



Bayesian Variable Selection Under Sample Selection and Model Misspecification

Adam Iqbal^{*}, Emmanuel O. Ogundimu^{*}, and F. Javier Rubio[†]

Abstract. Sample selection bias arises when missingness in the outcome of interest correlates with the outcome itself, leading to non-randomly selected samples. A common approach to correct bias from sample selection is to use *sample selection models* that jointly model the selection mechanism and the outcome of interest. Formulating these models typically rely on exclusion restrictions (variables that are predictors of selection but not appearing in the outcome equation) to ensure identifiability of the parameters. However, the choice of exclusion restrictions often depends on heuristics or expert judgment, potentially leading to the inclusion of irrelevant variables or the omission of important ones. Additionally, distributional misspecification and omitted variable bias are frequent challenges in this framework. To formally address these issues, we propose a Bayesian variable selection (BVS) methodology that incorporates both local priors (LPs) and non-local priors (NLPs), enabling the identification of variables with predictive power for the outcome and selection processes. We develop computational tools to conduct BVS in sample selection models based on a Laplace approximation of the marginal likelihood, and characterize the resulting Bayes factor rates under model misspecification. We establish model selection consistency for both classes of priors, showing that the proposed methodology correctly identifies active variables for both the selection process and outcome process asymptotically. The priors are calibrated to account for the possibility of distributional misspecification and omitted variable bias. We present a simulation study and real-data applications to explore the finite-sample effects of model misspecification on BVS. We compare the performance of the proposed methodology against BVS based on spike-and-slab (SS) priors and the Adaptive LASSO (ALASSO), an adaptive weighting of the least absolute shrinkage and selection operator (LASSO).

Keywords: Laplace approximation, local priors, non-local priors, sample selection models.

1 Introduction

Missing data poses critical challenges across scientific disciplines, as the manner in which data become missing can strongly influence statistical inferences (Molenberghs et al., 2014). In particular, missingness not at random (MNAR) arises when the probability of an observation being missing depends on the unobserved value itself (possibly in combination with observed covariates). Selection (or sample selection) models (Heckman, 1976, 1979; Diggle and Kenward, 1994) constitute a widely used framework for handling

^{*}Department of Mathematical Sciences, University of Durham, adam.iqbal@durham.ac.uk; emmanuel.ogundimu@durham.ac.uk

[†]Department of Statistical Science, University College London, f.j.rubio@ucl.ac.uk

data that are missing not at random: they decompose the joint distribution of the outcome and the indicator of missingness (or “selection”) into (i) the marginal distribution of the outcome (the “outcome equation”) and (ii) the conditional distribution of the selection process given the outcome (the “selection equation”). This two-equation perspective provides a principled way to analyze situations where data might be systematically missing based on unobserved values. A classical example is the Heckman selection model, which typically assumes that the error terms in the outcome and selection equations follow a bivariate normal distribution. Correctly specifying the covariates in each equation remains a central challenge in applied settings. Exclusion restrictions, that is, covariates that are allowed to influence the selection equation but are constrained to have zero (direct) effect in the outcome equation, are often imposed for practical identifiability. While model identifiability can, in principle, be ensured through appropriate distributional assumptions alone (Leung and Yu, 2000; Miao et al., 2016), in practice, omitting relevant covariates can lead to substantial bias (Certo et al., 2016; Puhani, 2000). Conversely, including extraneous covariates solely to achieve identification may introduce overfitting, multicollinearity, and unstable estimates (Sartori, 2003; Genbäck et al., 2015). Although recent work has proposed methods to relax exclusion restrictions (Honoré and Hu, 2020), broadly applicable solutions remain elusive.

The assumption of bivariate normality in sample selection models is very restrictive, and violation of this assumption leads to a type of misspecification known as distributional misspecification. Robustness to lack of normality, especially when sample selection is not severe, is one of the motivations behind the two-step estimator (Heckman, 1979), and due to the commonality of bivariate normality being violated, there have also been many formulations of sample selection models with other choices of error distributions. For instance, Marchenko and Genton (2012) extended sample selection models to cases where the errors follow a bivariate t -distribution, and Ogundimu and Hutton (2016) considered the case where the errors follow a bivariate skew-normal distribution, which accounts for departures from symmetry. Van-Hasselt (2011) proposed a fully Bayesian approach which allows for mixture-of-Gaussian distributions as error distributions. More recently, de Souza Bastos et al. (2022) proposed modeling dispersion and correlation parameters in the Heckman selection model using appropriately transformed linear predictors. This approach allows selection and dispersion to vary across covariate values but requires specifying which covariates enter each linear predictor.

To circumvent the need for exclusion restrictions or *ad hoc* rules for specifying covariates in each equation, formal variable selection methods are increasingly being employed in sample selection models. Ogundimu (2022) formulates an Adaptive LASSO penalty to select the variables that enter both the selection and outcome models. Iqbal et al. (2023) extended the Gibbs sampler of Van-Hasselt (2011) to use spike-and-slab priors on the parameters, allowing for a fast Bayesian approach to variable selection in sample selection models. These methods assume bivariate normality and do not address distributional misspecification, nor the case when a variable is omitted from the procedure entirely. Wiemann et al. (2022) introduced a flexible distributional regression scheme for sample selection models, modeling the relationship between the two equations as a copula while allowing for Bayesian variable selection priors as a special case. However, their

implementation relies on a computationally intensive Metropolis-within-Gibbs sampler and requires careful tuning of the priors.

We develop methodology for Bayesian variable selection in sample selection models using local and non-local priors (Johnson and Rossell, 2010) and explore the effects of different types of model misspecification. Our approach uses Laplace approximations to compute marginal likelihoods, and uses an accept-reject scheme to perform Gibbs sampling over the space of possible models that is based on previous proposals (Johnson and Rossell, 2012). Unlike the spike-and-slab approach proposed by Iqbal et al. (2023), we derive asymptotic results demonstrating model selection consistency and characterizing the rates of convergence of Bayes factors under both local and non-local priors. We conduct a simulation study which highlights the empirical performance of non-local priors under model misspecification.

The rest of the paper is structured as follows. Section 2 briefly introduces sample selection models. Section 3 defines the notation for Bayesian variable selection and introduces the local and non-local priors we use throughout the paper. Section 4 presents the Laplace approximation and the Gibbs sampler we use, while Section 5 covers theoretical results for our approach. The theoretical results in Section 5 emphasize that neglecting sample selection introduces a type of model misspecification that may affect both sensitivity and specificity. Section 6 compares the performance of local and non-local priors against appropriate competitors in a simulation study. Section 7 evaluates the performance of the proposed BVS methodology using well-studied Ambulatory Expenditures dataset and data from the RAND Health Insurance Experiment (RAND HIE), which are each suspected to have sample selection bias and exhibit departures from normality (Marchenko and Genton, 2012; de Souza Bastos et al., 2022), alongside a sensitivity analysis for the non-local prior calibration. Section 8 concludes with a discussion and possible extensions. The supplementary material contains expressions for the gradient and Hessian of the log-likelihood, an explicit algorithm for the sampling scheme in Section 4.2, proofs of the results in this paper, three additional simulation scenarios briefly described in Section 6.5 and further details on the sensitivity analysis. The code and data used in this study can be found at <https://github.com/adam-iqbal/bvsss-nlp>.

2 Sample selection models

Consider the setting where the outcome of interest, denoted y_i^* , is assumed to be linearly related to the covariates \mathbf{x}_i through a multiple regression model

$$y_i^* = \beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_{1,i}, \quad (1)$$

where $\mathbf{x}_i = (x_{i,1}, \dots, x_{i,p})^\top \in \mathbb{R}^p$, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top \in \mathbb{R}^p$ are the corresponding regression coefficients, β_0 is the intercept, and $i = 1, \dots, n$. Missing data may arise in this context through missingness in the outcome of interest or in the covariates. The data is said to be missing completely at random (MCAR) if the missingness mechanism is independent of both observed and unobserved data, while it is said to be missing at random (MAR) if the missingness mechanism does not depend on unobserved data (but may depend on observed data). If neither of these hold, the data is said to be missing

not at random (MNAR). Sample selection models with continuous outcomes (Heckman, 1976, 1979) are a case of MNAR outcomes such that the missing outcomes are directly correlated with their unobserved values. In this case, the multiple regression model (1) is supplemented by a selection (missingness) process given by

$$s_i^* = \alpha_0 + \mathbf{w}_i^\top \boldsymbol{\alpha} + \epsilon_{2,i}, \quad (2)$$

where $\mathbf{w}_i = (w_{i,1}, \dots, w_{i,q})^\top \in \mathbb{R}^q$, and $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_q)^\top \in \mathbb{R}^q$ are the regression coefficients (α_0 is the intercept). In the classical sample selection model, the errors are assumed to be distributed according to a bivariate normal distribution

$$\begin{pmatrix} \epsilon_{1,i} \\ \epsilon_{2,i} \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \sigma\rho \\ \sigma\rho & 1 \end{pmatrix} \right), \quad (3)$$

where $\sigma > 0$ and $\rho \in (-1, 1)$. We set the variance of the selection equation to 1 as this is the standard assumption for probit regression. The observed data consist of $\{(y_i, s_i) : i = 1, \dots, n\}$, where

$$s_i = 1\{s_i^* > 0\} \quad \text{and} \quad y_i = \begin{cases} y_i^* & \text{if } s_i = 1, \\ \text{not observed} & \text{if } s_i = 0. \end{cases}$$

Previous literature has shown that, under some regularity conditions and correct model specification, the maximum likelihood estimators of this model are consistent and asymptotically normal (de Souza Bastos et al., 2022).

