Bayesian Analysis (2025) **TBA**, Number TBA, pp. 1–73

Model Uncertainty and Missing Data: An Objective Bayesian Perspective

Gonzalo García-Donato*, María Eugenia Castellanos†, Stefano Cabras‡, Alicia Quirós§, and Anabel Forte¶

Abstract. The interplay between missing data and model uncertainty—two classic statistical problems—leads to primary questions that we formally address from an objective Bayesian perspective. For the general regression problem, we discuss the probabilistic justification of Rubin's rules applied to the usual components of Bayesian variable selection, arguing that prior predictive marginals should be central to the pursued methodology. In the regression settings, we explore the conditions of prior distributions that make the missing data mechanism ignorable, provided that it is missing at random or completely at random. Moreover, when comparing multiple linear models, we provide a complete methodology for dealing with special cases, such as variable selection or uncertainty regarding model errors. In numerous simulation experiments, we demonstrate that our method outperforms or equals others, in consistently producing results close to those obtained using the full dataset. In general, the difference increases with the percentage of missing data and the correlation between the variables used for imputation. Finally, we summarize possible directions for future research.

Keywords: Bayes factor, g-priors, ignorability, objective prior distribution, Rubin's rules.

MSC2020 subject classifications: Primary 62C10; secondary 62D10.

1 Introduction

Model uncertainty is a broad term for situations where a true data-generative model is assumed unknown. Paradigmatic model uncertainty problems include model choice, hypothesis testing, variable selection (VS), and model averaging. From a Bayesian perspective, a formal tool for addressing such problems is the posterior distribution over the model space. It assigns, conditionally on the data, the probability of each model and constitutes a comprehensive tool that is the basis for addressing *all* types of questions in model uncertainty scenarios.

Obtaining the posterior distribution, from straight probability arguments, entails severe difficulties of quite a different nature, particularly from an objective perspective

arXiv: 2410.05893

*Department of Economy and Finance, University of Castilla-La Mancha, gonzalo.garciadonato@uclm.es

†Department of Informatics and Statistics, Rey Juan Carlos University, maria.castellanos@urjc.es

‡Department of Statistics, University of Carlos III de Madrid, stefano.cabras@uc3m.es

§Department of Mathematics, Universidad de León, alicia.quiros@unileon.es

¶Department of Statistics and OR, University of Valencia, anabel.forte@uv.es

 https://doi.org/10.1214/25-BA1531

(see Berger, 2006, for a detailed discussion of objectivism in Bayesian statistics). Many of these challenges have to do with the conditions that prior distributions must satisfy for the Bayes factors (BF) to be well-defined (Jeffreys, 1961; Kass and Raftery, 1995; Berger and Pericchi, 2001), multiplicity issues (Scott and Berger, 2005), and numerical problems (not only for the computation of the marginal distribution for each model but also for sampling strategies when the model space is very large). Motivated by these, the field of model uncertainty has received considerable attention in recent decades and has acquired high levels of maturity (see Bayarri et al. (2012) for a pioneering attempt to standardize good practices and Tadesse and Vanucci (2022) for a collection of contemporaneous techniques in the field). Unfortunately, for problems with missing data, many of the proposed solutions do not apply directly, and their bases must be carefully reconsidered. A distinguished case is that of g-priors (Zellner, 1986), and the enormous number of generalizations they inspired (see, for example, Liang et al., 2008, for an extremely popular reference), which dependence on a complete fixed design matrix makes them useless in the case of missing observations.

1.1 Goals and structure

In this study, we focus on deriving reliable objective posterior distributions for model uncertainty with missing data in light of the standards in Bayarri et al. (2012) and all the references there compiled. For this task, Section 2 presents the problem from a broad perspective, emphasizing the interplay between Rubin's rules and posterior model probabilities.

The remainder of the article is organized as follows. Section 3 derives several equivalent expressions for the prior predictive marginals (the key ingredients of the posterior model probabilities and the BF) in regression models. We then establish the conditions for the ignorability of the missing mechanism and propose numerical strategies for marginal computation using simulation methods. We conclude this section with general considerations regarding the assignment of objective prior distributions on model parameters. Section 4 derives a complete methodology for VS in linear models with Gaussian errors based on a new prior distribution that extends g-priors, in a way that the dependence on the missed values of covariates is circumvented. Section 5 addresses a non-nested situation that has received very little attention in the literature. The considered linear models disagree on the distribution of errors, which we illustrate by comparing the different forms of the error covariance matrix. Section 6 evaluates the performance of the model uncertainty procedure and compares it with results obtained with the fully observed dataset (oracle); listwise deletions and some procedures proposed in the literature. To this purpose, we employ several simulated and real datasets with varying levels of missing data, some of which have been included in the supplementary material (García-Donato et al., 2025) as extra experiments. Finally, Section 7 concludes the paper by describing several directions for future research.

The Supplementary material is structured as follows. Section A provides a detailed overview of the main notation used throughout the paper. Section B contains the proofs of key theoretical results, including propositions and identities referenced in the paper. Section C presents additional experiments designed to further illustrate the performance

of the proposed methodology. Finally, Section D discusses the choice of prior distributions in the general modeling framework.

1.2 A brief review of the literature

Research on imputation methods *per se* is a classical topic in Bayesian literature on missing data. Recent studies on this topic include those of Xu et al. (2016); Mostafa et al. (2020); Gomez-Rubio (2020) and Aßmann et al. (2023). Other projects have focused on estimating a fixed model in the presence of missing data without considering model uncertainty, such as Ibrahim et al. (2002) for generalized linear models, Erler et al. (2016) for longitudinal models, and Erler (2019) for epidemiological modelling with time-varying covariates.

Within model uncertainty problems, several Bayesian researchers have responded to the difficulties of obtaining a sensible posterior distribution in the presence of missing data by proposing alternative criteria for model selection. This is the path taken by Ibrahim et al. (2002), who introduced a new criterion similar to the Bayesian Information Criterion (BIC), and Celeux et al. (2006) and Ibrahim et al. (2006), who extended DIC for missing data models or when missing data were present. Cohen and Berchenko (2021) proposed a normalized version of Akaike's Informatio Criterion (AIC) and BIC that allows the selection of variables without providing the full model uncertainty quantification based on the model posterior distribution. Similarly, Daniels et al. (2012) proposed a model choice measure based on a posterior predictive distribution. However, these measures do not provide any uncertainty regarding the model selection question and have complicated interpretability.

In the context of VS, methods based on the direct use of a posterior distribution have been published by Yang et al. (2005), Bozigar et al. (2020), and Storlie et al. (2020). The last two emphasize imputation methods with specific applications in mind. By contrast, Yang et al. (2005) is more general and has developed a full methodology to define and implement the computation of posterior distributions for VS with missing data. These studies consist of excellent deployments of Bayesian machinery to impute missing data. However, aspects that govern the essential properties of the resulting methods in relation to their model uncertainty are essentially unnoticed. For instance, vague priors are used—despite the many warnings advising against it—and there is no discussion regarding multiplicity issues, thus increasing the chances of reporting false positives, which could be inadvertently caused by a casual choice of initial probabilities assigned over the model space. In Section 6, we reproduce the simulation experiment in Yang et al. (2005) and show that their results are significantly outperformed by the posterior distribution we derive.

One work that connects, in spirit, to ours is Hoijtink et al. (2019). These authors argued that research on BF with missing data has received no attention in the literature and proposed easily implementable strategies to combine software for multiple imputation and BF calculations. Their study is limited to Bayesian testing; hence, it is strictly included in our study.

2 Notation and posterior probabilities

2.1 Notation

Following the convention in Little and Rubin (2020), let $\mathbf{d}_{(0)}$ identify the available values in a dataset and let $\mathbf{d}_{(1)}$ denote the missed observations. In model uncertainty there are several models under consideration, that we denote γ. This discrete parameter takes values from the set of possible alternatives Γ (also called the model space). The Bayesian framework transforms the prior distribution $p(\gamma)$ into its posterior distribution based on the available data, which is denoted by $p(\gamma \mid \mathbf{d}_{(0)})$. We adopt the Γ-closed perspective and assume that one of the models in the model space is the true model.

Regarding the remainder of the notation, the letter f denotes the density function for unknown but potentially observable random variables and vectors. The distribution of the parameters within each model, either a priori or a posteriori, is denoted as π and the marginal distributions, where the parameters have been integrated out with respect to π, are labeled m. The particular form of any of these functions under a given model γ is identified by the corresponding sub-index, and, for instance, f_γ is the form of f proposed under model γ.

2.2 Rubin's rules and model posterior probabilities

What has been termed "the key Bayesian motivation for multiple imputations" (Rubin, 1996, p.476) is a simple probabilistic identity that has greatly influenced the area of statistical methods to handle missingness. In the context of model uncertainty, this can be written as

$$p(\gamma \mid \mathbf{d}_{(0)}) = \int p(\gamma \mid \mathbf{d}_{(0)}, \mathbf{d}_{(1)})\, m(\mathbf{d}_{(1)} \mid \mathbf{d}_{(0)})\, \mathrm{d}\mathbf{d}_{(1)} \tag{1}$$

where $p(\gamma \mid \mathbf{d}_{(0)}, \mathbf{d}_{(1)})$ is the posterior probability of γ given the completed dataset and $m(\mathbf{d}_{(1)} \mid \mathbf{d}_{(0)})$ is the posterior predictive distribution for $\mathbf{d}_{(1)}$.

This identity suggests a possible strategy for approximating $p(\gamma \mid \mathbf{d}_{(0)})$ by creating multiple imputations of the dataset and then reporting the mean of the model's posterior probabilities over the completed datasets. This procedure aligns with Rubin's rules and can be easily implemented with specific software for imputation (such as `mice` by van Buuren and Groothuis-Oudshoorn, 2011) properly combined with software for posterior model probabilities (such as `BayesVarsel` by García-Donato and Forte, 2018). However, a close examination of the posterior predictive distribution reveals that it depends on the posterior probability of the model, that is

$$m(\mathbf{d}_{(1)} \mid \mathbf{d}_{(0)}) = \sum_{\gamma \in \Gamma} m_\gamma(\mathbf{d}_{(1)} \mid \mathbf{d}_{(0)})\, p(\gamma \mid \mathbf{d}_{(0)}). \tag{2}$$

The fact that the target probability $p(\gamma \mid \mathbf{d}_{(0)})$ appears on both sides of Equation (1) hampers the formulation of the mentioned strategies, *a la* Rubin's rules, simply because the distribution for imputation (2) is unknown. Unavoidably, a single model must be used for the imputation step (consciously or unconsciously), leading to a methodology

that is not endorsed by the probabilistic equation in (1). This is the basis of the "Impute Then Select" method in Yang et al. (2005).

Alternatively, as is routinely performed in estimation problems, we can envisage a Gibbs sampling algorithm where $(\gamma, \mathbf{d}_{(1)} \mid \mathbf{d}_{(0)})$ are jointly drawn from their full conditional distributions $(\gamma \mid \mathbf{d}_{(1)}, \mathbf{d}_{(0)})$ and $(\mathbf{d}_{(1)} \mid \gamma, \mathbf{d}_{(0)})$. This is the basis of the method "Simultaneously Impute And Select" (SIAS) proposed in Yang et al. (2005). Implementing such a strategy requires supplementing the main simulation algorithm with embedded steps that sample the model-specific unknown parameters from their posterior distributions. A complicating factor is that the model changes from iteration to iteration, potentially resulting in a simulation scheme with poor mixing properties. A theoretical formalization of this conjecture goes beyond the scope of this paper, but provides a possible explanation for the clear inferiority of SIAS shown in the numerical study in Section 6.1.

The starting point of our research follows directly from the application of Bayes' theorem:

$$p(\gamma \mid \mathbf{d}_{(0)}) = \frac{m_\gamma(\mathbf{d}_{(0)})p(\gamma)}{\sum_{\gamma' \in \Gamma} m_{\gamma'}(\mathbf{d}_{(0)})p(\gamma')}, \tag{3}$$

where $m_\gamma(\mathbf{d}_{(0)})$ is the density of the observations under model γ (i.e. the normalizing constant of model γ posterior parameters or the integrated likelihood of model parameters with respect to their prior distribution) and $p(\gamma)$ is the prior probability of model γ. Obviously, (1) and (3) are equivalent, but the latter focuses on marginals—which are key quantities in our research—and, in principle, does not have an interpretation in terms of Rubin's rules.

3 Model choice in regression settings with missing data

3.1 Notation and first considerations

In a regression problem, the data consist of samples of size n of a dependent variable, and k explanatory variables collected in corresponding vectors $\boldsymbol{y}$ and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$. We assume that all variables are quantitative and denote x_{ij}, $1 \leq i \leq n$, $1 \leq j \leq k$ the i-th component of vector $\boldsymbol{x}_j$. The case with qualitative regressors is discussed in Section 7. To identify missing values, we introduce the binary matrix $M = (M_{ij})$ of dimension $n \times (k+1)$. A value of $M_{ij} = 1$ for $j \leq k$ indicates that x_{ij} is unavailable (to the analyst), whereas $M_{ij} = 0$ indicates that it is available. In this notation, column $k+1$ of M represents the missingness of the response variable $\boldsymbol{y}$. Following this (0)/(1) notation, we denote $\boldsymbol{x}_{(0)} = \{x_{ij} : M_{ij} = 0, \text{ for } 1 \leq i \leq n, 1 \leq j \leq k\}$ and $\boldsymbol{y}_{(0)} = \{y_i : M_{i,k+1} = 0, \text{ for } 1 \leq i \leq n\}$ and, similarly, $\boldsymbol{x}_{(1)} = \{x_{ij} : M_{ij} = 1, \text{ for } 1 \leq i \leq n, 1 \leq j \leq k\}$ and $\boldsymbol{y}_{(1)} = \{y_i : M_{i,k+1} = 1, \text{ for } 1 \leq i \leq n\}$. The length of $\boldsymbol{y}_{(0)}$ is labeled n_0. The correspondence of this notation with the one introduced in the previous—more general—section is $\mathbf{d}_{(0)} \equiv (\boldsymbol{y}_{(0)}, \boldsymbol{x}_{(0)}, M)$ and the non-observed, random values as $\mathbf{d}_{(1)} \equiv (\boldsymbol{y}_{(1)}, \boldsymbol{x}_{(1)})$.

Additionally, we let X be the $n \times k$ matrix with entries x_{ij} and $\overline{X}$ the matrix with the same entries centered around their column means (i.e. with elements $x_{ij} -$

$n^{-1}\sum_{i=1}^{n} x_{ij}$). Without missing data and after data collection, these matrices will be known. In contrast, with missing data, some cells in X and/or $\overline{X}$ will remain random. Finally, we define X_{n_0} as the $n_0 \times k$ matrix with entries x_{ij}, but only for those i where the dependent variable has been observed, and a similar notation for $\overline{X}_{n_0}$ (the matrix with entries $x_{ij} - n_0^{-1}\sum_{i=1,M_{i,k+1}=0}^{n_0} x_{ij}$).

In this setting, the competing models are indexed by $\boldsymbol{\gamma} \in \Gamma$, where $\boldsymbol{\gamma}$ is a binary vector of dimension k. Each component of $\boldsymbol{\gamma}$ takes the value 1 if the corresponding covariate is included in the model, and 0 otherwise. The bold notation emphasizes its vector nature. We then assume that models $\boldsymbol{\gamma} \in \Gamma$ can be expressed as:

$$
\begin{aligned}
f_{\boldsymbol{\gamma}}(\boldsymbol{y}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k, M \mid \boldsymbol{\alpha}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\nu}, \boldsymbol{\psi}) &= f_{\boldsymbol{\gamma}}(\boldsymbol{y} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k, \boldsymbol{\alpha}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}) && (4)\\
&\times \prod_{i=1}^{n} f(x_{i1}, x_{i2}, \ldots, x_{ik} \mid \boldsymbol{\nu}) && (5)\\
&\times f(M \mid \boldsymbol{y}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k, \boldsymbol{\psi}). && (6)
\end{aligned}
$$

Notice that competing models only differ on (4), that is, on how the covariates influence y. This conditional distribution is presented in a general manner to cover a wide range of situations and, in particular, the two that will be treated in detail in Section 4 (variable selection) and Section 5 (concerning competing models agreeing on the relevant covariates but disagreeing on the density).

