X_1 and X_2 , and one auxiliary variable as predictors. This figure displays the MSE ratio. In the first row panel of Figure 9, with a sample size of 100 and a predictor correlation of .10, the spike-and-slab prior provided more accurate estimates (i.e., MSE ratios closer to 1) compared to the global-shrinkage priors for the regression slope estimates, residual variance, and intercept.

The differences in accuracy between the methods were particularly pronounced when examining the slope estimates for the variables that were included as predictors of missingness in the data generating model. For example, MSE for the spike-and-slab method was 1.75 times larger than the MSE from the complete-data analysis for the slope estimates of variables X_1 and X_2, while the average MSE ratio for the global-shrinkage priors was 2.25. On the other hand, the bias values for the variables not included in the data-generating model (X_3 through X_6) did not show a substantial difference between the spike-and-slab and global-local methods, although the spike-and-slab approach still remained the more accurate method. Regarding the residual variance and intercept estimates, once again, the spike-and-slab method exhibited a clear advantage. The spike-and-slab approach yielded, on average, a ratio of 2.13 between the MSE and the complete-data MSE for the residual variance and a MSE ratio of 1.73 for the intercept parameter. In comparison, the global-shrinkage priors displayed much higher MSE ratios for the residual variance (13.23) and the intercept (5.37).

In the second row panel of Figure 9, a larger sample size of 400 was used, enabling a comparison of the impact of sample size with the first row panel. The increased sample size affected the accuracy of the intercept parameter and the residual variance for the global-local shrinkage prior methods. On the other hand, for the spike-and-slab approach, the sample size only influenced the accuracy of the intercept. With the global-local shrinkage priors, the accuracy of the residual variance estimates improved as the sample size increased. In the condition with *N* = 400, the MSE ratio was 6.50, indicating greater accuracy compared to the *N* = 100 condition where the MSE ratio was 13.23. The results for the intercept parameter showed the opposite trend. Specifically, at *N* = 400, the MSE of the global-local shrinkage prior intercept was 10.97 times larger than the complete-data MSE. In contrast, at *N* = 100, this MSE was only about 5.37 times larger. For the spike-and-slab approach, the MSE ratio was 1.73 for the *N* = 100 condition and 2.69 for the *N* = 400 condition.

Figure 9's third row panel displays the outcomes for a stronger predictor correlation of .40 and a sample size of 100. This permits a comparison between the impact of predictor correlation ($\rho =$ .10 vs $\rho =$ .40) with the panel in the first row. The stronger predictor intercorrelation moderately influenced the slope coefficients X_1 and X_2 of the global-local shrinkage priors. Under the .10 predictor intercorrelation condition, the average MSE ratio for the global-local method's slope estimates X_1 and X_2 was 2.25. However, with a stronger predictor intercorrelation of .40, the accuracy improved, resulting in a MSE ratio of 2.03. In terms of the slope coefficients X_3 through X_6, residual variance, and intercept, increasing the predictor correlation did not lead to a proportional enhancement in accuracy compared to the true values.

Like the panels in the first and second rows, the spike-and-slab method generated more accurate estimates for all the slope coefficients and significantly more accurate estimates for both the residual variance and intercept parameters.

The fourth row panel in the trellis plot portrays a scenario with a predictor correlation of .40 and a larger sample size of 400. This setup allows for an examination of whether the impact of predictor correlations varies based on the sample size. If there is an interaction between predictor correlation and sample size, the disparity observed between the first and third row panels would not be the same as the difference between the second and fourth row panels. Conversely, if there is no interaction, the distinction between the first and second row panels would be approximately similar to the difference between the third and fourth row panels. However, the results depicted in Figure 9 indicate that there is no interaction between predictor correlations and sample size conditions, because the difference between first and second row panels are similar to the difference between the third and fourth row panels.

Lastly, we can examine the effect of varying complexity levels in the missingness model on bias by comparing three conditions: complexity 4, complexity 3, and complexity 2. Notably, there is a distinction between complexity 4 and complexity 2 regarding the behavior of slope coefficients within a row panel. In complexity 4, these coefficients exhibit similar MSE ratios, whereas in complexity 2, there is a clear pattern where slope coefficients X_1 and X_2 generally have lower accuracy compared to slope coefficients X_3 through X_6. This accuracy difference arises because X_3 through X_6 are not missingness predictors in the data-generating model. When comparing complexity 3 and 2, a noticeable difference in accuracy is observed for the variables included as missingness predictors in the data generating model (X_1 through X_4 in complexity 3 and X_1 and X_2 in complexity 2). For example, when comparing the first row panel of Figure 6 and Figure 9, the average MSE ratio for slope coefficients X_1 through X_4 in complexity 3 was 1.57 for the spike-and-slab and 1.91 for the global-local shrinkage priors. In contrast, the average MSE ratio for slope coefficients X_1 and X_2 is less accurate, with a spike-and-slab ratio of 1.75 and a global-local shrinkage prior ratio of 2.25 (Figure 9). This trend is consistent across different sample sizes and intercorrelations.

Summary of Results for Complexity 2

The results of Complexity 2, depicted in Figures 7-9, shed light on a scenario where the true missingness model did not include four variables (X_3 through X_6) from the substantive model. Figures 7 and 8 illustrate the percent and standardized bias, respectively, while Figure 9 presents the MSE ratio for Complexity 2. Overall, the spike-and-slab method provided more accurate and less biased estimates compared to the global-local shrinkage priors. Across all figures, notable differences were observed in bias and MSE ratios between variables that acted as predictors of missingness (X_1 and X_2) and those that did not (X_3 through X_6) in the data-generating model. Generally, the slope estimates for X_1 and X_2 exhibited lower accuracy and higher bias compared to X_3 through X_6. When comparing complexity 2 with complexity 3, it was evident that the average MSE ratios and percentage bias for the substantive variables that were also missingness predictors in the data-generating model (X_1 and X_2) were less accurate compared to the corresponding variables in complexity 3. This highlights the impact of the predictors included in the data-generating model on accuracy and bias.

Complexity 1

Figures 10 to 12 illustrate the condition where the true missingness model includes only the outcome variable and one auxiliary variable as predictors. In the complexity 1 condition, all substantive variables (X_1 through X_6) are not missingness predictors in the data-generating model. Again, all BVS methods are fitting the most complex missingness model and employing variable selection to establish exclusion criteria.

Percent and Standardize Bias

In the first row panel of Figures 10 and 11, the sample size was set to 100, and the predictors had an intercorrelation of .10. Across various regression slope estimates and intercepts, the spike-and-slab method demonstrated less bias in its estimates compared to the global-local shrinkage priors. However, unlike the previous complexity conditions, the differences in bias between the two methods were relatively small. Specifically, for slope estimates, the spike-and-slab method had a bias of -9%, while the global-shrinkage priors showed an average bias of -14%. In Figure 11, the standardized bias for intercept estimates was -0.24 for the spike-and-slab method and -0.33 for the global-shrinkage priors.

Regarding the residual variance and *R*-squared parameters, the differences between the global-local shrinkage priors and spike-and-slab method were consistent with the results from complexity conditions 2 to 4. The global-local shrinkage priors exhibited less bias in the *R*-squared parameter, with a bias of 47% compared to the 72% bias from the spike-and-slab approach. The spike-and-slab method showed almost no bias in the residual variance estimates, while the global-shrinkage priors had a bias of 44%. In general, both Figures 10 and 11 displayed a consistent pattern where the spike-and-slab method yielded less bias compared to the horseshoe prior and Bayesian LASSO for most parameters.

In the second row panel of Figure 10 the sample size increased to 400. Increasing the sample size did not lead to a substantial decrease in bias for the slope estimates. Specifically, for the slope estimates in the *N* = 400 condition, the spike-and-slab and global-local shrinkage methods had -6% and -10% bias, respectively, compared to -9% and 14% bias in the *N* = 100 condition. The same pattern is found in the standardized bias in the second row panel of Figure 11. It can be observed that the mean slope estimates from the two sample size conditions fell approximately at the same position in their respective complete-data sampling distributions. For the residual variance estimates, the spike-and-slab prior showed a percent bias of -5%, while the global-local shrinkage priors demonstrated a substantial reduction in bias from 44% in the N = 100 condition to 12% in the N = 400 condition. The R-squared estimate also exhibited a decrease in bias, with the spike-and-slab prior having an 16% biased estimate, while the global-local shrinkage priors had an average bias of 1%.

In the trellis plot shown in Figures 10 and 11, the third row panel investigated a stronger intercorrelation (.40) between predictors, using a sample size of 100. This allowed for a comparison of the impact of predictor correlation ($\rho = 0.1$ vs. $\rho = 0.4$) with the first row panel. The figures illustrate that increasing the predictor intercorrelation did not significantly increase bias in the slope estimates, residual variance, intercept, or R-squared estimates.

In the fourth row panel of Figure 10 and Figure 11, a higher sample size of 400 was used along with a predictor correlation of .40, to examine if the effect of predictor correlations varied with sample size. However, both Figure 10 and Figure 11 showed that the difference between the intercorrelation conditions remained consistent regardless of sample size. These findings suggest that there is no interaction between predictor correlations and sample size conditions.

To evaluate the impact of varying complexity levels in the true missingness model on bias, the different complexity conditions can be compared. Complexity 1 and 4 exhibit a similar pattern, with all slope coefficients in a row panel showing comparable bias values. In contrast, complexity 2 and 3 demonstrate differences in bias between variables that were and were not predictors of missingness in the data generating model. When comparing complexity 1 (the least complex model) to complexity 4 (the most complex model), it can be observed that the differences in bias between the BVS methods are more pronounced in the most complex conditions. In the least complex conditions, the bias values are relatively comparable across methods, except for the residual variance parameter, where the spike-and-slab method consistently outperforms the others.

MSE Ratio

Figure 12 illustrate the condition where the true missingness model was comprised of the outcome variable and one auxiliary variable as predictors. This figure displays the MSE ratio. In the first row panel of Figure 12, where the sample size was 100 and the predictor intercorrelation was 0.1, there was not a substantial difference in accuracy between the spike-and-slab and global-local methods for the slope coefficients. However, the spike-and-slab approach still showed a slight advantage. When it comes to estimating the residual variance and intercept, the spike-and-slab method clearly outperformed the global-local method. The spike-and-slab approach resulted in an average ratio of 2.49 between the MSE and the complete-data MSE for the residual variance, and a ratio of 2.25 for the intercept parameter. On the other hand, the global-local method exhibited much higher MSE ratio for the residual variance (14.87) and the intercept (6.05).

In the second row panel of Figure 12, a larger sample size of 400 was utilized, allowing for a comparison of the sample size's impact with the first row panel. The increased sample size resulted in improved accuracy of the slope coefficients for all BVS methods. Specifically, for the slope estimates in the N = 400 condition, the spike-and-slab and global-local shrinkage methods exhibited an MSE ratio of 1.35 and 1.49, respectively, compared to 1.55 and 1.70 in the N = 100 condition. Similarly, the accuracy of the residual variance estimation also improved as the sample size increased. In the N = 400 condition, the MSE ratio for the global-local shrinkage priors was 7.62, indicating greater accuracy compared to the N = 100 condition where the MSE ratio was 14.87. However, the results for the intercept parameter showed the opposite trend. Specifically, at N = 400, the MSE of the global-local shrinkage prior intercept was 12.45 times larger than the complete-data MSE. In contrast, at N = 100, this MSE was only about 6.05 times larger. For the spike-and-slab approach, the MSE ratio was 2.25 for the N = 100 condition and 4.32 for the N = 400 condition.

The third row panel of Figure 12 presents the results for a stronger predictor correlation of .40 and a sample size of 100. This allows for a comparison of the impact of predictor correlation ($\rho = .10$ vs. $\rho = .40$) with the first row panel. Surprisingly, the stronger predictor intercorrelation did not have a significant effect on the slope coefficients, residual variance, or intercept. Increasing the predictor correlation did not result in a proportional improvement in accuracy compared to the true values. Similar to the first row panel, the spike-and-slab method produced slightly more accurate estimates for all the slope coefficients and significantly more accurate estimates for both the residual variance and intercept parameters.

In the fourth row panel of the trellis plot, a scenario is depicted with a predictor correlation of .40 and a larger sample size of 400. This configuration enables an investigation into whether the impact of predictor correlations varies depending on the sample size. If there is an interaction between predictor correlation and sample size, the difference observed between the first and third row panels would not be the same as the difference between the second and fourth row panels. However, the results illustrated in Figure 12 suggest that there is no interaction between predictor correlations and sample size conditions. The disparities observed between the different predictor correlation conditions remain consistent across both the smaller and larger sample size conditions.

Lastly, the impact of varying complexity levels in the missingness model on the accuracy and bias of the estimates can be evaluated. It is worth noting a similarity between complexity 4, the most complex true missingness model, and complexity 1 in terms of the behavior of slope coefficients within a row panel. In complexity 1 and 4, there is less variability observed between MSE ratios. On the other hand, in complexity 2 and 3, there is a clear pattern where slope coefficients for predictors of missingness in the data-generating model generally exhibit lower accuracy compared to the other slope coefficients. Additionally, the influence of predictor intercorrelation on the accuracy of the estimates is less pronounced in the complexity 1 condition when compared to complexity 2 through 4.

Summary of Results for Complexity 1

The results of Complexity 1, depicted in Figures 10 to 12, shed light on a scenario where the true missingness model did not include any of the predictors from the substantive model. Figures 10 and 11 illustrate the percent and standardized bias, respectively, while Figure 12 presents the MSE ratio for Complexity 1. At the N = 100, the spike-and-slab method provided more accurate and less biased slope estimates compared to the global-local shrinkage priors, although this advantage was not substantial. In the N = 400 condition, all method had comparable slope estimates in terms of bias and accuracy. There was minimal influence of predictor intercorrelation on the accuracy or bias of the estimates in complexity 1.

Conclusion of Bayesian Variable Selection Methods Comparison

As stated in the introduction, the construction of a missingness model poses a significant practical challenge when applying MNAR selection models. Determining which variables should be included in the model is difficult, and statistical challenges arise due to the strict distribution assumptions required for estimation (Ibrahim et al., 2005). Relying solely on the available data may not provide sufficient information for accurate estimation. To address this, one approach is to identify Type C auxiliary variables that exhibit correlation with the missing data indicator but not with the variables in the substantive model (Collins et al., 2001). This process often involves dealing with a high-dimensional selection problem, as researchers have access to numerous measured variables that are not included in the model. Another way to improve estimation is by removing overlapping variables from the focal and missingness models. Since BVS methods have not been explored in this context, the focus is on the more specific problem involving overlapping predictors.

The simulations presented in this study, examined four missingness model complexity condition. At one end of the spectrum, the most complex true data-generating model (Complexity 4) included all variables, making it a more challenging problem for variable selection since all variables influenced missingness. At the other end, the complexity 1 condition included only the outcome variable Y and an auxiliary variable A_1 in the data-generating model, requiring the exclusion of all X predictors. Analyzing the simulation results revealed a consistent finding: when the missingness model was overfit and BVS methods were applied, there were positive effects on the slope estimates of the focal model for the non-overlapping predictors. This study did not explore the selection of the correct missingness model, as it is not of interest in this MNAR modeling framework. However, the reduction in bias and MSE for the focal model predictors suggests that the BVS methods were effectively "turning off" unnecessary predictors in the missingness model. Furthermore, the spike-and-slab prior consistently outperformed the global-local shrinkage priors, with few exceptions (the R-square parameter).

The following section aims to address the question of whether BVS offers improvements compared to analyses conducted without variable selection. This assessment includes an analysis under the assumption of MAR, where missing data assumptions are violated, as well as the correctly-specified selection model (true model), and the overfitting selection model that incorporates all focal variables in the missingness model (representing the correct specification for the complexity 4 condition). These approaches represent the currently utilized procedures available to researchers. The subsequent section of the results will compare the performance of the spike-and-slab method against these existing approaches.

Spike and Slab Comparison to Existing Methods

Figures 1 to 12 compared bias estimates and MSE ratios for the three Bayesian variable selection methods under different sample size, intercorrelation, and missing data complexity conditions. In general, the results showed that the spike-and-slab prior had less biased and more accurate estimates than the global-local shrinkage priors (horseshoe and Bayesian lasso) for the intercept, regression coefficients, and residual variance estimates. In the following sections, the spike-and-slab method will be compared with the full selection model, MAR model, and true data-generating model.

Complexity 4

Figures 13 to 15 displays complexity 4, which is the most complex true missingness model, including all variables from the focal regression and the auxiliary variable. Percent bias is shown in Figure 13, standardized bias is shown in Figure 14, while MSE ratio are shown in Figure 15. In the most complex missingness model condition (complexity 4), the full selection model and the true model are identical, as they both fit the most complex missingness model without utilizing BVS. Consequently, Figures 13 to 15 will exhibit overlapping results between full selection and true model. The spike-and-slab is also fitting the most complex missingness model and employing variable selection to establish exclusion criteria.

Percent and Standardize Bias

The first row panel of Figures 13 and 14, depict a sample size of 100 and a .10 predictor intercorrelation. The MAR model had an average bias of -23% across all six slope coefficients, while the spike-and-slab had an average bias of -13%. Additionally, the true and full selection models had a bias of -65% across all six slope coefficients. For the residual variance estimates bias, the MAR model displayed a -6% bias, while the true and full selection models had an 88% bias. The spike-and-slab model had minimal bias for the residual variance. As for the *R*-squared estimate, the spike-and-slab showed a biased estimate of 72%, and the MAR model exhibited a high bias percentage of 62%. However, the true and full selection models had a low bias of -2%. The standardized bias of the intercept parameter in Figure 14 was -0.19 for the spike-and-slab prior and -0.24 for the MAR model, while the true and full selection models had a higher standardized bias of -0.80. Overall, the MAR and spike-and-slab models had less bias than the full selection and true models for all parameters, except for the *R*-squared estimate. The spike-and-slab method, in general, exhibited the least biased estimates.

The second row panel of Figure 13, where the intercorrelation is .10 and the sample size increased to 400, shows that a larger sample size benefitted the estimation of the regression coefficients for the spike-and-slab, full selection model, and true model. The spike-and-slab prior displays the lowest average bias of -8% (-13% in the first row panel), while the true and full selection models exhibit an average bias of -46% (-65% in the first row panel) across slope coefficients. However, the MAR model did not show a significant reduction in bias in the N = 400 condition, with an average bias of -22%. This is expected because the MAR model estimates are not consistent under an MNAR process (i.e., do not achieve unbiasedness in large samples). The residual variance estimates for the spike-and-slab prior again had almost no bias, while the residual variance for the MAR model did not change significantly. On the other hand, the full selection model and the true model benefited from an increase in sample size as it reduced bias for the residual variance to a 32% bias (88% in the first row panel). In terms of the *R*-squared estimate, both the MAR model and the spike-and-slab prior improved with a higher sample size, with a 3% bias for the MAR model (62% in the first row panel) and a 10% biased estimate for the spike-and-slab prior (72% in the first row panel). In contrast, the full selection and true models showed an increase in bias to -48% as the sample size increased.

In the second row panel of Figure 14, the bias relative to the theoretical complete-data *SE* is depicted. Standardized bias exhibits a different trend concerning the impact of sample size compared to the percent bias shown in Figure 13. While the spike-and-slab model shows no significant difference in standardized bias, the regression coefficients for the MAR model, full selection model, and true model demonstrate higher standardized bias in the second row panel compared to the first row panel in Figure 14. For instance, the standardized bias for the MAR model across slope coefficients increased from -0.25 to -0.50. In contrast, the percent bias remains constant between the first and second row panels in Figure 13, indicating that the mean point estimates are essentially identical. Therefore, the increase in standardized bias for the MAR model in Figure 14 can be mainly attributed to the decrease in theoretical standard error (used as the denominator in the standardized bias formula) as the sample size increases.

The third row panel of the trellis plot in Figure 13 and 14 depicted a stronger predictor intercorrelation ($\rho = 0.4$) and a sample size of 100, allowing for a comparison of the simple effect of predictor correlation with the first-row panel ($\rho = 0.1$). As depicted in Figure 13, both the MAR model and the spike-and-slab method did not exhibit a noticeable increase in percentage bias compared to the first-row panel. The spike-and-slab prior had an average bias of -17% (-13% in the first row panel) for the substantive model coefficients, while the MAR model showed an average bias of -26% (-23% in the first row panel). In contrast, the full selection and true models exhibited a 10% increase in bias under the .40 predictor correlation condition for the regression coefficients. This highlights that high predictor correlation severely affected bias in the estimates of the regression coefficients for the true and full selection models, while it did not cause significant issues for the spike-and-slab and MAR models. This pattern aligns with expectations for the full selection model, as increasing predictor intercorrelation diminishes the non-overlapping variation between the selection and missingness models, resulting in weakened identification and support from the data. There were no major differences between residual variance and *R*-squared between the first and third row panels for all models, showing that high correlation between the predictors did not affected bias in the residual variance and *R*-squared for all models.

The fourth row panel in Figure 13 and 14 depicts a scenario where a larger sample size of 400 was used, with a higher correlation between the substantive regression coefficients. The fourth row panel allows us to investigate whether the intercorrelation differences observed in the first and third row panels change with sample size compared to the second and fourth row panels. Both Figure 13 and Figure 14 demonstrate that the intercorrelation differences remain constant across both sample size conditions. The bias discrepancies between the first and third row panels indicate that the spike-and-slab, full selection model, and true model had more biased estimates in the .40 intercorrelation condition than the .10 intercorrelation condition, while the MAR model's bias estimates did not change. This pattern was also observed in the N = 400 condition between the second and fourth row panels, indicating no interaction between predictor correlation and sample size.

The fourth row panel in Figure 14 further highlights that the spike-and-slab method has the least bias among the models considered. The spike-and-slab prior displayed a bias of -0.24 standard error units for the intercept parameter, whereas the MAR model showed a standardized bias of -0.48, and the true and full selection models had a standardized bias of -1.15. The bias for the regression coefficients in the fourth row panel in Figure 14 show a similar trend as to that observed in Figure 13. However, there is a significant difference in the standardized and percentage bias for the residual variance parameter in the MAR model. The percentage bias metric shows a small bias of 9% compared to 1.25 standardize bias.

MSE ratio

Figure 15 illustrate MSE ratios for the complexity 4 conditions, which is the most complex true missingness model, including all variables from the focal regression and the auxiliary variable. The first row panel in Figure 15, depicts a sample size of 100 and a weaker intercorrelation of predictors ($\rho = .10$). In this scenario, both the spike-and-slab prior and the MAR model demonstrated superior accuracy in estimating regression slope, residual variance, and intercept when compared to the true and full-selection model. The disparities in accuracy between the spike-and-slab and MAR models were minimal for the slope estimates. However, the MAR model outperformed the spike-and-slab model in estimating residual variance, while the spike-and-slab model excelled in estimating the intercept.

Examining the slope estimates, MSE of the MAR model and spike-and-slab model were 1.50 and 1.58 times larger than the MSE of the complete-data analysis for all six predictors, respectively. In contrast, the MSE ratio for the true and full-selection model was, on average, 2.91. The spike-and-slab methods yielded residual variance estimates that were, on average, 2.20 times larger than the MSE from the complete-data analysis, while the MAR model was more accurate with an MSE that was 1.57 times larger. For the intercept, the spike-and-slab model had an MSE ratio of 1.74, while the MAR model had an MSE ratio of 2.39. The true and full-selection model exhibited much higher MSE ratios for the residual variance (47.01) and the intercept (24.51).

In Figure 15, the second row panel presents a larger sample size of 400, allowing for a comparison of sample size impact with the first row panel. The increased sample size affected the accuracy of the intercept parameter and the residual variance, but it did not influence the slope estimates. The accuracy of the residual variance estimates improved as the sample size increased in both the true and full-selection models. The MSE ratio decreased from 47.01 in the N = 100 condition to 27.21 in the N = 400 condition. On the other hand, for the intercept parameter, increasing the sample size in the true and full-selection model did not result in a proportional improvement in accuracy compared to the true values. The MSE ratio in the N = 400 condition was 53.56, while it was 24.51 in the N = 100 condition. In terms of the MSE model, increasing the sample size did not lead to a proportional improvement in accuracy for the intercept and residual variance compared to the true values. The MSE ratio for the residual variance and intercept was 1.57 and 2.39, respectively, in the N = 100 condition. However, in the N = 400 condition, these ratios increased to 3.12 and 7.78 for the residual variance and intercept, respectively.

In Figure 15, the third row panel displays a sample size of 100 and a stronger predictor intercorrelation ($\rho = .40$). This setup allowed for a comparison of the impact of predictor correlation with the first-row panel ($\rho = .10$). The increased predictor correlation had a noticeable effect on the slope coefficients of the true and full-selection models. When the predictor intercorrelation was .10, the slope estimates had an average MSE ratio of 2.91. However, with a stronger predictor intercorrelation of .40, the accuracy of the slope estimates improved proportionally to the MSE from complete-data analysis, resulting in a MSE ratio of 2.61. It is worth noting that the spike-and-slab approach and the MAR model consistently outperformed other methods in terms of accuracy across all parameters.

The fourth row panel in the trellis plot of Figure 15 showcases a higher sample size of 400 and a strong predictor intercorrelation ($\rho = .40$). This configuration allows for an examination of whether the impact of predictor correlations varies depending on the sample size. If there is an interaction between predictor correlation and sample size, the difference observed between the first and third row panels would not be the same as the difference between the second and fourth row panels. In other words, the discrepancy between the first and second row panels would be approximately equivalent to the difference between the third and fourth row panels. The findings presented in Figure 15 indicate the presence of an interaction between predictor intercorrelation and sample size conditions. Specifically, increasing the sample size resulted in more accurate slope estimates, proportional to the MSE from complete-data analysis, in the .40 intercorrelation condition (difference between the third and fourth row panels), in contrast to the .10 intercorrelation conditions (difference between the first and second row panels).

Summary of Results for Complexity 4

The findings of Complexity 4 are presented in Figures 13 to 15, showcasing the outcomes of the most complex missingness model condition. These figures display the bias, standardized bias, and MSE ratio. In this study, I conducted a comparison of the performance of a selection model with a spike-and-slab prior against an misspecified MAR model, full-selection model, and the true data-generating model. Looking at bias as an outcome, the results demonstrate that the spike-and-slab prior and MAR model consistently outperforms the full-selection and true models across regression coefficients, residual variance, and intercept estimates. The results also revealed that lower sample sizes and highly correlated predictors can increase bias in all the models.