The density of the sample-selection model is composed of a continuous component corresponding to the conditional density $f(y | s = 1)$ and a discrete component $\Pr(s | w)$, which is a probit model. Thus, the conditional density of the observed data is given by

$$f(y_i | s_i = 1) = \frac{1}{\sigma} \phi \left(\frac{y_i - (\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta})}{\sigma} \right) \frac{\Phi \left\{ \frac{\rho}{\sqrt{1-\rho^2}} \left(\frac{y_i - (\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta})}{\sigma} \right) + \frac{(\alpha_0 + \mathbf{w}_i^\top \boldsymbol{\alpha})}{\sqrt{1-\rho^2}} \right\}}{\Phi(\alpha_0 + \mathbf{w}_i^\top \boldsymbol{\alpha})}, \quad (4)$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the standard normal density and distribution function, respectively. Equation (4) follows from the fact that

$$\Pr(s_i = 1 | y_i, \mathbf{x}_i, \mathbf{w}_i) = \Phi \left(\alpha_0 + \mathbf{w}_i^\top \boldsymbol{\alpha} + \rho \sigma^{-1} \frac{(y_i - (\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}))}{(1 - \rho^2)^{1/2}} \right).$$

We use the reparametrization $\tau = 1/\sigma$, $\boldsymbol{\theta} = \boldsymbol{\beta}/\sigma$, and $\theta_0 = \beta_0/\sigma$, and denote $\boldsymbol{\eta} = (\theta_0, \alpha_0, \boldsymbol{\theta}^\top, \boldsymbol{\alpha}^\top, \tau, \rho)^\top$. Under this formulation, the log-likelihood of the parameters for this model is given by

$$\begin{aligned} \ell(\boldsymbol{\eta}) &= \sum_{\{i:s_i=0\}} \log \Phi(-\alpha_0 - \mathbf{w}_i^\top \boldsymbol{\alpha}) \\ &+ \sum_{\{i:s_i=1\}} \log \Phi \left(\frac{\alpha_0 + \mathbf{w}_i^\top \boldsymbol{\alpha} + \rho(\tau y_i - \theta_0 - \mathbf{x}_i^\top \boldsymbol{\theta})}{\sqrt{1 - \rho^2}} \right) \end{aligned}$$

$$- \sum_{\{i:s_i=1\}} \left[\frac{1}{2} \log(2\pi) - \log(\tau) + \frac{1}{2} (\tau y_i - \theta_0 - \mathbf{x}_i^\top \boldsymbol{\theta})^2 \right].$$

This new parametrization allows for easily identifying that the log-likelihood function is concave with respect to $(\theta_0, \alpha_0, \boldsymbol{\theta}^\top, \boldsymbol{\alpha}^\top, \tau)^\top$ for a fixed value of ρ , a property that has been reported in previous literature (Olsen, 1982). In general, the log-likelihood is not globally concave (Olsen, 1982), which can lead to difficulties when finding maximum likelihood estimators for certain data sets, such as convergence to local maxima, or cases where a global maximum does not exist.

3 Bayesian variable selection

3.1 Formulation

Define $\boldsymbol{\omega} = (\boldsymbol{\theta}^\top, \boldsymbol{\alpha}^\top)^\top \in \mathbb{R}^{p+q}$, $\mathbf{Z} = (\mathbf{X}, \mathbf{W})$ the full design matrix, and $d = p + q$ the total model size. We formulate the variable selection problem as selecting the indicators

$$\gamma_j = \begin{cases} 0 & \text{if } \omega_j = 0, \\ 1 & \text{if } \omega_j \neq 0, \end{cases} \quad (5)$$

which determine which variables enter into the model, with $j = 1, \dots, d$. Define $\boldsymbol{\gamma}^O$ and $\boldsymbol{\gamma}^S$ as the indicator variables that determine which covariates from \mathbf{X} and \mathbf{W} , respectively, enter the model. The aim is to select $\boldsymbol{\gamma} = ((\boldsymbol{\gamma}^O)^\top, (\boldsymbol{\gamma}^S)^\top)^\top$. The variable selection is unrestricted, so that a variable can be entered in either or both equations without restriction.

Further define $\boldsymbol{\theta}_{\boldsymbol{\gamma}^O}$ and $\boldsymbol{\alpha}_{\boldsymbol{\gamma}^S}$ as the subvectors of $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ only containing variables active in $\boldsymbol{\gamma}^O$ and $\boldsymbol{\gamma}^S$ respectively, and define $\boldsymbol{\eta}_{\boldsymbol{\gamma}} = (\theta_0, \alpha_0, \boldsymbol{\theta}_{\boldsymbol{\gamma}^O}^\top, \boldsymbol{\alpha}_{\boldsymbol{\gamma}^S}^\top, \tau, \rho)^\top$. Given a prior $\pi(\boldsymbol{\gamma})$ on the model space, we can obtain the posterior probability of each model as

$$\pi(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{s}) = \frac{p(\mathbf{y}, \mathbf{s} \mid \boldsymbol{\gamma}) \pi(\boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma}'} p(\mathbf{y}, \mathbf{s} \mid \boldsymbol{\gamma}') \pi(\boldsymbol{\gamma}')} = \left(1 + \sum_{\boldsymbol{\gamma}' \neq \boldsymbol{\gamma}} B_{\boldsymbol{\gamma}', \boldsymbol{\gamma}} \frac{\pi(\boldsymbol{\gamma}')}{\pi(\boldsymbol{\gamma})} \right)^{-1}, \quad (6)$$

where $B_{\boldsymbol{\gamma}', \boldsymbol{\gamma}} = p(\mathbf{y}, \mathbf{s} \mid \boldsymbol{\gamma}') / p(\mathbf{y}, \mathbf{s} \mid \boldsymbol{\gamma})$ denotes the Bayes factor between two models, and $p(\mathbf{y}, \mathbf{s} \mid \boldsymbol{\gamma})$ denotes the likelihood of the data with respect to a given model, marginalized over the parameter prior $\pi(\boldsymbol{\eta}_{\boldsymbol{\gamma}} \mid \boldsymbol{\gamma})$,

$$p(\mathbf{y}, \mathbf{s} \mid \boldsymbol{\gamma}) = \int p(\mathbf{y}, \mathbf{s} \mid \boldsymbol{\eta}_{\boldsymbol{\gamma}}) \pi(\boldsymbol{\eta}_{\boldsymbol{\gamma}} \mid \boldsymbol{\gamma}) d\boldsymbol{\eta}_{\boldsymbol{\gamma}}. \quad (7)$$

The term $p(\mathbf{y}, \mathbf{s} \mid \boldsymbol{\gamma})$ is also called the marginal likelihood. With the posterior probabilities $\pi(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{s})$, one may choose the model attaining the largest value of $\pi(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{s})$ or the model consisting of variables with high marginal posterior probabilities $\pi(\gamma_j \neq 0 \mid \mathbf{y}, \mathbf{s})$. When the goal is prediction, one option is to use Bayesian model averaging, weighting models according to $\pi(\boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{s})$; alternatively, one can select a sparse model that provides similar predictive performance (Hahn and Carvalho, 2015; Forte et al., 2018; Tadesse and Vannucci, 2021).

3.2 Priors on the model parameters

Since the parameter prior $\pi(\boldsymbol{\eta}_\gamma \mid \gamma)$ appears in both the numerator and denominator of (6), its choice has a persistent effect on variable selection even asymptotically. Therefore, selecting an appropriate parameter prior is crucial in Bayesian variable selection. We adopt a product prior structure $\pi_k(\boldsymbol{\eta}_\gamma) = \pi(\theta_0)\pi(\alpha_0)\pi_k(\boldsymbol{\theta}_{\gamma^O})\pi_k(\boldsymbol{\alpha}_{\gamma^S})\pi(\tau)\pi(\rho)$, where $k = L, M$. For the parameter priors $\pi_k(\boldsymbol{\theta}_{\gamma^O})$ and $\pi_k(\boldsymbol{\alpha}_{\gamma^S})$, we present two concrete options. The first ($k = L$) represents the (local) g -Prior (Zellner, 1986; Liang et al., 2008), and the second ($k = M$) represents the non-local MOM prior (Johnson and Rossell, 2012; Rossell et al., 2013), adapted to our context. Specifically,

$$\begin{aligned}\pi_L(\boldsymbol{\theta}_{\gamma^O}) &= N\left(\boldsymbol{\theta}_{\gamma^O}; 0, g_L^O n_O \left(\mathbf{X}_{\gamma^O}^\top \mathbf{X}_{\gamma^O}\right)^{-1}\right), \\ \pi_L(\boldsymbol{\alpha}_{\gamma^S}) &= N\left(\boldsymbol{\alpha}_{\gamma^S}; 0, g_L^S n \left(\mathbf{W}_{\gamma^S}^\top \mathbf{W}_{\gamma^S}\right)^{-1}\right), \\ \pi_M(\boldsymbol{\theta}_{\gamma^O}) &= \prod_{\gamma_j^O=1} \frac{\theta_j^2}{g_M^O} N(\theta_j; 0, g_M^O), \\ \pi_M(\boldsymbol{\alpha}_{\gamma^S}) &= \prod_{\gamma_j^S=1} \frac{\alpha_j^2}{g_M^S} N(\alpha_j; 0, g_M^S),\end{aligned}$$

where $g_L^O, g_L^S, g_M^O, g_M^S \in \mathbb{R}_+$ are given dispersion parameters, for which we propose default values in Section 3.3 and $n_O = \sum_{i=1}^n s_i$. For $\pi_L(\boldsymbol{\theta}_{\gamma^O})$, we use n_O instead of n as only non-missing outcomes are informative about the outcome equation. For the model prior, we define a Beta-Binomial prior as in Scott and Berger (2010). That is,

$$\pi(\gamma) = \frac{\text{Beta}(d_\gamma + a_2, d - d_\gamma + b_2)}{\text{Beta}(a_2, b_2)},$$

with $d = p + q$ the total number of variables and $d_\gamma = \sum_{j=1}^{p+q} \gamma_j$ the number of active variables in γ .

We adopt commonly used priors for the remaining parameters, which appear in all models. For $\pi(\tau)$ we consider a Gamma prior distribution with parameters (a_0, b_0) . For $\frac{\rho+1}{2}$, we assign a Beta(a_1, b_1) prior. We additionally assign $\pi(\theta_0) = N(\theta_0; 0, v_\theta^2)$ and $\pi(\alpha_0) = N(\alpha_0; 0, v_\alpha^2)$, normal distributions with large variances.

The prior distribution encodes beliefs about sparsity and penalizes model complexity, directly influencing model selection (Tadesse and Vannucci, 2021). Local priors penalize model complexity, however, they do not strictly penalize inclusion of variables with small, potentially spurious effects. This can result in models that retain noise predictors, especially in high-dimensional settings. Non-local priors (Johnson and Rossell, 2010, 2012) offer a sharper separation between signal and noise, helping to ensure that only covariates with sufficient evidence are included and reducing false positives. Conceptually, the difference between local and non-local priors is that even under an alternative hypothesis $\gamma_j \neq 0$ the local prior assigns non-negligible prior mass around

the null hypothesis $\gamma_j = 0$. The g -prior π_L is a normal distribution with mean zero, so that a non-negligible proportion of the prior mass is concentrated around the null hypothesis. Non-local priors, on the other hand, assign very little mass around the null, with π_M approaching zero when any component of θ_{γ^O} or α_{γ^S} approaches zero. Thus the Bayes factors between the true model and overfitted models converges faster for non-local priors than they do for local priors. We formalize this intuition in Section 5.