Regarding the regression parameters, we denote those appearing in all competing models (if any) as $\boldsymbol{\alpha}$ (e.g., the intercept) and those specific to $\boldsymbol{\gamma}$ (such as the regression parameter corresponding to an included variable) as $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$. In this sense, we will refer to the *null model* as the model that depends only on $\boldsymbol{\alpha}$, since none of the covariates are included.

Notice that in the introduction of the models, we have opted for a nonstandard notation in which all covariates appear in the conditioning. This does not necessarily imply an effective dependence of y on *all* k covariates on all models in Γ (usually certain models in the model space will not depend on any covariate, e.g. a model with only the intercept, and even certain covariates in the database will not appear in any competing model). The reason for this additional complexity becomes clear when the missing data problem develops. However, we anticipate that variables that do not appear in any competing model for y may still be valuable in the imputation process (5).

Much of the literature assumes that the values of the covariates are known (either because the data come from a designed experiment or as a simplification), in which case, (5) won't be considered. The marginal to be inserted in (3) would be

$$
m_{\boldsymbol{\gamma}}(\boldsymbol{y} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k) = \int f_{\boldsymbol{\gamma}}(\boldsymbol{y} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k, \boldsymbol{\alpha}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}) \pi_{\boldsymbol{\gamma}}(\boldsymbol{\alpha}, \boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k) \mathrm{d}[\boldsymbol{\alpha}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}], \quad (7)
$$

where $\pi_{\boldsymbol{\gamma}}(\boldsymbol{\alpha}, \boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$ is a prior based on a fixed design matrix; for example, g-Zellner type priors (Zellner, 1986; Zellner and Siow, 1980; Bayarri et al., 2012; Liang et al., 2008; Fernández et al., 2001), spike and slab priors (Ishwaran and Rao, 2005), and non-local priors (Johnson and Rossell, 2010), among others.

However, for the case with missing data, it is customary to consider the covariates as random as it is explicitly assumed with (5) (sometimes referred to as the imputation model). It plays a central role in dealing with missing data, and there is substantial literature on imputation models to accommodate different types of variables, as we mentioned in the introduction. Any of these can be used as desired without affecting the methods in this paper, with the only condition of the existence of the mean vector

$$\boldsymbol{\mu} \equiv \boldsymbol{\mu}(\boldsymbol{\nu}) = E\{(x_1, \ldots, x_k)^t \mid \boldsymbol{\nu}\}, \tag{8}$$

and the variance-covariance matrix

$$\boldsymbol{\Sigma} \equiv \boldsymbol{\Sigma}(\boldsymbol{\nu}) = E\big\{[\big((x_1, \ldots, x_k)^t - \boldsymbol{\mu}\big)\big((x_1, \ldots, x_k)^t - \boldsymbol{\mu}\big)^t \mid \boldsymbol{\nu}\big\}. \tag{9}$$

We will use a multivariate normal imputation model in the section devoted to experiments.

Finally, to complete the probabilistic structure of the competing models, we must specify how the missing observations occur. The generally accepted framework that we adopt was introduced in Little and Rubin (2020) and assumes that originally data are fully observed but some observations are hidden for the analyst. The process by which some observations are hidden is unknown, which leads us to consider M as a random matrix whose modeling is specified in (6) where $\boldsymbol{\psi}$ are the parameters governing the missing mechanism.

For every model $\boldsymbol{\gamma} \in \Gamma$—defined in (4)–(6)—and given a prior distribution for the unknown parameters $\pi_{\boldsymbol{\gamma}}(\boldsymbol{\alpha}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\nu}, \boldsymbol{\psi})$, we can now proceed to obtain the posterior probability for each model using Equation (3) where

$$\begin{aligned} m_{\boldsymbol{\gamma}}(\mathbf{d}_{(0)}) = m_{\boldsymbol{\gamma}}(\boldsymbol{y}_{(0)}, \boldsymbol{x}_{(0)}, M) = \\ \int f_{\boldsymbol{\gamma}}(\boldsymbol{y}_{(0)}, \boldsymbol{y}_{(1)} \mid \boldsymbol{x}_{(0)}, \boldsymbol{x}_{(1)}, \boldsymbol{\alpha}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}) f(\boldsymbol{x}_{(0)}, \boldsymbol{x}_{(1)} \mid \boldsymbol{\nu}) f(M \mid \boldsymbol{y}_{(0)}, \boldsymbol{y}_{(1)}, \boldsymbol{x}_{(0)}, \boldsymbol{x}_{(1)}, \boldsymbol{\psi}) \\ \times \pi_{\boldsymbol{\gamma}}(\boldsymbol{\alpha}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\nu}, \boldsymbol{\psi}) \mathrm{d}[\boldsymbol{\alpha}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\nu}, \boldsymbol{\psi}, \boldsymbol{x}_{(1)}, \boldsymbol{y}_{(1)}]. \end{aligned} \tag{10}$$

3.2 Ignorability of the missing data mechanism

We inspect the conditions under which Bayes factor, and hence the posterior probabilities of models, remains unaffected by the specific form of the missing data mechanism (6).

According to van Buuren (2018), missing at random (MAR) holds if

$$f(M \mid \boldsymbol{y}_{(0)}, \boldsymbol{x}_{(0)}, \boldsymbol{y}_{(1)}, \boldsymbol{x}_{(1)}, \boldsymbol{\psi}) = f(M \mid \boldsymbol{y}_{(0)}, \boldsymbol{x}_{(0)}, \boldsymbol{\psi}), \tag{11}$$

while missing completely at random (MCAR) holds if the distribution of M does not depend on the observed data, i.e.

$$f(M \mid \boldsymbol{y}_{(0)}, \boldsymbol{x}_{(0)}, \boldsymbol{y}_{(1)}, \boldsymbol{x}_{(1)}, \boldsymbol{\psi}) = f(M \mid \boldsymbol{\psi}).$$

In the next result we prove that, under the MAR assumption (or the more restrictive MCAR), the distribution of M is ignorable.

Proposition 1. If we assume MAR, and for all $\boldsymbol{\gamma} \in \Gamma$ the prior distribution satisfies:

$$\pi_{\boldsymbol{\gamma}}(\boldsymbol{\alpha}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\nu}, \boldsymbol{\psi}) = \pi_{\boldsymbol{\gamma}}(\boldsymbol{\alpha}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\nu})\pi(\boldsymbol{\psi}). \tag{12}$$

then

$$m_{\boldsymbol{\gamma}}(\boldsymbol{y}_{(0)}, \boldsymbol{x}_{(0)}, M) \propto m_{\boldsymbol{\gamma}}(\boldsymbol{y}_{(0)}, \boldsymbol{x}_{(0)}), \tag{13}$$

where

$$m_{\boldsymbol{\gamma}}(\boldsymbol{y}_{(0)}, \boldsymbol{x}_{(0)}) = \int f_{\boldsymbol{\gamma}}(\boldsymbol{y}_{(0)} \,|\, \boldsymbol{x}_{(0)}, \boldsymbol{x}_{(1)}, \boldsymbol{\alpha}, \boldsymbol{\beta}_{\boldsymbol{\gamma}})\, f(\boldsymbol{x}_{(0)}, \boldsymbol{x}_{(1)} \,|\, \boldsymbol{\nu})\pi_{\boldsymbol{\gamma}}(\boldsymbol{\alpha}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\nu})\mathrm{d}[\boldsymbol{\alpha}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\nu}, \boldsymbol{x}_{(1)}]. \tag{14}$$

Proof. The proof can be found in Section B.1 of the Supplementary material. □

Note that missing values in the response, $\boldsymbol{y}_{(1)}$, do not affect the marginal but their observed covariates still affect the marginal and thus the model selection problem.

Condition (12) requires that the parameters in the regression and imputation components of the model are independent (a priori) of the parameters governing the missing data mechanism. In what follows in this paper, we assume MAR and that this condition holds. This combination exempts us from specifying (6) and $\pi(\boldsymbol{\psi})$ in virtue of Proposition 1. For a missing not at random mechanism, in which M depends on unobserved data, the proposition does not hold, and a model for M would have to be specified. We will revisit this hypothesis in the discussion in the concluding section.

3.3 A recognizable expression for marginals with missing data

Normally, procedures for dealing with missing data follow the logic of being extensions of a complete data method, with missing values replaced by some type of imputation. At first glance, the relevant marginal $m_{\boldsymbol{\gamma}}(\boldsymbol{y}_{(0)}, \boldsymbol{x}_{(0)})$ defined in (14), shows no evidence of this logic. Next, we derived an equivalent expression interpreted in this manner.

Result 1. Up to a proportionality constant common to all models and factorizing (12) as

$$\pi_{\boldsymbol{\gamma}}(\boldsymbol{\alpha}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\nu}, \boldsymbol{\psi}) = \pi_{\boldsymbol{\gamma}}(\boldsymbol{\alpha}, \boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \boldsymbol{\nu})\pi(\boldsymbol{\nu})\pi(\boldsymbol{\psi}), \tag{15}$$

an equivalent expression for (14) is

$$m_{\boldsymbol{\gamma}}(\boldsymbol{y}_{(0)}, \boldsymbol{x}_{(0)}) \propto \int \mathfrak{m}_{\boldsymbol{\gamma}}(\boldsymbol{y}_{(0)} \mid \boldsymbol{x}_{(0)}, \boldsymbol{x}_{(1)}, \boldsymbol{\nu})\pi(\boldsymbol{x}_{(1)}, \boldsymbol{\nu} \mid \boldsymbol{x}_{(0)})\, \mathrm{d}[\boldsymbol{x}_{(1)}, \boldsymbol{\nu}] \tag{16}$$

where $\pi(\boldsymbol{x}_{(1)}, \boldsymbol{\nu} \mid \boldsymbol{x}_{(0)})$ is the posterior distribution of $(\boldsymbol{x}_{(1)}, \boldsymbol{\nu})$ given $\boldsymbol{x}_{(0)}$, and

$$\mathfrak{m}_{\boldsymbol{\gamma}}(\boldsymbol{y}_{(0)} \mid \boldsymbol{x}_{(0)}, \boldsymbol{x}_{(1)}, \boldsymbol{\nu}) = \int f_{\boldsymbol{\gamma}}(\boldsymbol{y}_{(0)} \mid \boldsymbol{x}_{(0)}, \boldsymbol{x}_{(1)}, \boldsymbol{\alpha}, \boldsymbol{\beta}_{\boldsymbol{\gamma}})\pi_{\boldsymbol{\gamma}}(\boldsymbol{\alpha}, \boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \boldsymbol{\nu})\mathrm{d}[\boldsymbol{\alpha}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}]. \tag{17}$$

If model $\boldsymbol{\gamma}$ has only common parameters, then the expression becomes:

$$\mathfrak{m}_{\boldsymbol{\gamma}}(\boldsymbol{y}_{(0)} \mid \boldsymbol{x}_{(0)}, \boldsymbol{x}_{(1)}, \boldsymbol{\nu}) = \int f_{\boldsymbol{\gamma}}(\boldsymbol{y}_{(0)} \mid \boldsymbol{x}_{(0)}, \boldsymbol{x}_{(1)}, \boldsymbol{\alpha})\pi_{\boldsymbol{\gamma}}(\boldsymbol{\alpha} \mid \boldsymbol{\nu})\mathrm{d}\boldsymbol{\alpha}. \tag{18}$$

Proof. The proof can be found in Section B.2 of the Supplementary material. □

In Equation (17), $\mathfrak{m}_{\boldsymbol{\gamma}}$—a function of $\boldsymbol{\nu}$ and $\boldsymbol{x}_{(1)}$—is the "missing data" counterpart of the corresponding marginal used in the full data case for the calculation of BF (cf. Equation 7), that is $m_{\boldsymbol{\gamma}}(\boldsymbol{y} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$. Thus, Equation (16) states that $m_{\boldsymbol{\gamma}}(\boldsymbol{y}_{(0)}, \boldsymbol{x}_{(0)})$ is the expected value of such "missing data marginal" with respect to the posterior distribution $\pi(\boldsymbol{x}_{(1)}, \boldsymbol{\nu} \mid \boldsymbol{x}_{(0)})$ (which does not depend on $\boldsymbol{\gamma}$ and that only involves observed covariates).

For a generic model $\boldsymbol{\gamma}$ with no missing data in the corresponding included covariates, it is clear that $\mathfrak{m}_{\boldsymbol{\gamma}}(\boldsymbol{y}_{(0)} \mid \boldsymbol{x}_{(0)}, \boldsymbol{x}_{(1)}, \boldsymbol{\nu}) \equiv \mathfrak{m}_{\boldsymbol{\gamma}}(\boldsymbol{y}_{(0)} \mid \boldsymbol{x}_{(0)}, \boldsymbol{\nu})$. Therefore, after integrating $\boldsymbol{x}_{(1)}$ into Equation (16), a simpler expression for $m_{\boldsymbol{\gamma}}(\boldsymbol{y}_{(0)}, \boldsymbol{x}_{(0)})$ can be derived as

$$m_{\boldsymbol{\gamma}}(\boldsymbol{y}_{(0)}, \boldsymbol{x}_{(0)}) \propto \int \mathfrak{m}_{\boldsymbol{\gamma}}(\boldsymbol{y}_{(0)} \mid \boldsymbol{x}_{(0)}, \boldsymbol{\nu})\, \pi(\boldsymbol{\nu} \mid \boldsymbol{x}_{(0)})\, \mathrm{d}\boldsymbol{\nu}. \tag{19}$$

3.4 Computing the marginal by simulation

If a manageable expression for $\mathfrak{m}_{\boldsymbol{\gamma}}(\boldsymbol{y}_{(0)} \mid \boldsymbol{x}_{(0)}, \boldsymbol{x}_{(1)}, \boldsymbol{\nu})$ is available, the marginal $m_{\boldsymbol{\gamma}}(\boldsymbol{y}_{(0)}, \boldsymbol{x}_{(0)})$ can be approximated with a Monte Carlo-based routine, as follows:

For $j = 1, \ldots, J$:

Step 1: Draw $\boldsymbol{\nu}^{(j)} \sim \pi(\boldsymbol{\nu} \mid \boldsymbol{x}_{(0)})$,

Step 2: Draw $\boldsymbol{x}_{(1)}^{(j)} \sim f(\boldsymbol{x}_{(1)} \mid \boldsymbol{x}_{(0)}, \boldsymbol{\nu}^{(j)})$,

Step 3: Calculate $\mathfrak{m}^{(j)} = \mathfrak{m}_{\boldsymbol{\gamma}}(\boldsymbol{y}_{(0)} \mid \boldsymbol{x}_{(0)}, \boldsymbol{x}_{(1)}^{(j)}, \boldsymbol{\nu}^{(j)})$,

then compute $m_{\boldsymbol{\gamma}}(\boldsymbol{y}_{(0)}, \boldsymbol{x}_{(0)}) \approx J^{-1} \sum \mathfrak{m}^{(j)}$. The implementation of Step 2 can be approached with standard augmented Gibbs schemes (see, for instance, Hoff, 2009).

In connection with the comment made about the possible poor mixing properties of methods like SIAS (Yang et al., 2005), note that in our approach the simulations are conditional on a given model, so the possibility of poor mixing disappears.

3.5 Objective prior distributions on model parameters: general considerations

The standard Bayesian method for addressing the absence of prior information uses improper distributions. In estimation problems (the model is fixed), the impropriety of priors does not imply any additional difficulty as long as the posterior is proper. There is a large body of literature regarding which priors are best suited to different models (consult the catalogue Yang and Berger, 1997). Many of these can be obtained with mathematical rules (like Jeffreys' priors or reference priors; see Kass and Wasserman,

1996). We refer to such (objective for estimation and usually improper non-informative) priors with the superindex N.