When comparing the spike-and-slab model with the MAR model, the spike-and-slab consistently exhibits less bias across various conditions. This bias advantage becomes more pronounced in conditions with larger sample sizes. When assessing precision using the MSE ratio, the results indicate that both the MAR and spike-and-slab models perform similarly in terms of slope estimates across all conditions. This comparison highlights distinct strengths for each model. The spike-and-slab model's low bias suggests that its posterior distribution, which represents the uncertainty of the model's parameters, is centered around the true population parameter. On average, across many samples, the spike-and-slab model's estimates converge to the true values in the population. However, the approximately equal MSE ratio indicates that, within any given sample, both the spike-and-slab and MAR estimates deviate from the true values by the same amount. This implies that although the MAR model may have some bias, it has lower variance compared to the spike-and-slab model. As a result, in any individual sample, the performance of the MAR model is comparable to that of the spike-and-slab model. Overall, while the spike-and-slab model excels in reducing bias, the MAR model demonstrates lower variance. These factors contribute to their respective strengths, and understanding the trade-off between bias and variance is crucial in evaluating the performance of these models.

Complexity 3

Figures 16 to 18 illustrate the percent and standardized bias, and MSE ratio, respectively, for the condition where the true missingness model was comprised of the outcome variable, slope coefficients X_1 through X_4, and one auxiliary variable as predictors. As a reminder, the spike-and-slab model and the full-selection model fitted the most complex missingness model. In contrast, the MAR model fitted variables X_1 through X_6, and one auxiliary variable as predictors. Additionally, the true data generating model fitted outcome variable Y, slope coefficients X_1 through X_4, and auxiliary variable A_1 as predictors. Complexity 3 is compared to the complexity 4 condition, which includes not only these variables but also the slope coefficients X_5 and X_6 as predictors of missingness in the data-generating model.

Percent and Standardize Bias

Upon analyzing Figures 16 and 17, notable differences in bias were observed between the slope coefficients that acted as predictors of missingness in the data-generating model and those that were not involved in the missingness prediction. As a result, I will provide distinct explanations for the average bias regarding the slope coefficients that served as predictors (X_1 through X_4) of missingness in the data-generating model and those that were not predictors (X_5 and X_6).

In Figures 16 and 17, specifically in the first row panel, where the sample size is 100 and the predictors have an intercorrelation of .10, the estimates obtained from the spike-and-slab prior demonstrated substantially lower bias compared to the full-selection and true models. This observation applies to the regression slope estimates, residual variance, and intercept. Furthermore, the spike-and-slab exhibited less bias the MAR model, however this difference was small for all parameters.

Examining the slope estimates X_1 through X_4 (predictors of missingness in data-generating model) in the first row panel of Figure 16, the spike-and-slab method exhibited an average bias of -18%, the MAR model showed an average bias of -22%, the full-selection model displayed an average bias of -82%, and the true model had an average bias of -71%. These results highlight the substantial advantage in bias achieved by the MAR model and spike-and-slab method over the true and full-selection models. For the slope estimates of $\boldsymbol{X_5}$ and $\boldsymbol{X_6}$ (predictors that were not included in the data-generating model), all methods showed less bias when compared to the X_1 through X_4 slope estimates. On average, the spike-and-slab method had a bias of -5%, the MAR model exhibited a bias of -11%, the full-selection model showed a bias of -42%, and the true model had a bias of -28%.

Regarding the estimates of residual variance, the spike-and-slab method demonstrated virtually no bias, the MAR model showed a bias of 9%, while the full-selection and true models exhibited biases of nearly 85% and 72%, respectively. Thus far, the MAR and spike-and-slab models have shown an advantage in terms of bias. However, the *R*-squared parameter displayed a slightly different pattern. In this case, the true and full-selection models had the least bias compared to the MAR and spike-and-slab models. The spike-and-slab and MAR models yielded highly biased estimates of 71% and 62%, respectively, while the true and full-selection models had biases of 11% and 2%, respectively. Similar to the complexity 4 condition, the inflated *R*-squared estimate for the spike-and-slab and MAR models can be attributed to an excessively small residual variance.

Moving on to Figure 17, the standardized bias for the intercept estimates was -0.18 and -0.22 for the spike-and-slab and MAR models, respectively, while the true and full-selection models had biases of -0.73 and -0.82, respectively. Generally, both Figure 16 and 17 exhibited the same pattern, with the spike-and-slab method demonstrating less bias than the other models for almost all parameters, except for the R-squared estimate.

In the second row panel of Figure 16, as the sample size increased to 400, the bias percentages in the slope coefficients X_1 through X_4 decreased for all methods except for the MAR model. In this scenario, the spike-and-slab model exhibited the greatest advantage in terms of bias compared to the MAR model. While the bias in the MAR model remained unchanged, the spike-and-slab, true model, and full-selection model demonstrated a decrease in bias as the sample size increased. Specifically, in the *N* = 400 condition, the spike-and-slab, true, and full-selection models had biases of -10%, -29%, and -52%, respectively, for the X_1 through X_4 slope estimates. This is in contrast to biases of -18%, -71%, and -82% observed in the *N* = 100 condition. On the other hand, the MAR model did not display a significant change in bias, with biases of 22% in the *N* = 100 condition and 20% in the *N* = 400 condition for the X_1 through X_4 slope estimates.

When considering the slope estimates for $\boldsymbol{X_5}$ and $\boldsymbol{X_6}$, increasing the sample size only resulted in reduced bias in the true and full-selection models. The biases decreased from -28% and -42% in the *N* = 100 condition to -12% and -22% in the *N* = 400 condition. As for the estimates of residual variance, only the true and full-selection models exhibited a change in bias in the *N* = 400 condition. Both models showed a substantial reduction in bias, decreasing from 71% and 85% in the *N* = 100 condition to 11% and 32% in the *N* = 400 condition, respectively. The third row panel of the trellis plot in Figure 16 presented a stronger predictor intercorrelation ($\rho = .40$) , along with a sample size of 100, allowing for a comparison of the simple effect of predictor correlation with the first-row panel ($\rho = .10$). As depicted in Figure 16, the average bias across the slope estimates X_1 through X_4 substantially increased for the full-selection and true models. The full-selection model exhibited a bias of -86% (-71% in the first-row panel), while the true model demonstrated an average bias of -95% (-82% in the first-row panel). The mean bias across the X_1 through X_4 slope coefficients for the MAR and spike-and-slab did not change dramatically between the .10 and .40 predictor correlation conditions. Moving on to the slope estimates $\boldsymbol{X_5}$ and $\boldsymbol{X_6}$, residual variance, and *R*-squared, the presence of highly correlated predictors did not influence changes bias for those parameters.

The fourth row panel in Figure 16 and Figure 17, displayed a predictor correlation of .40 and a higher sample size of 400, enabling us to investigate whether the effect of the predictor correlations varied as a function of sample size. If there was an interaction between predictor correlation and sample size, a difference between the first and third row panel would not be the same as the difference between the second and fourth row panel. Both Figure 16 and 17 demonstrated that the difference between intercorrelation conditions remained uniform across both sample size conditions. These results indicate that there is no interaction between predictor correlations and sample size conditions.

Finally, the impact of missingness model complexity between the complexity 4 and complexity 3 conditions can be compared. One notable difference between the two is that in the complexity 4 condition, all slope coefficients within a row panel display similar bias values. However, in the complexity 3 condition, there is a distinct pattern where slope coefficients X_1 through X_4 generally exhibit higher bias compared to slope coefficients $\boldsymbol{X_5}$ and $\boldsymbol{X_6}$. This distinction between slope coefficients in complexity 3 arises because $\boldsymbol{X_5}$ and $\boldsymbol{X_6}$ are not predictors of missingness in the data generating model.

MSE Ratio

Figure 18 illustrate the MSE ratios for the complexity 3 conditions. Under this complexity condition, the true data generating missingness model was comprised of the outcome variable, slope coefficients X_1 through X_4, and one auxiliary variable as predictors. This figure displays the MSE ratio.

The top row panel in Figure 18, illustrates the outcomes achieved with a sample size of 100 and a weaker intercorrelation of predictors ($\rho = .10$). In this scenario, both the spike-and-slab prior and the MAR model demonstrated superior accuracy in estimating regression slope, residual variance, and intercept when compared to the true and full-selection model. The disparities in accuracy between the spike-and-slab and the MAR model were negligible for the slope estimates. However, when it came to estimating residual variance, the MAR model outperformed the spike-and-slab model, whereas the spike-and-slab model excelled in estimating the intercept. When examining the slope estimates for predictors X_1 through X_6, the MSE of the MAR model and spike-and-slab model were 1.51 and 1.56 times larger than the MSE of the complete-data analysis for all six predictors, respectively. In contrast, the MSE ratio for the true and full-selection model averaged 2.89 and 3.38, respectively.

Moving on to the residual variance, the spike-and-slab methods produced estimates that were, on average, 2.14 times larger than the MSE from the complete-data analysis, while the MAR model was more accurate with an MSE that was 1.57 times larger. Concerning the intercept, the spike-and-slab model had an MSE ratio of 1.72, while the MAR model had an MSE ratio of 2.27. In comparison, the true and full-selection model exhibited significantly higher MSE ratios for the residual variance (33.96 and 44.34, respectively) and the intercept (20.11 and 24.06, respectively).

In the second row panel of Figure 18, a larger sample size of 400 was used, enabling a comparison of the impact of sample size with the first row panel. The increased sample size affected the accuracy of the intercept parameter and the residual variance for all methods, and only influenced the X_1 through X_4 slope estimates of the true model and the X_5 and X_6 slope estimates of both the true and full selection models. Examining the slope estimates for predictors X_1 through X_4, the MSE of the true model was 2.14 times larger than the MSE of the complete-data analysis in the N = 400 condition, whereas the MSE ratio was 2.89 in the N = 100 condition. This shows an improvement in accuracy of the X_1 through X_4 slope estimates when sample size increase for the true model. Moving to the slope estimates for predictors X_5 and X_6, the true model MSE was 1.87 times larger than the MSE from the complete-data analysis and the full-selection model had a MSE ratio of 1.91, for both models bias decreased with the increase in sample size. The accuracy of the residual variance estimates improved as the sample size increased in both the true and full-selection models, however for the MAR model the residual variance did not result in a proportional improvement in accuracy compared to the true values.

In Figure 18, the third row panel presents the results for a stronger predictor intercorrelation ($\rho = .40$) and a sample size of 100. This allows for a comparison of the impact of predictor correlation with the first-row panel ($\rho = .10$). The stronger predictor intercorrelation had a moderate effect on the slope coefficients X_1 through X_4 of the true and full-selection model. When the predictor correlation condition was .10, the slope estimates X_1 through X_4 had an average MSE ratio of 2.89 and 3.38, respectively, however, with a stronger predictor intercorrelation of .40, the accuracy improved, resulting in a MSE ratio of 2.53 and 2.60. Regarding the slope-coefficients X_5 and X_6, residual variance, and intercept, increasing the predictor correlation did not lead to a proportional improvement in accuracy compared to the true values.

In the fourth row panel of the trellis plot, a predictor correlation of .40 and a higher sample size of 400 are depicted. This allows for an investigation of whether the effect of predictor correlations varies depending on the sample size. If there is an interaction between predictor correlation and sample size, the difference observed between the first and third row panels would not be the same as the difference between the second and fourth row panels. Alternatively, if there is no interaction, the difference between the first and second row panels would be approximately equivalent to the difference between the third and fourth row panels. However, the findings in Figure 18 indicate that there is no interaction between predictor correlations and sample size conditions.

Lastly, the impact in accuracy of true missingness model complexity between the complexity 4 and complexity 3 conditions can be compared. One notable difference between the two is that in the complexity 4 condition, all slope coefficients within a row panel display similar values for the MSE ratio. However, in the complexity 3 condition, there is a distinct pattern where slope coefficients X_1 through X_4 generally exhibit less accuracy than the slope coefficients X_5 and X_6. Again, this distinction between slope coefficients in complexity 3 arises because X_5 and X_6 are not predictors of missingness in the data-generating model.

Summary of Results for Complexity 3

The findings from Complexity 3, as shown in Figures 15 to 17, shed light on a scenario where two variables from the substantive model are not predictor of missingness in the true data-generating model. The results, measured by percent and standardized bias, demonstrate that the spike-and-slab prior and MAR model consistently outperforms the full-selection and true models across regression coefficients, residual variance, and intercept estimates.

When comparing the spike-and-slab model with the MAR model, the spike-and-slab prior is less biased than the MAR model, especially in the higher sample size condition. To evaluate the variation in the estimates, the MSE ratio is also examined. When assessing precision using the MSE ratio, the results indicate that both the MAR and spike-and-slab models perform similarly in terms of slope estimates across all conditions. This comparison highlights distinct strengths for each model. The spike-and-slab model's low bias suggests that its posterior distribution, which represents the uncertainty of the model's parameters, is centered around the true population parameter. On average, across many samples, the spike-and-slab model's estimates converge to the true values in the population. However, the approximately equal MSE ratio indicates that, within any given sample, both the spike-and-slab and MAR estimates deviate from the true values by the same amount. This implies that although the MAR model may have some bias, it has lower variance compared to the spike-and-slab model. As a result, in any individual sample, the performance of the MAR model is comparable to that of the spike-and-slab model. Overall, while the spike-and-slab model excels in reducing bias, the MAR model demonstrates lower variance. These factors contribute to their respective strengths, and understanding the trade-off between bias and variance is crucial in evaluating the performance of these models.

Complexity 2

Figures 19 to 21 show the percent bias, standardized bias, and MSE ratio, respectively. Complexity 2 represents the condition where the true data-generating model was comprised of the outcome variable, variables X_1 and X_2, and one auxiliary variable as predictors. The complexity 2 condition does not include variables X_3 through X_6 as predictors of missingness in the data-generating model. As a reminder, the spike-and-slab model and the full-selection model fitted the most complex missingness model. In contrast, the MAR model fitted variables X_1 through X_6, and one auxiliary variable as predictors. Additionally, the true data generating model fitted outcome variable Y, slope coefficients X_1 through X_2, and auxiliary variable A_1 as predictors.

Percent and Standardize Bias

Similar to the complexity 3 condition, Figures 19 and 20 showed notable differences in bias between the slope coefficients that serve as predictors and those that do not contribute to the data-generating model. As a result, I will provide separate descriptions of the average bias for slopes that are predictors (X_1 and X_2) and those that are not predictors (X_3 through X_6) of missingness in the data-generating model. In Figures 19 and 20, specifically in the first row panel, where the sample size is 100 and the predictors have an intercorrelation of .10, the estimates obtained from the spike-and-slab prior demonstrated substantially lower bias compared to the full-selection and true models. This observation applies to the regression slope estimates, residual variance, and intercept. Furthermore, the bias differences between the spike-and-slab prior and the MAR model were minimal for all parameters.

Examining the slope estimates X_1 and X_2 (predictors that were included in the data-generating model) in the first row panel of Figure 19, the spike-and-slab method exhibited an average bias of -26%, the MAR model showed an average bias of -32%, the full-selection model displayed an average bias of -108%, and the true model had an average bias of -69%. These results highlight the substantial advantage in bias achieved by the MAR model and spike-and-slab method over the true and full-selection models. For the slope estimates of X_3 through

X_6 (predictors not included in the data-generating model), all methods showed less bias when compared to the X_1 and X_2 slope estimates. On average, the spike-and-slab method had a bias of -6%, the MAR model exhibited a bias of -12%, the full-selection model showed a bias of -35%, and the true model had a bias of -28%. Regarding the estimates of residual variance, the spike-and-slab method demonstrated virtually no bias, the MAR model showed a bias of -7%, while the full-selection and true models exhibited biases of nearly 85% and 53%, respectively.

Thus far, the MAR and spike-and-slab models have shown an advantage in terms of bias. However, the *R*-squared parameter displayed a slightly different pattern. In this case, the true and full-selection models had the least bias compared to the MAR and spike-and-slab models. The spike-and-slab and MAR models yielded biased estimates of 73% and 62%, respectively, while the true and full-selection models had biases of 30% and 4%, respectively. Similar to the complexity 3 and 4 condition, the inflated *R*-squared estimate for the spike-and-slab and MAR models can be attributed to the narrow distribution of the residual variance. Moving on to Figure 20, the standardized bias for the intercept estimates was -0.19 and -0.24 for the spike-and-slab and MAR models, respectively, while the true and full-selection models had biases of -0.56 and -0.80, respectively. Generally, both Figure 19 and 20 exhibited the same pattern, with the spike-and-slab method demonstrating less bias than the other models for almost all parameters, except for the *R*-squared estimate.

In the second row panel of Figure 19, as the sample size increased to 400, the bias percentages in the slope coefficients X_1 and X_2 decreased for all methods except for the MAR model. In this scenario, the true model has the greatest advantage exhibited in terms of bias compared to the full-selection and MAR model, and it bias estimates are comparable to the spike-and-slab method. Specifically, in the *N* = 400 condition, the spike-and-slab, true, and full-selection models had biases of -18%, -11%, and -67%, respectively, for the X_1 through X_2 slope estimates. This is in contrast to biases of -26%, -69%, and -108% observed in the *N* = 100 condition. When considering the slope estimates for X_3 through X_6, increasing the sample size resulted in a substantial reduction bias for the true and full-selection models. The biases decreased from -28% and -35% in the *N* = 100 condition to -1% and -21% in the *N* = 400 condition, for the true and full-selection model, respectively. As for the estimates of residual variance, only the true and full-selection models exhibited a change in bias in the *N* = 400 condition. Both models showed a reduction in bias, decreasing from 53% and 85%, in the *N* = 100 condition to 1% and 29% in the *N* = 400 condition, for the true and full selection models respectively.

The third row panel of the trellis plot in Figure 19 presented a stronger predictor intercorrelation of .40, along with a sample size of 100, allowing for a comparison of the simple effect of predictor correlation ($\rho = .10$ vs. $\rho = .40$) with the first-row panel. The average bias across the slope estimates X_1 and X_2 substantially increased for the full-selection and true models. The full-selection model exhibited a bias of -135% (-108% in the first-row panel), while the true model demonstrated an average bias of -93% (-69% in the first-row panel). The mean bias across the X_1 and X_2 slope coefficients for the MAR and spike-and-slab did not change dramatically between the .10 and .40 predictor correlation conditions. Moving on to the slope estimates X_3 through X_6, residual variance, and *R*-squared, the presence of strongly intercorrelated predictors did not influence changes bias for those parameters.

The fourth row panel in Figure 19 and Figure 20, displayed a predictor intercorrelation of .40 and a higher sample size of 400, enabling us to investigate whether the effect of the predictor intercorrelations varied as a function of sample size. If there was an interaction between predictor correlation and sample size, a difference between the first and third row panel would not be the same as the difference between the second and fourth row panel. Both Figure 19 and 17 demonstrated that the difference between intercorrelation conditions remained uniform across both sample size conditions. These results indicate that there is no interaction between predictor correlations and sample size conditions.

Finally, the influence of different complexity levels on bias can be examined by comparing three conditions: complexity 4, complexity 3, and complexity 2. Complexity 2 and 3 exhibit a similar trend, where the substantive variables X_3 through X_6 generally display lower bias compared to the substantive variables (X_1 and X_2) that are predictors of missingness in the data-generating model. When comparing complexity 3 and 2, a noticeable difference in bias is observed for the true model. As the missingness model becomes less complex, the true model exhibits lower bias in its estimates. For instance, in the second row panel of Figure 16 (complexity 3), the true model had a -29% bias for the variables were predictors of missingness in the data-generating model (X_1 through X_4 in complexity 3), while in the second row panel of Figure 19 (complexity 2), there was an -11% bias for the variables included as predictors of missingness in the data-generating model (X_1 and X_2 in complexity 2).

MSE Ratio

Figure 21 illustrate the results of condition where the true missingness model was comprised of the outcome variable, slope coefficients X_1 and X_2 , and one auxiliary variable as predictors. This figure displays the MSE ratio. The top row panel in Figure 21, illustrates the outcomes achieved with a sample size of 100 and a .10 intercorrelation of predictors. In this scenario, both the spike-and-slab prior and the MAR model demonstrated superior accuracy in estimating regression slope, residual variance, and intercept when compared to the true and full-selection model. The disparities in accuracy between the spike-and-slab and MAR models were negligible for the slope estimates. However, when it came to estimating residual variance, the MAR model outperformed the spike-and-slab model, whereas the spike-and-slab model excelled in estimating the intercept.

When examining the slope estimates for predictors X_1 and X_2, the MSE of the MAR model and spike-and-slab model were 1.60 and 1.74 times larger than the MSE of the complete-data analysis. In contrast, the MSE ratio for the true and full-selection model averaged 2.65 and 4.25, respectively. Same trend occurred in the slope estimates X_3 through X_6, the spike-and-slab and MAR models were more accurate than the true and full-selection model. In addition, the slope estimates X_1 and X_2 were considerably more biased than the slope estimates X_3 through X_6, for only the true and full-selection models. Moving on to the residual variance, the spike-and-slab methods produced estimates that were, on average, 2.16 times larger than the MSE from the complete-data analysis, while the MAR model was more accurate with an MSE that was 1.55 times larger than the MSE from the complete-data analysis. Concerning the intercept, the spike-and-slab model had an MSE ratio of 1.71, while the MAR model had an MSE ratio of 2.35. In comparison, the true and full-selection model exhibited significantly higher MSE ratios for the residual variance (22.86 and 44.78, respectively) and the intercept (23.66 and 14.07, respectively).

In the second row panel of Figure 21, a larger sample size of 400 was used, enabling a comparison of the impact of sample size with the first row panel. The increased sample size affected the accuracy of the intercept parameter and the residual variance for all methods, and only influenced the X_1 and X_2 slope estimates of the true and MAR model and the X_3 through X_6 slope estimates of both the true and full selection models. Overall, under these conditions the true model had MSE estimates that were comparable to the spike-and-slab method, with the exception of the intercept.

When examining the slope estimates for predictors X_1 and X_2, it was found that in the *N* = 400 condition, the MSE of the true model was 1.67 times larger than the MSE of the complete-data analysis. In contrast, the MSE ratio was 2.65 in the *N* = 100 condition. On the other hand, the MAR model had a MSE ratio of 1.90 in the *N* = 400 condition, while it was 1.60 in the *N* = 100 condition. These findings indicate that as the sample size increased, the true model improved the accuracy of its slope estimates compared to the true values, whereas the MAR model did the opposite. This outcome is expected since the mechanism of the true missingness model is MNAR. Moving on to the slope estimates for predictors X_3 through X_6, it was observed that the true model had a MSE that was 1.46 times larger than the MSE from the complete-data analysis. Meanwhile, the full-selection model exhibited a MSE ratio of 1.87. As the sample size increased, bias decreased for both models. Additionally, the accuracy of the residual variance estimates improved with increasing sample size in both the true and full-selection models. However, for the MAR model, the improvement in accuracy of the residual variance estimates did not match the proportionate improvement observed in the true values.

The third row panel in Figure 21 displays the findings obtained from a larger predictor correlation of .40 and a sample size of 100. This allows for a comparison of the impact of predictor correlation ($\rho = .10$ vs $\rho = .40$) with the panel in the first row. The heightened predictor correlation moderately influenced the slope coefficients X_1 and X_2 in both the true and full-selection models. Under the predictor correlation condition of .10, the MSE ratios for the slope estimates X_1 and X_2 were 2.65 and 4.12 for the true and full-selection models, respectively. However, with an increased predictor correlation of .40, the accuracy improved, resulting in

MSE ratios of 2.37 and 3.20 for the true and full-selection models, respectively. Regarding the slope coefficients X_3 through X_6, residual variance, and intercept, enhancing the predictor correlation did not lead to a proportionate improvement in accuracy when compared to the true values.

The fourth row panel in the trellis plot illustrates a scenario where a predictor intercorrelation of .40 and a larger sample size of 400 are utilized. This setup allows for an examination of whether the impact of predictor correlations varies based on the sample size. If there is an interaction between predictor correlation and sample size, the disparity observed between the first and third row panels would differ from the difference between the second and fourth row panels. Conversely, if there is no interaction, the distinction between the first and second row panels would be roughly equivalent to the difference between the third and fourth row panels. However, the findings presented in Figure 21 indicate that there is no interaction between predictor correlations and sample size conditions.

Lastly, the impact of varying complexity levels in the missingness model on accuracy can be assessed by comparing three conditions: complexity 4, complexity 3, and complexity 2. Notably, there is a distinction between complexity 4 and complexity 2 regarding the behavior of slope coefficients within a row panel. In complexity 4, these coefficients exhibit similar MSE ratios, whereas in complexity 2, there is a clear pattern where slope coefficients X_1 and X_2 generally have lower accuracy compared to slope coefficients X_3 through X_6. This accuracy difference arises because X_3 through $X_6$are not predictors of missingness in the data-generating model.

Summary of Results for Complexity 2

The findings from Complexity 2, as depicted in Figures 19 to 21, provide insights into a scenario where the true missingness model from the data-generating model did not incorporate four variables (X_3 through X_6) from the substantive model. Figures 19 and 20 visualize the percent and standardized bias, respectively, while Figure 21 presents the MSE ratio for Complexity 2. When comparing the spike-and-slab model with the MAR model, the spike-and-slab prior is less biased than the MAR model, especially in the higher sample size condition. To evaluate the variation in the estimates, the MSE ratio is also examined.

When assessing precision using the MSE ratio, the results indicate that both the MAR and spike-and-slab models perform similarly in terms of slope estimates across all conditions. This comparison highlights distinct strengths for each model. The spike-and-slab model's low bias suggests that its posterior distribution, which represents the uncertainty of the model's parameters, is centered around the true population parameter. On average, across many samples, the spike-and-slab model's estimates converge to the true values in the population. However, the approximately equal MSE ratio indicates that, within any given sample, both the spike-and-slab and MAR estimates deviate from the true values by the same amount. This implies that although the MAR model may have some bias, it has lower variance compared to the spike-and-slab model. As a result, in any individual sample, the performance of the MAR model is comparable to that of the spike-and-slab model.

Complexity 1

Figures 22 to 24 provide visualizations of the percent bias, standardized bias, and MSE ratio, respectively. Conplexity1 represents the condition where the true missingness model includes only the outcome variable and one auxiliary variable as predictors. In the complexity 1 condition, all substantive variables (X_1 through X_6) are not missingness predictors in the data-generating model. As a reminder, the spike-and-slab model and the full-selection model fitted the most complex missingness model. In contrast, the MAR model fitted variables X_1 through X_6, and one auxiliary variable as predictors. Additionally, the true data generating model fitted outcome variable Y and auxiliary variable A_1 as predictors.

Percent and Standardize Bias

In Figures 22 and 23, specifically in the top row panel, where the sample size is 100 and the predictors have a correlation of 0.1, the spike-and-slab prior yielded estimates that were significantly less biased than the full-selection approach for all parameters except R-squared. Moreover, the bias differences between the spike-and-slab prior, the MAR model, and the true model were minimal for the intercept and slope estimates. Looking at the slope estimates in the top row panel of Figure 22, the spike-and-slab method had an average bias of -10%, the MAR model had an average bias of -12%, the full-selection model had an average bias of -36%, and the true model had an average bias of -13%. These results indicate that under a less complex missingness model, the MAR, spike-and-slab method, and true models had similar levels of bias for the slope estimates, while the full-selection model exhibited higher bias compared to the other three models.