3.3 Prior elicitation

For an effect size considered small, say m^O , and some small number ϵ , we choose g_M^O such that $\mathbb{P}(|\theta_j| < m^O) = \epsilon$. That is, the probability of an effect size being smaller than m^O *a priori* is equal to ϵ . For g_M^S , the interpretation changes slightly because the observed s_i is binary. For a given small probability m^S , an upper bound on the change in probability resulting from a unit change in a coefficient of magnitude $\Phi^{-1}(0.5 + m^S)$ is m^S , where Φ^{-1} is the inverse of the standard normal cumulative density function. So we choose g_M^S such that, *a priori*, $\mathbb{P}(|\alpha_j| < \Phi^{-1}(0.5 + m^S)) = \epsilon$, where m^S is considered a small change in probability.

As default choices, we suggest a “small effect size” of $m^O = 0.1$ for the outcome equation, and for the selection equation, we say that a change in probability of 0.05 is small, so $m^S = 0.05$. We additionally recommend $\epsilon = 0.05$, which corresponds to $g_M^O \approx 0.0284$ and $g_M^S \approx 0.0449$. This choice is calibrated with potential model misspecification mind. Empirically, we found this choice to perform reasonably well under correct model specification, while losing less power under model misspecification, as we show in the simulation study in Section 6. To illustrate how the choice of prior parameters affects the resulting sample, we perform a sensitivity analysis in Section 7.3, first comparing to a small perturbation by setting $\epsilon = 0.04$, and then to a choice of parameters that induce significantly more sparsity by setting $\epsilon = 0.02$. We see that enforcing significantly more sparsity can change the median model and inference, but small perturbations do not have a significant effect on the variable selection. For local priors, we suggest $g_L^S = g_L^O = 1$ as in the unit information prior. See (Liang et al., 2008; Li and Clyde, 2018) for further discussion on prior elicitation for local priors. It should be emphasized that while the choice above is a reasonable default, Bayesian variable selection can be sensitive to the choice of prior parameters, and there is a compromise between including small effects and excluding spurious effects. The hyperparameters should be chosen according to the nature of the problem and the effect sizes of interest. It should also be noted that the elicitation depends on the scale of the predictors, and our elicitation above assumes the predictors have been standardized. For the remaining hyperparameters, we use standard elicitations. $a_0 = b_0 = 1$ for the gamma prior on τ , $a_1 = b_1 = 1/2$ for the Beta prior on ρ , and $a_2 = b_2 = 1$ for the model prior (so the induced prior on model size is uniform). Empirically we found that performance is not sensitive to the prior parameters of τ .

4 Computations

4.1 Laplace approximation

Obtaining the posterior probabilities requires computing the marginal likelihoods, as shown in (7), but due to the given form of the likelihood, this integral does not have a closed-form solution. We instead propose approximating it using the Laplace approximation

$$\hat{p}(\mathbf{y}, \mathbf{s} \mid \gamma) = \exp\{\ell(\tilde{\boldsymbol{\eta}}_\gamma) + \log \pi(\tilde{\boldsymbol{\eta}}_\gamma)\} (2\pi)^{d_\gamma/2} |H(\tilde{\boldsymbol{\eta}}_\gamma) + \nabla^2 \log \pi(\tilde{\boldsymbol{\eta}}_\gamma)|^{-1/2},$$

where $\tilde{\boldsymbol{\eta}}_\gamma = \arg \max_{\boldsymbol{\eta}_\gamma} \{\ell(\boldsymbol{\eta}_\gamma) + \log \pi(\boldsymbol{\eta}_\gamma)\}$ are the maximum a posteriori (MAP) parameters under prior $\pi(\boldsymbol{\eta}_\gamma)$, and H is the Hessian matrix of the log-likelihood function. Thus, the Bayes factor is approximated as

$$\hat{B}_{\gamma, \gamma^*} = \frac{\hat{p}(\mathbf{y}, \mathbf{s} \mid \gamma)}{\hat{p}(\mathbf{y}, \mathbf{s} \mid \gamma^*)}, \quad (8)$$

and we obtain the approximate posterior $\hat{\pi}(\gamma \mid \mathbf{y}, \mathbf{s}) = \left(1 + \sum_{\gamma' \neq \gamma} \hat{B}_{\gamma', \gamma} \frac{\pi(\gamma')}{\pi(\gamma)}\right)^{-1}$.

In principle, any optimization method could be used to obtain $\tilde{\boldsymbol{\eta}}_\gamma$. Since we have access to the gradient and Hessian of the log-likelihood analytically, we use Newton-Raphson to find $\tilde{\boldsymbol{\eta}}_\gamma$. This is fast for cases where the number of active variables is small or moderate, but in cases with a large number of active variables a first order gradient descent or co-ordinate descent method could be used to avoid expensive matrix inversions.

Due to the lack of global log-concavity, the Laplace approximation may be less accurate. However, this deviation from log-concavity affects only the parameter ρ , and empirically, the approximation remains sufficiently accurate for our purposes. Theoretical studies have been recently done to provide bounds on the error of the Laplace approximation when the likelihood is not log-concave (Kasprzak et al., 2025). In the simulation study presented in Section 6, we demonstrate that this approximation performs well in our context.

4.2 Model space exploration

If there are $d = p + q$ variables in total, the model space has cardinality 2^d , making it infeasible to compute the posterior model probabilities $\hat{\pi}(\gamma \mid \mathbf{y}, \mathbf{s})$ for all γ . Consequently, we formulate a Gibbs sampler (George and McCulloch, 1993; Johnson and Rossell, 2012) to perform posterior influence and sample models from regions of the space with high posterior probability. At each step, we consider a single change to the current model - adding (or removing) a single variable. The acceptance probability for a proposed model γ_{prop} versus the current model γ_{curr} is $\frac{\hat{\pi}(\gamma_{prop} \mid \mathbf{y}, \mathbf{s})}{\hat{\pi}(\gamma_{prop} \mid \mathbf{y}, \mathbf{s}) + \hat{\pi}(\gamma_{curr} \mid \mathbf{y}, \mathbf{s})}$. We iterate this over every variable in both equations, and after d accept-reject steps, the resulting model is the sample γ_n , so that for a sample of size N there are $d \times N$ accept-reject steps. The procedure is also detailed in Section 2 of the Appendix (Iqbal et al.,

2025). To reduce computational time, we store the marginal likelihoods $\hat{p}(\mathbf{y}, \mathbf{s} \mid \boldsymbol{\gamma})$ after computing them for each visited model. These stored values are then re-used if the same model is proposed or sampled again. To provide a warm start to the optimization routine at each step, we propose two sets of starting parameters:

1. The MAP parameters from the most recently accepted model, $\tilde{\boldsymbol{\eta}}_{\boldsymbol{\gamma}_{curr}}$, with some variables removed or added (and newly added ones set to small values away from zero) to match $\boldsymbol{\gamma}_{prop}$;
2. The parameters are obtained using the two-step estimator described in Heckman (1979), where the maximum likelihood estimates for the probit component are computed via the Newton–Raphson algorithm.

We compare the log-posterior at both sets of parameters and choose the parameters that attain larger log-posterior. Empirically, we have found that the strategy outlined above for a warm start works well enough for our purposes.

5 Theoretical results

In this section, we present the asymptotic Bayes factor rates for local and non-local priors using Laplace-approximated marginal likelihoods, as well as results on model selection consistency. We assume the existence of a “true generating mechanism” or “true model”. This true model may coincide with the sample selection model under consideration (correct specification) or it may differ from it (misspecification). For example, the true model may involve additional covariates, alternative functional forms for the covariates, different distributional assumptions, or an entirely different regression structure. In this line, define $\mathbf{u} = (y, s, \mathbf{x}^\top, \mathbf{w}^\top)^\top$ as the random variable consisting of a single observation, and F_0 its underlying distribution. Since we allow for misspecification, the marginal distribution of (y, s) may not be as in Section 2. The log-likelihood of a single observation \mathbf{u} under model specification (3) with parameters $\boldsymbol{\eta}_\gamma$ is

$$\begin{aligned} m(\boldsymbol{\eta}_\gamma; \mathbf{u}) &= (1-s) \log \Phi(-\alpha_0 - \mathbf{w}^\top \boldsymbol{\alpha}_{\gamma s}) \\ &+ s \log \Phi\left(\frac{\alpha_0 + \mathbf{w}^\top \boldsymbol{\alpha}_{\gamma s} + \rho(\tau y - \theta_0 - \mathbf{x}^\top \boldsymbol{\theta}_{\gamma o})}{\sqrt{1-\rho^2}}\right) \\ &- s \left[\frac{1}{2} \log(2\pi) - \log(\tau) + \frac{1}{2} (\tau y - \theta_0 - \mathbf{x}^\top \boldsymbol{\theta}_{\gamma o})^2 \right]. \end{aligned}$$

Define now the expected log-likelihood $M(\boldsymbol{\eta}_\gamma) = E_{F_0}[m(\boldsymbol{\eta}_\gamma)]$, where the expectation is taken with respect to the joint distribution of \mathbf{u} , F_0 . This is,

$$\begin{aligned} M(\boldsymbol{\eta}_\gamma) &= P(s=0)E_{F_0}[\log \Phi(-\alpha_0 - \mathbf{w}^\top \boldsymbol{\alpha}_{\gamma s}) \mid s=0] \\ &+ P(s=1)E_{F_0}\left[\log \Phi\left(\frac{\alpha_0 + \mathbf{w}^\top \boldsymbol{\alpha}_{\gamma s} + \rho(\tau y - \theta_0 - \mathbf{x}^\top \boldsymbol{\theta}_{\gamma o})}{\sqrt{1-\rho^2}}\right) \mid s=1\right] \end{aligned}$$

$$- P(s=1) \left[\frac{1}{2} \log(2\pi) - \log(\tau) + E_{F_0} \left[\frac{1}{2} (\tau y - \theta_0 - \mathbf{x}^\top \boldsymbol{\theta}_{\gamma^o})^2 \mid s=1 \right] \right]. \quad (9)$$

Let us define $\boldsymbol{\eta}_\gamma^* = \arg \max_{\boldsymbol{\eta}_\gamma} M(\boldsymbol{\eta}_\gamma)$. This value can be interpreted as the choice of parameters that minimize the Kullback-Leibler divergence between the true data-generating distribution and the fitted sample selection model under γ . If the model is correctly specified and the true model is γ , then $\boldsymbol{\eta}_\gamma^*$ will be the true parameter values. We assume $\boldsymbol{\eta}_\gamma^*$ is unique, and denote by γ^* the most parsimonious¹ model such that $M(\boldsymbol{\eta}_{\gamma^*}^*)$ is maximized among all parameter values.