In the case of model uncertainty, the situation is quite different, and priors need to be carefully specified. In the Supplementary material, Section D, we discuss in depth about the structure of the prior, finally recommending:

$$\pi_{\boldsymbol{\gamma}}(\boldsymbol{\alpha}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\nu}) = \pi^N(\boldsymbol{\alpha})\pi^N(\boldsymbol{\nu})\pi_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \boldsymbol{\nu}, \boldsymbol{\alpha}), \tag{20}$$

reducing to

$$\pi_{\boldsymbol{\gamma}}(\boldsymbol{\alpha}, \boldsymbol{\nu}) = \pi^N(\boldsymbol{\alpha})\pi^N(\boldsymbol{\nu}) \tag{21}$$

for models with only common parameters. Above, the only ingredient that remains unspecified is $\pi_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \boldsymbol{\nu}, \boldsymbol{\alpha})$ which must be proper. We determine this distribution for the two problems considered in this paper in the following sections.

4 Variable selection in the general linear model

4.1 Model comparison and Zellner's g-prior

In variable selection, the standard notation used to index models identifies which covariates are active and which are not. We also follow this convention and define $\boldsymbol{\gamma}$ (note the bold symbol) as a k-dimensional binary vector. Then, for $\boldsymbol{\gamma} \neq \mathbf{0}$ let $X_{\boldsymbol{\gamma}}$ (and $\overline{X}_{\boldsymbol{\gamma}}$, $X_{\boldsymbol{\gamma},n_0}$, $\overline{X}_{\boldsymbol{\gamma},n_0}$) be the sub-matrix of X (respectively $\overline{X}$, X_{n_0} and $\overline{X}_{n_0}$) containing the $k_{\boldsymbol{\gamma}} = \sum_{j=1}^{k} \gamma_j$ columns corresponding to the ones in $\boldsymbol{\gamma}$.

Let's consider the problem of selecting between two models of the form (4)–(6) where $\Gamma = \{\mathbf{0}, \boldsymbol{\gamma}\}$ with

$$f_0(\boldsymbol{y} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k, \beta_0, \sigma) = N_n(\boldsymbol{y} \mid \beta_0 \mathbf{1}, \sigma^2 I)$$

and

$$f_{\boldsymbol{\gamma}}(\boldsymbol{y} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k, \beta_0, \sigma, \boldsymbol{\beta}_{\boldsymbol{\gamma}}) = N_n(\boldsymbol{y} \mid \beta_0 \mathbf{1} + X_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2 I),$$

where $\mathbf{1}$ and I represent the vector of ones and the identity matrix with conformably dimensions. Notice here that the common parameters among all models are the intercept, β_0, and the standard deviation, σ, i.e. $\boldsymbol{\alpha} = (\beta_0, \sigma)$.

For the traditional case where X was considered non-random (and obviously there were no-missing values), Zellner (1986) proposed the prior $N_{k_{\boldsymbol{\gamma}}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \mathbf{0}, g\,\sigma^2\,[\overline{X}_{\boldsymbol{\gamma}}^t \overline{X}_{\boldsymbol{\gamma}}]^{-1})$ with $g = n$. This prior, that has been also called the unit information prior (Kass and Wasserman, 1995), leads to a Bayes factor of $M_{\boldsymbol{\gamma}}$ to M_0 (see e.g. Liang et al., 2008):

$$\left(1 + n\frac{S_{\boldsymbol{\gamma}}}{S_0}\right)^{-(n-1)/2} \left(1 + n\right)^{(n-k_{\boldsymbol{\gamma}}-1)/2}, \tag{22}$$

where $S_{\boldsymbol{\gamma}}$ is the residual sum of squared errors for $\boldsymbol{\gamma}$.

4.2 Imputation g'-prior

Definition 4.1. We define the imputation g'-prior as

$$\pi_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \beta_0, \sigma, \boldsymbol{\nu}) = N_{k_{\boldsymbol{\gamma}}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \mathbf{0}, g'\,\sigma^2\,\Sigma^{-1}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}), \tag{23}$$

where $\Sigma_{\boldsymbol{\gamma}\boldsymbol{\gamma}}$ denotes the corresponding $k_{\boldsymbol{\gamma}} \times k_{\boldsymbol{\gamma}}$ block matrix from Σ—the variance covariance matrix of the covariates under the imputation model defined in (9).

The arguments that lead to the definition of the imputation g'-prior are elaborated in Section 4.3.

The prior just introduced (despite the label 'imputation' in its name) does not strictly require that the data are affected by missingness. It is conceived for when the covariates are assumed to be random (a hypothesis that can arguably be considered as more realistic than the standard one of a fixed design). In fact, g-Zellner prior can be interpreted as an empirical version of the imputation g'-prior with $g' = 1$, as the variance-covariance matrix in the g-prior converges to that of the imputation g'-prior as n grows. This limiting agreement justifies our conventional choice $g' = 1$ that we use in our numerical experiments. Alternatively, g' can be seen as a hyperparameter—as in e.g. Liang et al. (2008) or Bayarri et al. (2012)—with prior distribution $g' \sim \pi(g')$ leading to an hyper-g' imputation prior. We leave the exploration of this further generalization for a future research.

Proposition 2. The imputed g'-Bayes factor—corresponding to the prior under the scheme (20), (21) and (23) and $\pi^N(\beta_0, \sigma) = \sigma^{-1}$— $B_{\boldsymbol{\gamma}0}(\boldsymbol{y}_{(0)}, \boldsymbol{x}_{(0)})$ is

$$E\ \Big\{\ \Big[\frac{S_0}{S_0 - \boldsymbol{y}^t_{(0)}\overline{X}_{\boldsymbol{\gamma},n_0}\big(\overline{X}^t_{\boldsymbol{\gamma},n_0}\overline{X}_{\boldsymbol{\gamma},n_0} + \frac{1}{g'}\Sigma_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\big)^{-1}\overline{X}^t_{\boldsymbol{\gamma},n_0}\boldsymbol{y}_{(0)}}\Big]^{(n_0-1)/2} \times \Big|\overline{X}^t_{\boldsymbol{\gamma},n_0}\overline{X}_{\boldsymbol{\gamma},n_0}\ \Sigma^{-1}_{\boldsymbol{\gamma}\boldsymbol{\gamma}} + \frac{I}{g'}\Big|^{-1/2}\Big\}, \tag{24}$$

where the expectation is with respect to the posterior distribution $\pi^N(\boldsymbol{x}_{(1)}, \boldsymbol{\nu} \mid \boldsymbol{x}_{(0)})$, and S_0 is $n_0 - 1$ times the sample variance of $\boldsymbol{y}_{(0)}$.

Proof. The proof can be found in Section B.3 of the Supplementary material. □

The imputed g'-Bayes factor generalizes the one obtained with g-Zellner prior to the case where the covariates are considered random and data is fully observed. In fact, for the complete data case, Zellner's-g Bayes factor (cf. equation 22) can be seen as an *empirical* version of $B_{\boldsymbol{\gamma}0}(\boldsymbol{y}_{(0)}, \boldsymbol{x}_{(0)})$. This interpretation holds because (24) reduces to (22) if $g' = 1$ and $\Sigma_{\boldsymbol{\gamma}\boldsymbol{\gamma}}$ is replaced by its *sample version* $n^{-1}\overline{X}^t_{\boldsymbol{\gamma}}\overline{X}_{\boldsymbol{\gamma}}$ (for details, see Result B.3 in Supplementary material, Section B.3). Asymptotically, the *empirical* interpretation can be made *empirical Bayes*. This is because, under very mild conditions on the prior $\pi^N(\boldsymbol{\nu})$, the matrix $n^{-1}\overline{X}^t_{\boldsymbol{\gamma}}\overline{X}_{\boldsymbol{\gamma}}$ and the posterior mean of $\Sigma_{\boldsymbol{\gamma}\boldsymbol{\gamma}}$ (with respect to the posterior distribution $\pi^N(\boldsymbol{\nu} \mid \boldsymbol{x})$) asymptotically coincide.

4.3 The construction of the imputation g'-prior

With respect to the common parameters (β_0, σ), as discussed in the Supplementary material, the use of the same prior in both models is reasonable if these parameters represent similar magnitudes in both models requiring a reparameterization in the model. In particular, we need to reparameterize the intercept to justify the assumption of a similar meaning. The idea is to transfer the mean of $\boldsymbol{x}$ to the intercept such that it has zero mean, as follows:

$$\begin{aligned}\beta_0\mathbf{1}+X_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}} &= \beta_0\mathbf{1}+X_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}+\boldsymbol{\mu}_{\boldsymbol{\gamma}}^t\boldsymbol{\beta}_{\boldsymbol{\gamma}}\mathbf{1}-\boldsymbol{\mu}_{\boldsymbol{\gamma}}^t\boldsymbol{\beta}_{\boldsymbol{\gamma}}\mathbf{1}=(\beta_0+\boldsymbol{\mu}_{\boldsymbol{\gamma}}^t\boldsymbol{\beta}_{\boldsymbol{\gamma}})\mathbf{1}+(X_{\boldsymbol{\gamma}}-\mathbf{1}\boldsymbol{\mu}_{\boldsymbol{\gamma}}^t)\boldsymbol{\beta}_{\boldsymbol{\gamma}}= \\ &\stackrel{\text{def}}{=} \beta_0^\star\mathbf{1}+(X_{\boldsymbol{\gamma}}-\mathbf{1}\boldsymbol{\mu}_{\boldsymbol{\gamma}}^t)\boldsymbol{\beta}_{\boldsymbol{\gamma}},\end{aligned}$$

where $\boldsymbol{\mu}\equiv\boldsymbol{\mu}(\boldsymbol{\nu})=E\{\boldsymbol{x}\mid\boldsymbol{\nu}\}$ is the mean of $\boldsymbol{x}$ as obtained from the imputation model (5) and $\boldsymbol{\mu}_{\boldsymbol{\gamma}}$ is the corresponding elements of $\boldsymbol{\mu}$ in model $\boldsymbol{\gamma}$. With this reparameterization, the model $\boldsymbol{\gamma}$ is redefined as:

$$f_{\boldsymbol{\gamma}}(\boldsymbol{y}\mid\boldsymbol{x}_1,\ldots,\boldsymbol{x}_k,\beta_0^\star,\boldsymbol{\beta}_{\boldsymbol{\gamma}},\sigma)=N_n(\boldsymbol{y}\mid\beta_0^\star\mathbf{1}+(X_{\boldsymbol{\gamma}}-\mathbf{1}\boldsymbol{\mu}_{\boldsymbol{\gamma}}^t)\boldsymbol{\beta}_{\boldsymbol{\gamma}},\sigma^2I). \tag{25}$$

Now, the parameter $\beta_0^\star$ (in $\boldsymbol{\gamma}$) is the mean of y when the values of the covariates coincide with their expectations, which aligns with the meaning of β_0 in the null model (which represents the mean of y regardless of the values of the covariates). This justifies using the same prior distribution (informative or objective) for β_0 and $\beta_0^\star$. Note that this result is achieved when the columns of $X_{\boldsymbol{\gamma}}$ are centered with respect to their expectations, which is the counterpart to centering with respect to their sample means, as is routinely done in the literature.

The above argument is rather informal but was used in the early literature on Bayesian testing, such as Jeffreys (1961) or Zellner and Siow (1980). More recently, Kass and Raftery (1995) worked on formalizing the concept of common parameters with similar meanings. They reasoned that such an assumption is sensible when the common and new parameters are orthogonal (i.e., the expected Fisher information matrix is block diagonal). In this case, the common parameters represent the same quantities, opening the possibility of using the same prior for both. When the covariates are random, the expected Fisher information matrix, $\mathfrak{I}$, for the parameters involved in the regression component of the model $\boldsymbol{\gamma}$ (after the integration of $\boldsymbol{y}$) is obtained over the imputation model of the covariates. In particular,

$$\mathfrak{I}=\frac{1}{\sigma^2}E\left\{\begin{array}{ccc} n & \mathbf{1}^t(X_{\boldsymbol{\gamma}}-\mathbf{1}\boldsymbol{\mu}_{\boldsymbol{\gamma}}^t) & 0 \\ (X_{\boldsymbol{\gamma}}-\mathbf{1}\boldsymbol{\mu}_{\boldsymbol{\gamma}}^t)^t\mathbf{1} & (X_{\boldsymbol{\gamma}}-\mathbf{1}\boldsymbol{\mu}_{\boldsymbol{\gamma}}^t)^t(X_{\boldsymbol{\gamma}}-\mathbf{1}\boldsymbol{\mu}_{\boldsymbol{\gamma}}^t) & \mathbf{0} \\ 0 & \mathbf{0}^t & 2n \end{array}\right\}=\frac{n}{\sigma^2}\big(1\oplus\Sigma_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\oplus 2\big), \tag{26}$$

where $\Sigma_{\boldsymbol{\gamma}\boldsymbol{\gamma}}$ denotes the $k_{\boldsymbol{\gamma}}\times k_{\boldsymbol{\gamma}}$ block diagonal from $\Sigma\equiv\Sigma(\boldsymbol{\nu})=V(\boldsymbol{x}\mid\boldsymbol{\nu})$ corresponding to the active variables in $\boldsymbol{\gamma}$. We conclude that $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and $(\beta_0^\star,\sigma)$ are orthogonal, and that if $\pi_0(\beta_0,\sigma)$ is used for the null model, we can use $\pi_{\boldsymbol{\gamma}}(\beta_0^\star,\sigma)=\pi_0(\beta_0^\star,\sigma)$ for the alternative model. Note that this orthogonality does not hold for the original parameterization (β_0,σ). In the absence of prior information, the obvious choice in this case is to use the

reference priors $\pi_0(\beta_0, \sigma \mid \boldsymbol{\nu}) = \sigma^{-1}$ and $\pi_{\boldsymbol{\gamma}}(\beta_0^\star, \sigma \mid \boldsymbol{\nu}) = \sigma^{-1}$, which do not depend on the parameters of the distribution for the covariates, $\boldsymbol{\nu}$.

Once we have established the prior for the common parameters, we now determine the prior $\pi_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \beta_0^\star, \sigma, \boldsymbol{\nu})$. The extensive literature on g-priors agrees that we should use a $k_{\boldsymbol{\gamma}}$-multivariate normal density (perhaps mixed to obtain flat tails) centered at zero and with a unitary covariance matrix. This matrix is defined as the block corresponding to the inverse of the Fisher information matrix multiplied by sample size, n. For a complete dataset, this route leads to the use of $n\sigma^2\,(\overline{X}_{\boldsymbol{\gamma}}^t \overline{X}_{\boldsymbol{\gamma}})^{-1}$ (where $\overline{X}_{\boldsymbol{\gamma}}$ has columns centered around the sample mean), as proposed in Zellner and Siow (1980) and unanimously followed in the related research (see Bayarri et al., 2012, and references therein).

Mimicking this path in the case of missing data (or more in general for random covariates) is straightforward because we now have the expected Fisher information matrix. Furthermore, obtaining the inverse is rather simple because the matrix is block diagonal as a consequence of reparameterization (cf. Equation 26), leading to $n\frac{\sigma^2}{n}\Sigma_{\boldsymbol{\gamma}\boldsymbol{\gamma}}^{-1} = \sigma^2\,\Sigma_{\boldsymbol{\gamma}\boldsymbol{\gamma}}^{-1}$. Remarkably, the sample size does not enter in the expression leading to (23) with $g' = 1$, as proposed. Finally, putting it all together, we arrive at the prior

$$\pi_{\boldsymbol{\gamma}}(\beta_0^\star, \sigma, \boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \boldsymbol{\nu}) = \sigma^{-1} \times N_{k_{\boldsymbol{\gamma}}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \mathbf{0}, g'\sigma^2\Sigma_{\boldsymbol{\gamma}\boldsymbol{\gamma}}^{-1}).$$

It is straightforward to prove that this prior, expressed in the original parameterization of the model, leads to

$$\pi_{\boldsymbol{\gamma}}(\beta_0, \sigma, \boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \boldsymbol{\nu}) = \sigma^{-1} N_{k_{\boldsymbol{\gamma}}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \mathbf{0}, g'\sigma^2\Sigma_{\boldsymbol{\gamma}\boldsymbol{\gamma}}^{-1}),$$

since the Jacobian corresponding to the change of variables $\{\beta_0, \sigma, \boldsymbol{\beta}_{\boldsymbol{\gamma}}\} \rightarrow \{\beta_0^\star, \sigma, \boldsymbol{\beta}_{\boldsymbol{\gamma}}\}$—where recall $\beta_0^\star = \beta_0 + \boldsymbol{\mu}_{\boldsymbol{\gamma}}^t\boldsymbol{\beta}_{\boldsymbol{\gamma}}$—is 1.