Regarding the estimates of residual variance, the spike-and-slab and MAR models had biases of 3% and 5%, respectively. On the other hand, the full-selection and true models exhibited higher bias. The full-selection model had a bias of 76%, while the true model had a bias of 20%. As for the R-squared parameter, the full-selection and true models had the least bias compared to the MAR and spike-and-slab models. The spike-and-slab and MAR models produced biased estimates of 76% and 72%, respectively, while the true and full-selection models had biases of 60% and 18%, respectively. Similar to the complexity 3 and 4 condition, the inflated *R*-squared estimate for the spike-and-slab and MAR models can be attributed to a narrow distribution of the residual variance parameter.

Moving on to Figure 23, the standardized bias for the intercept estimates was -0.19 and -0.24 for the spike-and-slab and MAR models, respectively, while the true and full-selection models had biases of -0.56 and -0.80, respectively. Overall, both Figure 22 and 23 exhibited the same pattern, with the spike-and-slab method showing less bias than the other models for almost all parameters, except for the *R*-squared estimate.

In the second row panel of Figure 22, as the sample size increased to 400, the bias percentages in the slope coefficients decreased for all methods except for the MAR model. In this scenario, the true model had the least bias compared to all other models for all parameters except for the *R*-squared. However, both the spike-and-slab and true models exhibited biases within the 10% threshold for all parameters except *R*-squared. Specifically, in the N = 400 condition, the spike-and-slab, true, and full-selection models had biases of -5%, -1%, and -17%, respectively, for the slope estimates. This is in contrast to biases of -10%, -13%, and -36% observed in the N = 100 condition. Regarding the estimates of residual variance, only the true and full-selection models showed a change in bias in the N = 400 condition. Both models exhibited a reduction in bias, decreasing from 20% and 76% in the N = 100 condition to -1% and 13% in the N = 400 condition, respectively. Moving on to the *R*-squared, bias was reduced for all models as the sample size increased. In the N = 400 condition, the biases were 17%, 13%, 24%, and -9% for the spike-and-slab, MAR, true, and full-selection models, respectively.

In the third row panel of the trellis plot in Figure 22, a stronger predictor intercorrelation $(\rho = .40)$ was presented alongside a sample size of 100, allowing for a comparison of the effect of predictor correlation with the first-row panel ($\rho = .10$). The presence of strongly intercorrelated predictors did not result in changes in bias for any of the parameters. Moving on to the fourth row panel in Figure 22 and Figure 23, a predictor correlation of .40 and a higher sample size of 400 were displayed. This setup allowed us to investigate whether the effect of predictor intercorrelations varied depending on the sample size. If there was an interaction between predictor correlation and sample size, the difference between the first and third row panels would not be the same as the difference between the second and fourth row panels. However, both Figure 22 and 23 demonstrated that the difference between the intercorrelation conditions remained consistent across both sample size conditions. These results indicate that there is no interaction between predictor correlations and sample size conditions.

To evaluate the impact of varying complexity levels in the missingness model on bias, the different complexity conditions can be compared. Complexity 1 and 4 exhibit a similar pattern, with all slope coefficients in a row panel showing comparable bias values. In contrast, complexity 2 and 3 demonstrate differences in bias for the variables included in the missingness model. When comparing complexity 1 (the least complex model) to complexity 4 (the most complex model), the differences in bias between the BVS methods are more pronounced in the most complex conditions. In the least complex conditions, the bias values are relatively comparable across methods, except for the residual variance parameter, where the spike-and-slab and MAR model consistently outperforms the others.

MSE Ratio

Figure 24 illustrate the results of condition where the true missingness model was comprised of the outcome variable and one auxiliary variable as predictors. This figure displays the MSE ratio.

The top row panel in Figure 24, illustrates the outcomes achieved with a sample size of 100 and a .10 intercorrelation of predictors. In this scenario, both the spike-and-slab prior and the MAR model demonstrated superior accuracy in estimating regression slope, residual variance, and intercept when compared to the true and full-selection model. The disparities in accuracy between the spike-and-slab and MAR models were negligible for the slope estimates. However, when it came to estimating residual variance and the intercept, the MAR model outperformed the spike-and-slab model.

Examining the slope estimates, the MSE of the MAR model and spike-and-slab model were 1.45 and 1.55 times larger than the MSE of the complete-data analysis. In contrast, the MSE ratio for the true and full-selection model averaged 1.77 and 2.39, respectively. Moving on to the residual variance, the spike-and-slab methods produced estimates that were, on average, 2.48 times larger than the MSE from the complete-data analysis, while the MAR model was more accurate with an MSE that was 1.55 times larger. Concerning the intercept, the spike-and-slab model had an MSE ratio of 2.24, while the MAR model had an MSE ratio of 2.06. In comparison, the true and full-selection model exhibited significantly higher MSE ratios for the residual variance (7.54 and 37.53, respectively) and the intercept (5.03 and 20.24, respectively).

In the second row panel of Figure 24, a larger sample size of 400 was used, enabling a comparison of the impact of sample size with the first row panel. The true model’s intercept parameter had a better accuracy in the N = 400 sample size condition with a MSE ratio of 1.34 (5.03 in the N = 100 condition). However, the intercept results for the full-selection, MAR, and spike-and-slab models showed the opposite trend. Specifically, at N = 400, the MSE ratio of the MAR and full-selection model intercepts were 6.59 and 24.14 times larger than the complete-data MSE, respectably. In contrast, at N = 100, the MSE for the MAR and full-selection models were about 2.06 and 20.24 times larger. For the spike-and-slab approach, the MSE ratio was 2.25 for the N = 100 condition and 4.32 for the N = 400 condition.

Moving to the slope estimates in the second row panel, the increased sample size resulted in improved accuracy for all methods, except for the MAR model. Specifically, for the slope estimates in the N = 400 condition, the spike-and-slab, full-selection, and true models exhibited MSE ratios of 1.35, 1.71, and 1.46, respectively. Similarly, the accuracy of the residual variance estimation also improved as the sample size increased for all methods, except for the MAR model which did not show any significant changes. In the N = 400 condition, the MSE ratios for the spike-and-slab, full-selection, and true models were 2.15, 11.12, and 1.76, respectively, indicating greater accuracy compared to the N = 100 condition.

The third row panel of Figure 24 displays the findings obtained from a larger predictor correlation of .40 and a sample size of 100. This allows for a comparison of the impact of predictor intercorrelation ($\rho = .10$ vs $\rho = .40$) with the panel in the first row. The .40 predictor correlation moderately only influenced the slope coefficients in the full-selection model. Under the predictor correlation condition of .10, the average MSE ratio for the slope estimates was 2.39. However, with an increased predictor correlation of .40, the accuracy improved, resulting in a average MSE ratio of 2.07 for the full-selection model. Regarding the residual variance and intercept, increasing the predictor correlation did not lead to a proportionate improvement in accuracy when compared to the true values.

The fourth row panel in the trellis plot illustrates a scenario where a predictor correlation of .40 and a larger sample size of 400 are utilized. This setup allows for an examination of whether the impact of predictor correlations varies based on the sample size. If there is an interaction between predictor correlation and sample size, the disparity observed between the first and third row panels would differ from the difference between the second and fourth row panels. Conversely, if there is no interaction, the distinction between the first and second row panels would be roughly equivalent to the difference between the third and fourth row panels. However, the findings presented in Figure 24 indicate that there is no interaction between predictor correlations and sample size conditions.

Finally, it can be evaluated how the different levels of complexity in the missingness model affect the accuracy and bias of the estimates. It's important to note a similarity between complexity 4, which represents the most complex true missingness model, and complexity 1 in terms of the behavior of slope coefficients within a row panel. In both complexity 1 and 4, there is less variability observed among MSE ratios. Conversely, in complexity 2 and 3, a clear pattern emerges where slope coefficients that predict missingness in the data-generating model generally exhibit lower accuracy compared to the other slope coefficients. Additionally, it's worth mentioning that at lower levels of complexity, the true model performs better in terms of both bias and accuracy outcomes.

Summary of Results for Complexity 1

The findings from Complexity 1, as presented in Figures 22 to 24, provide insights into a scenario where the true missingness model did not incorporate any predictors from the substantive model. Figures 22 and 23 visualize the percentage and standardized bias, respectively, while Figure 24 displays the MSE ratio for Complexity 1. In the N = 100 case, the spike-and-slab, MAR, and true models exhibited comparable bias in the slope estimates, but the spike-and-slab and MAR models maintained an advantage over the true model in terms of bias in residual variance. In the N = 400 condition, the true model slightly outperformed the MAR and spike-and-slab models across all parameters. In terms of accuracy, in the N = 100 scenario, the spike-and-slab model and MAR model demonstrated the most precise slope estimates compared to the true and full-selection models. However, in the N = 400 condition, this advantage vanished, and accuracy was similar among the spike-and-slab, MAR, and the true model. Overall, the full-selection approach performed the poorest in both bias and accuracy outcomes. There was minimal influence of predictor correlation on the accuracy or bias of the estimates in complexity 1.

EMPIRICAL EXAMPLE

In this section, I illustrate the practical value and implementation of selection models with Bayesian shrinking priors within the context of a clinical-trial study. To exemplify this, I used the dataset from a study conducted by Ray et al. (2021). The study aimed to assess the effectiveness of combining varenicline, a partial agonist of the α4β2 nicotinic acetylcholine receptor, and naltrexone, an antagonist of opioid receptors, in enhancing smoking cessation and reducing alcohol consumption among heavy drinking smokers. The objective was to compare these results with the outcomes of using varenicline alone as the sole pharmacological treatment. In this context, it is plausible that an MNAR mechanism exists, as individuals with higher substance use rates might be more likely to drop out, leading to missing data.

Data Preparation

For this empirical illustration, I used data from all participants in the clinical trial (N=165) which is consistent with the sample size condition used in the simulation. I focused on the outcome "drinks per drinking day" at the eight-week follow-up ($DPDD_8$), as it had a missing data rate of 20%, which closely resembles the missing data rate in the simulation study. The treatment condition (TX) and “drinks per drinking day” at baseline ($DPDD_0$) served as the predictors in the substantive model, as represented in Equation 77.

$$DPDD_{8i} = \beta_0 + \beta_1 DPDD_{0i} + \beta_2 TX_i + \varepsilon_i \qquad (77)$$

The empirical example assumed a MNAR process, where an individual's unobserved "drinks per drinking day" value at week eight predicts missingness. To address this, a selection model for handling missing data was used, which combined the substantive model in Equation 77 with a missingness model.

The missingness model's predictors consisted of all the variables from the substantive model, "drinks per drinking day" at both week eight and baseline, and the treatment condition (TX). Additionally, Type B and C auxiliary variables were included in the model. As a reminder from the introduction section, Type B auxiliary variables are correlated with the analysis but not with the missingness indicator, while Type C auxiliary variables correlate with the missing data indicator but not with the substantive model variables. To determine the potential Type B and C auxiliary variables for inclusion in the missingness model, I tested various background variables from Table 1 of Ray et al. (2021) study that were considered potential correlates of treatment compliance. Each potential auxiliary variable needed to exhibit a correlation or Cohen's d effect size greater than .10 to be included as predictors in the subsequent missingness model.

Two Type B auxiliary variables, cotinine at baseline (Cot_0) and drinking days at baseline (DD_0), were selected based on their correlations with the outcome variable "drinks per drinking day" measured at week eight. Variables Cot_0 and DD_0 showed correlations with the outcome variable of .11 and .16, respectively. For the potential Type C auxiliary variables, effect sizes were computed. First, a missing data indicator was created for the outcome variable $DPDD_{8i}$, and independent t-tests were conducted with each of the background variables as the outcomes.

The effect sizes obtained from these tests helped determine which variables to include in the missingness model. After examining nine possible variables from Table 1, age, gender, and the Fagerström Test of Nicotine Dependence score ($FTND$) were selected based on their effect sizes with the missingness indicator.

Next the complete missingness model is presented. The outcome of this model is a value from a normally distributed latent variable for individual i represented as M_{8i}^* in the equation below:

$$\begin{aligned} M_{8i}^* = \gamma_0 + \gamma_1 DPDD_{8i} + \gamma_2 TX_i + \gamma_3 DPDD_{0i} + \gamma_4 Cot_{0i} + \gamma_5 DD_{0i} \\ + \gamma_6 Age_i + \gamma_7 Gender_i + \gamma_8 FTND_i + r_i \\ M_{8i}^* \sim N(E(M_{8i}^* | \cdot), \sigma_r^2)\ I(Q_i) \end{aligned} \tag{78}$$

The term $I(\cdot)$ is an indicator function, Q_i is either equal to $\{M_{8i}^* > \varphi\}$ or $\{M_{8i}^* \leq \varphi\}$ corresponding to $M_{8i} = 1$ or $M_{8i} = 0$. Where M_{8i} = 0 if the week eight “drinks per drinking day” score is observed and M_{8i} = 1 if it is missing, for individual i. The following regression models correspond to the conditional distributions of the Type B auxiliary variables.

$$\begin{aligned} Cot_{0i} = \ \gamma_{02} + \gamma_{12} DPDD_{8i} + \gamma_{22} DPDD_{0i} + \gamma_{32} TX_i + r_{2i} \\ DD_{0i} = \ \gamma_{01} + \gamma_{11} DPDD_{8i} + \gamma_{21} DPDD_{0i} + \gamma_{31} TX_i + \gamma_{41} Cot_{0i} + r_{1i} \end{aligned} \tag{79}$$

Finally, the sample characteristics at baseline for the selected variables by treatment conditions are presented in Table 2. All variables were standardized prior to analyses.

Data Analysis and Results

For the analysis, various models were applied, including a full-selection model, a model assuming a MAR process, and all three BVS adapted selection models as described in the simulation section. Additionally, a model assuming a focused MNAR process was also fitted:

$$M_{8i}^{*} = \gamma_0 + \gamma_1 DPDD_{8i} + r_i \tag{80}$$

where missingness depends on $DPDD$ scores at the eighth week assessment, some of which are missing. This model was fitted for comparison. The analyses were performed using Blimp for the MAR and focused selection models and R for the BVS selection models, with identical hyperpriors as in the simulation. Estimates of substantive model's intercept and treatment effect can be found in Table 3.

As mentioned in the previous section, all variables were standardized. Although this standardization is not natural for binary variables like treatment condition (TX), it was done to adhere to the norm in the literature. Consequently, the intercept represents the mean of the varenicline-only group on a standardized metric, and the regression coefficient for TX reflects the treatment effect on the same standardized metric. From the results presented in Table 3, it can be observed that all BVS, full-selection, and focused models produced similar outcomes for the intercept and treatment effect. On the other hand, the MAR model yielded substantially different results for these parameters.

To offer more insights into the BVS methods, we visually represent the posterior distribution of the regression coefficients for the missingness model. Figures 25 to 33 showcase the resulting posterior distributions of the regression slopes for the spike-and-slab, horseshoe prior, and Bayesian LASSO. These figures demonstrate the distributions of the independent variables from the missingness model, each displaying varying degrees of shrinkage. For illustration purposes, I will focus on two of these figures. First, the posterior distribution of the outcome variable $DPDD_8$ and then the posterior distribution for the baseline drinking days (DD_0) will be examined.

Figure 26 displays the posterior distributions of the regression coefficient for the outcome variable for different selection models. The distributions for the full selection model, horseshoe model, and Bayesian LASSO model appear left-skewed, with medians close to -6.38, -5.23, and -4.22, respectively. In contrast, the spike-and-slab model exhibit a more symmetric and narrower posterior distribution of the outcome variable. Its variance is approximately half that of the other three selection models, and it has a median of -2.49, indicating the most substantial shrinkage of the outcome variable regression coefficient out of the three BVS priors. Despite all BVS methods reducing the outcome coefficient toward zero, none of their distributions contain zero as a value, demonstrating that $DPDD_8$ is an important predictor in the missingness model.

Figure 30 displays the posterior distribution of the auxiliary variable baseline drinking days (DD_0). The posterior distribution of DD_0 for the full selection model, horseshoe model, and Bayesian LASSO model appear more symmetrical in comparison to Figure 26, with medians close to 1.46, 1.13, and 0.83, respectively. In contrast, the spike-and-slab model has a median value of 0.10, showing the most substantial shrinkage among all the BVS methods. This model's posterior distribution in Figure 30 shows a mixture of two distributions: one centered at zero (the "spike") and the other centered at a non-zero value (the "slab"). This unique characteristic is a result of the spike-and-slab prior effectively "turning off" the regression coefficient for DD_0 approximately half of the time, leading to the observed bimodal distribution.

In general, all BVS methods led to parameter estimates being shrunk compared to the full-selection model. Among the BVS methods, the spike-and-slab approach appeared to be the most conservative in terms of model complexity. Specifically, the spike-and-slab posterior distribution for the less informative predictors were the closest to zero and had the narrowest variance when compared to the horseshoe and Bayesian LASSO methods. In addition, the Bayesian LASSO model demonstrated more substantial shrinkage in its regression coefficients compared to the horseshoe model.

In summary, the empirical example aimed to determine if BVS could protect against overfitting in a clinical trial with a small sample size. The empirical analysis included one model that assumed an MAR process and five model that assume MNAR. Among the MNAR models, there were three selection models with different BVS adaptations, a full-selection model, and a focus selection model. Remarkably, all MNAR models produced similar estimates for the substantive model, while the MAR model produced different estimates for the substantive model estimates in comparison. It's important to note that there is no definitive way to choose the true model or determine which results are more accurate, as the differences in the results reflect the treatment effect under two different assumptions about missing data.

DISCUSSION

When missing data is MNAR, it's important for researchers to model the missingness and the substantive equation together to avoid biases in parameter estimates and standard errors (Little, 2008; Rubin, 1976). One commonly used method for this is selection models, which simultaneously model both the missingness and substantive models (Du et al., 2021; Heckman, 1976, 1979; Michiels et al., 1999). When constructing a missingness model, researchers need to consider whether to include predictors that overlap with the substantive analysis. A recommended approach is to remove redundant predictors shared between the two models using exclusion restrictions (Collins et al., 2001; Galimard et al., 2018; Sartori, 2003). This can be done by using a subset of predictors from the substantive model or by adding unrelated auxiliary variables to the missingness model. It's also important not to include too many variables in the missingness model to avoid convergence issues (Du et al., 2021; Ibrahim et al., 2005). However, specifying and estimating the missingness equation accurately is challenging, especially when researchers don't have a solid basis for constructing the missingness model.

The present study is the first to apply BVS procedures for the purposes of helping to establish identification for selection models. Three approaches were investigated: Bayesian LASSO, horseshoe prior, and spike-and-slab prior. The goal was to evaluate how well these methods can remove unnecessary components from the missingness model, thereby reducing biases in the parameters of the substantive model. Custom R functions were developed for estimation and analysis using the Bayesian LASSO, horseshoe, and spike-and-slab selection models since there is currently no existing software that handles missing data in selection models with a BVS adaptation.

The study examined bias and MSE for true data-generating models that varied the complexity of the missingness model. Complexity 4 represented the most complex missingness model scenario where every variable from the focal model also appeared in the missingness model, while Complexity 3, 2, and 1 represented scenarios with decreasing levels of complexity. Considering the focal model parameters, the comparison among BVS methods showed that the spike-and-slab prior method had less bias compared to the global-local shrinkage priors (horseshoe and Bayesian LASSO) for intercept, slope coefficients, and residual variance estimates across all complexity conditions except for Complexity 1, where the advantage was not substantial. However, bias alone does not capture the variation in estimates, so the MSE was also examined. The results consistently demonstrated that the spike-and-slab prior outperformed the global-shrinkage priors in terms of accuracy (lower MSE) for slope estimates, residual variance, and intercept across all complexity conditions. Particularly, the spike-and-slab method produced substantially more accurate estimates for residual variance and intercept compared to the global-local prior methods.

After selecting the spike-and-slab as the top performer among the three BVS methods, it was compared to the MAR model, full-selection model, and the true data-generating model. Across complexities 2 to 4, the spike-and-slab model consistently outperformed the competing model in terms of bias for regression coefficients, residual variance, and intercept estimates. The MSE results indicated that the MAR and spike-and-slab models performed similarly in terms of slope estimates across all conditions. This implies that, although the MAR model may be biased, it has lower variance (i.e., greater precision) compared to the spike-and-slab model. Consequently, in individual samples, the performance of the MAR model is comparable to that of the spike-and-slab model for the slope coefficients.

The comparable performance of the MAR model and the spike-and-slab model can be attributed to the bias-variance tradeoff. The MAR model, although biased, is efficient because it estimates fewer parameters compared to selection models with or without BVS. Additionally, the parameters estimated by the MAR model have strong support from the observed data. On the other hand, fitting selection models is challenging, particularly when dealing with datasets containing MNAR missing data. The observed data in such cases contains limited information about the missingness model parameters, leading to increased noise and imprecision when estimating the missingness model. Consequently, selection models usually require a large sample size to stabilize the model estimation, as the limited information becomes sufficient only with a substantial sample size.

The study results also emphasized the dependency of sample size for the performance of the MAR and selection models. For instance, in the second row panel of Figure 19, increasing the sample size did not improve bias for the MAR model, indicating that the estimator is not consistent and does not converge on the population parameter as the sample size increases. However, based on the results of this study, we cannot determine whether any of the BVS methods would gain an advantage in precision with a larger sample size. This highlights the need for further research and investigation into the behavior of BVS methods under different sample size conditions.

Additionally, it is important to highlight that the spike-and-slab model consistently outperformed the MAR model in terms of MSE ratio for the residual variance and intercept estimates. The current simulation involved a focal model with continuous predictors, where the intercept might not have been the primary focus. In a different scenario such as the clinical trial analysis in the empirical example, the intercept would represent a critical parameter (e.g., the placebo group average). In such cases, the accuracy advantage provided by the spike-and-slab model becomes practically significant and could have profound implications for result interpretation.

Limitations

This study has several limitations, with the first one being the use of a relatively small range of sample sizes. It is recommended to have a larger sample size to reduce noise and improve precision in selection models. However, in this study, the inclusion of larger sample size conditions was not feasible due to computational limitations, which resulted in excessively long running times. Nevertheless, the study's primary objective was to obtain a detailed snapshot of a small parameter space rather than to determine the sample size at which BVS methods achieve their asymptotic optimal performance. The simulated sample size conditions were realistic for many studies where one might apply this method, such as clinical trials. However, the limited range of sample sizes used in the study does not allow for a conclusive comparison of the performance of the MAR model and the spike-and-slab model under larger sample sizes. Future research should aim to replicate the study with larger sample sizes to assess whether the MAR model remains comparable to the spike-and-slab model or if the latter provides more precise slope coefficients under such conditions.

Another limitation is that the selection model only accounted for missingness in the outcome variable, and missing data was not generated for any of the covariates. While having missing data only on the outcome variable can be representative of certain studies (e.g., clinical trials with complete baseline measures and missing dependent variable), it does not reflect the majority of studies where missing data is present in both covariates and outcome measures. Future studies should include missing data in the predictor variables, as this would introduce additional biases and uncertainties and would likely require a larger sample size to accommodate the increased noise.

Moreover, the study has a limitation in terms of the generalizability of the results due to several fixed parameters. Both the number of predictors in the focal and missingness models and the effect sizes of these predictors remained constant throughout the simulation. Additionally, the missing data rate was also fixed at a specific value. Furthermore, the correlation among predictors was limited to only two values, and a single auxiliary variable was used in the missingness model. Lastly, the simulation solely focused on continuous variables, excluding other variable types. While these fixed values facilitated controlled comparisons and evaluations, they may not fully capture the variability and complexity encountered in real-world scenarios. Due to time constraints and computational intensity, the study prioritized exploring specific simulation conditions over examining a broader range of factors such as different numbers of predictors, variable types, effect sizes, and missing data rates. As a result, the findings applicability to real-world situations may be limited.

Future Studies

The findings from this study lay the groundwork for several potential directions of future research. A key finding from the simulation was that the spike-and-slab model exhibited lower bias and greater precision when estimating the intercept compared to all other methods, including the MAR model. However, it is worth noting that the estimated intercept for the focal model was not the primary focus of interest in this study. An important direction for future investigation would be to simulate a focal model where the intercept holds significance, similar to the empirical example presented. In such an analysis, the intercept would represent the control group average, and the slope would indicate the mean difference. If the intercept estimates show substantial improvement with BVS, the implication for the selection model with a BVS adaptation could be quite different. The same holds true for other models where mean parameters are of primary interest, such as growth models. Therefore, conducting a future study that emphasizes the intercept as a crucial parameter of the population model would be highly valuable in further understanding the performance of BVS in such scenarios.

Another promising area for future research is to assess how well the BVS methods perform under various hyperparameters. For the current simulation, I used recommended hyperparameters from existing literature as an initial step to explore BVS adaptation to selection models. Varying the prior hyperparameters was not given as much priority as other aspects of the design. However, it is vital for future investigations to explore diverse hyperparameter settings in order to evaluate the sensitivity and robustness of the BVS method. In both the spike-and-slab model used in the simulation and the empirical example study, the hyperparameter w, which represents the probability of predictor inclusion, was set to $1/p$, where p is the total number of predictors. Nevertheless, researchers have the flexibility to select any value for w based on their prior beliefs or with the goal of improving convergence. To conduct a comprehensive investigation, future studies should simulate a range of liberal and conservative probabilities of predictor inclusion, as optimizing the hyperparameters holds the potential to enhance the performance and convergence of all three BVS methods.

Another promising area for future research involves developing criteria to assess convergence in selection models with BVS adaptations. Although PSR is a widely used technique for evaluating the convergence of MCMC simulations, it may not be perfectly suitable for Bayesian variable selection methods. When applying these methods, some coefficients may be repeatedly pushed towards zero, leading to higher PSR values than the desired threshold, even if the chains have already converged to the same stable distribution. This happens because the within-chain variation of the parameter of interest becomes very small due to the coefficient taking on near-constant values within a very small range. Thus, even small between/mean difference variation among chains could look large relative to attenuated within-chain variation that occurs with BVS. Future studies can focus on identifying and assessing alternative convergence diagnostics specifically tailored to these types of models. This is a practical issue for users of these approaches as well as for methodologists studying BVS methods.

The focus of this study's simulation was to achieve exclusion criteria by eliminating overlapping variables from both the missingness model and the focal model. However, in certain situations, such as the empirical analysis presented in the previous section, the focal model might consist of only a few variables. Consequently, constructing a missingness model would require selecting from a substantial pool of candidate auxiliary variables that are not part of the focal analysis. This concept was demonstrated in a smaller scale through the empirical analysis example.