We make the following technical assumptions:

- C1.** The parameter space is a compact subset $\Delta_\gamma \subset \Gamma_\gamma = \mathbb{R}^{p+q+2} \times \mathbb{R}_+ \times (0, 1)$.
- C2.** There exists n_0 such that, for $n > n_0$, the matrices $\mathbf{X}_{\gamma^o}^\top \mathbf{X}_{\gamma^o}$ and $\mathbf{W}_{\gamma^s}^\top \mathbf{W}_{\gamma^s}$ are positive definite almost surely.
- C3.** The following conditions are satisfied.

- (a) $M(\boldsymbol{\eta}_\gamma) < \infty$ for all $\boldsymbol{\eta}_\gamma$, and $|m(\boldsymbol{\eta}_\gamma; \mathbf{u})| \leq \psi_0(\mathbf{u})$, for all $\boldsymbol{\eta}_\gamma \in \Delta_\gamma$, where

$$\int \psi_0(\mathbf{u}) dF_0(\mathbf{u}) < \infty.$$

- (b) The maximum $\boldsymbol{\eta}_\gamma^* = \arg \max_{\Delta_\gamma} M(\boldsymbol{\eta}_\gamma)$ is a unique interior point of Δ_γ .

- C4.** There exist functions $\psi_1(\mathbf{u})$ and $\psi_2(\mathbf{u})$ such that

$$\left| \frac{\partial m(\boldsymbol{\eta}_\gamma; \mathbf{u})}{\partial \boldsymbol{\eta}_{\gamma_i}} \cdot \frac{\partial m(\boldsymbol{\eta}_\gamma; \mathbf{u})}{\partial \boldsymbol{\eta}_{\gamma_j}} \right| \leq \psi_1(\mathbf{u}),$$

$$\left| \frac{\partial^2 m(\boldsymbol{\eta}_\gamma; \mathbf{u})}{\partial \boldsymbol{\eta}_{\gamma_i} \partial \boldsymbol{\eta}_{\gamma_j}} \right| \leq \psi_2(\mathbf{u}),$$

where

$$\int \psi_1(\mathbf{u}) dF_0(\mathbf{u}) < \infty,$$

$$\int \psi_2(\mathbf{u}) dF_0(\mathbf{u}) < \infty.$$

- C5.** The matrix $E_{F_0} [\nabla m(\boldsymbol{\eta}_\gamma^*; \mathbf{u}) \nabla m(\boldsymbol{\eta}_\gamma^*; \mathbf{u})^\top]$ is non-singular and the Hessian matrix $\nabla^2 M(\boldsymbol{\eta}_\gamma)$ has constant rank in a neighbourhood of $\boldsymbol{\eta}_\gamma^*$.

¹Most parsimonious in the sense that there are no variables such that $\eta_{\gamma_j^*}^* = 0$, where γ_j^* refers exclusively to those entries of $\boldsymbol{\eta}$ associated with regression effects, *i.e.*, the components of $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$, and not to the intercepts or additional parameters. That is, there are no spurious variables in γ^* . Adding such variables does not change the value $M(\boldsymbol{\eta}_\gamma^*)$ so that the KL divergence would still be minimized, hence we must specify this condition explicitly.

The assumptions are similar to those by White (1982), adapted to our context, which are used to establish consistency and asymptotic normality of maximum likelihood estimation under model misspecification. The same assumptions also prove sufficient for establishing the rates of convergence for the Bayes factor and model selection consistency. Assumption C1 is primarily made for theoretical convenience but, as noted by White (1981), it could feasibly be relaxed to allow for locally compact spaces. Nonetheless, compactness is sufficiently general for our purposes. For all practical applications, the parameters can be assumed to lie within a suitably large compact set, potentially as large as machine precision allows (and for ρ , bounded away from ± 1 by machine precision). C2 is an assumption on the distributions of \mathbf{x} and \mathbf{w} , ensuring the absence of collinearity in sufficiently large samples. C3(a) and C4 are moment conditions on the true data-generating process. Since this process involves the joint distribution of $(y, s, \mathbf{x}^\top, \mathbf{w}^\top)^\top$, these assumptions apply to both the response variables y and s , as well as implicitly to the covariates \mathbf{x} and \mathbf{w} . Assumptions C3(b) and C5 ensure the existence of a unique maximum and guarantee that the Hessian matrix of the second-order expansion around any point sufficiently close to this maximum is positive definite.

Our main result establishes the asymptotic Bayes factor rates using Laplace-approximated marginal likelihoods for both local and non-local priors, thereby characterizing the methodology introduced in this paper. Model selection consistency follows as a corollary.

Proposition 1. *Let $\hat{B}_{\gamma, \gamma^*}$ be the Bayes factor in (8) under either π_L or π_M . Assume that C1-C5 hold. Suppose that $(g_L^O, g_L^S, g_M^O, g_M^S)$ are non-decreasing in n .*

(i) *Overfitted models. If $\gamma^* \subset \gamma$, then*

$$\log \hat{B}_{\gamma, \gamma^*} = \frac{d_{\gamma^*} - d_\gamma}{2} \log(n) + \frac{q_{\gamma^*} - q_\gamma}{2} \log(a_n) + \frac{p_{\gamma^*} - p_\gamma}{2} \log(b_n) + \mathcal{O}_p(1),$$

where for the local prior, $a_n = g_L^S$ and $b_n = g_L^O$, and in the case of the MOM prior, $a_n = n^2(g_M^S)^3$ and $b_n = n^2(g_M^O)^3$.

(ii) *Non-overfitted models. If $\gamma^* \not\subset \gamma$, then*

$$\begin{aligned} \log \hat{B}_{\gamma, \gamma^*} &= -n [M(\boldsymbol{\eta}_{\gamma^*}^*) - M(\boldsymbol{\eta}_\gamma^*)] + \frac{d_{\gamma^*} - d_\gamma}{2} \log(n) \\ &\quad + \frac{q_{\gamma^*} - q_\gamma}{2} \log(a_n) + \frac{p_{\gamma^*} - p_\gamma}{2} \log(b_n) + \mathcal{O}_p(1), \end{aligned}$$

where for the local prior, $a_n = g_L^S$ and $b_n = g_L^O$, and in the case of the MOM prior, $a_n = (g_M^S)^3$ and $b_n = (g_M^O)^3$.

The proof of Proposition 1 can be found in Sections 3 and 4 of the Appendix (Iqbal et al., 2025).

For overfitted models, this result implies a faster Bayes factor rate under the use of non-local priors compared to local priors. A natural consequence is that non-local priors asymptotically remove spurious variables faster than local priors do (except if g_L^S

and g_L^O grow faster than both $\log(n)$ and g_M^S and g_M^O , which is in line with previous literature (Rossell and Rubio, 2018, 2023). For non-overfitted models, the Bayes factor rates contain the term $-n [M(\boldsymbol{\eta}_{\gamma^*}^*) - M(\boldsymbol{\eta}_{\gamma}^*)]$, which represents the difference in goodness of fit, and is $\mathcal{O}_p(n)$. This term implies that the power to detect active variables is affected by model misspecification, which includes departures from the assumption of normality, omitting important variables, non-linear effects or interactions.

The Bayes factor rates presented in Proposition 1 can also be viewed as an extension of the normal linear regression case discussed in (Rossell and Rubio, 2018). In the case of non-overfitted models, $\log \hat{B}_{\gamma, \gamma^*}$ is asymptotically driven by the difference between likelihoods for both Proposition 1 and the rates in Rossell and Rubio (2018). If there is underfitting in the selection equation under γ , the additional terms in Equation (9) may lead to larger differences and faster convergence compared to a standard linear regression log-likelihood. In the case of overfitted models, we can separate $d_{\gamma^*} - d_{\gamma}$ in Proposition 1 into $p_{\gamma^*} - p_{\gamma}$ and $q_{\gamma^*} - q_{\gamma}$, so that the rates in the overfitted case are affected separately by each of the selection and outcome equations. The terms corresponding to the outcome equation in Proposition 1 are the same as in Rossell and Rubio (2018) as long as g_M^O is constant. If there are variables in the selection equation under γ that are not in γ^* , then there is an extra nonzero term $\frac{3}{2} \log(n)(q_{\gamma^*} - q_{\gamma})$ present in $\log \hat{B}_{\gamma, \gamma^*}$, leading to faster convergence in Proposition 1 than those obtained in standard linear regression (Rossell and Rubio, 2018). Another key distinction is that the rates presented in Proposition 1 are not affected by sample selection as a form of model misspecification, as this aspect is explicitly accounted for within the model. Though this misspecification does not directly affect convergence rates asymptotically, the optimal model γ^* may not be under a normal linear regression specification the same as under a sample selection model specification. Indeed, we can observe that the conditional mean of the observed outcomes (Heckman, 1979) is

$$\mathbb{E}(y_i \mid s_i = 1, \mathbf{x}_i, \mathbf{w}_i) = \sigma\theta_0 + \sigma\mathbf{x}_i^\top \boldsymbol{\theta} + \rho\sigma\lambda(\alpha_0 + \mathbf{w}_i^\top \boldsymbol{\alpha}),$$

where $\lambda(t) = \phi(t)/\Phi(t)$ denotes the Inverse Mills' Ratio with respect to the standard normal distribution. Thus, if one neglects sample selection and conducts BVS on the observed outcomes only, any additional variables affecting the selection equation but not the outcome might still be selected as they affect the outcome in the Inverse Mill's ratio scale.

This result highlights a broader point. Ignoring sample selection and performing Bayesian variable selection using a normal linear regression on only the observed outcomes would introduce model misspecification. The effects of model misspecification in linear regression have been widely studied (Rossell and Rubio, 2018, 2023), primarily in terms of reduced power to detect small effects (sensitivity) due to inflated error variance and coefficient estimator variance. However, under sample selection, specificity may also be affected, as some variables influencing selection may not impact the outcome. Consequently, such variables could be mistakenly included in the model during variable selection in finite samples, or even asymptotically. The simulation study in Section 5 of the Appendix (Iqbal et al., 2025) shows this effect empirically - even as the sample size becomes very large, spurious variables will be included if they explain the

selection process. This highlights the relevance of our work: incorporating a correction model (Heckman, 1976) directly addresses model misspecification due to sample selection. However, other types of model misspecification may still remain. In the simulation study in Section 6, we examine the finite-sample effects of other types of model misspecification affecting the sample selection model. The results guide prior calibration in misspecified models.

The following result establishes model selection consistency under model misspecification. That is, under model misspecification and sample selection, the proposed methodology for Bayesian variable selection asymptotically recovers the model γ^* that minimizes the Kullback-Leibler divergence between the sample selection model under γ^* and the true data-generating mechanism. The model γ^* represents the projection (best approximation) of the true model onto the family of sample selection models considered.