4.4 Variable Selection

The basis for developing VS methods in the context of missing data is the two-model selection problem described in Section 4.1. In VS, the goal is to find which of the covariates $\{x_1, \ldots, x_k\}$ have a real effect on the response, y.

The list of possible models can be expressed using the binary parameter vector $\boldsymbol{\gamma}^t = (\gamma_1, \ldots, \gamma_k)$, where $\gamma_j = 1$ if the response depends on x_j and zero otherwise. For example, a model with only x_2 corresponds to $\boldsymbol{\gamma}^t = (0, 1, 0, \ldots, 0)$. The set of possible models is denoted by Γ and its cardinality is 2^k, considering only the main effects. The posterior probability of each model $\boldsymbol{\gamma}$, as shown in Equation (3), depends on the prior probabilities over the model space. Some objective prior proposals are uniform, $p(\boldsymbol{\gamma}) = 1/2^k$ for $\boldsymbol{\gamma} \in \Gamma$, or the hierarchical uniform prior discussed by Scott and Berger (2010): $p(\boldsymbol{\gamma}) \propto 1/\binom{k}{k_{\boldsymbol{\gamma}}}$,—recall $k_{\boldsymbol{\gamma}} = \sum_j \gamma_j$—, which is uniform in the model size. We strongly recommend the last prior because it accounts for the multiplicity of comparisons (Scott and Berger, 2010).

The model posterior distribution $p(\boldsymbol{\gamma} \mid \boldsymbol{y}_{(0)}, \boldsymbol{x}_{(0)})$ is the main tool for quantifying uncertainty in the VS problem and must be properly summarised to produce useful reports. Rather than selecting a single model, as in the case of model comparison, the posterior distribution offers an enormous variety of ways to gain insight into the primary question of measuring the effect of different covariates on the response. Common summaries are the highest probability model and its probability; the posterior inclusion probability of each individual variable, which for the jth covariate is $p(\gamma_j = 1 \mid \boldsymbol{y}_{(0)}, \boldsymbol{x}_{(0)}) = \sum_{\boldsymbol{\gamma} \in \Gamma : \gamma_j = 1} p(\boldsymbol{\gamma} \mid \boldsymbol{y}_{(0)}, \boldsymbol{x}_{(0)})$, and the median probability model, which includes covariates with inclusion probabilities larger than 0.5 (Barbieri et al., 2021; Barbieri and Berger, 2004).

Finally, the posterior distribution provides straightforward access to (Bayesian) Model Averaged estimations and predictions as described in Hoeting et al. (1999) or Steel (2020).

5 Uncertainty on the distributions of the errors

Model comparison Let X_1 be an $n \times k_1$ matrix containing certain subset of the covariates in X (possibly with missing cells). We consider the problem where competing models agree on the covariates but differ in the density assumed for the errors. Consequently, we have two candidate models of the form (4)–(6) where

$$f_\gamma(\boldsymbol{y} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k, \beta_0, \sigma, \boldsymbol{\beta}_1) = \sigma^{-n}\, h_\gamma\big(\frac{\boldsymbol{y} - \mathbf{1}\beta_0 - X_1\boldsymbol{\beta}_1}{\sigma}\big), \quad \gamma \in \Gamma = \{1, 2\}$$

and $h_1, h_2 : \mathbb{R}^n \to \mathbb{R}$ are known (multivariate) probability density functions symmetric about the origin (ie. $h_\gamma(\boldsymbol{v}) = h_\gamma(-\boldsymbol{v})$ for all $\boldsymbol{v} \in \mathbb{R}^n$). In this problem, there is uncertainty regarding the distribution of the errors (e.g., a multivariate normal versus a multivariate Student's t or, as in the accompanying example, testing a particular heteroscedastic form). This situation, assuming constant matrix X_1, was studied in Berger et al. (1998).

Following the arguments below, the priors we propose are

$$\pi_\gamma(\beta_0, \boldsymbol{\beta}_1, \sigma, \boldsymbol{\nu}) = \pi^N(\beta_0, \boldsymbol{\beta}_1, \sigma)\pi^N(\boldsymbol{\nu}) = \sigma^{-1}\pi^N(\boldsymbol{\nu}). \tag{27}$$

From these, the imputation Bayes factor is the ratio of marginals

$$B_{12}(\boldsymbol{y}_{(0)}, \boldsymbol{x}_{(0)}) = \frac{E\{\mathfrak{m}_1(\boldsymbol{y}_{(0)} \mid \boldsymbol{x}_{(0)}, \boldsymbol{x}_{(1)}, \boldsymbol{\nu})\}}{E\{\mathfrak{m}_2(\boldsymbol{y}_{(0)} \mid \boldsymbol{x}_{(0)}, \boldsymbol{x}_{(1)}, \boldsymbol{\nu})\}}, \tag{28}$$

where both expectations are with respect to $\pi^N(\boldsymbol{x}_{(1)}, \boldsymbol{\nu} \mid \boldsymbol{x}_{(0)})$ and

$$\mathfrak{m}_\gamma(\boldsymbol{y}_{(0)} \mid \boldsymbol{x}_{(0)}, \boldsymbol{x}_{(1)}, \boldsymbol{\nu}) = \int \sigma^{-n_0}\, h_\gamma\left(\frac{\boldsymbol{y}_{(0)} - \mathbf{1}\beta_0 - X_1\boldsymbol{\beta}_1}{\sigma}\right)\frac{1}{\sigma}\,\mathrm{d}[\beta_0, \boldsymbol{\beta}_1, \sigma],$$

for $\gamma = 1, 2$. Notice that, in this case, $\mathfrak{m}_\gamma$ does not depend on $\boldsymbol{\nu}$—only on $\boldsymbol{x}_{(1)}$—and hence the expectation in (28) is with respect to the *a posteriori* predictive distribution $\pi^N(\boldsymbol{x}_{(1)} \mid \boldsymbol{x}_{(0)})$.

The construction of the imputation prior. First notice that both competing models share a common group of invariance. More concisely, they are group-invariant with respect to transformations of type (see for example, Eaton, 1989): $\{\boldsymbol{y} \to c\boldsymbol{y} + [\mathbf{1}\, X_1]\, \boldsymbol{b},\ c \in \mathbb{R}; \boldsymbol{b} \in \mathbb{R}^m\}$.

The parameters $\beta_0, \boldsymbol{\beta}_1, \sigma$ have the same dimension and common meaning regarding their roles within the aforementioned shared invariance structure. For instance, σ acts as a scale parameter in both models, whereas β_0 is the location parameter. This provides a justification for using the *same* prior $\pi_1(\beta_0, \boldsymbol{\beta}_1, \sigma) = \pi_2(\beta_0, \boldsymbol{\beta}_1, \sigma)$. The common form is taken to be σ^{-1} because it is the right Haar measure associated with the said type of invariance. Remarkably, this informal reasoning was supported by formal arguments from Berger et al. (1998), who perhaps provided one of the most important results for objective priors within model uncertainty. These authors showed that the assumed symmetry of h_γ, the right Haar density provides an exact predictive match (see also Bayarri et al., 2012).

Note that the proposed prior (27) follows the general recommended scheme (21) where the parameters $\beta_0, \boldsymbol{\beta}_1, \sigma$ and $\boldsymbol{\nu}$ play the role of common parameters and there are no new parameters.

Example. In this example, we test for possible heteroscedasticity in the errors comparing

$$h_1(\boldsymbol{\varepsilon}) = N_n(\boldsymbol{\varepsilon} \mid \mathbf{0}, I),\ \ h_2(\boldsymbol{\varepsilon}) = N_n(\boldsymbol{\varepsilon} \mid 0, \Psi), \tag{29}$$

where Ψ is a known positive definite matrix. If there are no missing values for the dependent variable, in Section B.4 of the Supplementary material we show that $\mathfrak{m}_\gamma(\boldsymbol{y} \mid \boldsymbol{x}_{(0)}, \boldsymbol{x}_{(1)}, \boldsymbol{\nu})$, for $\gamma = 1, 2$, has a closed-form expression leading to

$$B_{12}(\boldsymbol{y}, \boldsymbol{x}_{(0)}) = \frac{E\{S_I^{-(n-k_1-1)/2} |(1\, X_1)^t (1\, X_1)|^{-1/2}\}}{E\{S_\Psi^{-(n-k_1-1)/2} |(1\, X_1)^t \Psi^{-1} (1\, X_1)|^{-1/2} |\Psi|^{-1/2}\}}, \tag{30}$$

where

$$S_\Psi = \boldsymbol{y}^t \Big(\Psi^{-1} - \Psi^{-1}(1X_1)\big((1X_1)^t \Psi^{-1}(1X_1)\big)^{-1}(1X_1)^t \Psi^{-1} \Big) \boldsymbol{y}$$

(sum of the squared errors when regressing $\Psi^{-1/2}\boldsymbol{y}$ with the columns $\Psi^{-1/2}(1\, X_1)$), and

$$S_I = \boldsymbol{y}^t \Big(I - (1X_1)\big((1X_1)^t (1X_1)\big)^{-1}(1X_1)^t \Big) \boldsymbol{y},$$

the sum of squared errors when regressing $\boldsymbol{y}$ with $(1\, X_1)$.

If $\boldsymbol{y}$ had missing observations, the expressions would be similar, replacing $\boldsymbol{y}$ with $\boldsymbol{y}_{(0)}$ and n with n_0 and selecting the rows corresponding to the observed units in $(1\, X_1)$ and Ψ.

6 Numerical experiments

We conducted several experiments to shed light on the implications of missing observations in model uncertainty problems. This study attempts to fill a gap in the literature

where the evidence thus far is limited, especially from a Bayesian perspective. We performed five experiments based on the general linear model but of quite a different nature, ranging from highly controlled simulated cases to real datasets. The first four experiments considered the uncertainty of the regressors, while the fifth experiment questioned the structure of the error covariance. For comparisons, in all cases we have access to the full dataset (before missingness occurs).

In all cases, we use a multivariate normal imputation model. That is, (5) is

$$(x_{i1}, x_{i2}, \ldots, x_{ik}) \mid \boldsymbol{\nu} \sim N_k(\boldsymbol{\mu}, \Sigma)$$

where $\boldsymbol{\nu} = (\boldsymbol{\mu}, \Sigma)$. Some of our experiments are based on real data with covariates far from being normal (see Experiment E2 of Supplementary material), hence allowing to analyze the effect of a bad imputation model in the posterior distribution. The reference prior that corresponds to the multivariate normal distribution is derived in Chang and Eaves (1990): $\pi^N(\boldsymbol{\mu}, \Sigma) = |\Sigma|^{-(k+1)/2}|I + \Sigma * \Sigma^{-1}|^{-1/2}$, where $*$ denotes the Hadamard product (component by component). The corresponding posterior distribution has no closed form, but it can be sampled easily using the simple rejection algorithm described in Sun and Berger (2006).

Experiments 1, 2, E1 and E2 concern Section 4 and we refer to the *oracle g-BF* to the Bayes factor (22)—corresponding to the g-prior with $g = n$—using the full dataset. Similarly, the same Bayes factor applied to the dataset resulting after listwise deletion is termed as *listwise deletion g-BF*. Finally, our proposed Bayes factor, which utilizes all the available data by means of (24) is the *imputed g'-Bayes factor*.

The corresponding software can be found as a shiny application,[1] and the core code is available on github[2] along with the other pieces of code mentioned below.

6.1 Experiment 1. Variable selection

In this section, we reproduce the simulated experiment of Yang et al. (2005) to compare our results with SIAS (see Section 2.2), which showed the best performance among the methods compared in that paper. For a comprehensive comparative study, we added the results for the full dataset (referred to as the oracle) and listwise deletion.

The experiment consisted of $k = 10$ potential explanatory variables, $x_1, \ldots, x_{10}$, simulated independently of a multivariate normal, where the off-diagonal elements of the correlation matrix were $\rho \in \{0.1, 0.5\}$ (defining two different scenarios). This is combined with two ignorable missing data mechanisms: the MCAR mechanism, where values are randomly dropped from x_j, $j = 1, \ldots, 10$ independently with a probability of either 5% or 10%, resulting in a global missing percentage (i.e., the proportion of individuals with at least one missing value in any covariate) of 40% and 65%, respectively, and an MAR, where $x_1, \ldots, x_5$ are fully observed, while amputation (i.e. the process where missingness is induced in complete data) is performed over x_j, $j = 6, \ldots, 10$, with the same overall percentages of missing data as before, that is, 40% and 65%. For the latter

[1] https://stefanocabras.shinyapps.io/muqmissing/
[2] https://github.com/scabras/muqmissing

scenario, we use the `ampute` function from the `mice` package in R, considering different missing patterns with 20% or 40% missing data for each variable to obtain the desired global missing percentages.

The response variable y was simulated using the following linear regression model:

$$y_i = x_{i1} + 2x_{i2} + x_{i6} + 2x_{i7} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2 = 2.5),$$

where $i = 1, \ldots, n = 100$. For each combination of $\rho \in \{0.1, 0.5\}$, the overall percentage of missingness (40%, 65%), and missing data mechanism (MCAR, MAR), 100 datasets were simulated.

We calculated the posterior inclusion probabilities based on the oracle g-BF, listwise deletion g-BF, and imputed g'-BF, to which we appended the results reported in Yang et al. (2005) corresponding to the SIAS method. Following Yang et al. (2005) and for comparison purposes, we also compute and adopt the same summary statistic, which they call the signal-to-noise ratio (SNR) of the inclusion probabilities, to compare the discriminatory power of the procedures, namely the ratio of the minimum inclusion probability (the "signal") for true covariates to the maximum for spurious predictors (the "noise"):

$$\text{SNR} = \frac{\min\limits_{\text{true cov.} \equiv j \in \{1,2,6,7\}} p(\gamma_j = 1 \mid \boldsymbol{y}_{(0)}, \boldsymbol{x}_{(0)})}{\max\limits_{\text{noise cov.} \equiv j \in \{3,4,5,8,9,10\}} p(\gamma_j = 1 \mid \boldsymbol{y}_{(0)}, \boldsymbol{x}_{(0)})}.$$

Table 1 shows the mean and standard deviation of each active variable's posterior inclusion probability and SNR for each combination of design elements. The conclusions drawn from this table are as follows:

The first conclusion is that listwise deletion performs competently and clearly outperforms SIAS. This is a surprising result, especially considering that this superiority occurs in all cases, both in the ability to preserve the strength of the true signals and in the discriminatory power (as measured by SNR). Furthermore, the differences between the two approaches are generally substantial. When comparing imputation and deletion, when the correlation between the covariates is small ($\rho = 0.1$), the two approaches behave similarly in terms of sensitivity (the ability to detect true positives). As ρ and the percentage of missingness increase, imputation outperforms listwise deletion, justifying the extra effort required in the procedure.