Further investigations should explore scenarios where researchers have access to a large pool of candidate auxiliary variables that could be used as predictors in the missingness model. The primary focus of this future study should be to assess whether BVS can satisfy the necessary exclusion restriction criteria by excluding or shrinking predictors from an expanded list of candidate variables, which includes type A, B, and C auxiliary variables (Collins et al., 2001). Examining BVS under high-dimensional conditions may reveal additional distinctions between BVS methods. For instance, prior studies have indicated that sampling from posteriors in high-dimensional regression cases requires computationally intensive procedures for the discrete variation of the spike-and-slab prior as opposed to global-local shrinkage priors (Bhadra et al., 2017; Mitchell & Beauchamp, 1988). Additionally, another study has demonstrated that the horseshoe prior outperforms Bayesian LASSO in high-dimensional settings (Yamaguchi & Zhang, 2023). Consequently, it is plausible that the outcomes under such conditions could differ significantly from those obtained in the current study.

Another important avenue for future research involves analyzing multi-level data structures by employing BVS-adapted selection models to handle missing data. Many real-world datasets exhibit hierarchical structures (e.g., a clinical trial where repeated measurements are obtained from each participant). In such contexts, MNAR models are commonly employed, especially in longitudinal studies (Enders, 2011; B. Muthén et al., 2011). The BVS methods we examined here all extend to the multilevel context without substantial modification. Thus, extending BVS to linear mixed models could be a useful future contribution.

Final Summary

The objective of the study was to examine the effectiveness of BVS procedures in establishing identification for selection models when dealing with MNAR data. Three BVS-adapted selection models, namely Bayesian LASSO, horseshoe prior, and spike-and-slab prior, were compared. The study also compared the spike-and-slab model to existing methods like an MAR model and full-selection model. Results indicated that the spike-and-slab prior consistently outperformed other BVS methods in terms of accuracy and bias for slope estimates, residual variance, and intercept. When comparing the full-selection, MAR, and spike-and-slab models, the MSE results showed that the spike-and-slab model performed better than the full-selection model across all parameters. The MAR and spike-and-slab models had similar performance for slope estimates, but the spike-and-slab consistently outperformed the MAR model in estimating residual variance and intercept.

The comparable performance of the MAR and spike-and-slab models can be attributed to the bias-variance tradeoff. The MAR model is biased but efficient because it estimates fewer parameters compared to selection models and those parameters estimates have a strong support from the observed data. On the other hand, the spike-and-slab model outperformed the full-selection model, even when the full-selection model matched the true data-generating model. The adaptation of BVS to selection models showed promising results, particularly the spike-and-slab method, which demonstrated unbiased estimates under most conditions. As it was a first foray into this topic, this study was necessarily limited in scope. Nevertheless, the results provide a foundation for numerous future studies on this topic.

Tables and Figures

Table 1

Convergence Rates

	N = 100		*N* = 200		*N* = 400	
	C.R	B.I.	C.R	B.I.	C.R	B.I.
			Complexity 4			
BLASSO	42%	120,000	44%	160,000	54%	220,000
Horseshoe prior	41%	120,000	42%	160,000	43%	220,000
MAR model	97%	6,000	100%	5,000	100%	4,000
Full selection	20%	170,000	27%	220,000	37%	280,000
Spike-and-slap	59%	80,000	54%	140,000	35%	220,000
True model	20%	170,000	28%	220,000	37%	280,000
			Complexity 3			
BLASSO	41%	120,000	46%	160,000	57%	220,000
Horseshoe prior	40%	120,000	40%	160,000	48%	220,000
MAR model	97%	6,000	100%	5,000	100%	4,000
Full selection	20%	170,000	27%	220,000	34%	280,000
Spike-and-slap	60%	80,000	54%	140,000	46%	220,000
True model	22%	170,000	28%	220,000	36%	280,000
			Complexity 2			
BLASSO	42%	120,000	46%	160,000	57%	220,000
Horseshoe prior	42%	120,000	44%	160,000	48%	220,000
MAR model	98%	6,000	100%	5,000	100%	4,000
Full selection	20%	170,000	27%	220,000	34%	280,000
Spike-and-slap	60%	80,000	56%	140,000	46%	220,000
True model	28%	170,000	29%	220,000	36%	280,000
			Complexity 1			
BLASSO	42%	120,000	47%	160,000	54%	220,000
Horseshoe prior	45%	120,000	48%	160,000	50%	220,000
MAR model	97%	6,000	100%	5,000	100%	4,000
Full selection	17%	170,000	22%	220,000	23%	280,000
Spike-and-slap	66%	80,000	68%	140,000	71%	220,000
True model	31%	170,000	31%	220,000	36%	280,000

Note. C.R. = convergence rate, B.I. = burn-in iterations.

Table 2

Sample Characteristics at Baseline by Treatment Conditions

Variable	*Varenicline + Placebo (n=82)*		*Varenicline + Naltrexone (n=83)*	
Demographic Characteristics				
	Mean	SD	Mean	SD
Age	42.07	11.75	41.24	12.42
	N	%	N	%
Gender				
Male	50	60.98	52	62.65
Female	32	39.02	31	37.35
Covariates Characteristics				
	Mean	SD	Mean	SD
Cotinine (ng/ml)	5.37	1.49	5.50	1.14
FTND Score	4.61	2.35	4.64	2.26
Drinking Days	19.63	7.92	20.71	8.08
Drinks per Drinking Day	6.44	3.76	6.50	4.65

Note. FTND = Fagerström Test of Nicotine Dependence

Table 3

Substantive Model Estimates

	Intercept		TX Effect	
	Median (SE)	95% CI	Median (SE)	95% CI
BLASSO	-0.40 (0.10)	-0.59, -0.20	-0.26 (0.10)	-0.46, -0.07
Horseshoe prior	-0.40 (0.09)	-0.59, -0.20	-0.26 (0.10)	-0.46, -0.07
MAR model	-0.03 (0.09)	-0.20, 0.14	-0.13 (0.09)	-0.30, 0.04
Full selection	-0.40 (0.10)	-0.60, -0.20	-0.26 (0.10)	-0.46, -0.06
Spike-and-slap	-0.39 (0.10)	-0.59, -0.19	-0.26 (0.10)	-0.46, -0.07
Focused selection	-0.41 (0.10)	-0.61, -0.21	-0.27(0.10)	-0.47, -0.08

Note.