Corollary 1. *For γ such that $\pi(\gamma) \neq 0$. Assume the conditions C1-C5 hold, with γ^* defined as above, and that $\pi(\gamma^*) \neq 0$. Then $\hat{\pi}(\gamma^* | \mathbf{y}, \mathbf{s}) \xrightarrow{P} 1$ as $n \rightarrow \infty$.*

6 Simulation study

In this section, we present a simulation study comparing the performance of the proposed methodology using local and non-local priors with that of competing approaches, such as spike-and-slab and Adaptive LASSO.

6.1 Experiment design

In the simulations, we consider three scenarios. In all three scenarios, we have $\sigma = 1$, $\rho = 0.5$ and sample size $n = 500$. We have 5 active variables in each equation, with the rest being spurious, and we test $p = q = 10, 25, 50$. The simulations are run over 1000 replicates. We assume $\mathbf{x} = \mathbf{w}$, with correlation described by $Cor(x_i, x_j) = 0.5^{|i-j|}$.

In Scenario 1, the model is correctly specified, with $\alpha_0 = 1.5$ (chosen such that roughly 30% of the data is missing), $\beta_0 = 0.5$, $\boldsymbol{\alpha} = (0.5, 1, 1, 1.5, 1.5, 0, \dots, 0) / \sqrt{2}$ and $\boldsymbol{\beta} = (0.25, 0.5, 0.5, 1, 1, 0, \dots, 0)$. The errors are assumed bivariate normal.

Scenarios 2 and 3 explore two kinds of misspecification. For each of these scenarios, we use Scenario 1 as a baseline, and introduce a modification to induce model misspecification. In Scenario 2, we test the performance of each method under distributional misspecification. The errors are generated from a contaminated normal distribution, with contamination probability 0.1 and scale multiplier 10. In Scenario 3, we test performance under omitted variable misspecification by excluding w_3 and x_3 from the model input while retaining them in the data generation process.

While, in theory, different variables may affect selection and outcome, in practice, the same variables often influence both processes. This overlap can lead to multicollinearity and identifiability challenges, which are commonly addressed through exclusion restrictions. Our additional simulations (in Section 6.5) demonstrate the performance of our methods in the presence of exclusion restrictions. In that section, we also provide a

scenario to observe the effects of misspecifying a sample selection model as a linear regression model, using the R package *mombf* (Rossell et al., 2025).

We compare four variable selection methods. The first is our proposed algorithm using non-local priors, using Newton-Raphson as the internal optimizer. For this, we use the gradient stopping condition with 10^{-6} as tolerance level, and use the warm start proposed in Section 4.2 as our initial values. We use Beta-Binomial(1, 1) as our model prior. We use the values of g^S, g^O described in Section 3.3, which leads to $g_M^S \approx 0.0449$ and $g_M^O \approx 0.0284$. The second variable selection method corresponds to the proposed algorithm, but with a local prior. We use $g_S = g_O = 1$ as in Section 3.3, but the specification is otherwise the same. The third variable selection method is the spike-and-slab sampler proposed in Iqbal et al. (2023). The parameters are the same as recommended in the paper, using hyperparameters $(1, p + q)$ for the model space prior. This choice induces heavy sparsity and is strictly decreasing with respect to model size. The fourth variable selection method is Adaptive LASSO as in Ogundimu (2022). We perform a grid search over 100 different values of λ , and use the choice of λ that attains the lowest Bayesian Information Criterion (BIC).

We compare performance by comparing the “chosen model” by each method to the true model used to generate the data. For the non-local and local prior methods, alongside the spike-and-slab, we select the median model, that is the model with all posterior inclusion probabilities greater than 0.5. For the Adaptive LASSO, we use the model with all non-zero covariates included. The performance metrics we use to compare these are sensitivity (Sens), specificity (Spec), True Model Rate (TMR) and Matthews Correlation Coefficient (MCC), the latter of which gives a single-number summary of sensitivity and specificity. Because the small effects are the hardest to recover, we also record the inclusion rates of the small effects α_1 and β_1 in the median model, denoted α_1 inc. and β_1 inc respectively.

6.2 Correct model specification

Table 1 shows the results for Scenario 1. As expected, under increasing sparsity, the small effects become increasingly difficult to detect, so the sensitivity and true model rate drop for all methods, but the specificity increases - aside for Adaptive LASSO in the selection equation, for which the specificity remains low and many spurious variables are included.

The differences are apparent in the sensitivity and specificity of the various methods. The non-local method has higher sensitivity and lower specificity than the other Bayesian methods, prioritizing correctly identifying the small effect, while the spike-and-slab method induces more sparsity. This sparsity induced by the spike-and-slab leads to slightly better performance for $p = 10$, similar performance for $p = 25$ and worse performance for $p = 50$, where heavy sparsity is already present and the non-local method is better at correctly identifying the small effect than the spike-and-slab or local methods are. As such, for these elicitation of the priors, the non-local method performs better in the presence of heavy sparsity. This is more evident for the selection equation than the outcome. While all the methods perform worse for the selection equation than the

Fit		Selection equation					Outcome equation				
		TMR	Sens	Spec	MCC	α_1 inc.	TMR	Sens	Spec	MCC	β_1 inc.
$p = 10$	Non-local	0.696	0.980	0.944	0.925	0.908	0.804	0.994	0.958	0.953	0.974
	Local	0.664	0.964	0.962	0.925	0.818	0.848	0.986	0.981	0.967	0.931
	Spike-and-slab	0.739	0.962	0.982	0.944	0.810	0.884	0.984	0.993	0.976	0.918
	ALASSO	0.606	0.968	0.940	0.909	0.839	0.831	0.976	0.990	0.966	0.881
$p = 25$	Non-local	0.673	0.955	0.992	0.953	0.791	0.798	0.980	0.994	0.971	0.905
	Local	0.545	0.927	0.994	0.938	0.643	0.753	0.964	0.996	0.967	0.819
	Spike-and-slab	0.599	0.933	0.996	0.946	0.668	0.784	0.965	0.998	0.972	0.825
	ALASSO	0.361	0.955	0.952	0.863	0.775	0.695	0.952	0.996	0.959	0.760
$p = 50$	Non-local	0.590	0.929	0.997	0.947	0.668	0.762	0.970	0.997	0.970	0.856
	Local	0.431	0.897	0.997	0.928	0.501	0.666	0.943	0.998	0.960	0.719
	Spike-and-slab	0.481	0.905	0.998	0.935	0.533	0.700	0.949	0.999	0.964	0.745
	ALASSO	0.147	0.956	0.937	0.746	0.781	0.529	0.918	0.998	0.944	0.592

Table 1: Scenario 1 (correctly specified case) results.

outcome, the difference in performance between the non-local methods and the other methods is bigger for the selection equation, especially in recovering the true model.

6.3 Misspecified error distribution

Table 2 shows the results of simulations on Scenario 2. Note that the effect of having contaminated observations is an increased estimate of the variance in both equations. In the selection equation, the variance is fixed at 1, so the relative “increase” in variance leads to smaller parameter estimates, leading to decreased sensitivity and increased specificity. For the outcome equation, it is apparent in the transformation $\beta \rightarrow \beta/\sigma$ that increased variance will lead to smaller parameter estimates, and similarly decrease sensitivity and increase specificity. This is reflected in the results, and all methods suffer in terms of sensitivity and recovering the small effect. Regardless, the non-local method performs better than all other methods across the board. Even for $p = 10$, the performance of the non-local prior is notably better than the other models when comparing the rate at which the true model is recovered. The heavier sparsity induced by the spike-and-slab with the Beta-Binomial elicitation $(1, p + q)$ is apparent when comparing the sensitivity and specificity of the given methods. Once again, the difference in performance is more apparent in the selection equation than the outcome equation. When comparing values of MCC attained by the non-local method and spike-and-slab method for different scenarios, the values are similar in the outcome equation, but there is a substantial difference in the selection equation, where the non-local method attains 0.019 and 0.021 higher MCC than the spike-and-slab method for $p = 25$ and $p = 50$ respectively.

6.4 Missing non-spurious covariate

Table 3 shows the results for Scenario 3. A consequence of missing a medium-size effect is an increase in the estimated variance, as some of the missing signal is attributed to the unknown variance. But due to the presence of correlation, some of this unknown signal is attributed to the other effects. Similarly to Scenario 2, all methods suffer from a decrease in sensitivity. When $p = 10$, the results for the non-local and spike-and-slab methods are similar to the correctly specified case, but the spike-and-slab method experiences a drop in true model rate for $p = 25$ and $p = 50$, while the non-local method

Fit		Selection equation					Outcome equation				
		TMR	Sens	Spec	MCC	α_1 inc.	TMR	Sens	Spec	MCC	β_1 inc.
$p = 10$	Non-local	0.661	0.971	0.941	0.913	0.855	0.655	0.961	0.957	0.918	0.813
	Local	0.540	0.923	0.970	0.894	0.633	0.543	0.922	0.977	0.901	0.632
	Spike-and-slab	0.585	0.923	0.987	0.912	0.627	0.588	0.925	0.986	0.913	0.641
	ALASSO	0.531	0.934	0.956	0.890	0.679	0.533	0.915	0.986	0.904	0.596
$p = 25$	Non-local	0.566	0.938	0.991	0.937	0.694	0.516	0.922	0.991	0.928	0.640
	Local	0.348	0.875	0.995	0.906	0.430	0.373	0.876	0.996	0.912	0.436
	Spike-and-slab	0.407	0.887	0.996	0.918	0.466	0.442	0.894	0.997	0.924	0.497
	ALASSO	0.278	0.924	0.961	0.861	0.635	0.388	0.894	0.993	0.915	0.495
$p = 50$	Non-local	0.467	0.909	0.996	0.930	0.563	0.418	0.896	0.996	0.921	0.523
	Local	0.227	0.843	0.998	0.898	0.310	0.266	0.850	0.998	0.905	0.331
	Spike-and-slab	0.297	0.861	0.998	0.909	0.367	0.341	0.870	0.998	0.917	0.398
	ALASSO	0.136	0.924	0.955	0.776	0.627	0.279	0.869	0.996	0.903	0.396

Table 2: Scenario 2 (contaminated normal errors) results.