The imputation SNR was considerably better than the other methods (persistent in all cases and quite pronounced in some cases). Obviously, this is essentially a better performance in terms of specificity (the ability to detect true negatives) simply because the inclusion probabilities of signals are very close to one in the vast majority of cases. In fact, a possible estimate for the average maximum inclusion probability of the spurious variables can be obtained by dividing the minimum of the average inclusion probability of the true signals by the SNR, implying that these estimates would be between 0.24 and 0.30 for the imputation g'-BF, 0.27 and 0.37 for deletion, while for SIAS, it falls within the range 0.29 and 0.46. Compared with listwise deletion, this is also explained by the

	$\rho = 0.1$					$\rho = 0.5$				
	x_1	x_2	x_6	x_7	SNR	x_1	x_2	x_6	x_7	SNR
oracle	1	1	1	1	3.8(.2)	1	1	1	1	4.8(.2)
40%-MCAR										
Imputed	1	1	1	1	3.5(.2)	1	1	1	1	4.2(.2)
Deletion	1	1	1	1	3.3(.1)	.99(.05)	1	.99(.06)	1	3.7(.2)
SIAS	.92	.99	.91	.99	2.9	.80	.99	.80	.99	2.4
65%-MCAR										
Imputed	1(.01)	1	1(.02)	1	3.4(.2)	1	1	1	1	3.9(.2)
Deletion	.98(.06)	1	.95(.13)	1	2.6(.1)	.86(.18)	1	.88(.18)	1(.02)	2.6(.1)
SIAS	.88	.99	.88	.99	3.0	.71	.99	.72	.99	2.3
40%-MAR										
Imputed	1(.02)	1	1(.01)	1	3.5(.2)	1(.01)	1	1(.01)	1	3.9(.2)
Deletion	1	1	1(.02)	1	3.0(.1)	.99(.05)	1	.99(.04)	1	3.7(.2)
SIAS	.90	.99	.90	.99	2.8	.88	.99	.77	.99	2.1
65%-MAR										
Imputed	.99(.07)	1	.93(.14)	1	3.2(.2)	.99(.03)	1	.92(.16)	1	3.1(.2)
Deletion	.97(.08)	1	.97(.08)	1	2.6(.2)	.87(.18)	1	.88(.17)	1	2.5(.1)
SIAS	.82	.99	.85	.99	2.4	.89	.99	.69	.98	1.5

Table 1: Experiment 1. Mean posterior inclusion probabilities for the truly active variables and mean signal-to-noise ratio of the inclusion probabilities. The number in parentheses corresponds to the standard deviation, reported only when ≥ 0.01. Values for the SIAS method are borrowed from Yang et al. (2005).

differences in the amount of sampling information used by each method. For example, consider the MCAR case with 65% missing data. Out of the $n \times (k+1) = 1100$ total observations used by the oracle, listwise deletion preserves $0.35 \times 1100 = 385$, whereas the results based on imputation use $1100 - 0.1 \times 1000 = 1000$ (as the response variable is not imputed). This corresponds to approximately 2.6 times more sampling information, which, when accompanied by reliable imputations, leads to a substantial increase in specificity and sensitivity simply because the sample size is much larger.

Experiment E1 of the Supplementary material, although considering a simpler design, aimed to analyze the performance of our method by confronting it with listwise deletion in a more extreme case of the strength of the signal.

6.2 Experiment 2. The Ozone dataset

We consider VS problems from popular real-world datasets in this and in Experiments E2 and E3 of the Supplementary material. The role of the distribution of the covariates, which is unknown in this case, is the main difference from previous simulated experiments. As a reminder, we assume a multivariate imputation model. Clearly, misspecification of this component does not affect listwise deletion procedures, but it is an essential part of all imputation methods. This observation is important for understanding the following results.

The Ozone datasets previously used in Garcia-Donato and Martinez-Beneito (2013), Casella and Moreno (2006), and Berger and Molina (2005) consisted of $n = 178$ measurements of atmospheric ozone concentration, along with several covariates. From the original 10 main effects, we only used seven with atmospheric relevance, which corresponds to the main effects in the `Ozone35` dataset from the `BayesVarSel` library in R, named x_4 to x_{10}. An initial examination of the data suggests that the assumption of normality is reasonable. For further details on these data, see Casella and Moreno (2006).

We introduced MAR missing values into variables x_6 to x_{10} by using the function `ampute` from the `mice` package in R. The percentage of missing values per variable was 10, 20, or 30%, resulting in a mean overall percentage of missingness of approximately 37%, 60%, and 74%, respectively. For each of these percentages, we considered 1000 replications where the variability was caused by the removed observations (that changed in the replicas). Figure 1 shows the variation in the inclusion probabilities for each variable obtained with the different Bayes factors.

The potential of the imputed g'-BF to preserve the evidence is shown in Figure 1. Its superiority over deletion was evident for all variables and levels of missingness. We also observed that the imputed g'-BF was less sensitive to variations in the removed observations, producing less variable results.

Experiment E2 of the Supplementary material uses the Boston dataset, which includes variables that do not follow the normality assumption. This experiment helped us assess the performance of the proposed method under more challenging conditions where some of the assumptions may not hold. Finally, Experiment E3 of the Supplementary material focused on comparing the distributions of errors to illustrate the methods discussed in Section 5.

7 Conclusions and future Work

This study presents a comprehensive approach for addressing model uncertainty when dealing with missing data in a regression framework. Through a series of experiments, we demonstrated the effectiveness of our proposed imputed g'-prior methodology compared to listwise deletion and the SIAS proposal from Yang et al. (2005), particularly regarding reduced variability and more accurate posterior inclusion probabilities. The proposed method is fully automatic and does not depend on hyperparameters such as the penalty parameter of lasso methods. Moreover, exploiting the analytical integration from the closed-form output of the "completed" predictive distribution makes the method much faster and more efficient than the alternative procedure of imputing and estimating the model's posterior probability. There are several directions for future research to further enhance the applicability and robustness of our approach.

1. *Large model spaces*: Our experiments were conducted in relatively small model spaces, which allowed for exhaustive enumeration. However, in many modern applications, large model spaces (i.e., large k) are common, such as those arising from variable selection with many covariates. Adapting our approach to handle these

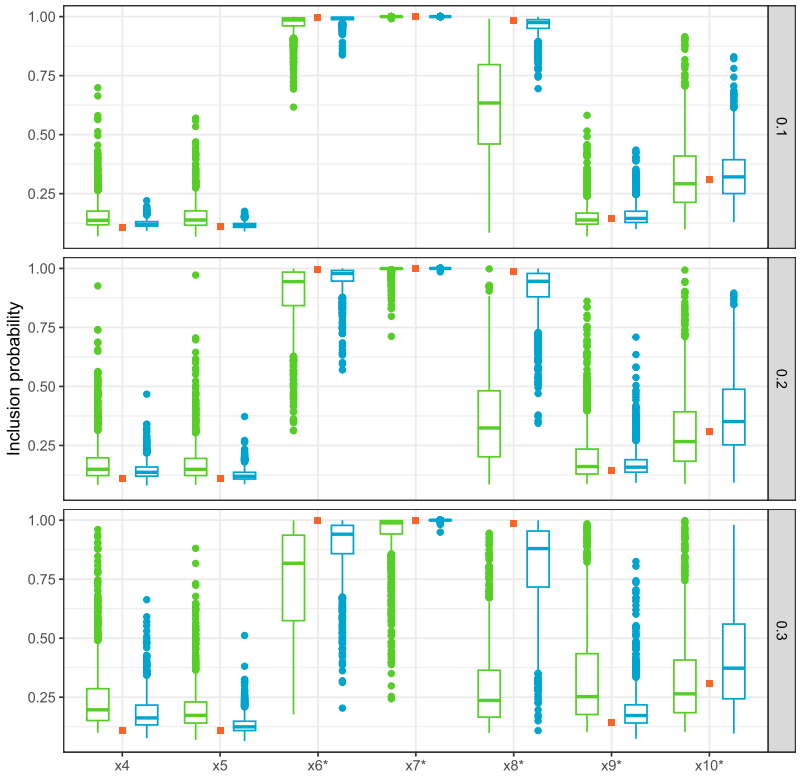

Figure 1: Boxplots of the inclusion probabilities for each variable using imputed g'-BF (blue) and listwise deletion g-BF (green), when considering 10 (top), 20 (middle) or 30% (bottom) of missing values per variable, for Ozone dataset. The corresponding oracle g-BF inclusion probabilities are depicted in red. The symbol $\star$ in variable names explicitly indicates the variables with missing data.

situations would require the development of numerical algorithms, such as "missing data" adaptations of Gibbs sampling methods (Garcia-Donato and Martinez-Beneito, 2013).

2. *Prior distribution for regression parameters*: A central question in this study has been the construction of objective prior distributions when the covariates are assumed to be random. Although arguably more realistic than the fixed design assumption, this perspective has barely received any attention in the literature despite the broad potential interest in such inferential objects. In this study, we followed Zellner (1986); Zellner and Siow (1980), in what has been called g-priors, constructed based on the expected Fisher information matrix. We have derived a new class of g-priors in which we have focused on fixed g, but extensions to random hyperparameter (of the type in Liang et al., 2008) are straightforward. The construction of priors following alternative procedures, assuming that the

covariates are random, opens up areas for future research that would extend the scope of non-local priors (Johnson and Rossell, 2010); modern spike and slab formulations (Bai et al., 2021); intrinsic priors (Berger and Pericchi, 1996; Moreno et al., 1998) or power expected posterior priors (Fouskakis and Ntzoufras, 2022), to mention a few.

3. *Other patterns of missingness*: Our procedure can incorporate other missing data patterns, making it possible, in principle, to test for different missing data mechanisms. Further work in this direction is needed to understand how to separate the comparison of the missing data mechanism from the models for observed variables, response and covariates.

4. *Qualitative covariates or factors*: Our work assumes that the explanatory variables are quantitative. For complete datasets, the problem of model selection with qualitative explanatory variables (factors) is treated in depth in Garcia-Donato and Paulo (2022). Their approach only depends on Bayes factors and, at least conceptually, is directly applicable to the Bayes factors with missing data developed in this paper. Obviously, putting into practice the resulting procedure entails difficulties (like the joint imputation model or strategies to impute qualitative variables) that are specific to the problem at hand and that may require additional insights.

Acknowledgments

The authors would like to thank the anonymous referees, an Associate Editor and the Editor for their constructive comments that improved the quality of this paper.

Funding

This work has been funded by MICIU/AEI/10.13039/501100011033 (Grant PID2022-138201NB-100) and by ERDF/EU.

The first author was also supported by Grant SBPLY/21/180501/000241 funded by JCCM/EU.

Anabel Forte is also partially funded by Dirección General de Ciencia e Investigación (Generalitat Valenciana) with the grant CIAICO/2022/165.

Supplementary Material

Model Uncertainty and Missing Data: An Objective Bayesian Perspective
(DOI: 10.1214/25-BA1531SUPP; .pdf). The Supplementary material is included in a single pdf file containing four sections.

- Section A: Notation. Detailed description of the notation used throughout the paper, including key symbols and definitions.
- Section B: Proofs. Proofs of key theoretical results, including Proposition 1, Result 1, Proposition 2, Result B.3 needed for the proof of the identity (21) and identity (30).

- Section C: Extra Experiments. Additional simulation experiments illustrating the performance of the proposed methodology. Includes model selection analyses, applications to the Boston dataset, and comparisons of error distributions.

- Section D: Discussing priors for the general case. Discussion on the choice of prior distributions in the general modeling framework, highlighting their implications and applicability.

References

Aßmann, C., Gaasch, J., and Stingl, D. (2023). "A Bayesian Approach Towards Missing Covariate Data in Multilevel Latent Regression Models." *Psychometrika*, 88: 1495–1528. MR4668577. doi: https://doi.org/10.1007/s11336-022-09888-0. 3

Bai, R., Rockova, V., and George, E. (2021). "Spike-and-slab meets LASSO: A review of the spike-and-slab LASSO." In *Handbook of Bayesian Variable Selection*. Chapman and Hall/CRC. MR4646623. doi: https://doi.org/10.1080/01621459.2022.2025815. 20

Barbieri, M. M. and Berger, J. O. (2004). "Optimal predictive model selection." *Annals of Statistics*, 32: 870–897. MR2065192. doi: https://doi.org/10.1214/009053604000000238. 14

Barbieri, M. M., Berger, J. O., George, E. I., and Ročková, V. (2021). "The Median Probability Model and Correlated Variables." *Bayesian Analysis*, 16(4): 1085–1112. MR4381128. doi: https://doi.org/10.1214/20-BA1249. 14

Bayarri, M. J., Berger, J. O., Forte, A., and García-Donato, G. (2012). "Criteria for Bayesian Model Choice with Application to Variable Selection." *The Annals of Statistics*, 40: 1550–1577. MR3015035. doi: https://doi.org/10.1214/12-AOS1013. 2, 6, 11, 13, 15

Berger, J. and Pericchi, L. (1996). "The Intrinsic Bayes Factor for Linear Models." In Bernardo, J. M., Berger, J. O., Dawid, A. P., and M., S. A. F. (eds.), *Bayesian Statistics 5*, 23–42. London: Oxford University Press. MR1425398. 20

Berger, J. O. (2006). "The Case for Objective Bayesian Analysis." *Bayesian Analysis*, 1(3): 385–402. MR2221271. doi: https://doi.org/10.1214/06-BA115. 2

Berger, J. O. and Molina, G. (2005). "Posterior Model Probabilities Via Path-Based Pairwise Priors." *Statistica Neerlandica*, 59(1): 3–15. MR2137378. doi: https://doi.org/10.1111/j.1467-9574.2005.00275.x. 18

Berger, J. O., Pericchi, L. R., and Varshavsky, J. (1998). "Bayes Factors and Marginal Distributions in Invariant Situations." *Sankhya: The Indian Journal of Statistics, Series A*, 60: 307–321. MR1718789. 14, 15

Berger, J. O. and Pericchi, R. L. (2001). "Objective Bayesian Methods for Model Selection: Introduction and Comparison (with discussion)." In Lahiri, P. (ed.), *Model Selection*, 135–207. Institute of Mathematical Statistics Lecture Notes- Monograph

Series, volume 38. MR2000753. doi: https://doi.org/10.1214/lnms/1215540968. 2

Bozigar, M., Lawson, A., Pearce, J., King, K., and Svendsen, E. (2020). "A geographic identifier assignment algorithm with Bayesian variable selection to identify neighborhood factors associated with emergency department visit disparities for asthma." *International Journal of Health Geographics*, 19(1): 9. 3

Casella, G. and Moreno, E. (2006). "Objective Bayesian Variable Selection." *Journal of the American Statistical Association*, 101(473): 157–167. MR2268035. doi: https://doi.org/10.1198/016214505000000646. 18, 19

Celeux, G., Forbes, F., Robert, C., and Titterington, D. (2006). "Deviance information criteria for missing data models (with discussion)." *Bayesian Analysis*, 1: 651–674. MR2282197. doi: https://doi.org/10.1214/06-BA122. 3

Chang, T. and Eaves, D. (1990). "Reference prior for the orbit in a group model." *The Annals of Statistics*, 18: 1595–1614. MR1074425. doi: https://doi.org/10.1214/aos/1176347868. 16

Cohen, N. and Berchenko, Y. (2021). "Normalized information criteria and model selection in the presence of missing data." *Mathematics*, 9(19): 2474. 3

Daniels, M. J., Chatterjee, A. S., and Wang, C. (2012). "Bayesian Model Selection for Incomplete Data Using the Posterior Predictive Distribution." *Biometrics*, 68(4): 1055–1063. MR3040012. doi: https://doi.org/10.1111/j.1541-0420.2012.01766.x. 3

Eaton, M. L. (1989). *Group Invariance Applications in Statistics*. Institute of Mathematical Statistics. MR1089423. 14

Erler, N. S. (2019). "Bayesian Imputation of Missing Covariates." Ph.D. thesis, Erasmus University Rotterdam. 3

Erler, N. S., Rizopoulos, J., Jaddoe, V. W., Franco, O. H., and Lesaffre, E. (2016). "Dealing with missing covariates in epidemiologic studies: a comparison between multiple imputation and a full Bayesian approach." *Statistics in medicine*, 35: 2955–2974. MR3528236. doi: https://doi.org/10.1002/sim.6944. 3

Fernández, C., Ley, E., and Steel, M. F. (2001). "Benchmark Priors for Bayesian Model Averaging." *Journal of Econometrics*, 100: 381–427. MR1820410. doi: https://doi.org/10.1016/S0304-4076(00)00076-2. 6

Fouskakis, D. and Ntzoufras, I. (2022). "Power-Expected-Posterior Priors as Mixtures of g-Priors in Normal Linear Models." *Bayesian Analysis*, 17(4): 1073–1099. URL https://doi.org/10.1214/21-BA1288 MR4506022. 20

García-Donato, G., Castellanos, M. E., Cabras, S., Quirós, A. and Forte, A. (2025). Supplementary Material for "Model Uncertainty and Missing Data: An Objective Bayesian Perspective." *Bayesian Analysis*. doi: https://doi.org/10.1214/25-BA1531SUPP. 2