Figure 1

Complexity 4 Percent Bias for BVS Methods

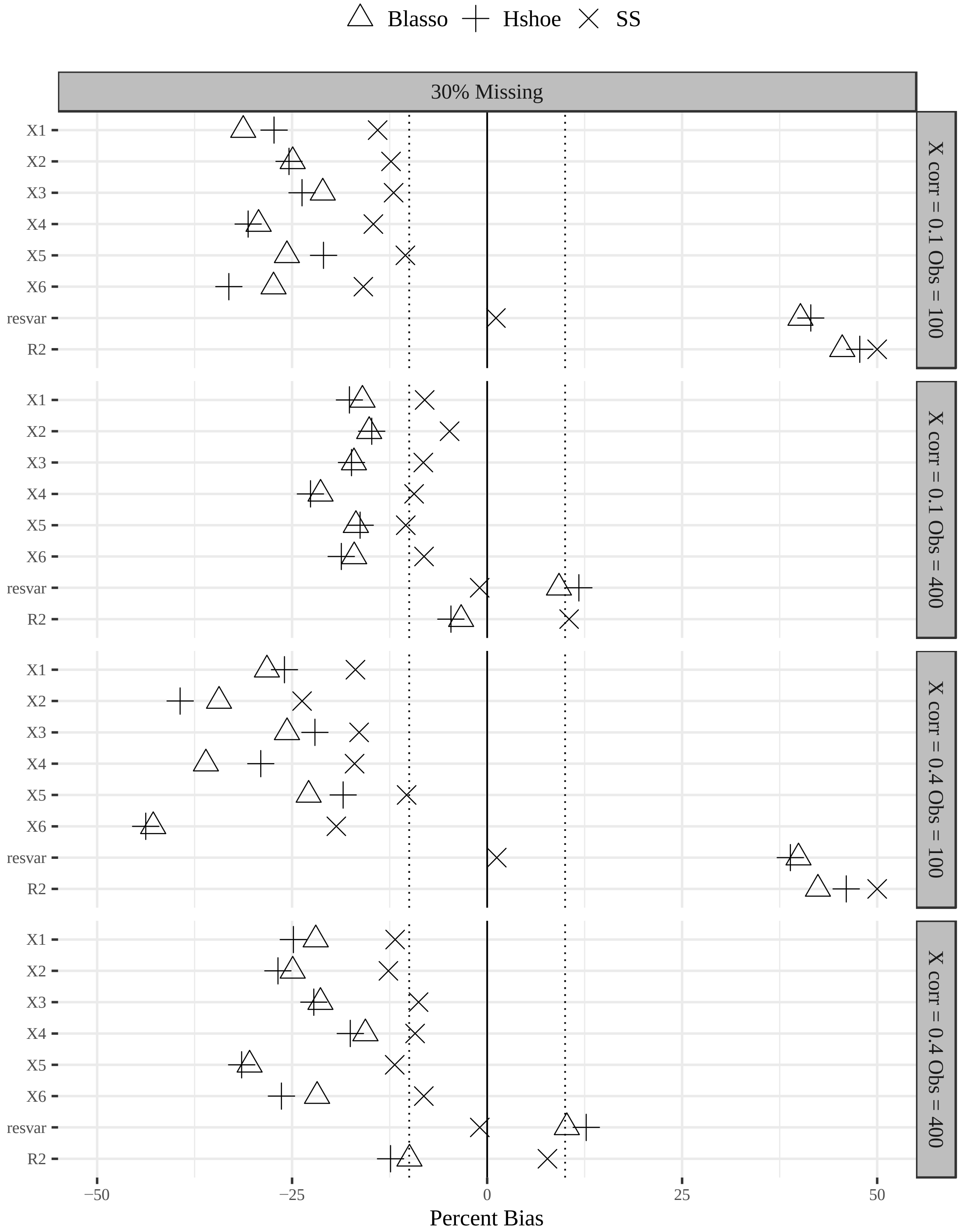


Figure 2

Complexity 4 Standardized Bias for BVS Methods

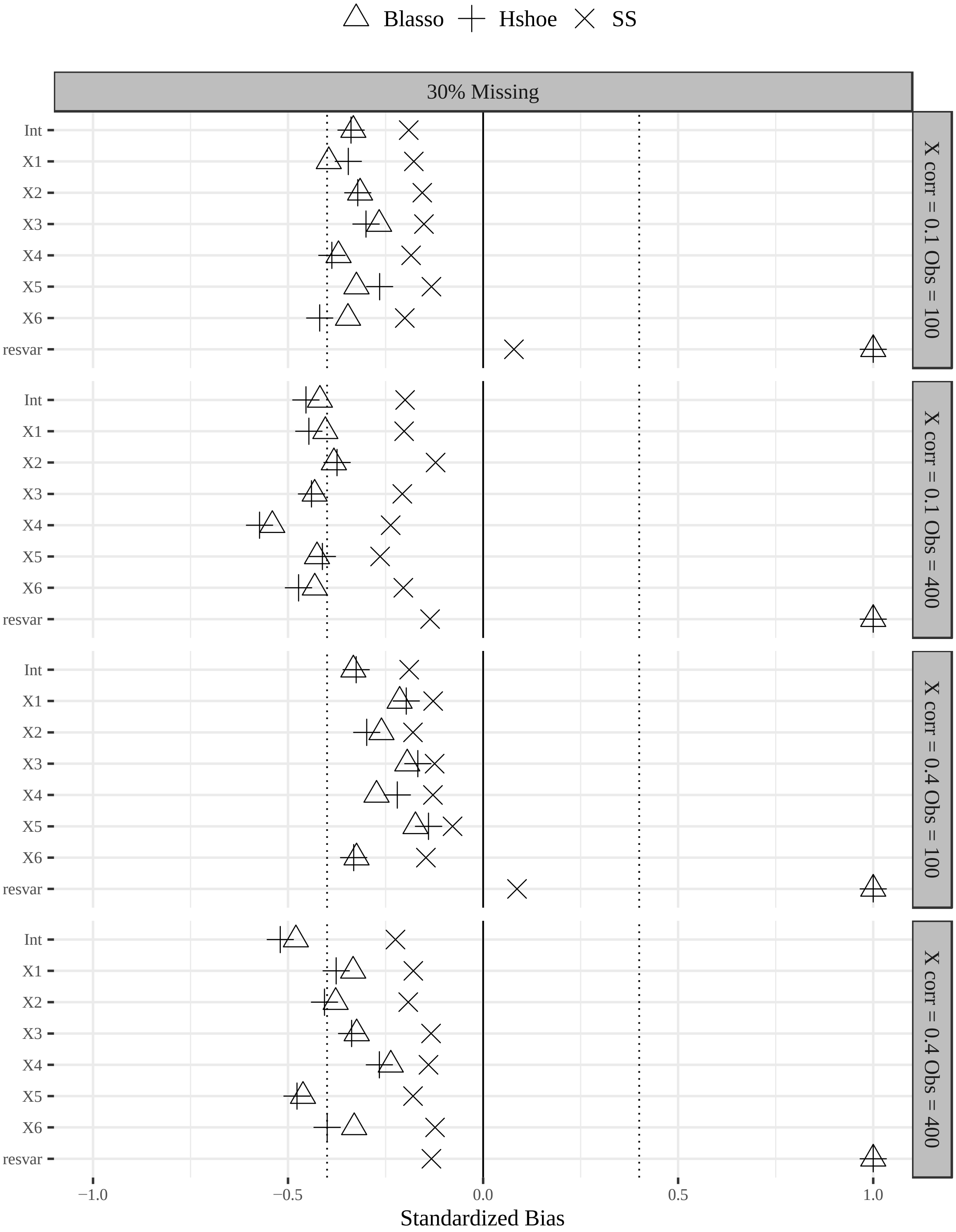


Figure 3

Complexity 4 MSE Ratios for BVS Methods

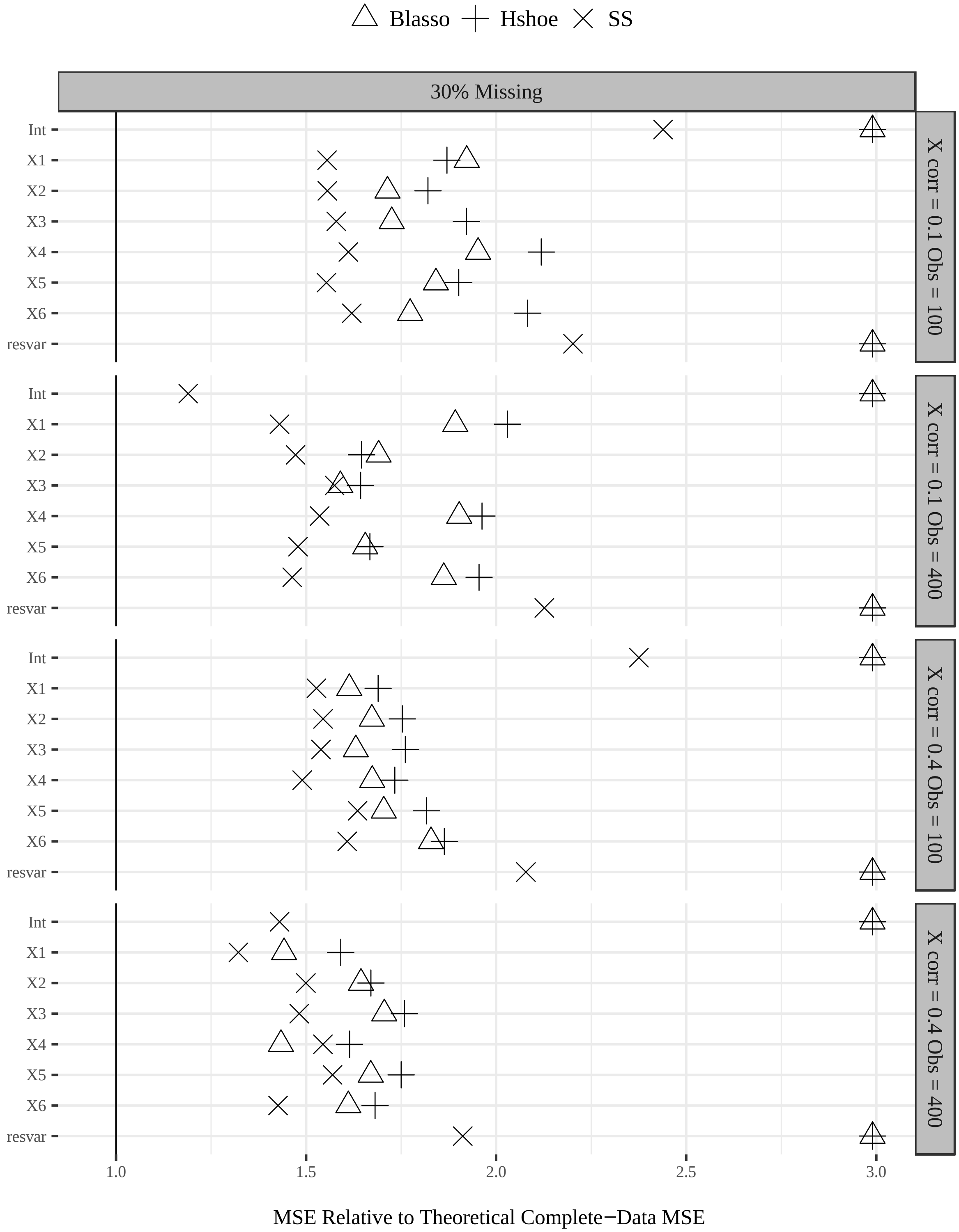


Figure 4
Complexity 3 Percent Bias for BVS Methods

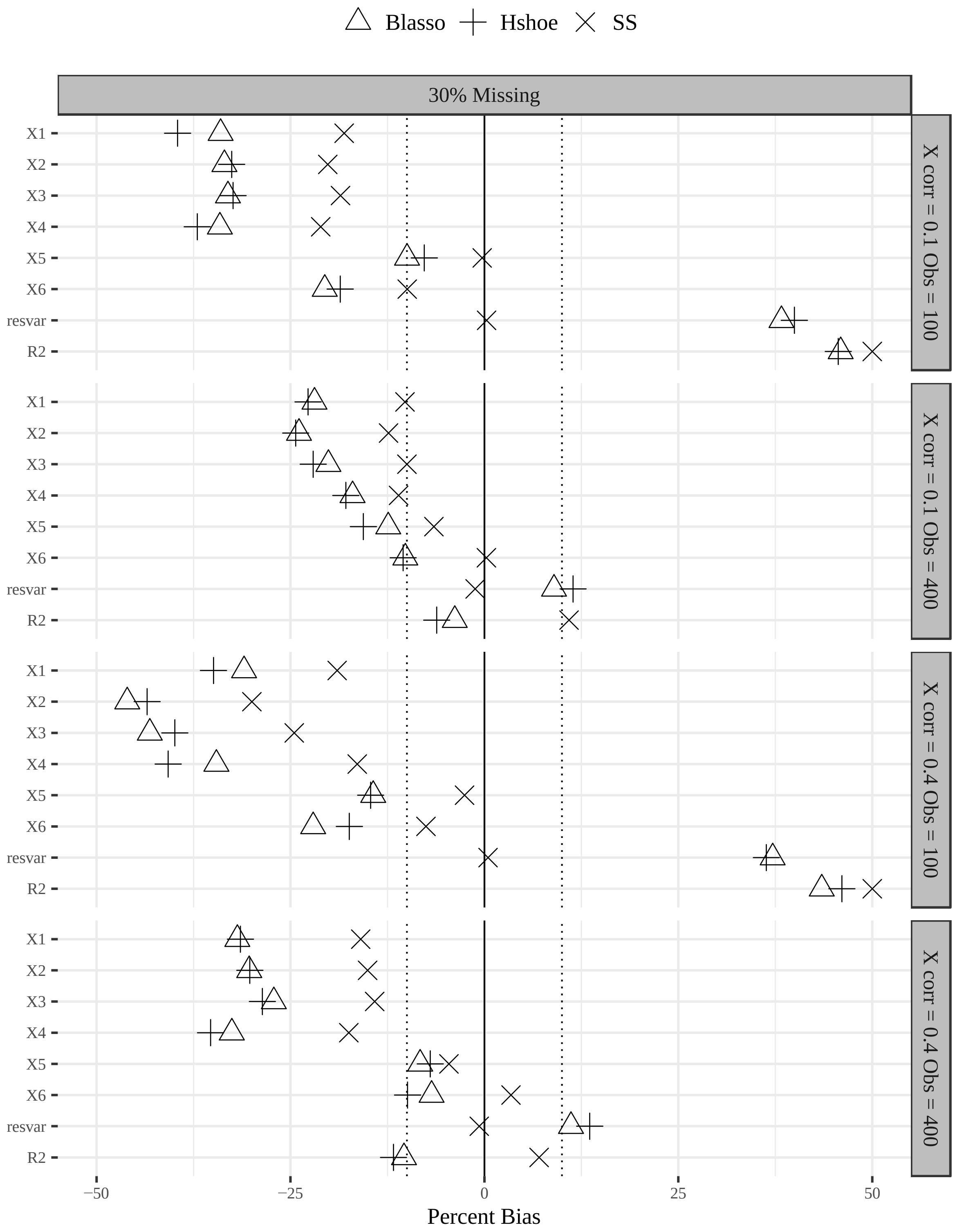


Figure 5

Complexity 3 Standardized Bias for BVS Methods

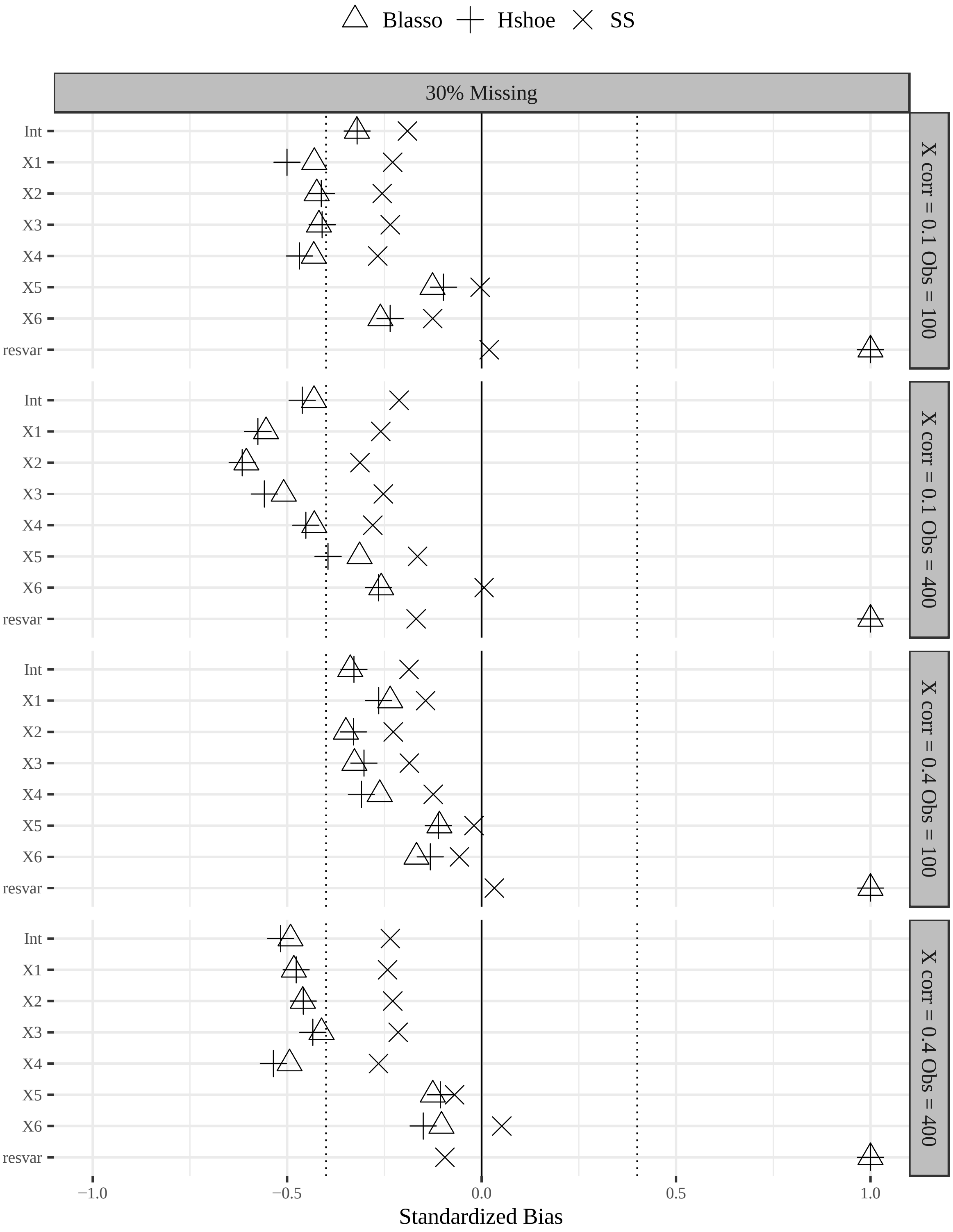


Figure 6

Complexity 3 MSE Ratios for BVS Methods

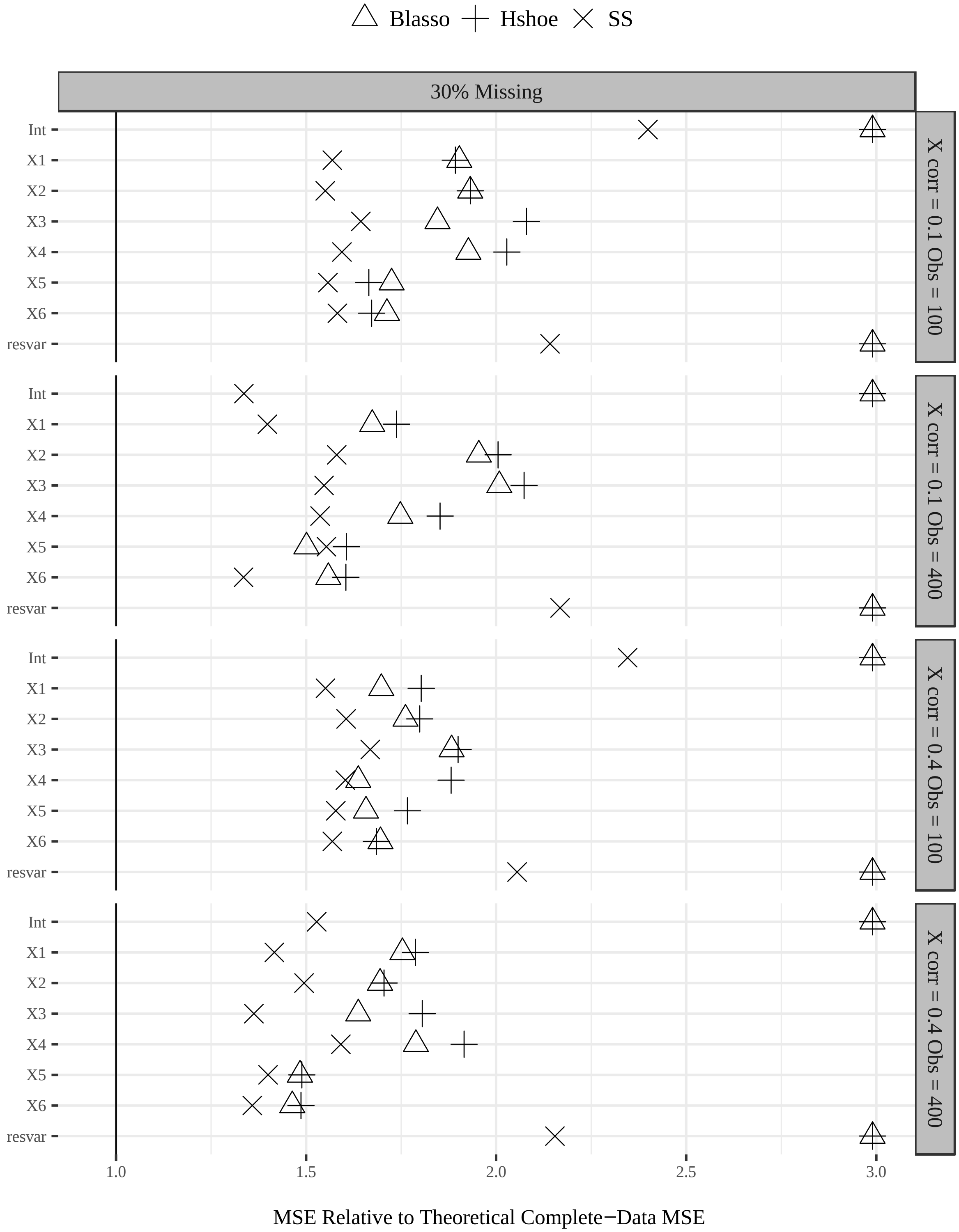


Figure 7

Complexity 2 Percentage Bias for BVS Methods

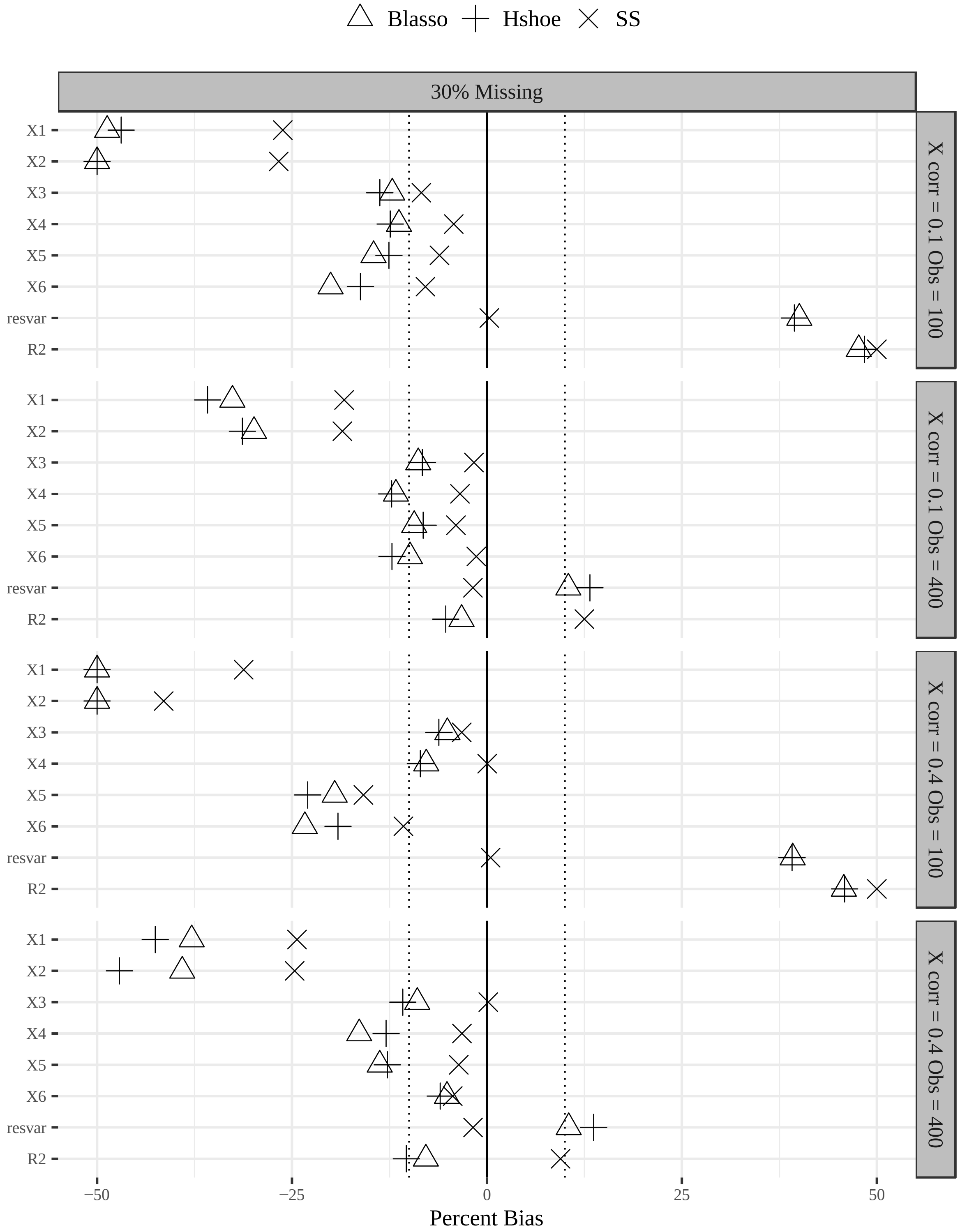


Figure 8

Complexity 2 Standardize Bias for BVS Methods

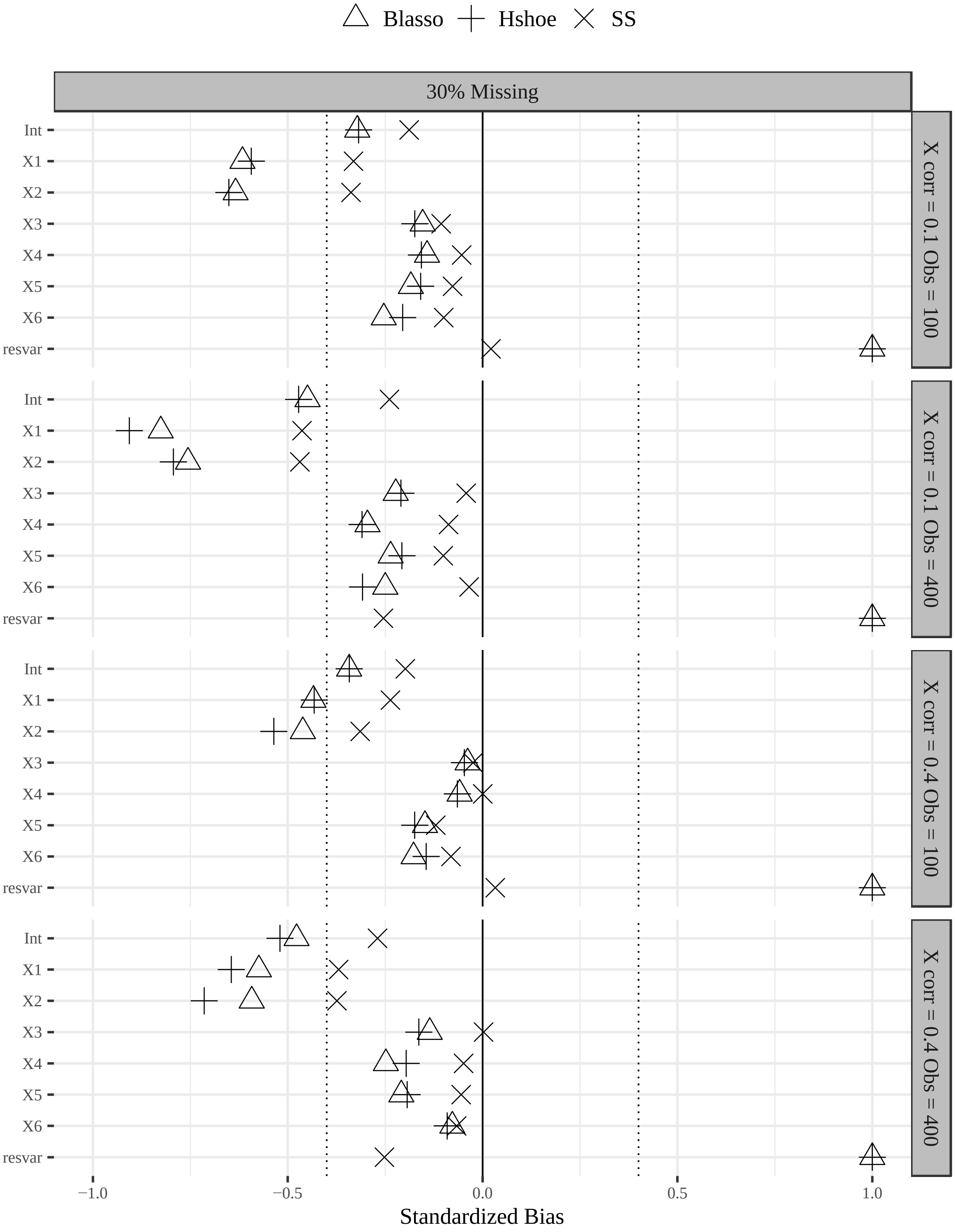


Figure 9

Complexity 2 MSE ratios for BVS Methods

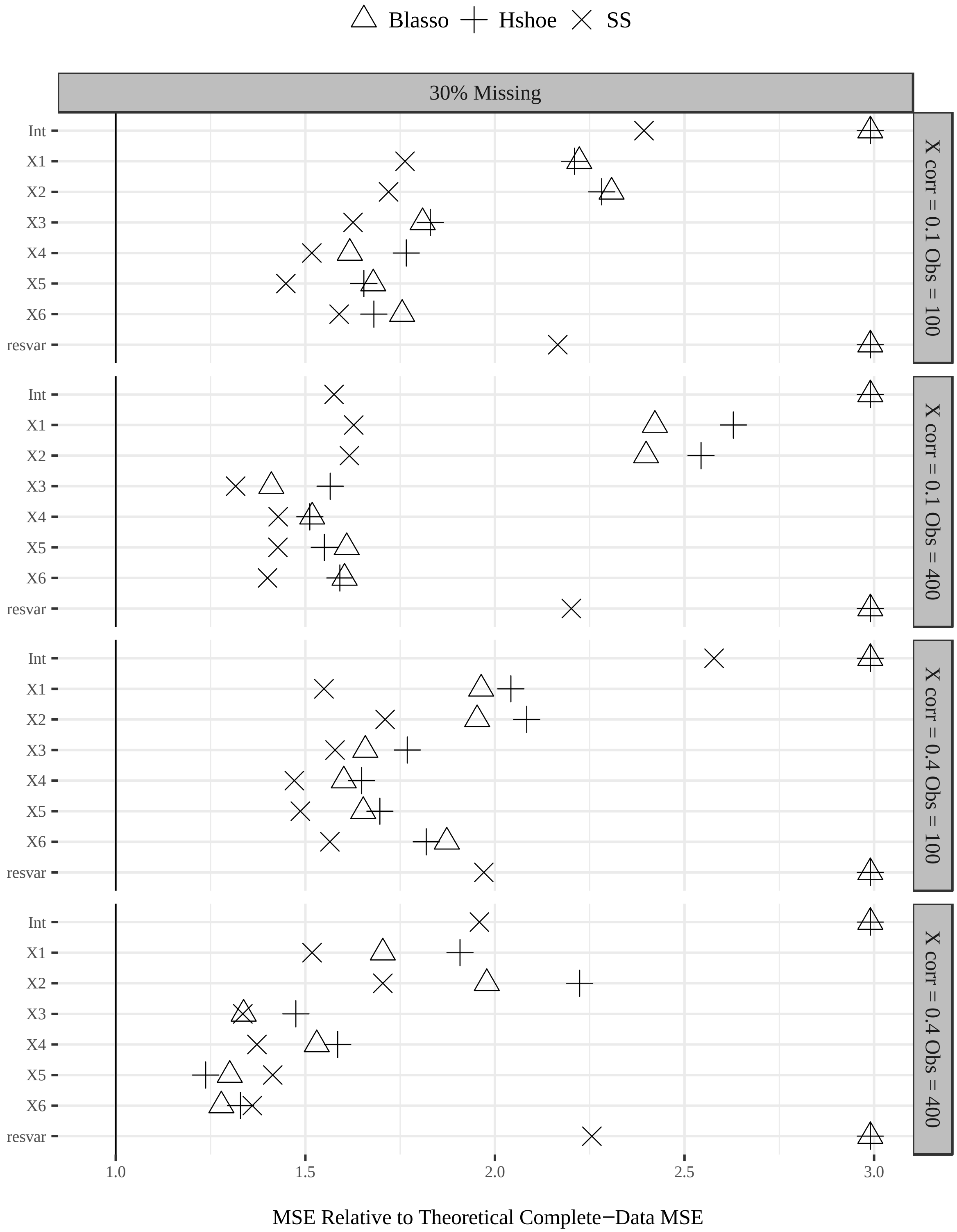


Figure 10

Complexity 1 Percentage Bias for BVS Methods

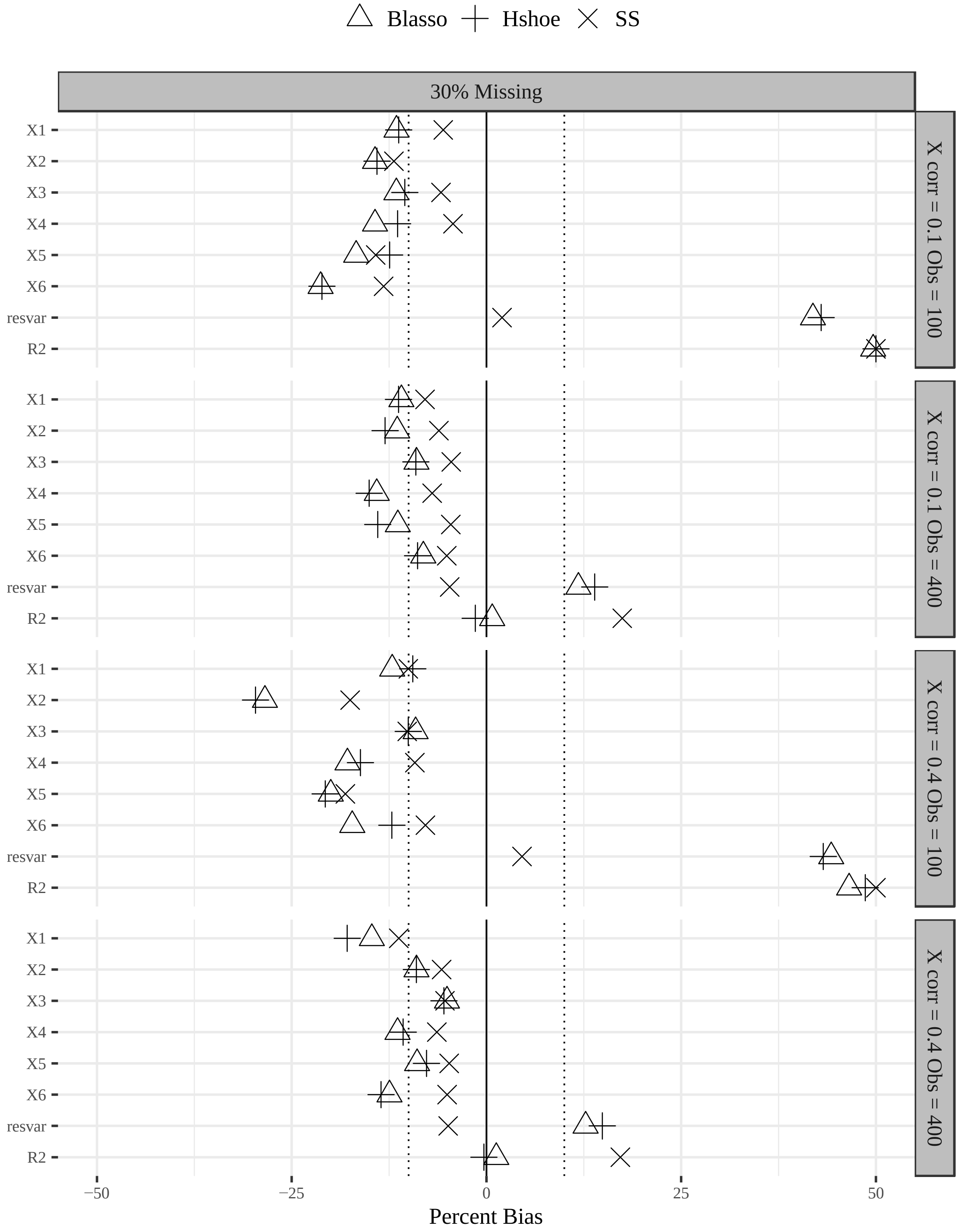


Figure 11

Complexity 1 Standardized Bias for BVS Methods

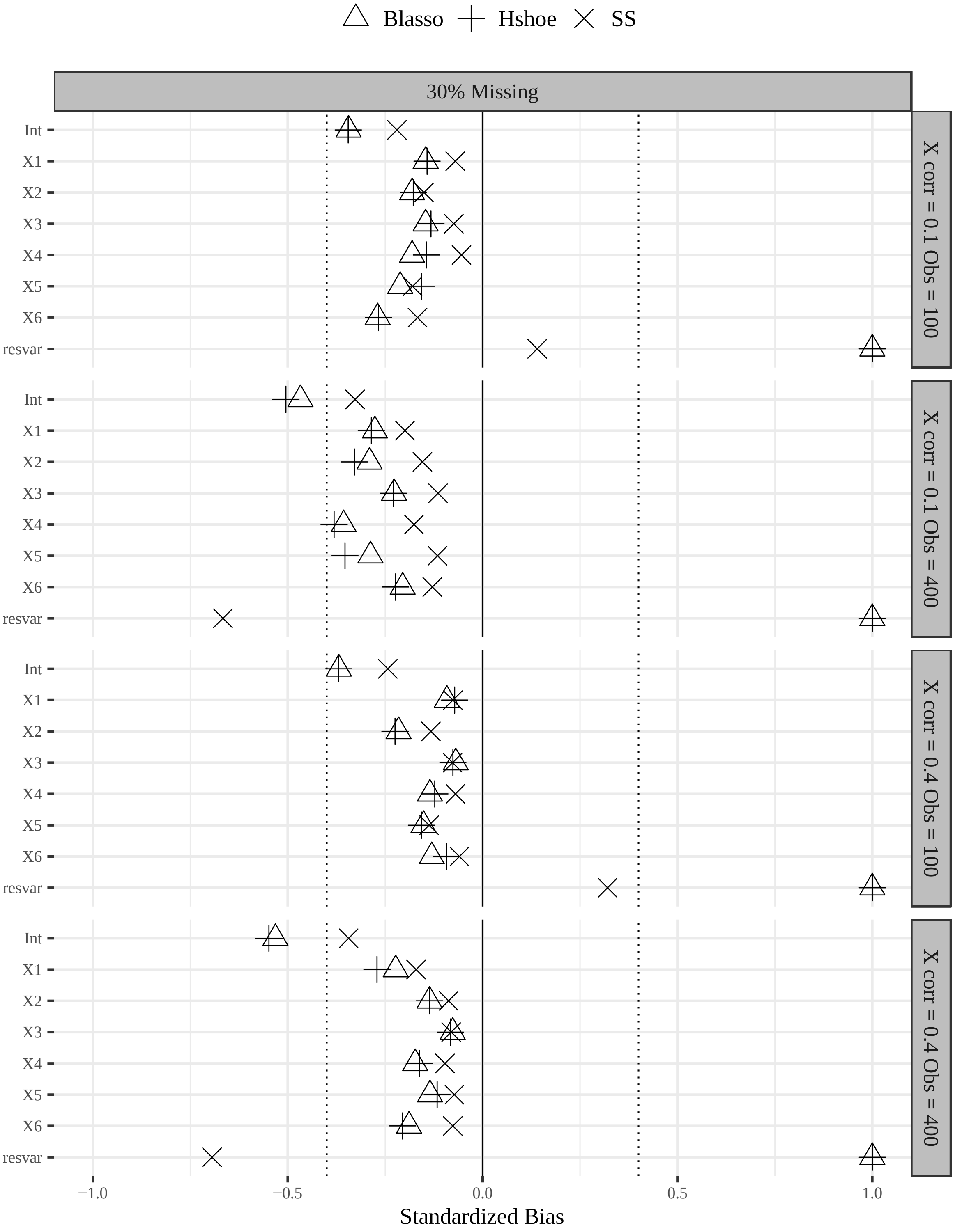


Figure 12

Complexity 1 MSE Ratios for BVS Methods

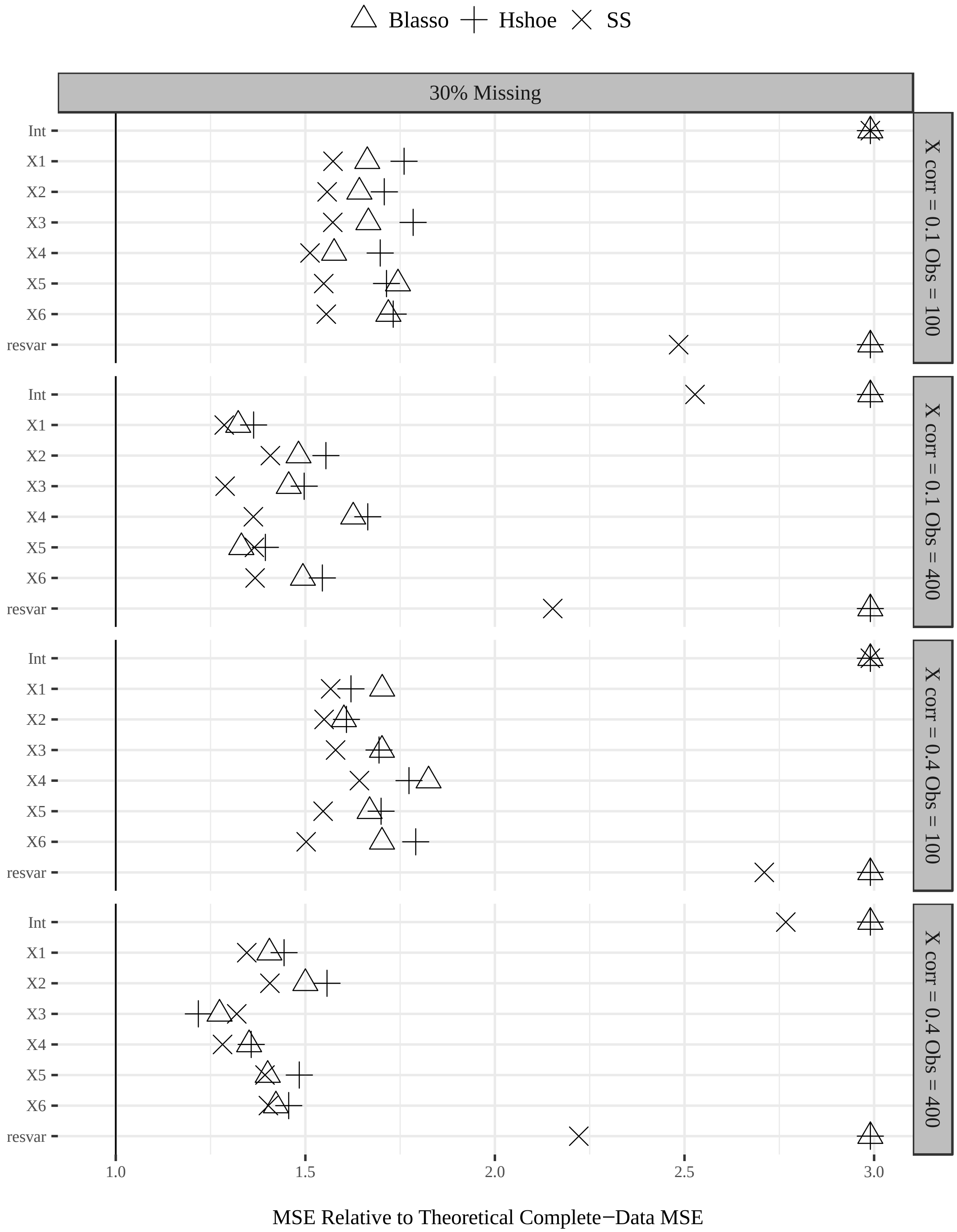


Figure 13

Complexity 4 Percentage Bias for Spike-and-Slab Prior and Standard Missing Data Methods