Fit		Selection equation					Outcome equation				
		TMR	Sens	Spec	MCC	α_1 inc.	TMR	Sens	Spec	MCC	β_1 inc.
$p = 10$	Non-local	0.739	0.978	0.956	0.931	0.911	0.818	0.988	0.969	0.954	0.959
	Local	0.633	0.933	0.975	0.912	0.733	0.802	0.966	0.986	0.953	0.862
	Spike-and-slab	0.690	0.933	0.989	0.928	0.732	0.834	0.966	0.993	0.962	0.865
	ALASSO	0.568	0.935	0.954	0.891	0.742	0.752	0.950	0.989	0.943	0.798
$p = 25$	Non-local	0.730	0.955	0.994	0.954	0.823	0.792	0.967	0.995	0.964	0.891
	Local	0.453	0.877	0.995	0.908	0.507	0.630	0.918	0.997	0.941	0.630
	Spike-and-slab	0.499	0.887	0.996	0.919	0.548	0.683	0.929	0.998	0.949	0.717
	ALASSO	0.322	0.913	0.966	0.851	0.653	0.585	0.914	0.995	0.932	0.658
$p = 50$	Non-local	0.647	0.929	0.998	0.947	0.717	0.736	0.949	0.998	0.959	0.814
	Local	0.354	0.849	0.998	0.902	0.396	0.553	0.897	0.999	0.934	0.587
	Spike-and-slab	0.400	0.858	0.999	0.911	0.430	0.600	0.906	0.999	0.941	0.626
	ALASSO	0.194	0.910	0.965	0.778	0.639	0.449	0.879	0.998	0.917	0.514

Table 3: Scenario 3 (omitted medium effect) results.

is not substantially impacted. In fact, the values of MCC in the selection equation are similar to the correctly specified case for the non-local method, while the spike-and-slab experiences a drop in MCC. In the outcome equation, there is a small decrease in performance for the non-local method, but the spike-and-slab suffers a more substantial drop in performance. When looking at the true model rates, aside from the outcome equation in $p = 10$, where the spike-and-slab attains similar true model rate, the non-local method attains considerably higher true model rates than the other methods, especially for the selection equation. As a result, the non-local method performs better than the alternative methods under this type of misspecification.

6.5 Additional simulation scenarios

We describe three additional simulation scenarios and their results. The corresponding tables and detailed analysis can be found in Sections 5, 6 and 7 of the Appendix (Iqbal et al., 2025).

In Section 5 of (Iqbal et al., 2025), we empirically illustrate the asymptotic properties of Corollary 1. We repeat Scenarios 1 (correct model specification) and 2 (contaminated normal errors) but for larger sample sizes $n = 1000$ and $n = 2000$. The true model rates are significantly improved for $n = 1000$ and further yet for $n = 2000$.

Scenario 0 studies the impact of misspecifying a sample selection model as a normal linear regression on variable selection. The data is generated as in Scenario 1, but with an additional large effect active only in the selection equation, that is $\alpha_8 = 1.5/\sqrt{2}$, and with $n = 5,000, 10,000$ and $20,000$ and $p = 10$ considered. We use a MOM non-local prior with `mombf` to perform the variable selection, assuming a normal linear regression model. Table 5 of the Appendix (Iqbal et al., 2025) shows that as n increases, β_8 is included with increasing probability, having inclusion rate 0.992 for $n = 20,000$. It is reasonable to conclude that not accounting for sample selection bias can lead to variable selection methods incorrectly including covariates that are only associated with the missingness mechanism and not the outcome of interest.

The second scenario we provide studies the effect of having an exclusion restriction variable. The data is generated as in Scenario 1 but we set $\beta_3 = 0$ while α_3 remains non-zero, so that w_3 acts as an exclusion restriction. We also only consider $p = 25$. Table 6 of the Appendix (Iqbal et al., 2025) shows that the performance of each variable selection method is similar to the case where β_3 is non-zero. There is slightly higher inclusion rate of β_3 for the non-local prior, but this is a consequence of inducing less sparsity, and the non-local prior still has higher sensitivity and MCC than other methods. Though it is of less practical interest, we provide results for the reverse case where $\alpha_3 = 0$ and $\beta_3 \neq 0$ in Table 7 of the Appendix (Iqbal et al., 2025), for which the performance of variable selection methods are again similar to that in Scenario 1. There is a moderate increase in sensitivity in the selection part of the model for all methods, but this is a result of the presence of less active variables in that part of the model, and as such is an artefact of the simulation scenario.

7 Applications

In this section, we present two real data applications that illustrate the use of the proposed Bayesian variable selection methodology. The first example provides an analysis of the well-studied ambulatory expenditures data (Cameron and Trivedi, 2010), and the second application analyzes a subset of data from the RAND Health Insurance Experiment (RAND HIE) as used in Cameron and Trivedi (2010). For each of these datasets, we compare the non-local prior, local prior, spike-and-slab prior and Adaptive LASSO. Because the non-local and local prior methods do not directly provide parameter estimates, we instead compare the posterior inclusion probabilities of each variable. This is not possible for the Adaptive LASSO, so for that method we only record whether a variable was included or not. For the non-local and local prior methods, we use the same hyperparameters as in the simulations, but instead take a chain length of 10,000, excluding an initial burn-in of 1,000 iterations. We take the spike-and-slab and Adaptive LASSO results directly from Iqbal et al. (2023) for comparison. Additionally, in Section 7.3 we provide a sensitivity analysis on the non-local prior parameters.

7.1 Ambulatory data

The ambulatory expenditures data studies the impact of several variables on ambulatory expenditures. For some patients, the expenditure is zero. It is suspected that the decision

to spend is correlated with cost, so that sample selection bias arises. For this dataset, 516 of the 3,328 observations (15.8%) have zero expenditure. As such, the dataset is commonly used to illustrate the performance of sample selection methods (Marchenko and Genton, 2012; Lachos et al., 2021; Ogundimu, 2022; Iqbal et al., 2023).

The predictors recorded in the dataset are `age`, `gender`, education status (`educ`), ethnicity (`blhisp`), number of chronic conditions (`totchr`), insurance status (`ins`) and `income`. In line with previous publications, we include all the variables in each equation, except for `income` which we take as an exclusion restriction. One motivation for the use variable selection methods in sample selection models is that we can judge whether this choice is necessary, and if other variables could also fulfill the exclusion restriction criterion. Additionally, the assumption of normality may be violated for this data set (Marchenko and Genton, 2012; Lachos et al., 2021).

Table 4 shows the posterior inclusion probabilities for ambulatory data. The results for the non-local and spike-and-slab are similar, with the inclusion probabilities only differing by up to 0.08, and selecting the same median model. Both the non-local and spike-and-slab priors exclude `income` from the selection equation in the median model, while the Adaptive LASSO and local prior methods do not exclude it. Because the inclusion probabilities for `income` are not close to zero for any method, the inclusion of it as an exclusion restriction could be considered. However, based on inclusion probabilities `educ` and `ins` could both be taken as exclusion restrictions as well, possibly better ones. Both are included in the median models by all methods in the selection equation (though `ins` has relatively smaller inclusion probability), but neither are included in the outcome equation, with very low inclusion probabilities across the board.

Variable	Posterior inclusion probability							
	Selection equation				Outcome equation			
	NLP	LP	SS	ALASSO	NLP	LP	SS	ALASSO
<code>educ</code>	1.000	1.000	1.000	1	0.081	0.328	0.116	0
<code>age</code>	0.976	0.976	0.949	1	1.000	1.000	1.000	1
<code>income</code>	0.270	0.584	0.349	1	-	-	-	-
<code>female</code>	1.000	1.000	1.000	1	1.000	1.000	1.000	1
<code>totchr</code>	1.000	1.000	1.000	1	1.000	1.000	1.000	1
<code>blhisp</code>	1.000	1.000	1.000	1	0.978	0.982	0.895	1
<code>ins</code>	0.625	0.810	0.571	1	0.001	0.137	0.033	0

Notes: NLP = non-local prior, LP = local prior, SS = spike-and-slab, ALASSO = Adaptive LASSO. For the Adaptive LASSO, we only record whether the variable was shrunk to zero or not.

Table 4: Ambulatory data results.

7.2 RAND data

We analyze a subset of data from the RAND Health Insurance Experiment (RAND HIE), which is a 15-year study of how randomized health insurance affects costs, utilization and outcomes in the United States (Cameron and Trivedi, 2010). Similarly to the ambulatory data, there is a decision to spend, leading to sample selection bias,

and a subset of this data has recently been studied in this context (Zhao et al., 2020; Lachos et al., 2021). In keeping consistent with previous work, we use the logarithm of medical expenditures per individual (`lnmeddol`) for the outcome, use the indicator variable `binexp` as the selection variable, and let $\mathbf{w} = \mathbf{x}$, so no exclusion restriction is present. The covariates in the outcome and selection models consist of the following: the logarithm of coinsurance rate plus 1 ($\text{logc} = \log(\text{coins} + 1)$), the dummy variable for individual deductible plan (`idp`), the logarithm of participation incentive payment (`lpi`), an artificial variable `fmde` that is 0 if `idp` = 1 and $\log(\max(1, \text{mde}/(0.01 * \text{coins}))$ otherwise (where `mde` is the maximum expenditure offer), physical limitations (`physlm`), the number of chronic diseases (`disea`), dummy variables for good (`hlthg`), fair (`hlthf`) and poor (`hlthp`) self-rated health (where the baseline is excellent self-rated health), the log of family income (`linc`), the log of family size (`lfam`), education of household head in years (`educdec`), age of individual in years (`xage`), a dummy variable for female individuals (`female`), a dummy variable for individuals younger than 18 years (`child`), a dummy variable for female individuals younger than 18 years (`fchild`), and a dummy variable for black household heads (`black`). We study a subset of the data consisting of the second year of observations with `educdec` not equal to “NA”. This subset has 1,293 missing expenditures of 5,574 total observations. Additionally, Lachos et al. (2021) notes that the selection-t model fits the data better than selection-normal, indicating that distributional misspecification may be present.

Table 5 shows the posterior inclusion probabilities for the RAND data. While the non-local and spike-and-slab priors lead to the same median model *a posteriori*, there are some differences in the posterior inclusion probabilities. For the variables that are not included in the median model, aside from `fmde` in the selection equation, the non-local prior shrinks posterior inclusion probabilities closer to zero than the spike-and-slab does. As such, there will be fewer models in the non-local sample including variables that are considered to be spurious, and the model samples will on average contain fewer variables. Of particular interest is the relationship between `xage` and `child`. These are the only two variables with correlation greater than 0.5 in magnitude, having correlation -0.804 . The non-local method has lower posterior probability for `xage` than `child`, while the spike-and-slab has slightly higher posterior probability for `xage`.