García-Donato, G. and Forte, A. (2018). "Bayesian Testing, Variable Selection and Model Averaging in Linear Models using R with BayesVarSel." *The R Journal*, 10(1): 155–174. 4

Garcia-Donato, G. and Martinez-Beneito, M. A. (2013). "On Sampling strategies in Bayesian variable selection problems with large model spaces." *Journal of the American Statistical Association*, 108(501): 340–352. MR3174624. doi: https://doi.org/10.1080/01621459.2012.742443. 18, 20

Garcia-Donato, G. and Paulo, R. (2022). "Variable selection in the presence of factors: a model selection perspective." *Journal of American Statistical Association*, 117(540): 1847–1857. MR4528475. doi: https://doi.org/10.1080/01621459.2021.1889565. 21

Gomez-Rubio, V. (2020). *Bayesian Inference with INLA*. Chapman and Hall/CRC. 3

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). "Bayesian Model Averaging: A Tutorial." *Statistical Science*, 14(4): 382–401. MR1765176. doi: https://doi.org/10.1214/ss/1009212519. 14

Hoff, P. (2009). *A First Course in Bayesian Statistical Methods*. Springer Texts in Statistics. Springer New York.
URL https://books.google.es/books?id=V8jT2SimGR0C MR2648134.
doi: https://doi.org/10.1007/978-0-387-92407-6. 9

Hoijtink, H., Gu, X., Mulder, J., and Rosseel, Y. (2019). "Computing Bayes factors from data with missing values." *Psychological Methods*, 24(2): 253–268. 3

Ibrahim, J., Chen, M., and Kim, S. (2006). "Bayesian variable selection for the Cox regression model with missing covariates." *Lifetime Data Analysis*, 14(4): 496–520. MR2464772. doi: https://doi.org/10.1007/s10985-008-9101-5. 3

Ibrahim, J., Chen, M.-H., and Lipsitz, S. (2002). "Bayesian methods for generalized linear models with covariates missing at random." *Canadian Journal of Statistics*, 30: 55–78. MR1907677. doi: https://doi.org/10.2307/3315865. 3

Ishwaran, H. and Rao, J. (2005). "Spike and slab variable selection: frequentist and Bayesian strategies." *The Annals of Statistics*, 33(2): 730–773. MR2163158. doi: https://doi.org/10.1214/009053604000001147. 6

Jeffreys, H. (1961). *Theory of Probability*. London: Oxford University Press. MR0187257. 2, 12

Johnson, V. E. 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.
URL http://dx.doi.org/10.1111/j.1467-9868.2009.00730.x MR2830762.
doi: https://doi.org/10.1111/j.1467-9868.2009.00730.x. 6, 20

Kass, R. E. and Raftery, A. (1995). "Bayes factors." *Journal of the American Statistical Association*, 90: 773–795. MR3363402. doi: https://doi.org/10.1080/01621459.1995.10476572. 2, 12

Kass, R. E. and Wasserman, L. (1995). "A Reference Bayesian Test for Nested Hypotheses and Its Relationship to the Schwarz Criterion." *Journal of the American Statistical Association*, 90: 928–934. MR1354008. 10

Kass, R. E. and Wasserman, L. (1996). "The Selection of Prior Distributions by Formal Rules." *Journal of the American Statistical Association*, 91: 1343–1369. 9

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). "Mixtures of g-Priors for Bayesian Variable Selection." *Journal of the American Statistical Association*, 103(481): 410–423. MR2420243. doi: https://doi.org/10.1198/016214507000001337. 2, 6, 10, 11, 20

Little, R. and Rubin, D. (2020). *Statistical Analysis with Missing Data*. Wiley, 3rd edition. MR1925014. doi: https://doi.org/10.1002/9781119013563. 4, 7

Moreno, E., Bertolino, F., and Racugno, W. (1998). "An intrinsic limiting procedure for model selection and hypothesis testing." *Journal of the American Statistical Association*, 93: 1451–1460. MR1666640. doi: https://doi.org/10.2307/2670059. 20

Mostafa, S. M., Eladimy, A. S., Hamad, S., and Amano, H. (2020). "CBRG: A Novel Algorithm for Handling Missing Data Using Bayesian Ridge Regression and Feature Selection Based on Gain Ratio." *IEEE Access*, 8: 216969–216985. 3

Rubin, D. (1996). "Multiple Imputation After 18+ Years." *Journal of American Statistical Association*, 91(434): 473–489. MR1294072. 4

Scott, J. and Berger, J. (2005). "An exploration of aspects of Bayesian multiple testing." *Journal of Statistical Planning and Inference*, 136: 2144–2162. MR2235051. doi: https://doi.org/10.1016/j.jspi.2005.08.031. 2

Scott, J. and Berger, J. (2010). "Bayes and Empirical-Bayes multiplicity adjustment in the variable-selection problem." *The Annals of Statistics*, 38: 2587–2619. MR2722450. doi: https://doi.org/10.1214/10-AOS792. 13

Steel, M. F. J. (2020). "Model Averaging and Its Use in Economics." *Journal of Economic Literature*, 58(3): 644–719.
URL https://www.aeaweb.org/articles?id=10.1257/jel.20191385 14

Storlie, C., Therneau, T., Carter, R., Chia, N., Bergquist, J., Huddleston, J., and Romero-Brufau, S. (2020). "Prediction and Inference With Missing Data in Patient Alert Systems." *Journal of the American Statistical Association*, 115: 32–46. MR4078443. doi: https://doi.org/10.1080/01621459.2019.1604359. 3

Sun, D. and Berger, J. O. (2006). "Objective Bayesian analysis for the multivariate normal model." In Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M. (eds.), *Proc. Valencia / ISBA 8th World Meeting on Bayesian statistics*. Oxford university Press. MR2433206. 16

Tadesse, M. G. and Vanucci, M. (eds.) (2022). *Handbook of Bayesian Variable Selection*. Chapman and Hall/CRC. 2

Van Buuren, S. (2018). *Flexible Imputation of Missing Data*. Chapman & Hall/CRC Interdisciplinary Statistics. CRC Press LLC. 7

Van Buuren, S. and Groothuis-Oudshoorn, K. (2011). "mice: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software*, 45(3): 1–67. URL https://www.jstatsoft.org/v45/i03/ 4

Xu, D., Daniels, M. J., and Winsterstein, A. G. (2016). "Sequential BART for imputation of missing covariates." *Biostatistics*, 17(3): 589–602. MR3603956. doi: https://doi.org/10.1093/biostatistics/kxw009. 3

Yang, R. and Berger, J. O. (1997). "A catalog of noninformative priors." Technical Report 97-42, ISDS Discusion paper. 9

Yang, X., Belin, T. R., and Boscardin, W. J. (2005). "Imputation and Variable Selection in Linear Regression Models with Missing Covariates." *Biometrics*, 61(2): 498–506. URL http://www.jstor.org/stable/3695970 MR2140922. doi: https://doi.org/10.1111/j.1541-0420.2005.00317.x. 3, 5, 9, 16, 17, 18, 20

Zellner, A. (1986). "On Assessing Prior Distributions and Bayesian Regression Analysis with g-prior Distributions." In Zellner, A. (ed.), *Bayesian Inference and Decision techniques: Essays in Honor of Bruno de Finetti*, 389–399. Edward Elgar Publishing Limited. MR0881437. 2, 6, 10, 20

Zellner, A. and Siow, A. (1980). "Posterior odds for selected regression hypotheses." In Bernardo, J. M., DeGroot, M. H., Lindley, D. V., and Smith, A. F. M. (eds.), *Bayesian Statistics*, 585–603. Valencia University Press. 6, 12, 13, 20

Invited Discussion

Adam Iqbal*, Emmanuel O. Ogundimu*, and F. Javier Rubio†

1 Comments

We congratulate the authors of García-Donato et al. (2025) on their thought-provoking contribution, which connects important topics such as objective Bayes, missing at random data, and Bayesian variable selection. Our discussion focuses on potential extensions to the missing not at random setting, particularly sample selection models with missing outcomes and covariates.

MNAR and sample selection. A key feature of the authors' framework is the assumption of ignorability of the missingness mechanism, specifically when data are missing completely at random (MCAR) or missing at random (MAR). In this setting, the analysis does not depend on the missing outcomes. However, this justification does not extend to situations where the data are missing not at random (MNAR). Such cases can arise in various forms, and both parametric and nonparametric models have been proposed to address particular instances of MNAR (Linero and Daniels, 2018).

Sample selection bias arises when missingness in the outcome of interest is correlated with the outcome itself, leading to non-randomly selected observations. Sample selection models aim to correct for this bias, most notably the Heckman selection model (Heckman, 1979), which is specified as a two-equation system. Suppose y_i^* is the outcome variable of interest, modelled by a multiple linear regression model with covariates $\boldsymbol{x}_i$. The Heckman model supplements this with a missingness process s_i^* that is modelled by a latent-index probit with covariates $\boldsymbol{w}_i$:

$$y_i^* = \beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_{1,i}, \quad s_i^* = \alpha_0 + \boldsymbol{w}_i^\top \boldsymbol{\alpha} + \epsilon_{2,i},$$

with the errors typically defined as $\begin{pmatrix}\epsilon_{1,i}\\ \epsilon_{2,i}\end{pmatrix} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma^2 & \sigma\rho\\ \sigma\rho & 1\end{pmatrix}\right)$. We fix $\mathrm{Var}(\epsilon_{2,i}) = 1$ to identify the scale of the selection index. If $\rho = 0$, then the missingness of y_i^* is MAR conditional on (x_i, w_i); if $\rho \neq 0$, it is MNAR. The observed data consist of $\{(y_i, s_i) : i = 1, \ldots, n\}$, where

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

Variable selection for sample selection models. The motivation for extending variable selection to this context is twofold. First, the variables that are relevant in each

*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

individual process are not usually known *a priori*. Second, practical non-identifiability (*i.e.*, for specific data sets) may arise in the Heckman model when the same variables appear in both equations. Such issues often lead practitioners to include unnecessary covariates or to exclude potentially relevant ones, which in turn may cause overfitting, multicollinearity, and unstable parameter estimates (Sartori, 2003). Variable selection can mitigate these problems by guiding practitioners in identifying the most relevant variables. There has been recent progress in extending variable selection methods to Heckman selection models. Ogundimu (2022) applies the Adaptive LASSO algorithm to sample selection models using a least squares approximation, additionally showing an oracle property for this estimator. From the Bayesian perspective, Iqbal et al. (2023) developed a closed-form Gibbs sampling algorithm to incorporate spike-and-slab priors, allowing for variable selection in sample selection models. Of further interest is the work of Iqbal et al. (2025), which develops Bayesian variable selection under g-priors and non-local priors in the sample selection framework, and shows that ignoring sample selection can lead to the inclusion of spurious variables if they explain the selection process. As the marginal likelihood for the Heckman model is unavailable in closed form, Laplace approximations are employed in its place.

Missing covariates. To address missing covariates, García-Donato et al. (2025) adopt a general framework in which covariates are treated as random variables. Both Iqbal et al. (2023) and Iqbal et al. (2025) assume fully observed covariates, so the advancements of García-Donato et al. (2025) could be used to extend these works to settings with missing covariates. Such extensions might take the form of outcomes modelled through a Heckman-type framework with MAR covariates, or could consider MNAR mechanisms for both outcomes and covariates.

References

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. 27, 28

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. 27

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. 28

Iqbal, A., Ogundimu, E., and Rubio, F. (2025). "Bayesian variable selection under sample selection and model misspecification." *Under Review*. 28

Linero, A. and Daniels, M. (2018). "Bayesian approaches for missing not at random outcome data: the role of identifying restrictions." *Statistical Science*, 33(2): 198. MR3797710. doi: https://doi.org/10.1214/17-STS630. 27

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. 28

Sartori, A. E. (2003). "An estimator for some binary-outcome selection models without exclusion restrictions." *Political Analysis*, 11(2): 111–138. 28

Invited Discussion

Marco A. R. Ferreira*

I congratulate the authors on producing a stimulating and inspiring paper. I was fortunate to discuss parts of this work as presented by María Eugenia Castellanos in an invited session during the Objective Bayes Meeting in Santa Cruz, California, in 2022. I was deeply saddened to learn of María Eugenia's recent passing, and I offer my heartfelt condolences to her family and coauthors. Her passing is a tremendous loss for the Bayesian community.

Missing data are ubiquitous in all areas of human endeavor. Thus, García-Donato and coauthors' objective Bayesian treatment of this subject is tremendously welcome. Of course, they had to start somewhere, and naturally they decided to consider the cases of missing at random (MAR) or missing completely at random (MCAR). I think the general framework laid out in this paper may be useful for the development of objective Bayesian methods for regression analysis of datasets with other types of missingness.

On another hand, practitioners wanting to apply the method proposed in this paper should be aware of its limitations. Here, I am going to focus on limitations resulting from two aspects of the proposed method: the MAR/MCAR assumptions, and the assumption that the true model is among the models being considered.

1 The MAR and MCAR assumptions

The MAR and MCAR assumptions imply that observations with missing dependent variables $y_{(1)}$ can be ignored. This is a simplifying assumption that may have severe consequences for applications of the proposed method in cases when the MAR/MCAR assumptions do not hold. Unfortunately, in several important practical problems the missingness mechanism is not of the MAR or MCAR types.

Consider the following simple example that illustrates some of the difficulties with the MAR or MCAR assumptions, and that may suggest a different path. Assume we have two regressors, x_{1i} and x_{2i}, $i = 1, \ldots, n$, that are iid $N(0, 1)$. Say that the dependent variable is related only to x_{1i} through the linear model $y_i = x_{1i} + \epsilon_i$, with $\epsilon_1, \ldots, n$ iid $N(0, 1)$. In addition, all values of x_{1i} and x_{2i} are reported, thus the regressors have no missing values. Further, because of some technical limitations, the measurement instrument reports values of y_i larger than 2 as missing values. Thus, the missingness mechanism is such that if $y_i > 2$ then y_i is not observed and, in the notation of the paper under discussion, $M_{i,k+1} = 1$. Hence, $P(M_{i,k+1} = 1|y_i) = 1$ if $y_i > 2$ and $P(M_{i,k+1} = 1|y_i) = 0$ otherwise. Therefore, $P(M|\mathbf{y}_{(0)}, \mathbf{x}_{(0)}, \mathbf{y}_{(1)}, \mathbf{x}_{(1)}, \boldsymbol{\psi}) = P(M|\mathbf{y}_{(0)}, \mathbf{y}_{(1)}, \boldsymbol{\psi})$ and both the MAR and MCAR conditions are violated. In this example, ignoring observations with missing dependent variables $\mathbf{y}_{(1)}$ may lead to potentially misleading inferences.

*Department of Statistics, Virginia Tech, Blacksburg, VA 24061, marf@vt.edu

An alternative to ignoring observations with missing dependent variables is, instead, to investigate if missingness of dependent variables in the problem at hand can be predicted with the observed covariates. In the case of the example above, $P(M_{i,k+1} = 1|x_{1i}, x_{2i}) = 1 - \Phi(2 - x_{1i})$, where Φ is the cumulative distribution function of the standard normal distribution. Hence, modeling the conditional relationship between $M_{i,k+1}$ and the regressors using a probit regression may partially inform about the missingness mechanism of dependent variables. This latter piece of information may then be incorporated in a method for missing data that goes beyond the MAR and MCAR assumptions.

The trouble with missing dependent variable data is that, more often than we Bayesian statisticians would like, the missingness is due to someone deciding the data points were not nice enough and deleting the data. Unfortunately, in that situation the MAR and MCAR assumptions do not hold. Beyond dealing with missing data, modeling the conditional relationship between $M_{i,k+1}$ and the regressors may provide information that supports investigations of fraud in data collection or data reporting. For example, say that in a given application data are being collected daily, but every month there are three or four days where all the regressors are obtained but the dependent variable is missing. Say that, incidentally, one of the regressors is a binary variable that measures something different than the missingness, but the dependent variable is missing every time that binary regressor is equal to one. This relationship would possibly be detected by a regression analysis of $M_{i,k+1}$ on the regressors, which could then support an investigation of data collection and reporting fraud.