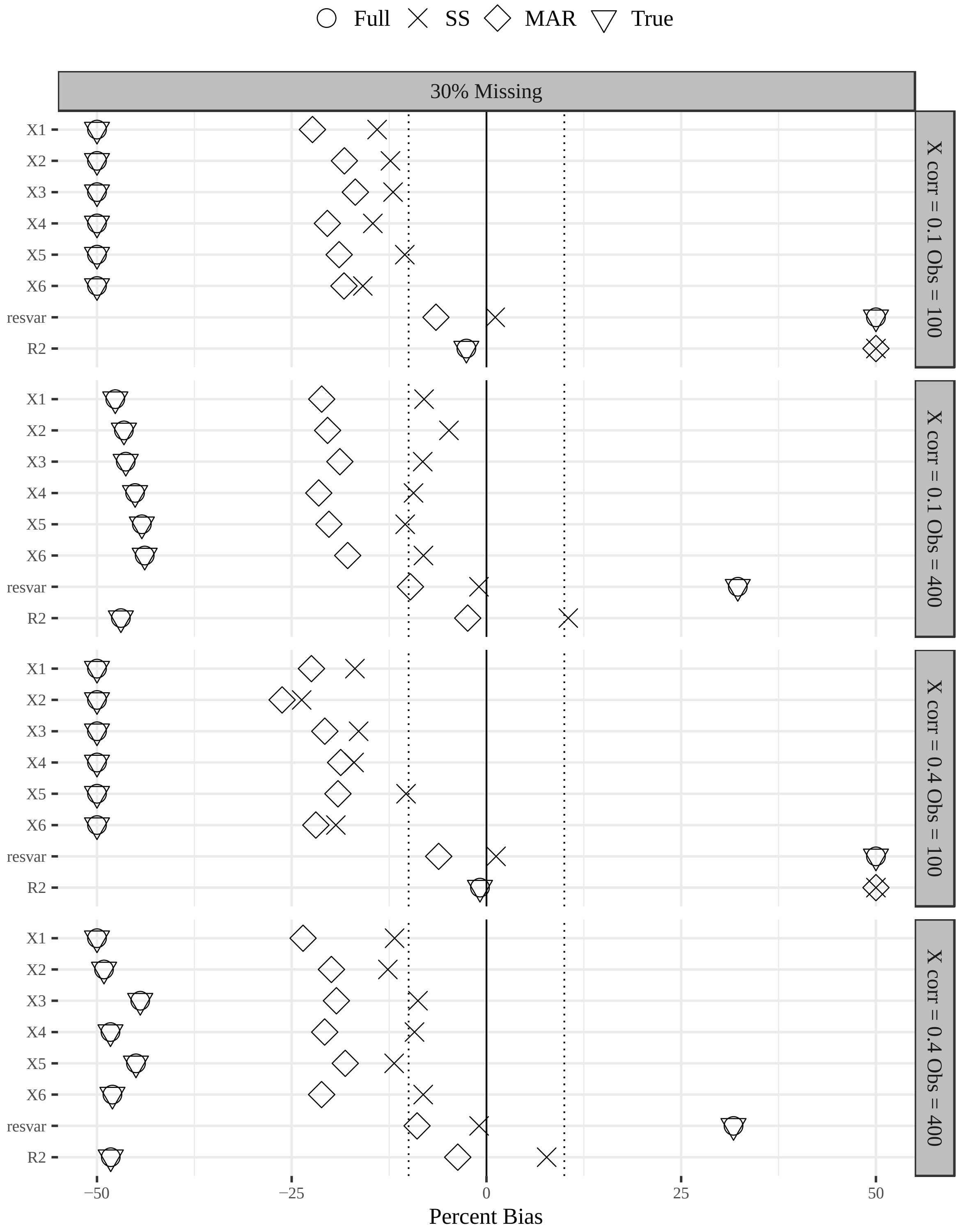


Figure 14

Complexity 4 Standardize Bias for Spike-and-Slab Prior and Standard Missing Data Methods

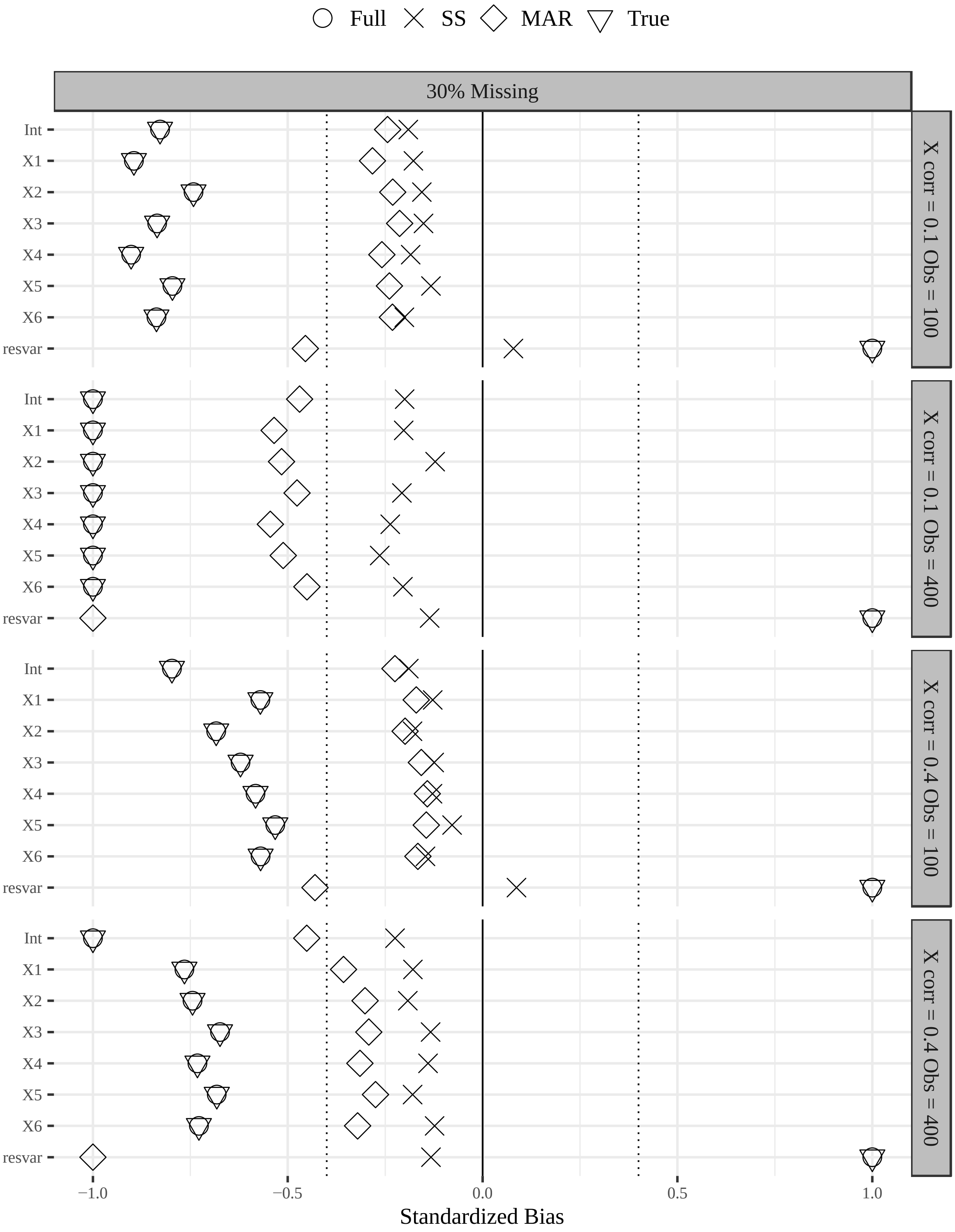


Figure 15

Complexity 4 MSE Ratios for Spike-and-Slab Prior and Standard Missing Data Methods

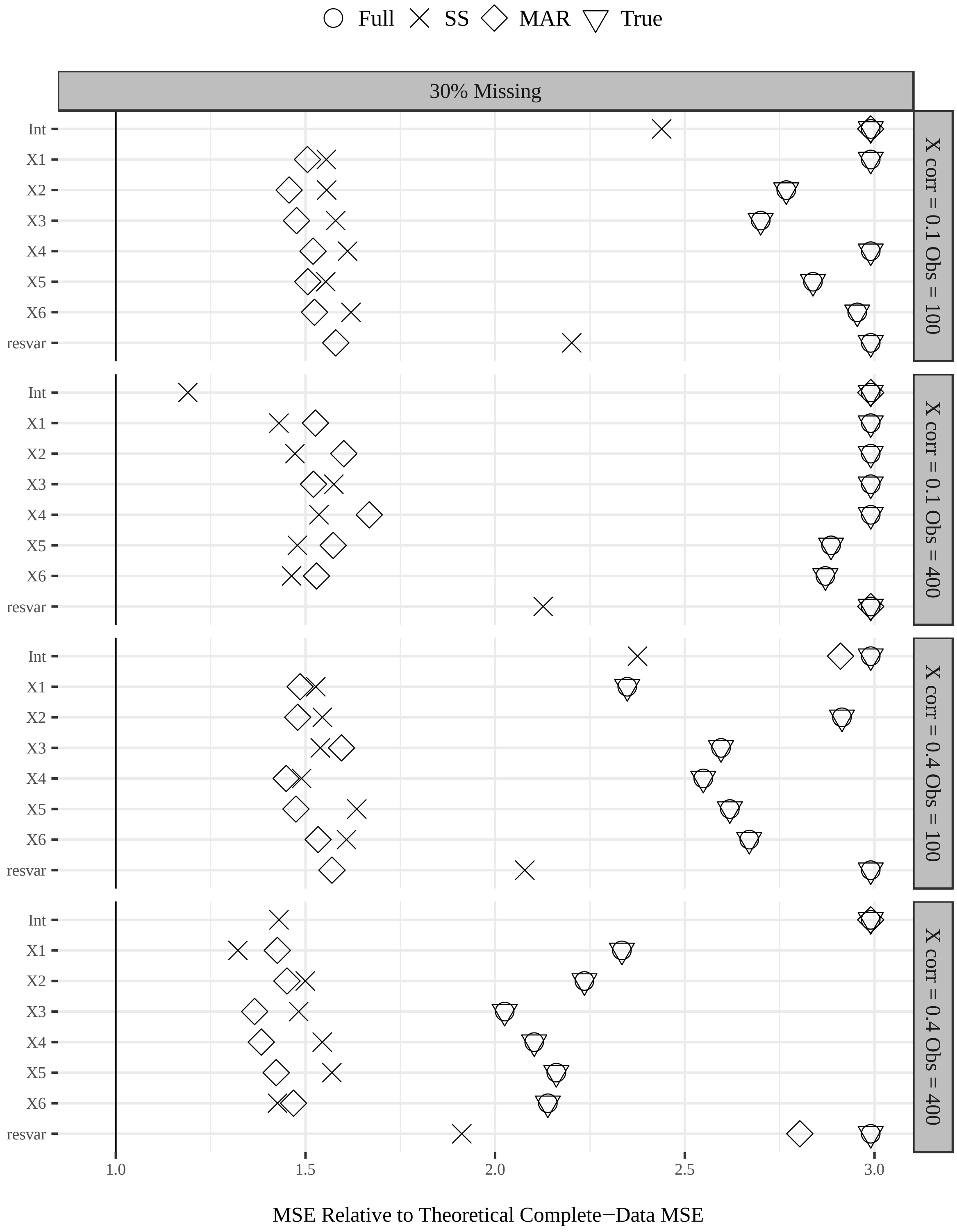


Figure 16

Complexity 3 Percent Bias for Spike-and-Slab Prior and Standard Missing Data Methods

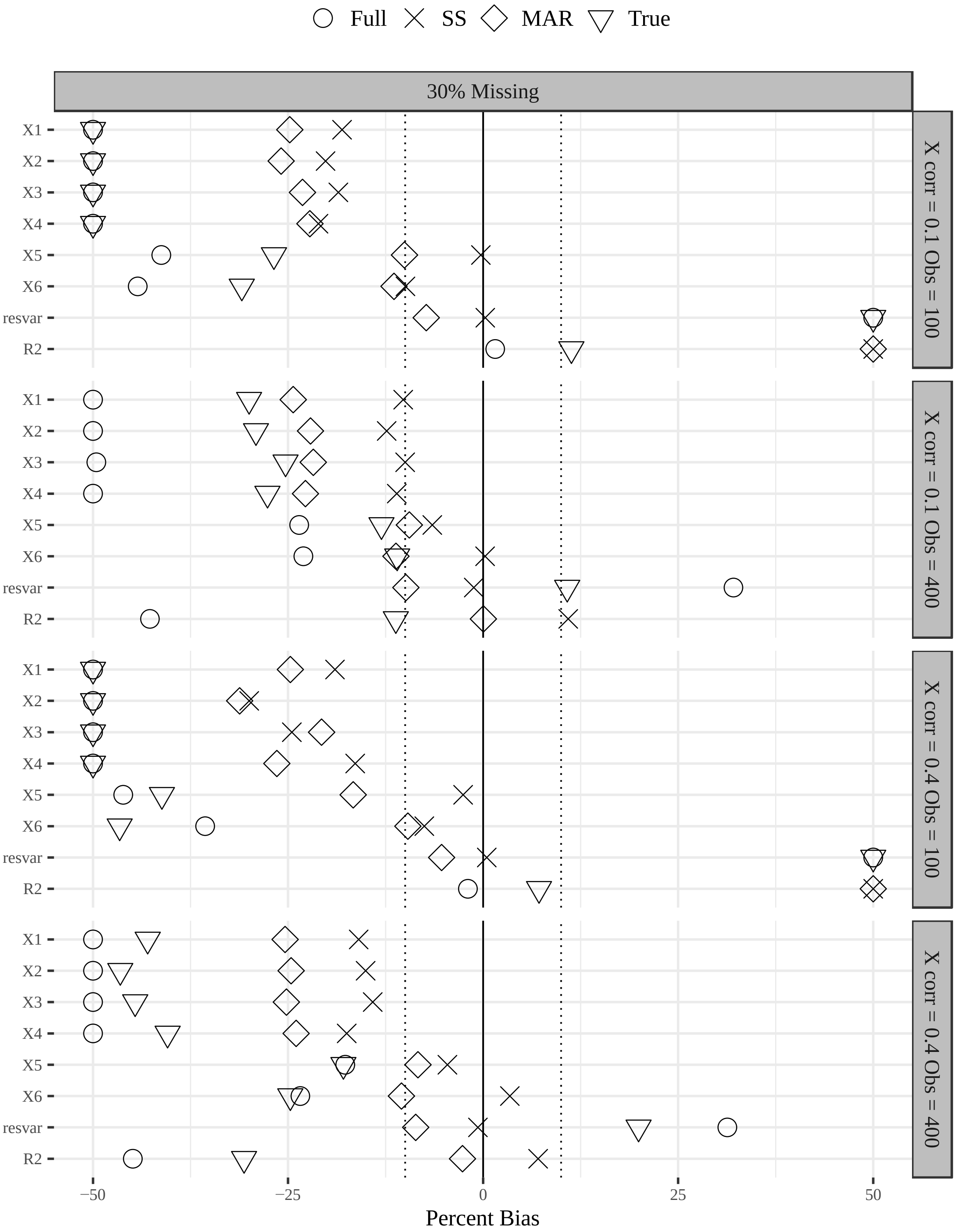


Figure 17

Complexity 3 Standardized Bias for Spike-and-Slab Prior and Standard Missing Data Methods

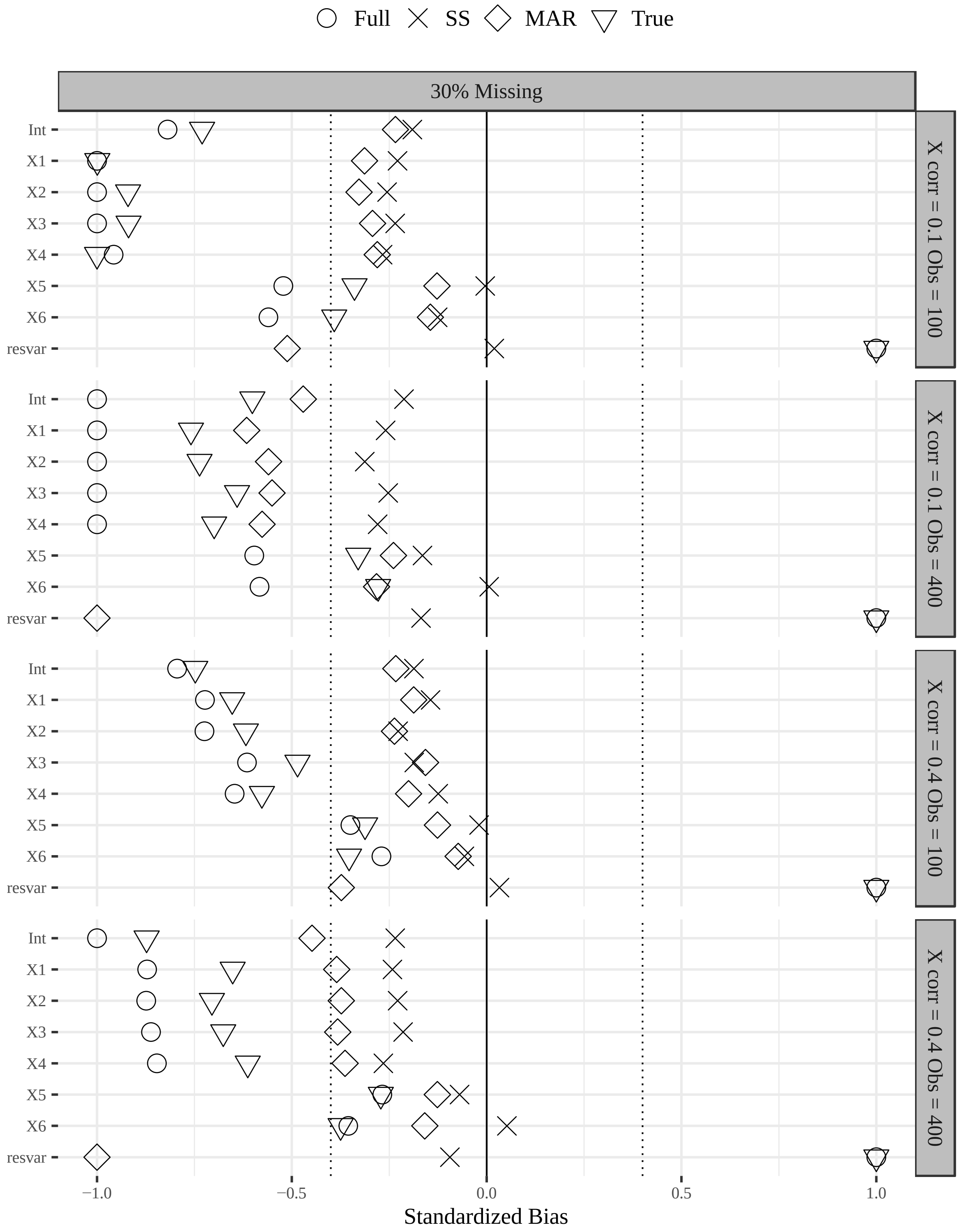


Figure 18

Complexity 3 MSE Ratios for Spike-and-Slab Prior and Standard Missing Data Methods

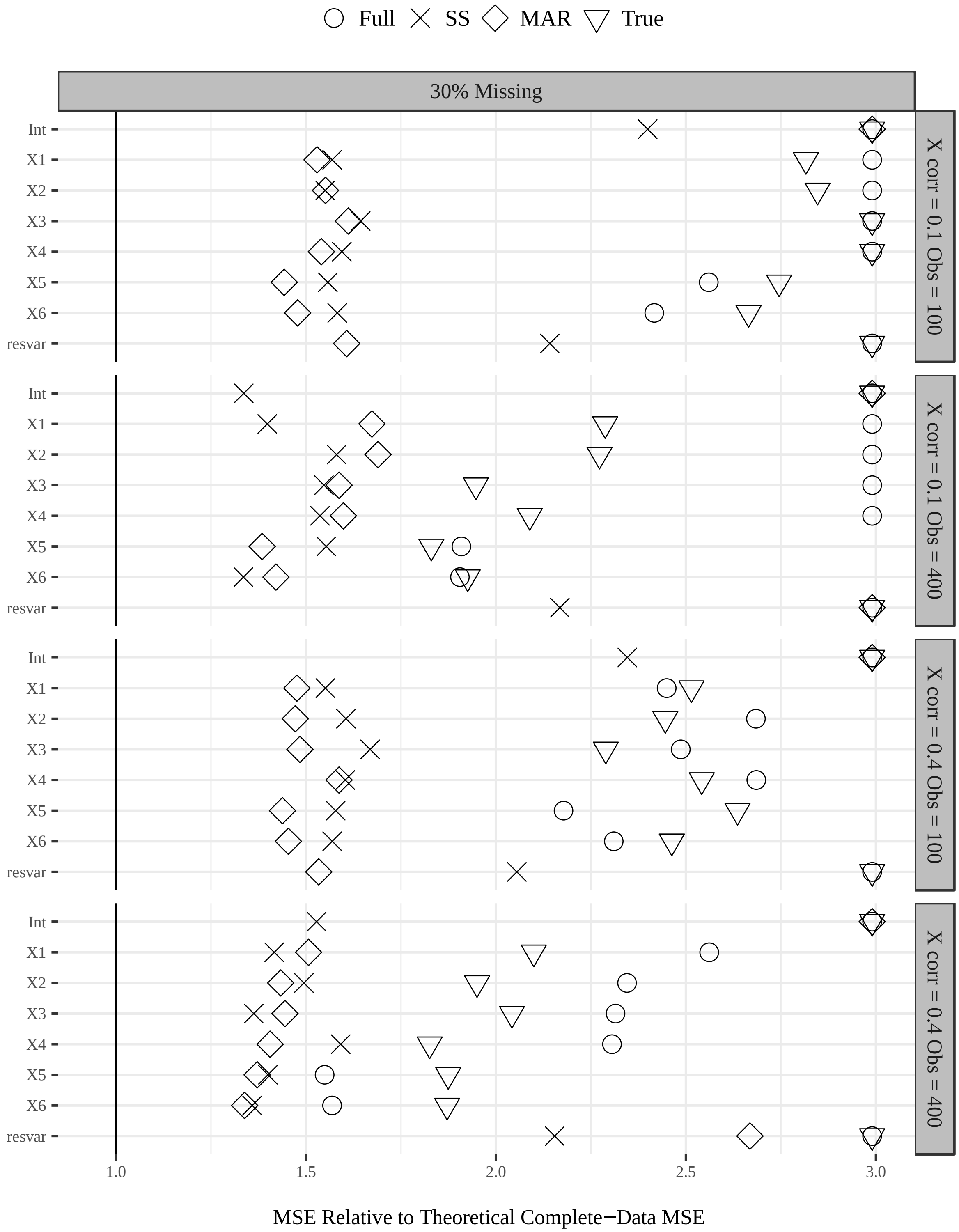


Figure 19

Complexity 2 Percent Bias for Spike-and-Slab Prior and Standard Missing Data Methods

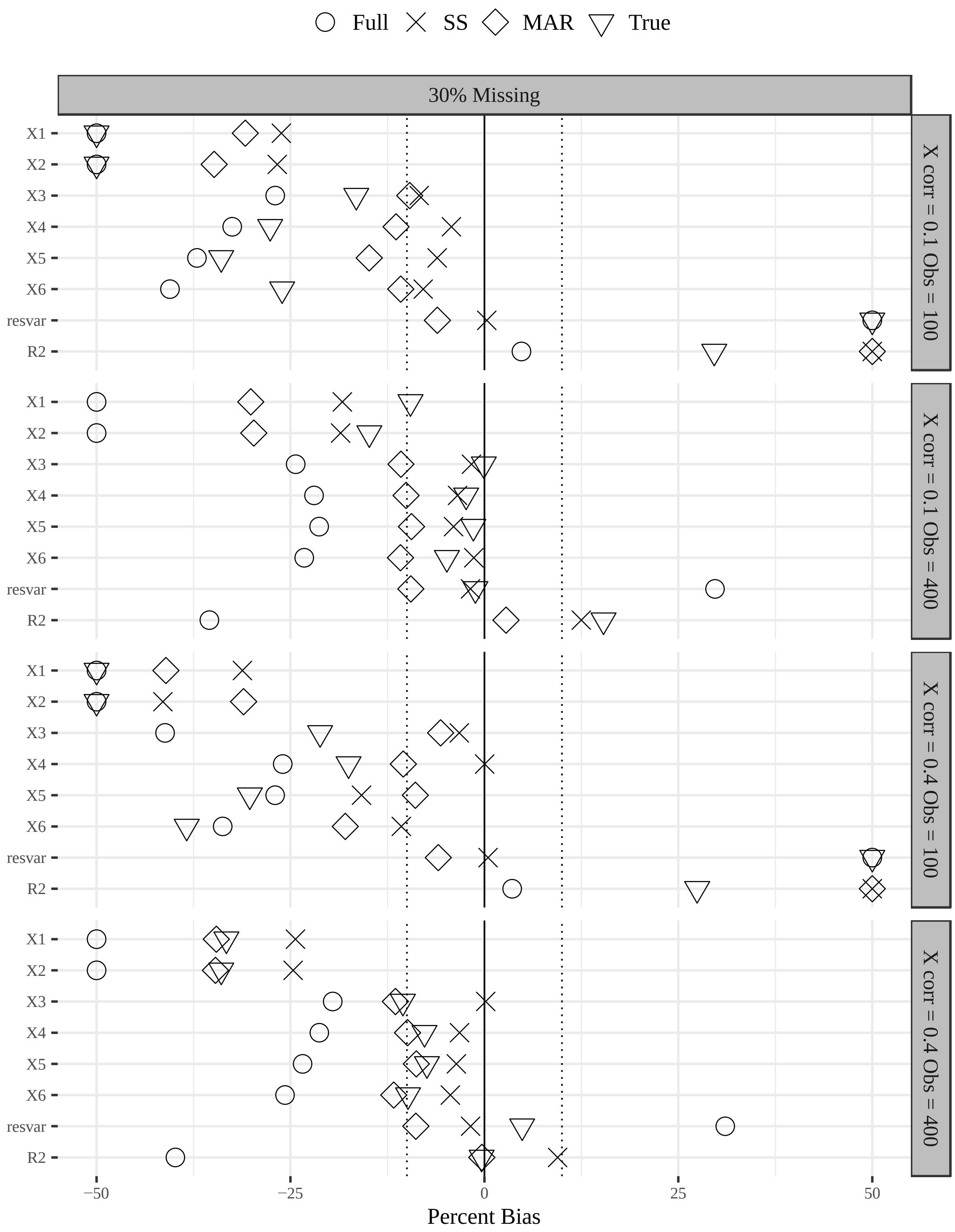


Figure 20

Complexity 2 Standardized Bias for Spike-and-Slab Prior and Standard Missing Data Methods