The posterior mode (the most sampled model) includes `child` but not `xage`, while the second most sampled model includes `xage` but not `child`. These models have posterior probability 0.0709 and 0.0600 respectively, so do not represent substantial posterior mass, but still indicate that the two variables may be carrying the same information, and the non-local prior tries to only include one at a time. That being said, the fact that both posterior inclusion probabilities are above 0.5 suggest that both may have useful information. To further investigate this, we used Stan to perform posterior inference on two models - one identical to the median model recovered by the non-local sampler, including both `xage` and `child` in the outcome equation, and the same model but with `xage` excluded from the outcome equation. We used $N(0, 1)$ priors on $\log(\tau)$ and $\text{atanh}(\rho)$ and $N(0, 10^2)$ priors on all other parameters, and compared them using the leave-one-out estimator of expected log-predictive density (`elpd_loo`) (Vehtari et al., 2017). The model including both had `elpd_loo` of -10211.9 , while the model excluding

Variable	Posterior inclusion probability							
	Selection equation				Outcome equation			
	NLP	LP	SS	ALASSO	NLP	LP	SS	ALASSO
logc	1.000	1.000	1.000	1	1.000	0.999	1.000	1
idp	0.014	0.194	0.094	0	0.002	0.070	0.033	0
lpi	0.831	0.929	0.919	1	0.006	0.064	0.046	0
fmde	0.096	0.092	0.090	0	0.001	0.035	0.037	0
physlm	0.995	0.996	0.998	1	0.999	0.999	1.000	1
disea	1.000	1.000	1.000	1	1.000	1.000	1.000	1
hlthg	0.000	0.043	0.020	0	0.119	0.396	0.236	1
hlthf	0.087	0.330	0.131	0	0.558	0.867	0.740	1
hlthp	0.890	0.953	0.844	1	0.959	0.995	0.964	1
linc	0.630	0.820	0.640	1	0.996	1.000	0.997	1
lfam	0.001	0.043	0.019	0	0.822	0.907	0.853	1
educdec	0.932	0.970	0.922	1	0.013	0.134	0.076	0
xage	0.004	0.055	0.031	0	0.531	0.693	0.749	1
female	1.000	1.000	1.000	1	1.000	1.000	1.000	1
child	0.011	0.078	0.057	0	0.682	0.686	0.670	1
fchild	1.000	1.000	1.000	1	1.000	1.000	1.000	1
black	1.000	1.000	1.000	1	1.000	0.999	1.000	1

Notes: For the Adaptive LASSO, we only record whether the variable was shrunk to zero or not.

Table 5: RAND data results.

`xage` had `elpd_loo` of -10213.9 . The difference between the two is 2.0, while the standard error of that difference between `elpd_loo` values was 2.8, so if there is any gain in predictive power by including `xage` in the outcome, it is negligible.

7.3 Sensitivity analysis

We provide a sensitivity analysis on both applications comparing to small perturbations of prior parameters and a much sparser elicitation. For the small perturbation, we elicit both priors as in Section 3.3, but with $\epsilon = 0.04$, which induces prior variances roughly 17% larger than the original elicitation. For the sparser elicitation, we elicit with $\epsilon = 0.02$. Tables and further details can be found in Section 8 of the Appendix (Iqbal et al., 2025).

Tables 8 and 9 in the Appendix (Iqbal et al., 2025) show the posterior inclusion probabilities under the original elicitation and small perturbation, for the ambulatory and RAND datasets respectively. With larger variances, the posterior inclusion probabilities generally decrease, but not significantly, and their values are similar across the two elicitations for both the ambulatory data and RAND data. As such, the non-local prior is stable under small perturbations of the prior parameters.

For the sparser elicitation, Tables 10 and 12 in the Appendix (Iqbal et al., 2025) shows the posterior inclusion probabilities under the original elicitation and the sparser elicitation for the ambulatory and RAND data respectively. For the ambulatory data,

inducing extra sparsity further shrinks inclusion probabilities, to the point where the median model is different. For the RAND data, the inclusion probabilities decrease for most variables, but there are some exceptions. For instance, the variable `xage` in the outcome equation, which is highly correlated with `child`, has its inclusion probability increased by 0.188, while `child` has its inclusion probability decreased by 0.274. As such, in the presence of highly correlated variables, inferences can change significantly under heavy sparsity.

Since the median models differ between the two elicitations for both datasets, we used Stan to sample from the median models in each case, using the same priors as described in Section 7.2. Tables 11 and 13 in the Appendix (Iqbal et al., 2025) provide posterior summaries from the Stan sampler for the ambulatory and RAND datasets respectively. For the ambulatory data, the parameter medians and 95% confidence intervals are similar across both elicitations, and the variable `ins` which is only included in the original elicitation has small parameter samples. The same is true for most variables in the RAND data, but there are some differences in which small effects have been kept and which have been shrunk. In particular, the exclusion of `child` has made a noticeable difference to the sampled parameters. Regardless, the predictive performance of the median models (measured using `elpd_loo`) are similar for each dataset, the estimates of ρ and σ are sensible and similar to those of the spike-and-slab (Iqbal et al., 2023).

8 Discussion

While Bayesian variable selection under missing at random (MAR) assumptions has been extensively studied (García-Donato et al., 2025), far less attention has been given to the case of data missing not at random (MNAR) and sample selection. In this paper, we propose local and non-local priors for Bayesian variable selection in sample selection models, allowing for model uncertainty to be quantified via posterior model probabilities, in contrast to frequentist approaches such as the LASSO. We characterize the asymptotic properties of the proposed methodology under potential model misspecification and show that it asymptotically identifies the variables that explain the outcome and the selection processes. On the computational side, we develop a Gibbs sampling scheme that incorporates Laplace approximations of the marginal likelihood at each step, enabling direct sampling from the posterior distribution over the model space. We have also highlighted the importance of conducting sensitivity analyses on prior calibration and demonstrated how a careful and principled prior specification can enhance the performance of Bayesian variable selection, particularly in the presence of model misspecification. We have shown that the proposed methodology scales well to the dimensions of interest in practice.

The simulation study shows that the proposed methodology, with default hyperparameters, is competitive with methods based on spike-and-slab priors (Iqbal et al., 2023) in correctly specified cases. Under model misspecification, our approach outperforms existing alternatives (Adaptive LASSO and spike-and-slab priors), particularly for the selection equation. In this context, the non-local prior methodology shows a greater advantage over the spike-and-slab prior in the selection equation than in the

outcome equation and is markedly better at recovering the optimal model in misspecified cases. Consistent with previous work on variable selection in sample selection models (Ogundimu, 2022), our results also highlight differing learning rates for identifying active variables in the outcome and selection equations. In particular, variable selection in the selection equation is performed through a probit link with a binary response variable, making identification of variables more difficult than in the linear regression model of the outcome equation.

In this paper, we employed proper priors for all parameters of the sample selection model. An Associate Editor noted that this naturally requires careful calibration of the prior for the parameter σ (or τ), and asked whether it would be possible to instead use the improper prior $\pi(\sigma) \propto \frac{1}{\sigma}$, which is commonly employed in linear regression to avoid manual calibration. The answer is positive, under some mild conditions presented in the following result.

Proposition 2. *Consider the sample selection model defined by equations (1)–(3), coupled with the prior structure:*

$$\pi(\beta, \alpha, \sigma, \rho) \propto \frac{1}{\sigma} \pi(\beta) \pi(\alpha) \pi(\rho), \quad (10)$$

where $\pi(\alpha)$ and $\pi(\rho)$ are proper priors, and $\pi(\beta)$ is bounded above. Then, the posterior distribution of $(\beta, \alpha, \sigma, \rho)$ is proper if $n_O = \sum_{i=1}^n s_i \geq p + 2$.

This implies that the posterior distribution associated with the prior structure (10) is proper, provided that the number of observed outcomes exceeds the number of covariates plus 2 (or the number of regression coefficients plus one). Under these conditions, the prior (10) may be used as a reference prior, thereby avoiding the need for calibration of the prior on σ . We implemented this prior and ran simulation Scenario 1 from Section 6 with a non-local prior and found that the performance was very similar to the Gamma prior on σ . The maximum difference in inclusion probabilities for any variable across the entire simulation was 0.04, and both priors attained the same median model in every Monte Carlo replicate. That said, because sample selection models have a more complex structure than linear regression (*e.g.*, the marginal variance of the outcome depends on ρ), the prior $\pi(\sigma) \propto \frac{1}{\sigma}$ cannot be strictly justified as an independence Jeffreys prior. Consequently, the investigation of “objective priors” in the context of sample selection models constitutes a promising direction for future research, and our result above indicates that it is possible to obtain proper posterior distributions under certain improper prior structures.

The Gibbs sampling scheme formulated in Section 4.2 works sufficiently well for the purposes of the paper, but it only consists of horizontal moves and iterating over every variable. Sampling efficiency and computational speed could be improved by considering more sophisticated approaches that adaptively sample from higher posterior density regions, such as neighbourhood-based sampling (Liang et al., 2023).

While we have observed the effects of distributional misspecification and shown that the proposed method outperforms alternatives under such conditions, it still assumes a

Heckman model with bivariate normal errors. A potential extension would be to use a more flexible model. In principle, if the gradient and Hessian of the log-likelihood are tractable for a given model, they could be incorporated into the sampling scheme in a similar manner, and the assumptions underlying the theoretical results may still hold. In practice, however, the additional parameters introduced by more flexible models could pose additional computational challenges. In this work, we have used a specific type of non-local prior, the moment prior (MOM). One could also consider the exponential moment (eMOM) and inverse moment (iMOM) priors (Rossell and Telesca, 2017), as well as other non-local priors. The implementation could be easily extended to any prior for which the gradient and Hessian can be expressed analytically, though establishing theoretical results would require additional work. Finally, an issue that has been reported in practice for certain data sets is the presence of flat ridges in the likelihood surface. This suggests practical non-identifiability and near-redundancy of parameters, which are challenges that arise in finite-sample inference (Cole, 2020). Although this was not observed in our simulations or case studies, a natural question arises: what should be done in such cases? If near-redundancy is detected for a model during model space exploration, would it be reasonable to assign zero probability to such models? These questions will be explored in future research.

Funding

Adam Iqbal was supported by the Heilbronn Institute for Mathematical Research via funding from the EPSRC grant “Additional Funding Programme for Mathematical Sciences” (EP/V521917/1).

Supplementary Material

Supplementary materials for “Bayesian variable selection under sample selection and model misspecification” (DOI: [10.1214/25-BA1567SUPP](https://doi.org/10.1214/25-BA1567SUPP); .pdf). Contains analytic expressions for gradients and Hessians, the explicit sampling algorithm, proofs of theoretical results, the three brief simulation scenarios described in Section 6.5 and further details on the sensitivity analysis.