2 Is the true model among the models being considered?

What if data on a regressor that is in the true model is missing for all observations? Such a case would violate the key assumption in the method proposed by García-Donato and coauthors that the true model is among the models being considered. Among other things, this latter assumption requires that all of the regressors present in the true model are among the regressors being considered. While this is reasonable to assume when the data come from experiments performed in a laboratory under strictly controlled conditions, such assumption is likely to be violated when data are obtained under less controlled conditions.

Consider for example a case where two genetic markers such as those considered in genome-wide association studies (Doerge, 2002; Li et al., 2009; Williams et al., 2022; Xu et al., 2023; Williams et al., 2023; Xu et al., 2025) are available for a sample of subjects. The objective is to decide which genetic marker is a better predictor of a continuous Gaussian phenotype of interest. But, unknown to the scientist conducting the study, the actual genetic variant that is causal to the phenotype of interest is equally distant from and — because of linkage disequilibrium — equally correlated to each of the genetic markers. Hence, each of the genetic markers is an equally good proxy of the causal genetic variant. A scientist would hope that an objective Bayesian approach would indicate that the data equally supports each of the two genetic markers.

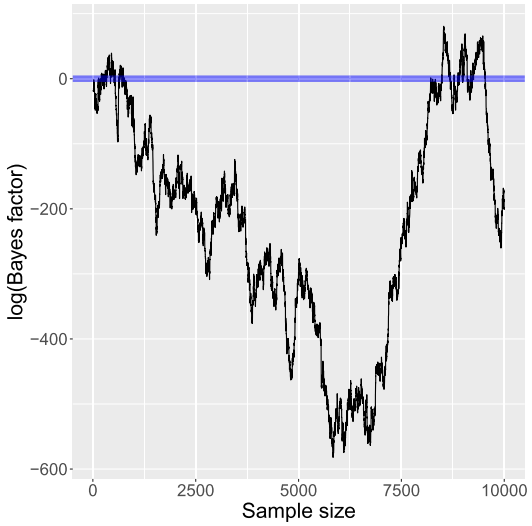

Figure 1: Logarithm of Bayes factor of a model with one proxy regressor versus a model with another equally good proxy regressor, as the sample size increases. Blue shade indicates region where evidence in favor of one of the two models is not very strong. Here, evidence in favor of one of the two models is undesirably very strong about 97% of the time.

Unfortunately, a Bayes factor — computed with usual objective Bayes model selection priors — to compare the two models would with high probability provide very strong evidence for one of the two genetic markers. That is because, in terms of Kullback-Leibler divergence, the two models would be equally distant from the true model, leading to the M-open multiple optima (MOMO) paradox (Ferreira et al., 2024). Here is a simple simulated example that illustrates the MOMO paradox. Consider a case with regressor x_i, $i = 1, \ldots, n$, iid $N(0, 1)$, and dependent variable $y_i = \beta x_i + \epsilon_i$, with $\epsilon_i \overset{iid}{\sim} N(0, 1)$. However, the scientist does not observe x_i. Instead, the scientist obtains observations for two proxy variables $x_{1i} = x_i + \xi_{1i}$ and $x_{2i} = x_i + \xi_{2i}$, with ξ_{1i} and ξ_{2i} iid $N(0, 1)$. To check which of the proxies is better, the scientist obtains a sample of size n=10,000 and uses a Bayes factor based on the Zellner-g prior to compare model $M_1 : \ y_i = \beta x_{1i} + \epsilon_i$ versus $M_2 : \ y_i = \beta x_{2i} + \epsilon_i$. Figure 1 shows the logarithm of the Bayes factor of M_1 versus M_2 as the sample size increases. Instead of converging to zero, because of the MOMO paradox, the logarithm of the Bayes factor follows a random walk and stays in the region of very strong evidence (Kass and Raftery, 1995) for one of the two models about 97% of the time.

Luckily, for this particular regression problem there is an easy fix. Instead of comparing just M_1 and M_2, the scientist should compare all four possible models using posterior probabilities. In that case, the posterior probability of the full model that contains both proxies quickly converges to 1 indicating that a better model for the dependent variable would combine both proxies.

References

Doerge, R. W. (2002). "Mapping and analysis of quantitative trait loci in experimental populations." *Nature Reviews Genetics*, 3(1): 43–52. 31

Ferreira, M. A. R., Jaramillo, M. A., and Hypólito, E. B. (2024). "A model selection paradox with implications to multiscale modeling." In Ferreira, M. A. R. (ed.), *Modeling Spatio-Temporal Data*, 191–216. Chapman and Hall/CRC. 32

Kass, R. E. and Raftery, A. E. (1995). "Bayes Factor." *Journal of the American Statistical Association*, 90: 773–795. MR3363402. doi: https://doi.org/10.1080/01621459.1995.10476572. 32

Li, Q., Zheng, G., Liang, X., and Yu, K. (2009). "Robust tests for single-marker analysis in case-control genetic association studies." *Annals of Human Genetics*, 73(2): 245–252. 31

Williams, J., Ferreira, M. A. R., and Ji, T. (2022). "BICOSS: Bayesian iterative conditional stochastic search for GWAS." *BMC Bioinformatics*, 23(1): 1–14. 31

Williams, J., Xu, S., and Ferreira, M. A. R. (2023). "BGWAS: Bayesian variable selection in linear mixed models with nonlocal priors for genome-wide association studies." *BMC Bioinformatics*, 24(1): 1–20. 31

Xu, S., Williams, J., and Ferreira, M. A. R. (2023). "BG2: Bayesian variable selection in generalized linear mixed models with nonlocal priors for non-Gaussian GWAS data." *BMC Bioinformatics*, 24(1): 343. MR3732038. 31

Xu, S., Williams, J., Tegge, A., and Ferreira, M. A. R. (2025). "Genome-wide iterative fine-mapping for non-Gaussian phenotypes." *Scientific Reports*, 15(1): 30080. 31

Contributed Discussion

Pericchi G. Luis R.*

1 Discussion

This article is to be welcomed for its practical and theoretical contributions. The ideas put forward may find use in a very different context, in addition to the practical problem of model selection under the situation of missing observations. My questions are as follows:

1. **First Question**: The use of equation (23), adds an interesting variability, motivated by the missingness of the data. This may bring advantages even in traditional cases where there is no missing data, but the regressors are not fixed and are not designed. Have you considered this possibility?

2. **Second question**: Related to the previous question, in prediction, Bayesian Model Averaging (BMA) is the optimal predictor for the usual utility functions, and this requires the posterior probabilities of the different models. But, a big "but", for which regressors? If it is a prediction and we cannot control the regressors, as in economic and social data, for instance, then the regressors should not be considered fixed. This is a substantive application well beyond the area of traditional missing observations. The methods suggested here may prove useful in the prediction of a future non-identical to the past.

3. **Third question**: Assuming $g' = 1$, seems to imply the use of a g-prior without any mixing. Does this entail that there is "information inconsistency", a type of inconsistency when the sample size is fixed, but the non-centrality parameter grows without a bound, as described in Berger and Pericchi (2001))?. Or the methods of this article avoid it?

4. **Final Question**: Can the assumptions of "Informative Missingness" versus "Non-Informative Missingness" etc, be framed as a model selection problem in this setup?

*Department of Mathematics, University of Puerto Rico Rio Piedras, San Juan, PR, luis.pericchi@upr.edu

Contributed Discussion

Nadja Klein* and Nicolas Bianco*

We congratulate the authors for their contribution to the field of objective Bayes, taking up the important interplay between missing data and model uncertainty in linear regression problems. The proposed imputation g'-prior extends classical g-priors (Zellner, 1986) to settings with incomplete covariates by incorporating the covariance structure from the imputation model. This prior enables Bayesian variable selection and model comparison under missingness, while maintaining the principles of objective Bayes (Berger, 2006). Their framework offers a fully automatic, principled alternative to existing imputation and selection methods, with demonstrable improvements in both sensitivity and specificity across a range of simulated and real-data scenarios.

MNAR in spatial models. In this discussion, we aim to highlight the consequences of missing data on variable selection in spatial regression models. The authors assume that data are missing at random (MAR) or, more restrictively, missing completely at random (MCAR), and that the prior distribution for the parameters governing the missingness mechanism is independent of the prior for the remaining model parameters. Under these assumptions, the authors establish in Proposition 1 that the distribution of the missingness indicator matrix is ignorable. While MCAR is rarely plausible in spatial applications, MAR may hold when missingness arises from known sampling designs, such as administrative boundaries (Diggle et al., 2013). However, missing not at random (MNAR) is often more realistic in spatial contexts (Banerjee et al., 2014), for instance, when data are missing due to inaccessible terrain, ecologically sensitive areas, or measurement limitations that correlate with the variable of interest (e.g., pollution levels or ice depth in climate-vulnerable regions). In such cases, the assumptions underlying Proposition 1 are violated. To explore the implications of this violation, we conduct the following small simulation experiment examining how MNAR affects model uncertainty through variable selection.

Simulation design. Let $p = 10$ covariates be generated from Gaussian processes (GPs) $X_j(s) \sim GP(0, C(\theta))$, where $C(\theta)$ is an exponential covariance function, $\theta = \{\eta^2, \phi\}$, with marginal variance $\eta^2 = 1$ and range $\phi = 1$. The response variable is $Y(s) = X(s)^\top \beta + \varepsilon(s)$, where $\varepsilon(s) \sim N(0, 1)$, and β is the same as in the authors' paper, that is, $\beta = (1, 2, 0, 0, 0, 1, 2, 0, 0, 0)^\top$. We generate 100 observations $\{s_i, Y(s_i), X(s_i)\}_{i=1}^{100}$ where $s_i \in [0, 1]^2$. We consider a scenario without missing values in the response, while we introduce MNAR by setting the covariate values $X_j(s)$ to missing if they exceed the α-quantile of the empirical sample distribution, for $\alpha \in \{0.8, 0.9\}$, for $j = 1, \ldots, 10$.

Following the authors' notation, we consider a data-imputation mechanism that exploits the properties of GPs to derive the posterior predictive distribution $p(X_{j,(1)}|X_{j,(0)})$ of missing values given the observed values, which is an n_1-dimensional

*Scientific Computing Center, Karlsruhe Institute of Technology, Germany, nadja.klein@kit.edu; nicolas.bianco@kit.edu

	Model	SNR	$P(\gamma_0\|X,Y)$
$\alpha = 0.9$	M0	3.47	0.33
	M1	2.61	0.24
	M2	3.20	0.29
	M3	3.28	0.30
	M4	3.96	0.33

	Model	SNR	$P(\gamma_0\|X,Y)$
$\alpha = 0.8$	M0	3.41	0.32
	M1	1.37	0.11
	M2	1.78	0.15
	M3	2.21	0.21
	M4	2.65	0.25

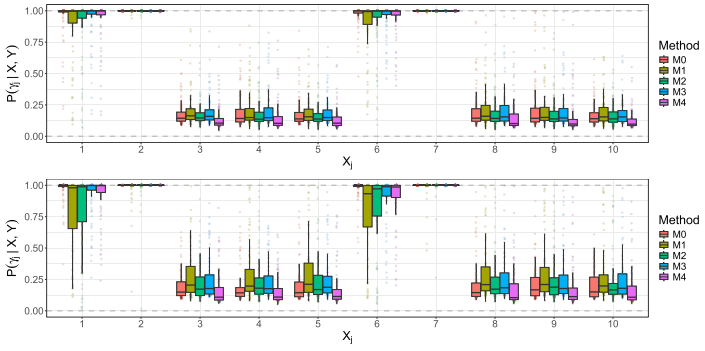

Figure 1: Averages over 100 replicates are shown for SNRs and posterior probability of the true model for the different methods (left) and distribution of the PIPs for each covariate (right). Top and bottom panels refer to $\alpha = 0.9, 0.8$, respectively. The true model is $\gamma_0 = (1,1,0,0,0,1,1,0,0,0)^\top$. Higher SNR values are preferable.

Gaussian distribution with mean $C_{10}(\theta)C_{00}^{-1}(\theta)X_{j,(0)}$ and covariance $C_{11}(\theta) - C_{10}(\theta)C_{00}^{-1}(\theta)C_{10}^{\top}(\theta)$.

We compare the variable selection performance based on the indicators γ_j for the following models. (M0) is the data-complete model. (M1) uses the objective imputation g'-prior proposed by the authors. (M2) assumes knowledge of the values of β by centering its prior around the true values and by treating g' as a tuning parameter that we fix at $g' = 0.05$. (M3) informs the prior on the imputation mechanism such that imputed values are above the respective α-quantile defined above. Finally, (M4) assumes an informed prior $p(\gamma) \propto 1/\binom{p}{p_\gamma} I(p_\gamma < 6)$, where $p_\gamma = \sum_{j=1}^{p} \gamma_j$. This gives zero prior probability to models of size greater than six.

Results and conclusion. In Figure 1, we report the posterior inclusion probabilities (PIPs) for each covariate, the signal-to-noise ratios (SNRs) of the PIPs as defined by the authors, and the posterior probability of the true model. The figure shows that subjective priors can help mitigate the effect of MNAR in spatial regression models. This observation has also been made in other contexts (e.g., Zhu et al., 2014; Alkan et al., 2017; Vera, 2023) and in non-Bayesian models via regularization (e.g., Tseng and Chen, 2019). The main advantage is the reduction in false negatives for the first and sixth covariate. This is even more evident when the percentage of missing values increases (here from 10% to 20% missing covariate values). A subjective prior $p(\gamma)$ on the model space in M4 seems to be the most effective specification in this scenario. Based on this small simulation experiment, we believe that further investigation into the impact of MNAR on variable selection in spatial models is a promising research avenue.

Funding

This work was partially supported through the TRR 391, Project A07. Nadja Klein acknowl-

edges funding through the Emmy Noether grant KL 3037/1-1. Both projects are funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation).

References

Alkan, N., Terzi, Y., and Cengiz, M. A. (2017). "Using informative priors for handling missing data problem in Cox regression." *Communications in Statistics - Simulation and Computation*, 46(10): 7614–7623. MR3764990. doi: https://doi.org/10.1080/03610918.2016.1248568. 36

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*. CRC Press, 2nd edition. MR3362184. 35

Berger, J. O. (2006). "The case for objective Bayesian analysis (with discussion)." *Bayesian Analysis*, 1: 385–402. MR2221271. doi: https://doi.org/10.1214/06-BA115. 35

Diggle, P. J., Moraga, P., Rowlingson, B., and Taylor, B. M. (2013). "Spatial and spatio-temporal log-Gaussian Cox processes: extending the geostatistical paradigm." *Statistical Science*, 28(4): 542–563. 35

Tseng, C.-H. and Chen, Y.-H. (2019). "Regularized approach for data missing not at random." *Statistical Methods in Medical Research*, 28(1): 134–150. URL https://journals.sagepub.com/doi/10.1177/0962280217717760 MR3894518. doi: https://doi.org/10.1177/0962280217717760. 36

Vera, J. D. (2023). "Bayesian selection model with shrinking priors for nonignorable missingness." Ph.D. dissertation, University of California, Los Angeles. URL https://escholarship.org/uc/item/5zz270bs MR4675793. 36

Zellner, A. (1986). "On assessing prior distributions and Bayesian regression analysis with g-prior distributions." In Goel, P. and Zellner, A. (eds.), *Bayesian Inference and Decision Techniques: Essays in Honour of Bruno de Finetti*, 233–243. MR0881437. 35

Zhu, H., Ibrahim, J. G., and Tang, N. (2014). "Bayesian sensitivity analysis of statistical models with missing data." *Statistica Sinica*, 24(2): 871–896. MR3235403. 36

Contributed Discussion

Christos Thomadakis* and Ioannis Ntzoufras*

Introduction We congratulate the authors on a substantial contribution to the Bayesian model selection literature. They propose a fully Bayesian framework for variable selection in linear models with missing values in both outcomes and covariates. Their method combines a covariate imputation model with a family of candidate regression models, under a missing at random (MAR) assumption, to derive posterior model probabilities and marginal likelihoods in a coherent and efficient way. Notably, they introduce the Imputation g'-Prior, which avoids dependence on incomplete design matrices and ensures intercept comparability across models. Their Monte Carlo scheme for estimating marginal likelihoods appears to outperform existing strategies such as "impute, then select" and "simultaneously impute and select" (SIAS) (Yang et al., 2005).