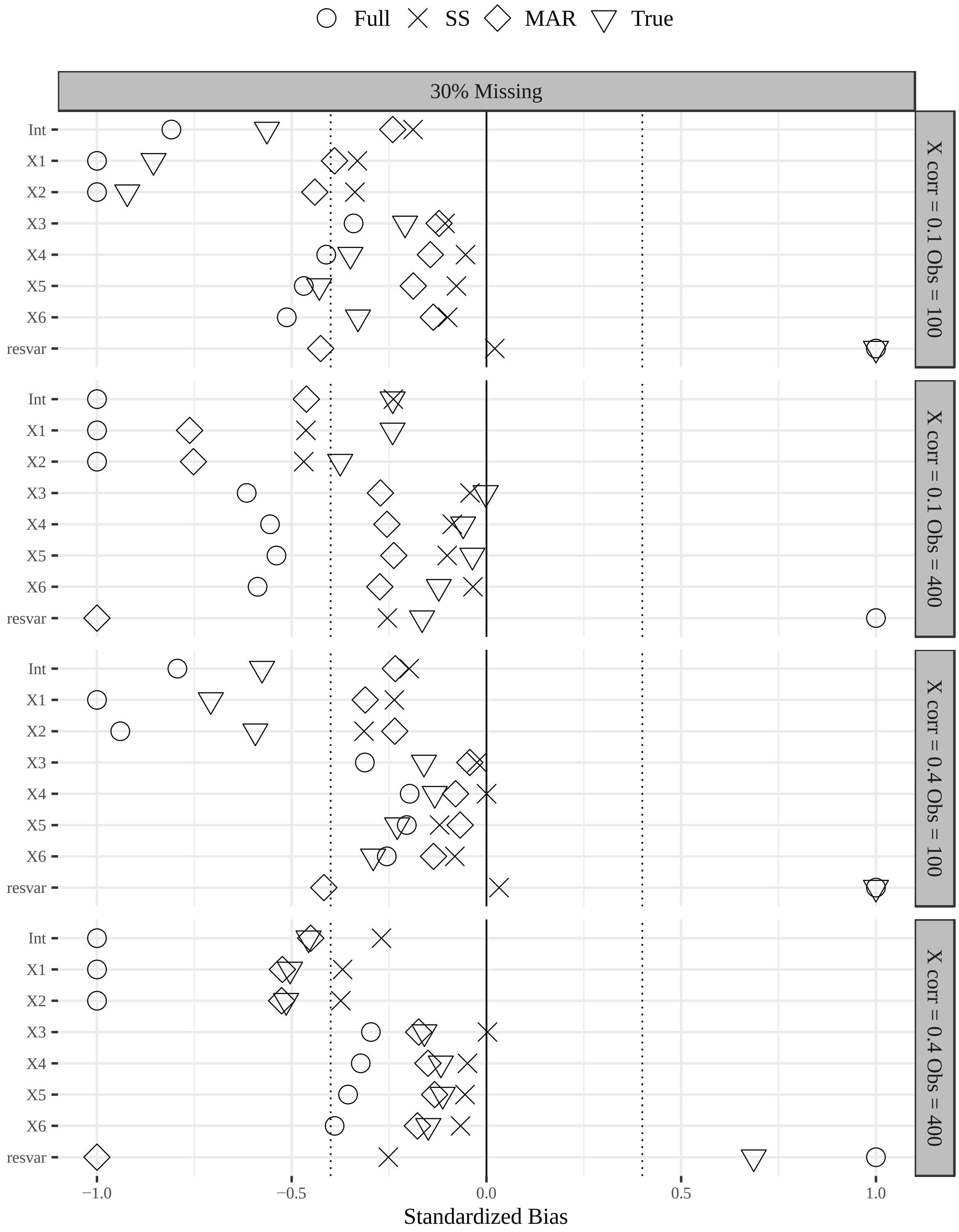


Figure 21

Complexity 2 MSE Ratios for Spike-and-Slab Prior and Standard Missing Data Methods

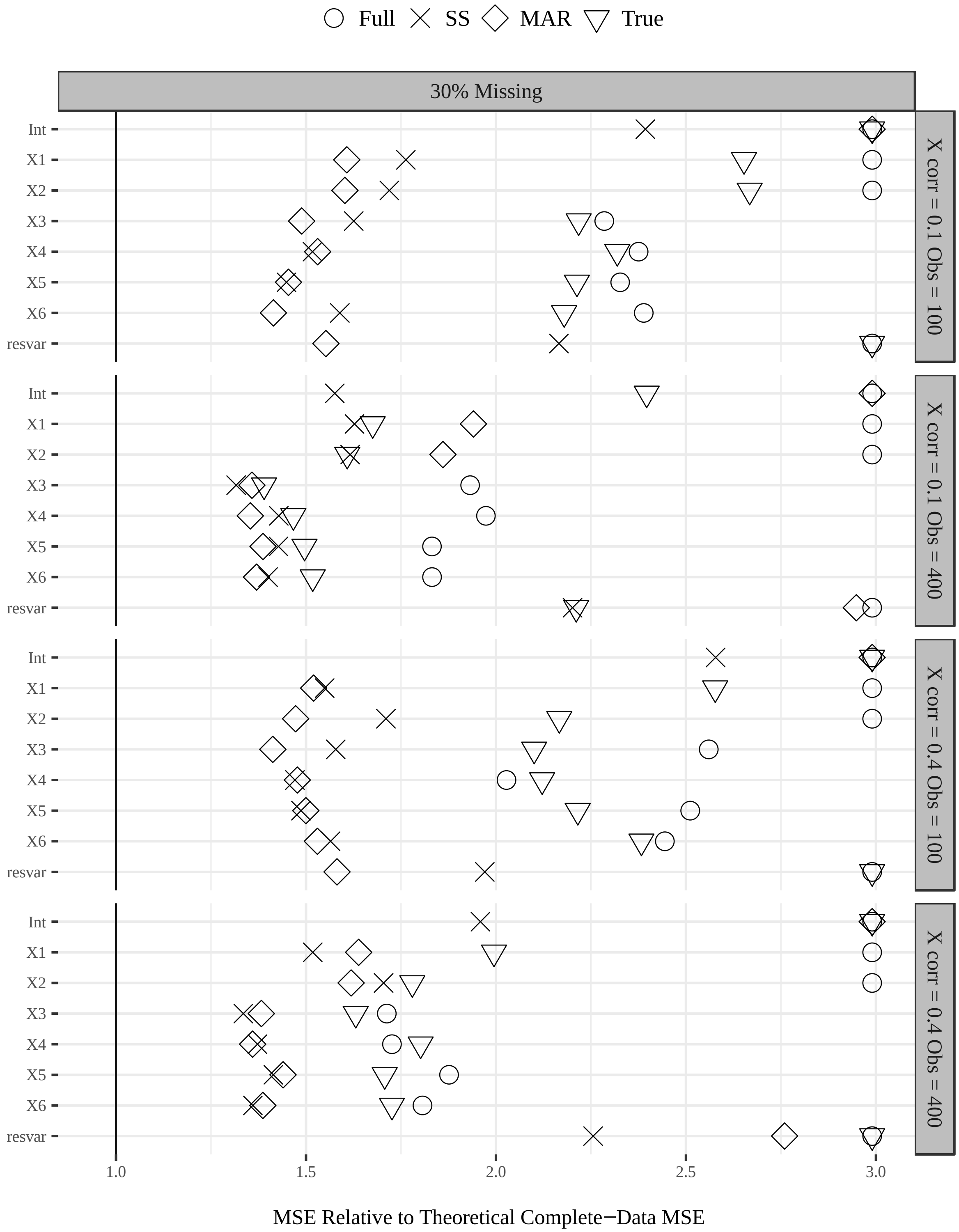


Figure 22

Complexity 1 Percent Bias for Spike-and-Slab Prior and Standard Missing Data Methods

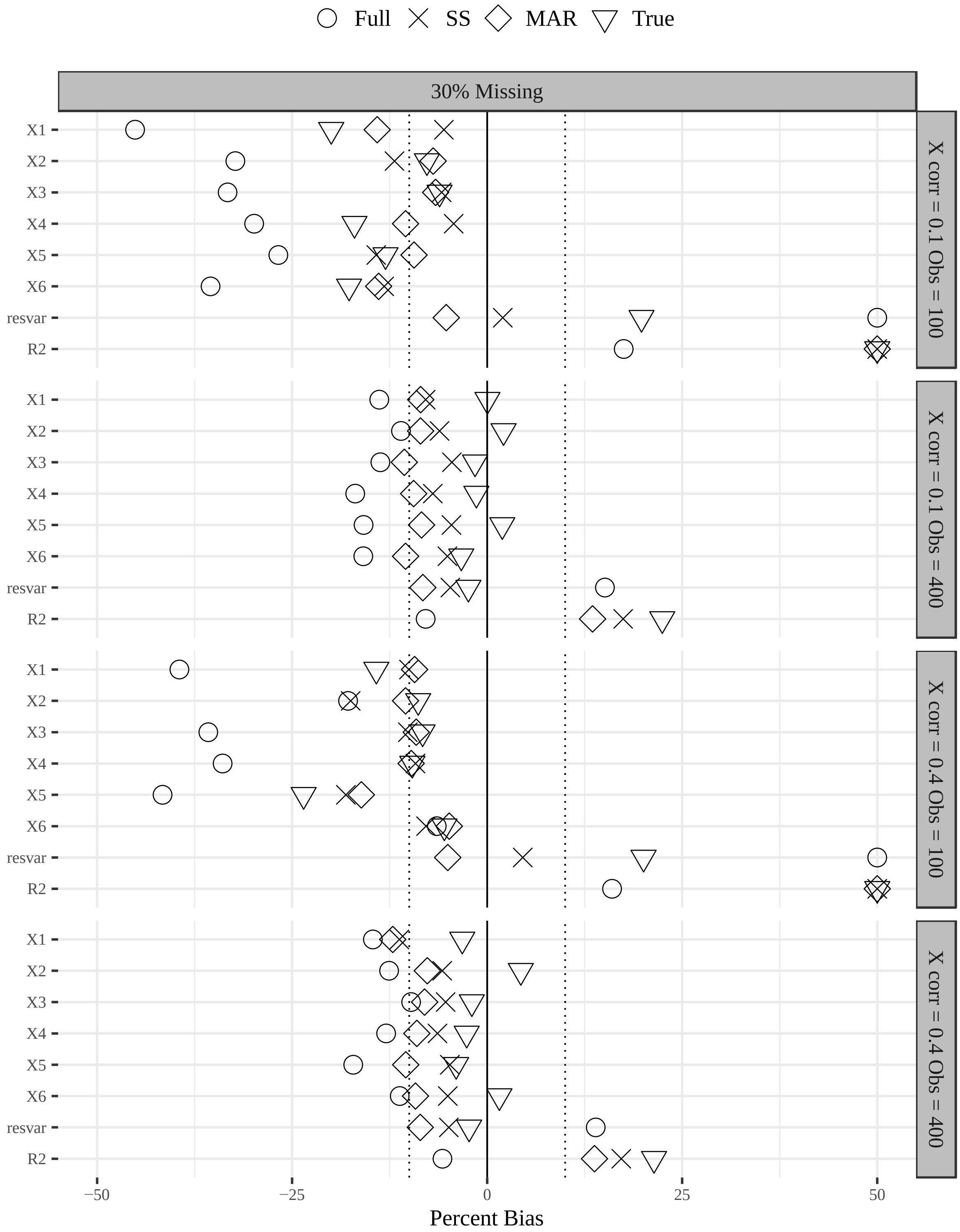


Figure 23

Complexity 1 Standardized Bias for Spike-and-Slab Prior and Standard Missing Data Methods

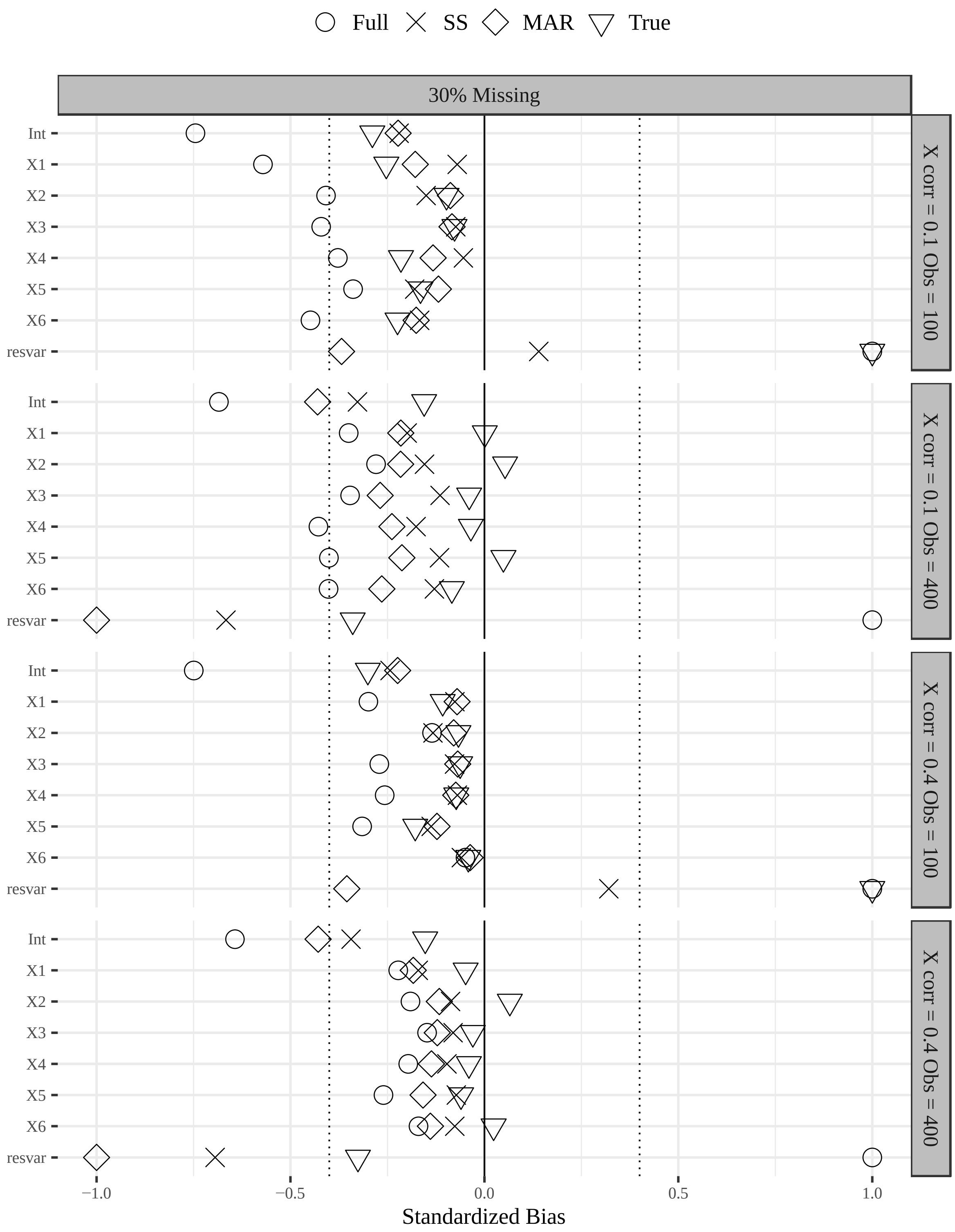


Figure 24

Complexity 1 MSE Ratios for Spike-and-Slab Prior and Standard Missing Data Methods