References

- Cameron, A. C. and Trivedi, P. (2010). *Microeconometrics using Stata*. College Station, TX: Stata Press, revised edition. [17](#), [19](#)
- Certo, S. T., Busenbark, J. R., Woo, H.-s., and Semadeni, M. (2016). “Sample selection bias and Heckman models in strategic management research.” *Strategic Management Journal*, 37(13): 2639–2657. [2](#)
- Cole, D. (2020). *Parameter Redundancy and Identifiability*. Boca Raton, FL: CRC Press. [23](#)
- de Souza Bastos, F., Barreto-Souza, W., and Genton, M. (2022). “A Generalized Heck-

- man Model With Varying Sample Selection Bias and Dispersion Parameters.” *Statistica Sinica*, 32: 1911–1938. [MR4478183](#). 2, 3, 4
- Diggle, P. and Kenward, M. G. (1994). “Informative drop-out in longitudinal data analysis.” *Journal of the Royal Statistical Society Series C: Applied Statistics*, 43(1): 49–73. [MR3405479](#). doi: <https://doi.org/10.1111/rssa.12132>. 1
- Forte, A., Garcia-Donato, G., and Steel, M. (2018). “Methods and tools for Bayesian variable selection and model averaging in normal linear regression.” *International Statistical Review*, 86(2): 237–258. [MR3852410](#). doi: <https://doi.org/10.1111/insr.12249>. 5
- García-Donato, G., Castellanos, M., Cabras, S., Quirós, A., and Forte, A. (2025). “Model Uncertainty and Missing Data: An Objective Bayesian Perspective.” *Bayesian Analysis*, 1(1): 1–26. 21
- Genbäck, M., Stanghellini, E., and de Luna, X. (2015). “Uncertainty intervals for regression parameters with non-ignorable missingness in the outcome.” *Statistical Papers*, 56: 829–847. [MR3369432](#). doi: <https://doi.org/10.1007/s00362-014-0610-x>. 2
- George, E. I. and McCulloch, R. E. (1993). “Variable Selection via Gibbs Sampling.” *Journal of the American Statistical Association*, 88(423): 881–889. URL <https://www.tandfonline.com/doi/abs/10.1080/01621459.1993.10476353> 8
- Hahn, P. and Carvalho, C. (2015). “Decoupling shrinkage and selection in Bayesian linear models: a posterior summary perspective.” *Journal of the American Statistical Association*, 110(509): 435–448. [MR3338514](#). doi: <https://doi.org/10.1080/01621459.2014.993077>. 5
- Heckman, J. (1979). “Sample selection bias as a specification error.” *Econometrica: Journal of the Econometric Society*, 153–161. [MR0518832](#). doi: <https://doi.org/10.2307/1912352>. 1, 2, 4, 9, 12
- 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*, 475–492. NBER. 1, 4, 13
- Honoré, B. E. and Hu, L. (2020). “Selection without exclusion.” *Econometrica*, 88(3): 1007–1029. [MR4102633](#). doi: <https://doi.org/10.3982/ecta16481>. 2
- Iqbal, A., Ogundimu, E., and Rubio, F. (2023). “Bayesian variable selection in sample selection models using spike-and-slab priors.” *arXiv preprint arXiv:2312.03538*. [MR4394864](#). doi: <https://doi.org/10.1007/s00362-021-01246-z>. 2, 3, 14, 17, 18, 21
- Iqbal, A., Ogundimu, E., and Rubio, F. (2025). “Supplement to “Bayesian variable selection under sample selection and model misspecification”.” *Bayesian Analysis*. doi: <https://doi.org/10.1214/25-BA1567SUPP>. 8, 11, 12, 16, 17, 20, 21
- Johnson, V. and Rossell, D. (2010). “On the use of non-local prior densities in Bayesian hypothesis tests.” *Journal of the Royal Statistical Society: Series B (Statistical*

- Methodology*, 72(2): 143–170. MR2830762. doi: <https://doi.org/10.1111/j.1467-9868.2009.00730.x>. 3, 6
- Johnson, V. and Rossell, D. (2012). “Bayesian model selection in high-dimensional settings.” *Journal of the American Statistical Association*, 107(498): 649–660. MR2980074. doi: <https://doi.org/10.1080/01621459.2012.682536>. 3, 6, 8
- Kasprzak, M. J., Giordano, R., and Broderick, T. (2025). “How good is your Laplace approximation of the Bayesian posterior? Finite-sample computable error bounds for a variety of useful divergences.” *Journal of Machine Learning Research*, 26(87): 1–81. URL <http://jmlr.org/papers/v26/24-0619.html> MR4923386. 8
- Lachos, V. H., Prates, M. O., and Dey, D. K. (2021). “Heckman selection-t model: Parameter estimation via the EM-algorithm.” *Journal of Multivariate Analysis*, 184. MR4233412. doi: <https://doi.org/10.1016/j.jmva.2021.104737>. 18, 19
- Leung, S. and Yu, S. (2000). “Collinearity and Two-Step Estimation of Sample Selection Models: Problems, Origins, and Remedies.” *Computational Economics*, 15(3): 173–199. 2
- Li, Y. and Clyde, M. (2018). “Mixtures of g-priors in generalized linear models.” *Journal of the American Statistical Association*, 113(524): 1828–1845. MR3902249. doi: <https://doi.org/10.1080/01621459.2018.1469992>. 7
- Liang, F., Paulo, R., Molina, G., Clyde, M., and Berger, J. (2008). “Mixtures of g priors for Bayesian variable selection.” *Journal of the American Statistical Association*, 103(481): 410–423. MR2420243. doi: <https://doi.org/10.1198/016214507000001337>. 6, 7
- Liang, X., Livingstone, S., and Griffin, J. (2023). “Adaptive MCMC for Bayesian Variable Selection in Generalised Linear Models and Survival Models.” *Entropy*, 25(9). URL <https://www.mdpi.com/1099-4300/25/9/1310> 23
- Marchenko, Y. V. and Genton, M. G. (2012). “A Heckman Selection-t Model.” *Journal of the American Statistical Association*, 107(497): 304–317. MR2949361. doi: <https://doi.org/10.1080/01621459.2012.656011>. 2, 3, 18
- Miao, W., Ding, P., and Geng, Z. (2016). “Identifiability of normal and normal mixture models with nonignorable missing data.” *Journal of the American Statistical Association*, 111(516): 1673–1683. MR3601726. doi: <https://doi.org/10.1080/01621459.2015.1105808>. 2
- Molenberghs, G., Fitzmaurice, G., Kenward, M., Tsiatis, A., and Verbeke, G. (2014). *Handbook of Missing Data Methodology*. CRC Press. MR3380562. 1
- Ogundimu, E. (2022). “Regularization and variable selection in Heckman selection model.” *Statistical Papers*, 63(2): 421–439. MR4394864. doi: <https://doi.org/10.1007/s00362-021-01246-z>. 2, 14, 18, 22
- Ogundimu, E. O. and Hutton, J. L. (2016). “A Sample Selection Model with Skew-normal Distribution.” *Scandinavian Journal of Statistics*, 43(1): 172–190. MR3467000. doi: <https://doi.org/10.1111/sjos.12171>. 2

- Olsen, R. (1982). “Distributional tests for selectivity bias and a more robust likelihood estimator.” *International Economic Review*, 223–240. MR0655711. doi: <https://doi.org/10.2307/2526473>. 5
- Puhani, P. (2000). “The Heckman Correction for Sample Selection and Its Critique.” *Journal of Economic Surveys*, 14(1): 53–68. 2
- Rossell, D., Cook, J. D., Telesca, D., Roebuck, P., Abrli, O., and Torrens, M. (2025). *mombf: Model Selection with Bayesian Methods and Information Criteria*. R package version 4.0.0, commit e4d6dda7592041d48ded50f0ea95d8dbd0f1d7c3. URL <https://github.com/davidrusi/mombf> 14
- Rossell, D. and Rubio, F. (2018). “Tractable Bayesian variable selection: beyond normality.” *Journal of the American Statistical Association*, 113(524): 1742–1758. MR3902243. doi: <https://doi.org/10.1080/01621459.2017.1371025>. 12
- Rossell, D. and Rubio, F. (2023). “Additive Bayesian variable selection under censoring and misspecification.” *Statistical Science*, 38: 13–29. MR4534642. doi: <https://doi.org/10.1214/21-sts846>. 12
- Rossell, D. and Telesca, D. (2017). “Nonlocal priors for high-dimensional estimation.” *Journal of the American Statistical Association*, 112(517): 254–265. MR3646569. doi: <https://doi.org/10.1080/01621459.2015.1130634>. 23
- Rossell, D., Telesca, D., and Johnson, V. (2013). “High-dimensional Bayesian classifiers using non-local priors.” In *Statistical Models for Data Analysis*, 305–313. Springer. 6
- Sartori, A. E. (2003). “An estimator for some binary-outcome selection models without exclusion restrictions.” *Political Analysis*, 11(2): 111–138. 2
- Scott, J. and Berger, J. (2010). “Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem.” *The Annals of Statistics*, 38(5): 2587–2619. MR2722450. doi: <https://doi.org/10.1214/10-AOS792>. 6
- Tadesse, M. and Vannucci, M. (2021). *Handbook of Bayesian variable selection*. CRC Press. 5, 6
- Van-Hasselt, M. (2011). “Bayesian inference in a sample selection model.” *Journal of Econometrics*, 165(2): 221–232. MR2846646. doi: <https://doi.org/10.1016/j.jeconom.2011.08.003>. 2
- Vehtari, A., Gelman, A., and Gabry, J. (2017). “Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC.” *Statistics and Computing*, 27: 1413–1432. MR3647105. doi: <https://doi.org/10.1007/s11222-016-9696-4>. 20
- White, H. (1981). “Consequences and detection of misspecified nonlinear regression models.” *Journal of the American Statistical Association*, 76(374): 419–433. MR0624344. 11
- White, H. (1982). “Maximum likelihood estimation of misspecified models.” *Econometrica: Journal of the Econometric Society*, 1–25. MR0640163. doi: <https://doi.org/10.2307/1912526>. 11

- Wiemann, P. F. V., Klein, N., and Kneib, T. (2022). “Correcting for sample selection bias in Bayesian distributional regression models.” *Computational Statistics & Data Analysis*, 168: 107382. [MR4350999](#). doi: <https://doi.org/10.1016/j.csda.2021.107382>. 2
- Zellner, A. (1986). “On assessing prior distributions and Bayesian regression analysis with g-prior distributions.” *Bayesian Inference and Decision Techniques*. [MR0881437](#). 6
- Zhao, J., Kim, H. J., and Kim, H. M. (2020). “New EM-type algorithms for the Heckman selection model.” *Computational Statistics & Data Analysis*, 146. [MR4065386](#). doi: <https://doi.org/10.1016/j.csda.2020.106930>. 19