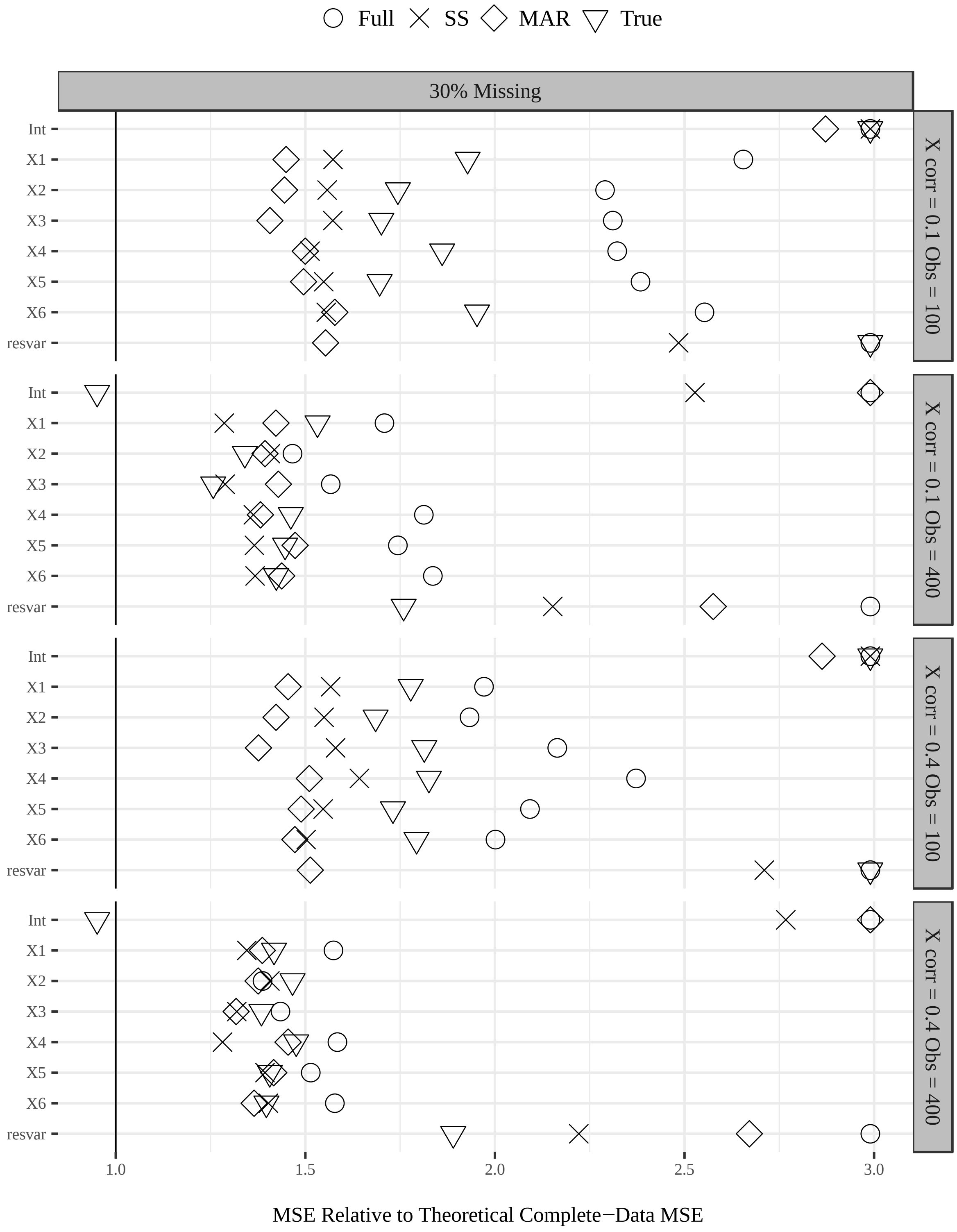


Figure 25

Posterior Distribution of the Missingness Model Intercept

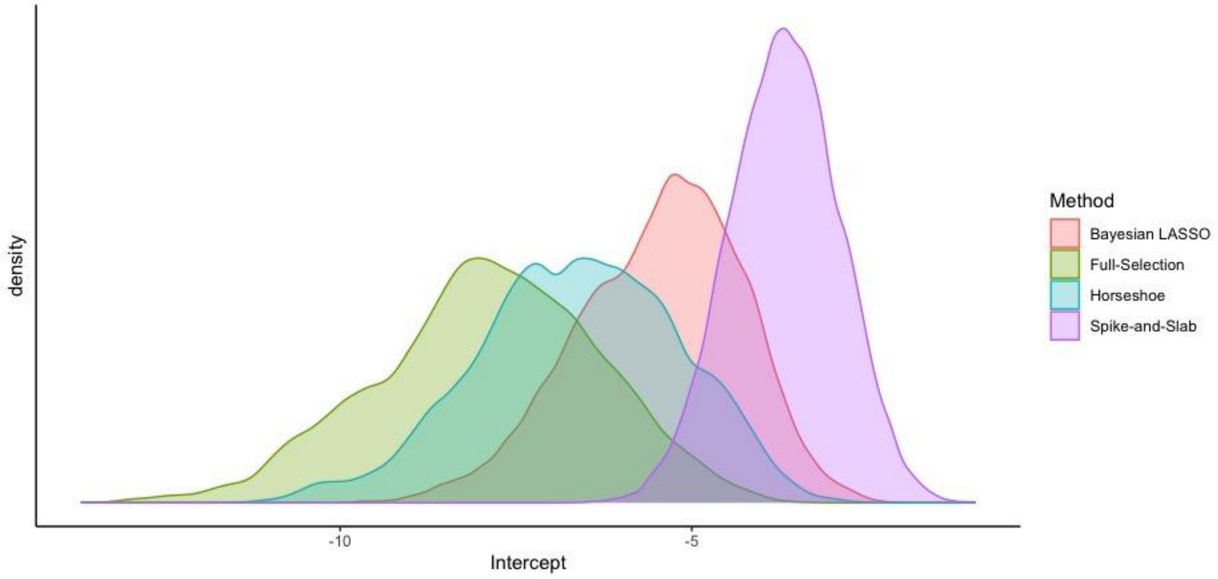


Figure 26

Posterior Distribution of Drink per Day at 8th Week

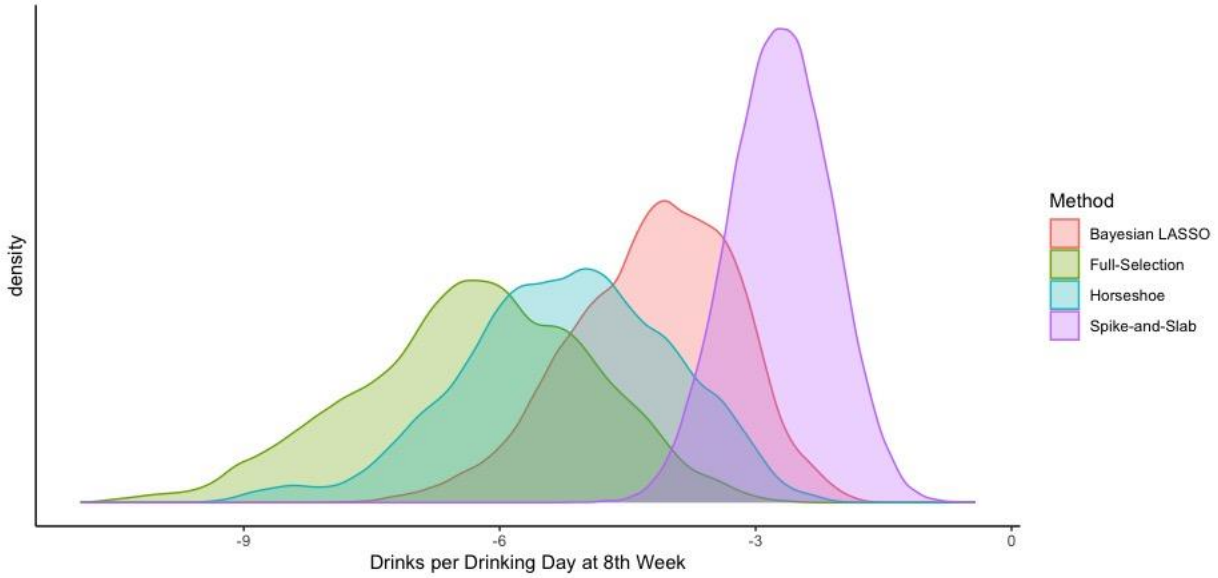


Figure 27

Posterior Distribution of Treatment Condition

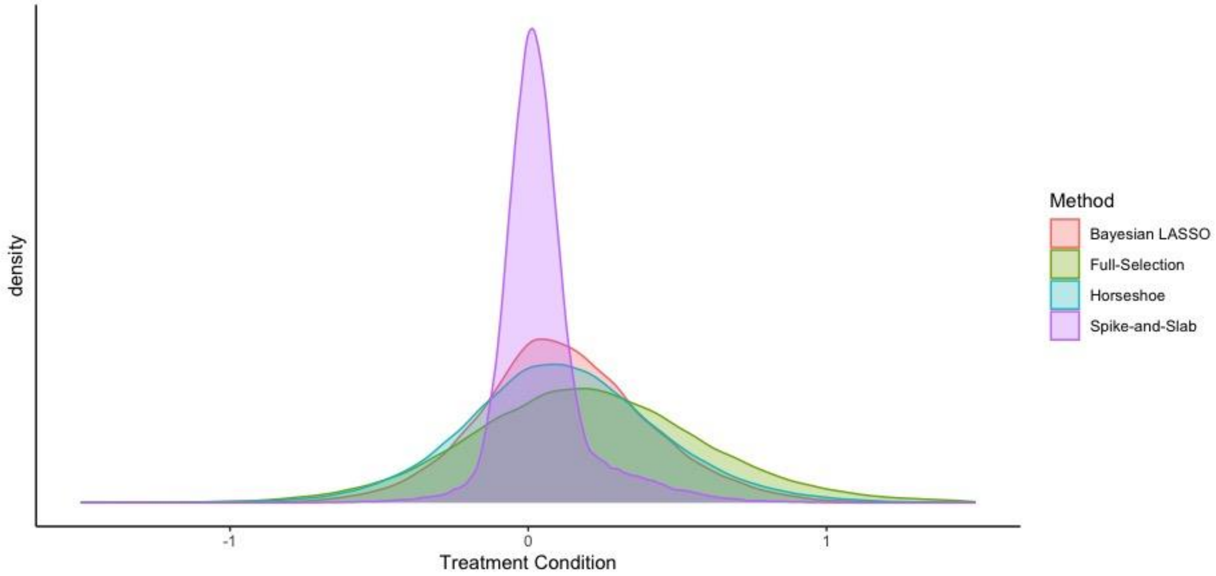


Figure 28

Posterior Distribution of Drinks per Drinking Day at Baseline

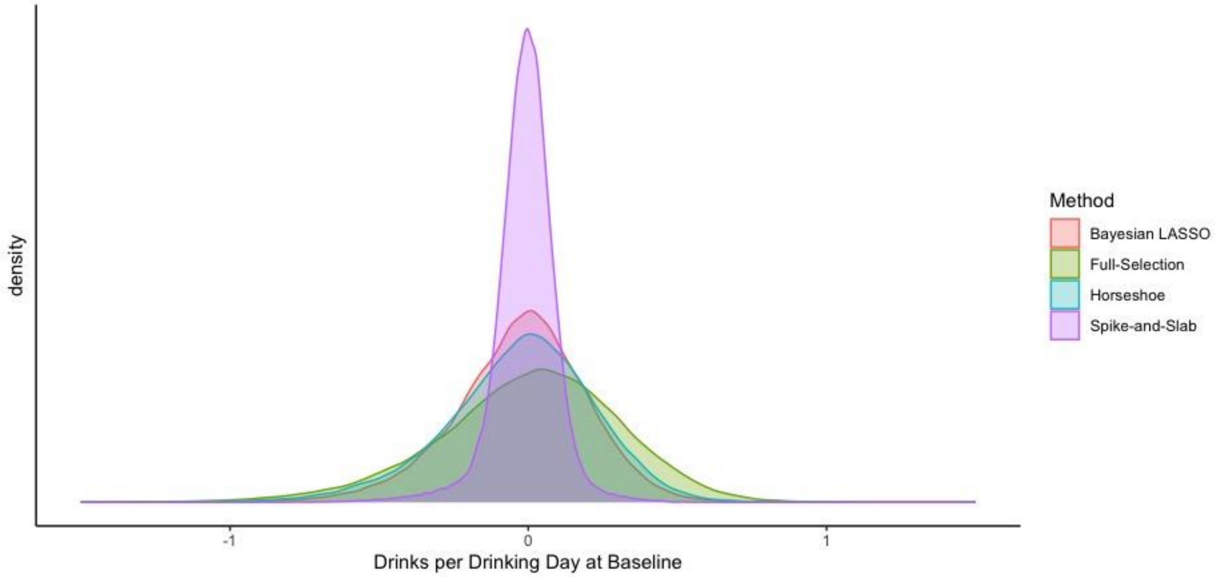


Figure 29

Posterior Distribution of Cotinine at Baseline

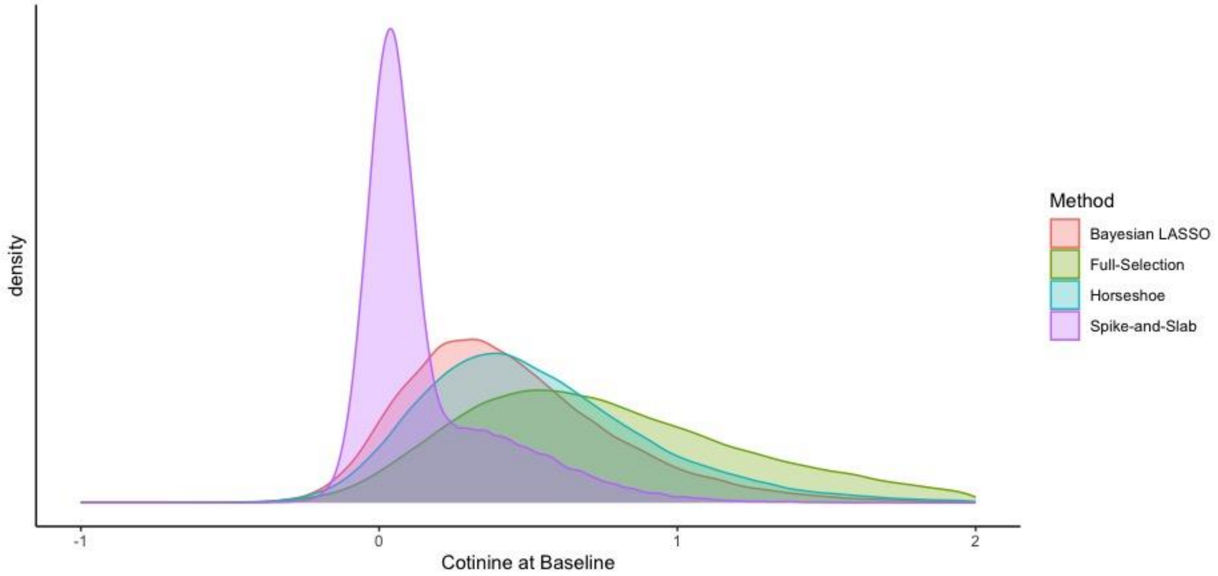


Figure 30

Posterior Distribution of Drinking Days at Baseline

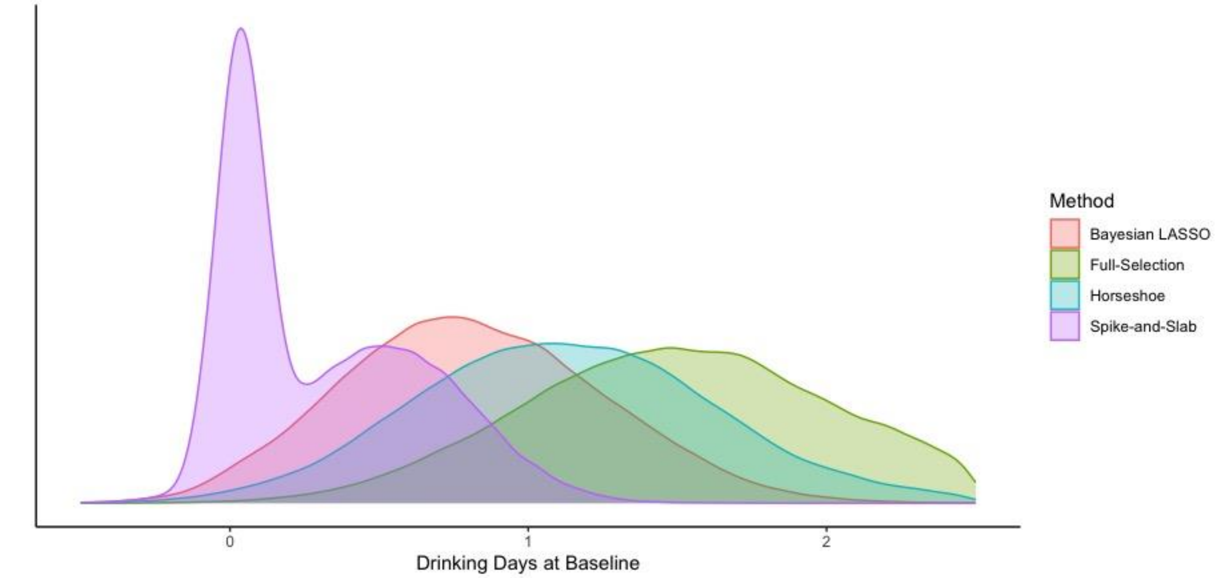


Figure 31

Posterior Distribution of Age

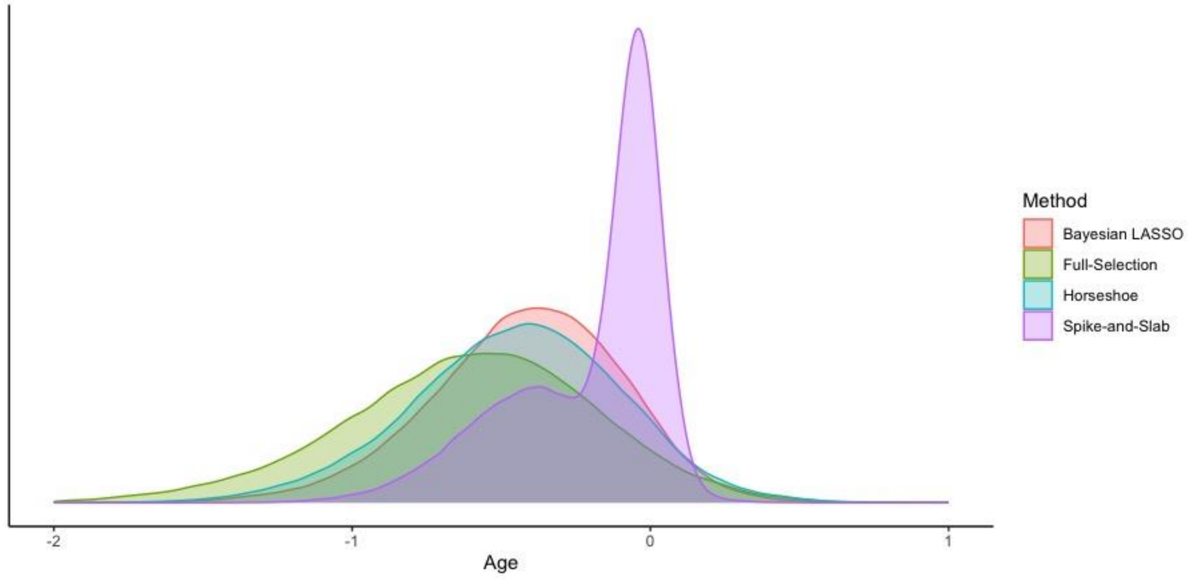


Figure 32

Posterior Distribution of Gender

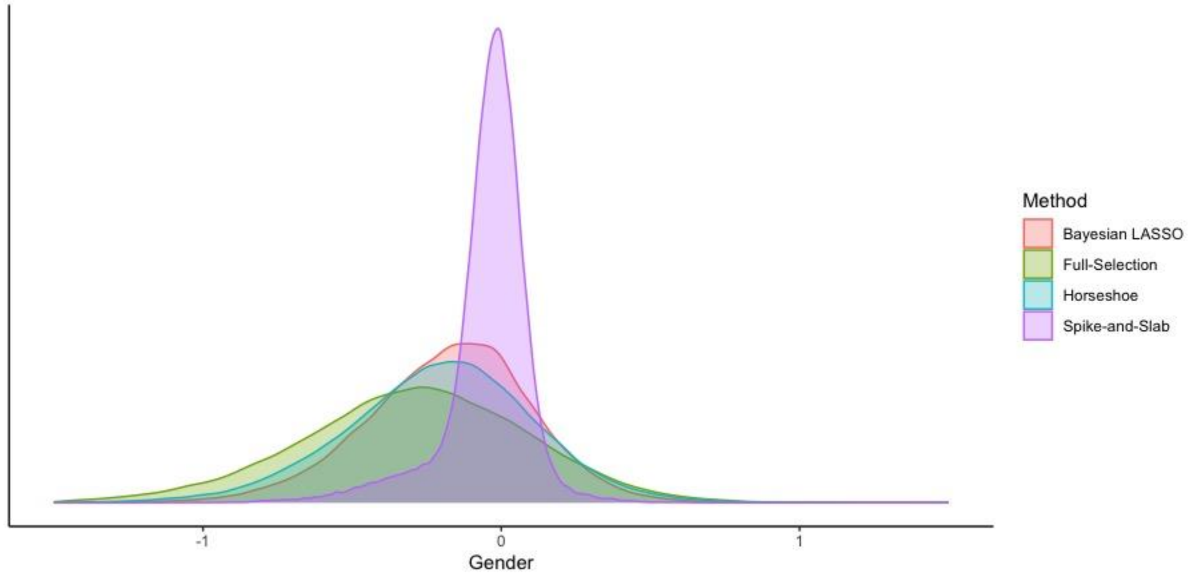
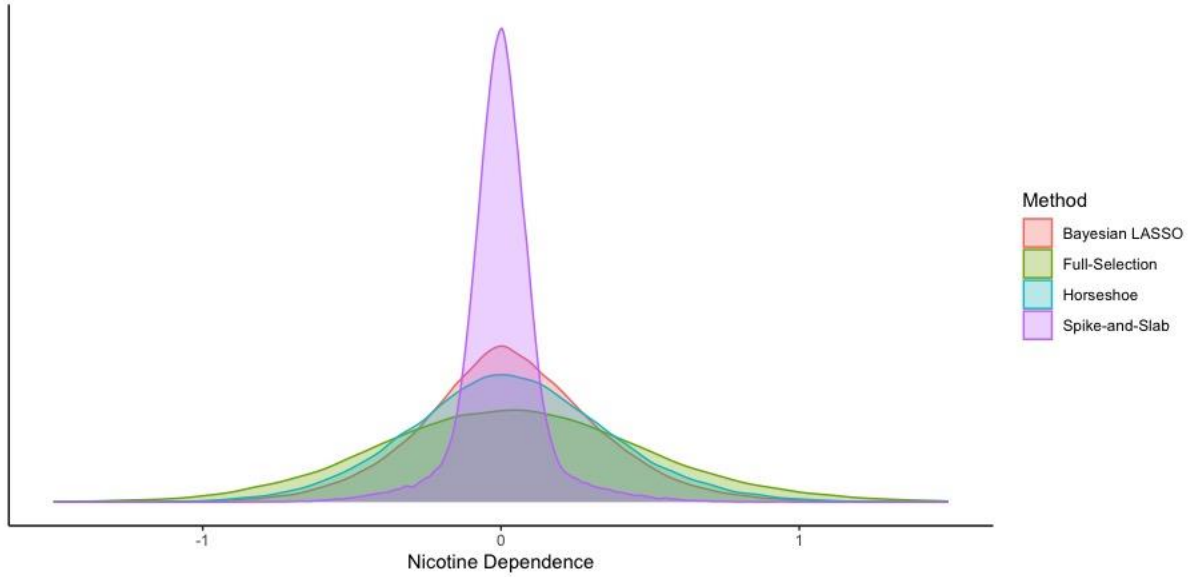


Figure 33

Posterior Distribution of Nicotine Dependence



REFERENCES

Agresti, A. (2003). *Categorical data analysis*. John Wiley & Sons.

Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, *88*(422), 669–679. http://dx.doi.org/10.1080/01621459.1993.10476321

Andrews, D. F., & Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society: Series B (Methodological)*, *36*(1), 99–102.

Bai, R., Ročková, V., & George, E. I. (2021). Spike-and-slab meets lasso: A review of the spike-and-slab lasso. In *Handbook of Bayesian Variable Selection* (1st ed., pp. 78–106). CRC Press.

Bainter, S. A., McCauley, T. G., Wager, T., & Losin, E. A. R. (2020). Improving practices for selecting a subset of important predictors in psychology: An application to predicting pain. *Advances in Methods and Practices in Psychological Science*, *3*(1), 66–80. http://dx.doi.org/10.1177/2515245919885617

Bärnighausen, T., Bor, J., Wandira-Kazibwe, S., & Canning, D. (2011). Correcting HIV prevalence estimates for survey nonparticipation using Heckman-type selection models. *Epidemiology*, *22*(1), 27–35.

Bartlett, J. W., Seaman, S. R., White, I. R., Carpenter, J. R., & Initiative*, A. D. N. (2015). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research*, *24*(4), 462–487.

Bhadra, A., Datta, J., Polson, N. G., & Willard, B. (2017). *The horseshoe+ estimator of ultra-sparse signals*.

Bhadra, A., Datta, J., Polson, N. G., & Willard, B. (2019). Lasso meets horseshoe: A survey. *Statistical Science*, *34*(3), 405–427.

Bhattacharya, A., Pati, D., Pillai, N. S., & Dunson, D. B. (2015). Dirichlet–Laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, *110*(512), 1479–1490.

Biswas, N., Mackey, L., & Meng, X.-L. (2022). Scalable Spike-and-Slab. *Proceedings of the 39th International Conference on Machine Learning*, 2021–2040. https://proceedings.mlr.press/v162/biswas22a.html

Borboudakis, G., & Tsamardinos, I. (2019). Forward-backward selection with early dropping. *The Journal of Machine Learning Research*, *20*(1), 276–314.

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, *97*(2), 465–480.

Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, *46*(3), 167–174.

Casella, G., Ghosh, M., Gill, J., & Kyung, M. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, *5*(2), 369–411.

Chen, S. M., Bauer, D. J., Belzak, W. M., & Brandt, H. (2022). Advantages of Spike and Slab Priors for Detecting Differential Item Functioning Relative to Other Bayesian Regularizing Priors and Frequentist Lasso. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(1), 122–139.

Chipman, H. (1996). Bayesian variable selection with related predictors. *Canadian Journal of Statistics*, *24*(1), 17–36.

Chow, S.-M., & Hoijtink, H. (2017). *Bayesian estimation and modeling: Editorial to the second special issue on Bayesian data analysis*. *22*(4), 609–615. https://doi.org/10.1037/met0000169

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330.

Du, H., Enders, C., Keller, B. T., Bradbury, T. N., & Karney, B. R. (2021). A Bayesian latent variable selection model for nonignorable missingness. *Multivariate Behavioral Research*, 1–49.

Eltoft, T., Kim, T., & Lee, T.-W. (2006). On the multivariate Laplace distribution. *IEEE Signal Processing Letters*, *13*(5), 300–303.

Enders, C. K. (2011). Missing not at random models for latent growth curve analyses. *Psychological Methods*, *16*(1), 1–16. https://doi.org/10.1037/a0022640

Enders, C. K. (2022). *Applied missing data analysis* (2nd ed.). Guilford Publications.

Enders, C. K., Du, H., & Keller, B. T. (2020). A model-based imputation procedure for multilevel regression models with random coefficients, interaction effects, and nonlinear terms. *Psychological Methods*, *25*(1), 88.

Erler, N. S., Rizopoulos, D., Rosmalen, J. van, Jaddoe, V. W., Franco, O. H., & Lesaffre, E. M. (2016). Dealing with missing covariates in epidemiologic studies: A comparison between multiple imputation and a full Bayesian approach. *Statistics in Medicine*, *35*(17), 2955–2974.

Fitzmaurice, G., Davidian, M., Verbeke, G., & Molenberghs, G. (2008). *Longitudinal data analysis*. CRC press.

Galimard, J.-E., Chevret, S., Curis, E., & Resche-Rigon, M. (2018). Heckman imputation models for binary or continuous MNAR outcomes and MAR predictors. *BMC Medical Research Methodology*, *18*(1), 1–13.

Galimard, J.-E., Chevret, S., Protopopescu, C., & Resche-Rigon, M. (2016). A multiple imputation approach for MNAR mechanisms compatible with Heckman's model. *Statistics in Medicine, 35*(17), 2907–2920.

Gao, D. (2018). *Bayesian Lasso Models – With Application to Sports Data*. https://library.ndsu.edu/ir/handle/10365/27949

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis, Third Edition*. CRC Press.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*(6), 721–741.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., Bornkamp, B., Maechler, M., Hothorn, T., & Hothorn, M. T. (2021). Package 'mvtnorm.' *J. Comput. Graph. Stat.*, *11*(1), 950–971.

George, E. I., & McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, *7*(2), 339–373.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. (1995). *Markov chain Monte Carlo in practice*. CRC press.

Glynn, R. J., Laird, N. M., & Rubin, D. B. (1986). Selection Modeling Versus Mixture Modeling with Nonignorable Nonresponse. In H. Wainer (Ed.), *Drawing Inferences from Self-Selected Samples* (pp. 115–142). Springer. https://doi.org/10.1007/978-1-4612-4976-4_10

Gomer, B., & Yuan, K.-H. (2021). Subtypes of the missing not at random missing data mechanism. *Psychological Methods*, *27*(5), 559–598. https://doi.org/10.1037/met0000377

Graham, J. W. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling*, *10*(1), 80–100.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*(1), 549–576.

Griffin, J., & Brown, P. (2017). Hierarchical shrinkage priors for regression models. *Bayesian Analysis*, *12*(1), 135–159.

Hans, C. (2009). Bayesian lasso regression. *Biometrika*, *96*(4), 835–845.

Hastings, W. K. (1970). *Monte Carlo sampling methods using Markov chains and their applications*. *57*(1), 97–109. https://doi.org/10.1093/biomet/57.1.97

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, *5*(4), 475–492.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the Econometric Society*, *47*(1), 153–161.

Hedeker, D., & Gibbons, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods*, *2*(1), 64.

Hesterberg, T., Choi, N. H., Meier, L., & Fraley, C. (2008). Least angle and ℓ1 penalized regression: A review. *Statistics Surveys*, *2*(1), 61–93. https://doi.org/10.1214/08-SS035

Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R., & Herring, A. H. (2005). Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, *100*(469), 332–346.

Ishwaran, H., & Rao, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics*, *33*(2), 730–773.

Jackman, S. (2000). Estimation and inference via Bayesian simulation: An introduction to Markov chain Monte Carlo. *American Journal of Political Science*, *44*(2), 375–404.

Johnson, V. E., & Albert, J. H. (2006). *Ordinal data modeling*. Springer Science & Business Media.

Keller, B. T., & Enders, C. K. (2021). Blimp user's guide (Version 3). *Blimp Software: Los Angeles, CA, USA*.

Kenward, M. G. (1998). Selection models for repeated measurements with non-random dropout: An illustration of sensitivity. *Statistics in Medicine*, *17*(23), 2723–2732.

Kruschke, J. K. (2011). Introduction to special section on Bayesian data analysis. *Perspectives on Psychological Science*, *6*(3), 272.

Lee, K. E., Sha, N., Dougherty, E. R., Vannucci, M., & Mallick, B. K. (2003). Gene selection: A Bayesian variable selection approach. *Bioinformatics*, *19*(1), 90–97.

Leung, S. F., & Yu, S. (2000). Collinearity and two-step estimation of sample selection models: Problems, origins, and remedies. *Computational Economics*, *15*(3), 173–199.

Levy, R., & Enders, C. K. (2021). Full conditional distributions for Bayesian multilevel models with additive or interactive effects and missing data on covariates. *Communications in Statistics-Simulation and Computation*, 1–225. https://doi.org/10.1080/03610918.2021.1921799

Li, Q., & Lin, N. (2010). The Bayesian elastic net. *Bayesian Analysis*, *5*(1), 151–170.

Little, R. J. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, *88*(421), 125–134.

Little, R. J. (2008). Selection and pattern-mixture models. In *Longitudinal data analysis* (pp. 423–446). Chapman and Hall/CRC.

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). John Wiley & Sons.

Little, R. J., & Wang, Y. (1996). Pattern-mixture models for multivariate incomplete data with covariates. *Biometrics*, 98–111.

Lüdtke, O., Robitzsch, A., & West, S. G. (2020). Regression models involving nonlinear effects with missing data: A sequential modeling approach using Bayesian estimation. *Psychological Methods*, *25*(2), 157.

Lynch, S. M. (2020). *Bayesian statistics*. SAGE Publications Limited.

Maity, A. K., Carroll, R. J., & Mallick, B. K. (2019). Integration of Survival and Binary Data for Variable Selection and Prediction: A Bayesian Approach. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *68*(5), 1577–1595. https://doi.org/10.1111/rssc.12377

Makalic, E., & Schmidt, D. F. (2016a). A Simple Sampler for the Horseshoe Estimator. *IEEE Signal Processing Letters*, *23*(1), 179–182. https://doi.org/10.1109/LSP.2015.2503725

Makalic, E., & Schmidt, D. F. (2016b). *High-Dimensional Bayesian Regularised Regression with the BayesReg Package* (arXiv:1611.06649). arXiv. https://doi.org/10.48550/arXiv.1611.06649

Marra, G., Radice, R., Bärnighausen, T., Wood, S. N., & McGovern, M. E. (2017). A simultaneous equation approach to estimating HIV prevalence with nonignorable missing responses. *Journal of the American Statistical Association*, *112*(518), 484–496.

McKelvey, R. D., & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology*, *4*(1), 103–120.

Mealli, F., & Rubin, D. B. (2016). Clarifying missing at random and related definitions, and implications when coupled with exchangeability. *Biometrika*, *103*(2), 491–491.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087–1092.

Michiels, B., Molenberghs, G., Bijnens, L., Vangeneugden, T., & Thijs, H. (2002). Selection models and pattern-mixture models to analyse longitudinal quality of life data subject to drop-out. *Statistics in Medicine*, *21*(8), 1023–1041. https://doi.org/10.1002/sim.1064

Michiels, B., Molenberghs, G., & Lipsitz, S. R. (1999). Selection models and pattern-mixture models for incomplete data with covariates. *Biometrics*, *55*(3), 978–983.

Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, *83*(404), 1023–1032.

Molenberghs, G., Thijs, H., Jansen, I., Beunckens, C., Kenward, M. G., Mallinckrodt, C., & Carroll, R. J. (2004). Analyzing incomplete longitudinal clinical trial data. *Biostatistics*, *5*(3), 445–464.

Muthén, B., Asparouhov, T., Hunter, A. M., & Leuchter, A. F. (2011). Growth modeling with nonignorable dropout: Alternative analyses of the STAR* D antidepressant trial. *Psychological Methods*, *16*(1), 17.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, *9*(4), 599–620.

Ogundimu, E. O. (2022). Regularization and variable selection in Heckman selection model. *Statistical Papers*, *63*(2), 421–439.

Ogundimu, E. O., & Collins, G. S. (2019). A robust imputation method for missing responses and covariates in sample selection models. *Statistical Methods in Medical Research*, *28*(1), 102–116.

O'Hara, R. B., & Sillanpää, M. J. (2009). A review of Bayesian variable selection methods: What, how and which. *Bayesian Analysis*, *4*(1), 85–117.

Park, T., & Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, *103*(482), 681–686.

Piironen, J., & Vehtari, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, *11*(2), 5018–5051.

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News*, *6*(1), 7–11.

Polson, N. G., & Scott, J. G. (2010). Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian Statistics*, *9*(501–538), 105.

Polson, N. G., & Scott, J. G. (2012). Local shrinkage rules, Lévy processes and regularized regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *74*(2), 287–311.

Puhani, P. (2000). The Heckman correction for sample selection and its critique. *Journal of Economic Surveys*, *14*(1), 53–68.

Ratitch, B., O'Kelly, M., & Tosiello, R. (2013). Missing data in clinical trials: From clinical assumptions to statistical analysis using pattern mixture models. *Pharmaceutical Statistics*, *12*(6), 337–347.

Ray, L. A., Green, R., Enders, C., Leventhal, A. M., Grodin, E. N., Li, G., Lim, A., Hartwell, E., Venegas, A., Meredith, L., Nieto, S. J., Shoptaw, S., Ho, D., & Miotto, K. (2021). Efficacy of Combining Varenicline and Naltrexone for Smoking Cessation and Drinking Reduction: A Randomized Clinical Trial. *The American Journal of Psychiatry*, *178*(9), 818. https://doi.org/10.1176/appi.ajp.2020.20070993

Rockova, V. (2013). *Bayesian Variable Selection in High-dimensional Applications*. https://repub.eur.nl/pub/51587/

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Russu, A., Malovini, A., Puca, A. A., & Bellazzi, R. (2012). Stochastic model search with binary outcomes for genome-wide association studies. *Journal of the American Medical Informatics Association*, *19*(e1), e13–e20. https://doi.org/10.1136/amiajnl-2011-000741

Sartori, A. E. (2003). An estimator for some binary-outcome selection models without exclusion restrictions. *Political Analysis*, *11*(2), 111–138.

Sterba, S. K., & Gottfredson, N. C. (2015). Diagnosing Global Case Influence on MAR Versus MNAR Model Comparisons. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(2), 294–307. https://doi.org/10.1080/10705511.2014.936082

Stolzenberg, R. M., & Relles, D. A. (1990). Theory testing in a world of constrained research design: The significance of Heckman's censored sampling bias correction for nonexperimental research. *Sociological Methods & Research*, *18*(4), 395–415.

Stolzenberg, R. M., & Relles, D. A. (1997). Tools for intuition about sample selection bias and its correction. *American Sociological Review*, *62*(3), 494–507. https://doi.org/10.2307/2657318

Terenin, A., Dong, S., & Draper, D. (2019). GPU-accelerated Gibbs sampling: A case study of the Horseshoe Probit model. *Statistics and Computing*, *29*(2), 301–310. https://doi.org/10.1007/s11222-018-9809-3

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Toomet, O., & Henningsen, A. (2008). Sample selection models in R: Package sampleSelection. *Journal of Statistical Software*, *27*(7), 1–23. https://doi.org/10.18637/jss.v027.i07

Van Erp, S., Oberski, D. L., & Mulder, J. (2019). Shrinkage priors for Bayesian penalized regression. *Journal of Mathematical Psychology*, *89*(1), 31–50. https://doi.org/10.1016/j.jmp.2018.12.004

Vella, F. (1998). Estimating models with sample selection bias: A survey. *Journal of Human Resources*, *33*(1), 127–169.

Wu, T. T., & Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, *2*(1), 224–244.

Yang, A., Jiang, X., Shu, L., & Liu, P. (2019). Sparse bayesian kernel multinomial probit regression model for high-dimensional data classification. *Communications in Statistics - Theory and Methods*, *48*(1), 165–176. https://doi.org/10.1080/03610926.2018.1463385

Zhang, Q., & Wang, L. (2017). Moderation analysis with missing data in the predictors. *Psychological Methods*, *22*(4), 649.