A practical concern with the imputation model While the theoretical framework is rigorous and resulting equations are valid, we raise a practical concern with its default implementation: excluding the outcome variable Y from the covariate imputation model, even when the missingness mechanism depends on observed Y, may induce bias in estimating the covariate distribution in real-world applications (McGowan et al., 2024). If Y is predictive of missing covariates and also informs their missingness probability, omitting it from the imputation model may violate the practical MAR conditions (van Buuren, 2018), rendering the mechanism effectively missing not at random (MNAR) and possibly affecting posterior model probabilities and selection.

Simulation study We simulated data with covariates $(X_1, X_2) \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \begin{pmatrix} 2.25 & 0.825 \\ 0.825 & 1.21 \end{pmatrix}$ and outcome $Y = 0.8X_1 + \epsilon$, with $\epsilon \sim N(0, 1)$. Missingness was introduced only in X_1 via a MAR mechanism dependent on Y: $P(M_1 = 1 \mid X_1, X_2, Y) = P(M_1 = 1 \mid Y) = (1 + e^{-1.5Y})^{-1}$, leading to 46% missingness.

In this setting, the authors' imputation model, which includes only covariates, implicitly assumes a stronger MAR condition: $P(M_1 = 1 \mid X_1, X_2) = P(M_1 = 1 \mid X_2)$, since X_2 is fully observed and X_1 may be missing in some units. However, under our setup,

$$P(M_1 = 1 \mid X_1, X_2) = \int P(M_1 = 1 \mid Y) f(Y \mid X_1) dY, \tag{1}$$

which depends on the possibly unobserved X_1, violating the previous condition and making the mechanism effectively MNAR.

We compared three methods: (1) observed likelihood estimation, (2) MICE with only X_1, X_2, and (3) MICE with X_1, X_2, Y. Approaches (1) and (2), which omit Y, yielded

*Department of Statistics, Athens University of Economics and Business, Athens, Greece, cthomadak@aueb.gr; ntzoufras@aueb.gr

biased estimates for $E(X_1)$, $\text{Var}(X_1)$, and $\text{Cov}(X_1, X_2)$, even though the marginal covariate model was correctly specified and the mechanism was theoretically MAR. In contrast, including Y (approach 3) produced nearly unbiased estimates (Table 1). These differences stem from the mismatch between theoretical and practical assumptions. We are worried about how these problems in the estimation could influence the proposed implemented Bayesian variable selection methodology. Hence, it would be informative to present results from simple simulation studies with different sample sizes (e.g., 100 and 500), where the probability of missing covariate data depends on observed Y's.

The use of a single imputation model The authors assume a fixed specification for the imputation model, using the full set of covariates. While this choice is justifiable and may perform well asymptotically, it could introduce bias or instability in finite samples, particularly if some covariates are weakly associated with those being imputed. On the other hand, selecting a parsimonious imputation model through variable selection would add considerable computational burden, especially in moderate- to high-dimensional settings. A practical compromise might be to pre-specify the imputation model based on external knowledge or a preliminary analysis, before applying the authors' methodology.

Scalability and computation when the marginal likelihood is not available Another concern is scalability. Even if the marginal likelihood of each candidate model is available in closed form for pseudo-complete data, computation becomes burdensome when the number of covariates is large, since the Monte Carlo scheme proposed in Section 3.4 must be implemented across 2^p models. Moreover, once we move beyond normal models, closed-form marginal likelihoods are rarely available. While Laplace approximations can be effective for a broad class of models, it is unclear how the presence of missingness affects their accuracy. A natural next step would therefore be to develop a Markov chain Monte Carlo (MCMC)-based method that can handle such cases more efficiently. It would also be valuable to hear the authors' views on extensions to high-dimensional settings and situations where the marginal likelihood is not available.

Computational details and comparison to SIAS The SIAS method (Yang et al., 2005) appears to be a reasonable approach for handling missing data, and one might therefore expect SIAS to perform equally well to the method proposed in this paper. Nevertheless, the authors' approach performs better, at least in the presented examples and simulations. The authors attribute this to computational limitations arising from the high dimensionality of the problem. However, they do not provide sufficient details on computation times or the number of iterations required in their examples. A fairer comparison would have been to report results under equal computational running times. Another open question is whether the two methods yield the same results as the number of iterations in SIAS tends to infinity. Furthermore, SIAS might perform better than the proposed method in scenarios with large p or in situations where the marginal likelihood is not available. We would greatly value the authors' insights on these issues.

Concluding remarks We thank the authors for their rigorous contribution to objective Bayesian model selection under missing data. Our remarks are intended to strengthen the practical applicability of their method. We highlight the potential importance of auxiliary variables such as Y in the imputation step, particularly under MAR mechanisms that depend on the outcome. We also raise questions regarding the interaction between model selection and imputation, computational details and scalability to high-dimensional data. These topics could generate further research on the topic in the near future.

Parameter	$E(X_1)$	$E(X_2)$	$Var(X_1)$	$Var(X_2)$	$Cov(X_1, X_2)$
True	0.000	0.000	2.250	1.210	0.825
(1) Obs. Likelihood (GLS)*	-0.587	-0.002	1.697	1.189	0.656
(2) MICE** (X_1, X_2)	-0.586	-0.002	1.790	1.201	0.664
(3) MICE (X_1, X_2, Y)	-0.001	-0.002	2.326	1.201	0.821

** Maximum observed likelihood estimates using `gls()` based on available data for each case;*
*** MICE: Multivariate Imputation by Chained Equations (*R *function* `mice`*).*

Table 1: Estimated means, variances, and covariance of (X_1, X_2) across different imputation strategies; results are based on 1000 simulated datasets with $n = 100$.

References

McGowan, L. D., Lotspeich, S. C., and Hepler, S. A. (2024). "The "Why" behind including "Y" in your imputation model." *Statistical Methods in Medical Research*, 33(6): 996–1020. PMID: 38625810.
URL https://doi.org/10.1177/09622802241244608 MR4755429. doi: https://doi.org/10.1177/09622802241244608. 38

van Buuren, S. (2018). *Flexible Imputation of Missing Data*. A Chapman & Hall book. CRC Press, Taylor & Francis Group.
URL https://books.google.gr/books?id=bLmItgEACAAJ 38

Yang, X., Belin, T. R., and Boscardin, W. J. (2005). "Imputation and Variable Selection in Linear Regression Models with Missing Covariates." *Biometrics*, 61(2): 498–506.
URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-0420.2005.00317.x MR2140922. doi: https://doi.org/10.1111/j.1541-0420.2005.00317.x. 38, 39

Contributed Discussion

Eric Chen*, Jim Griffin* and F. Javier Rubio*

1 Comments

We commend the authors of García-Donato et al. (2025) for their interesting publication, which provides a rigorous exploration of model uncertainty, variable selection, and missing data from an objective Bayesian perspective, highlighting how principled prior specification can lead to good performance in such challenging contexts.

Calibration of g' One of the main contributions in García-Donato et al. (2025) is the definition of the imputation g'-prior, which connects classical prior constructions, such as Zellner's g-prior, with the imputation model. Intuitively, the presence of missing values in the covariates increases model uncertainty. The definition of the g'-prior is rooted in its asymptotic connection to the g-prior, which, as noted by the authors, can be interpreted as an empirical version of the g'-prior. Under additional uncertainty (*e.g.*, from missing covariate values), a loss in the signal-to-noise ratio is expected in finite samples, potentially reducing the power to detect small effects or to distinguish them from spurious variables. A natural question arises: how should this additional uncertainty be incorporated into the choice of g'? The authors propose $g' = 1$, which corresponds to the limit of $g = n$, the choice used in the unit information g-prior. Given the logic of the unit information prior and the notion of the *fraction of missing information* (Carpenter et al., 2023), should reduced covariate information be incorporated into the calibration of the g'-prior?

A simple way to operationalise this idea, in the fixed-g' framework, is to incorporate information about the proportion of missingness into the choice of g'. To do so, define the $\boldsymbol{G}'$-prior $\pi_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \beta_0, \sigma, \boldsymbol{\nu}) = N_{k_{\boldsymbol{\gamma}}}\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid 0, \sigma^2 \boldsymbol{G} \Sigma_{\boldsymbol{\gamma}\boldsymbol{\gamma}}^{-1} \boldsymbol{G}\right)$, with

$$\boldsymbol{G}' = \begin{pmatrix} \varphi(n, n_{o1})^{1/2} & 0 & \dots & 0 \\ 0 & \varphi(n, n_{o2})^{1/2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \varphi(n, n_{op})^{1/2} \end{pmatrix},$$

where n_{oj} denotes the number of observed values for the jth variable; n is the total sample size. For instance, φ may be specified so that $0 < \varphi(n, n_{oj}) < 1$ for $n_{0j} < n$, while $\varphi(n, n) = 1$. This approach incorporates information about the number of missing values for each covariate by decreasing the value of the corresponding diagonal entry (marginal variance), while preserving the correlation, thereby inducing less sparsity and potentially aiding in the identification of small or moderate effects. In this framework, it is straightforward to define a mixture of $\boldsymbol{G}'$-priors by placing independent priors on each diagonal element of $\boldsymbol{G}'$, centred at the corresponding φ function.

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

Forms of misspecification The illustrations (particularly, the Boston housing data) show that a normal imputation model can lead to poorer performance if unsupported by the data. This raises questions about the trade-off between the specification of the regression and imputation models, which become increasingly difficult as the number of covariates increases, and variable selection performance. Within a standard Bayesian framework, Bayesian nonparametric models are natural, flexible choices to reduce misspecification of the imputation model but can be computationally demanding. The Weighted Bayesian Bootstrap (Newton et al., 2021) offers a way to build hybrid methods using non-Bayesian models (such as tree-based or deep learning models) to alleviate misspecification of the imputation model. Modular Bayesian inference cuts feedback from (misspecified) parts of the model to avoid distorting inference in a trusted (well-specified) part of the model. See Liu and Goudie (2025), and references within, for recent work. Roeling and Nicholls (2020) discuss an application to missing data in network autocorrelation models. Frazier and Nott (2025) discuss using generalized Bayes for inference on the untrusted part of the model. This offers a way of controlling the effects of misspecification of the imputation model on variable selection. Alternatively, the linear regression could be an approximation of the covariate-response relationship with a simple interpretation leading to distortion of the estimated imputation model. The use of modular methods would lead to something between the "Impute Then Select" and "Simultaneously Impute And Select".

References

Carpenter, J., Bartlett, J., Morris, T., Wood, A., Quartagno, M., and Kenward, M. (2023). *Multiple imputation and its application*. John Wiley & Sons, Ltd. 41

Frazier, D. T. and Nott, D. J. (2025). "Cutting Feedback and Modularized Analyses in Generalized Bayesian Inference." *Bayesian Analysis*, Advanced Publication. 42

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. 41

Liu, Y. and Goudie, R. J. B. (2025). "A general framework for cutting feedback within modularized Bayesian inference." *Journal of the Royal Statistical Society - Series B*, forthcoming. 42

Newton, M., Polson, N., and Xu, J. (2021). "Weighted Bayesian bootstrap for scalable posterior distributions." *Canadian Journal of Statistics*, 49(2): 421–437. MR4267927. doi: https://doi.org/10.1002/cjs.11570. 42

Roeling, M. P. and Nicholls, G. K. (2020). "Imputation of attributes in networked data using Bayesian autocorrelation regression models." *Social Networks*, 62: 24–32. 42

Contributed Discussion

Mark FJ Steel* and Gregor Zens†

1 Introduction

The authors are to be congratulated on tackling an often overlooked, yet very important, problem. Indeed, model uncertainty is typically more pronounced when the number of observations is limited and, thus, simply discarding incomplete data is far from an ideal strategy. Surprisingly, the previous literature has rarely addressed model uncertainty under missing data in a formal manner.

2 Imputation of the covariates

An important difference from the full-data approach is that we can no longer condition on fully observed covariates, but must assume a specific imputation mechanism for them. The authors use a joint multivariate Gaussian distribution for all covariates. While the method shows some robustness in cases where data are not really Gaussian, the multivariate normality assumption becomes quite implausible for highly skewed, binary, categorical, or small-count covariates. This issue is mentioned in the concluding section, but no clear avenues towards solving it are suggested. Since many real-world problems feature non-Gaussian covariates, this limitation is highly relevant in practice.

Could we perhaps build upon the idea of latent Gaussian models (Rue et al., 2009; Steel and Zens, 2025) to tackle this issue? In particular, assuming $\boldsymbol{z}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for an underlying p-variate latent variable with $i = 1, \ldots, n$, a covariate vector $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ could be imputed through either deterministic or stochastic mappings from the latent space. Deterministic functions $x_{ij} = f_j(z_{ij})$ can handle various covariate types, for instance $f_j(z_{ij}) = z_{ij}$ for roughly normal covariates, $f_j(z_{ij}) = \mathbb{I}(z_{ij} > 0)$ for binary (or binarized categorical) covariates, and $f_j(z_{ij}) = \frac{2z_{ij}}{\eta + \frac{1}{\eta}} \left(\frac{1}{\eta} \mathbb{I}(z_{ij} > 0) + \eta \mathbb{I}(z_{ij} \leq 0) \right)$ for skewed normal (as in Fernández and Steel, 1998) covariates, where $\eta > 0$ is a skewness parameter. Stochastic mappings specify conditional distributions, for example $x_{ij} \mid z_{ij} \sim \mathcal{P}(\exp(z_{ij}))$ for count-valued covariates.

Core components of the authors' approach, such as those given in (8) or (9), may still be computed in closed form in many cases. As an example, with one normal covariate $\boldsymbol{x}_{i1}$ and one binary covariate $\boldsymbol{x}_{i2}$, a natural imputation mechanism uses

$$\boldsymbol{z}_i = (z_{i1}, z_{i2})^\top \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & 1 \end{pmatrix},$$

*Dept. of Statistics, University of Warwick, m.steel@warwick.ac.uk
†International Institute for Applied Systems Analysis, zens@iiasa.ac.at

with $\sigma_{22} = 1$ being the typical probit normalization. Linking covariates via $x_{i1} = z_{i1}$ and $x_{i2} = \mathbb{I}(z_{i2} > 0)$ yields

$$\mathbb{E}[\boldsymbol{x}_i] = \begin{pmatrix} \mu_1 \\ \pi \end{pmatrix} \quad \mathbb{V}[\boldsymbol{x}_i] = \begin{pmatrix} \sigma_1^2 & \sigma_{12}\,\phi(\mu_2) \\ \sigma_{12}\,\phi(\mu_2) & \pi(1-\pi) \end{pmatrix}.$$

where $\pi = \Phi(\mu_2)$ and ϕ, Φ are the standard normal pdf and cdf. Similar expressions likely exist for count and skew-normal data. Such expressions for the covariance matrix could then directly be used in the imputation g'-prior on the regression coefficients in (23). The posterior of (some elements of) $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ might be more diffuse than with normally distributed covariates, but that merely reflects weaker likelihood information for some covariates (e.g., binary ones). In any case, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are nuisance parameters.

3 Distributional uncertainty

In addition to variable selection, the authors investigate model uncertainty limited to the error distribution specification. This leads to an improper uniform prior over all regression coefficients (now common to all models). One query and a (somewhat nitpicky) comment on this:

- In practical applications, both types of uncertainty likely occur simultaneously. Since regression coefficients would then no longer be common to all models, would the appropriate prior still be the one in (23)?
- One example mentioned is choosing between multivariate normal and multivariate Student-t distributions. However, under priors like (27), the Bayes factor between them collapses to a data-independent constant (see Rubio and Steel, 2018) because the multivariate t is just a single-scale scale-mixture of normals. In contrast, with i.i.d. Student-t errors (with *observation-specific* scale-mixture representation), the Bayes factor is data-dependent and can meaningfully discriminate.

Finally, we wish to thank the authors for a wonderful and thought-provoking paper, which we fully expect to inspire exciting new research directions.

References

Fernández, C. and Steel, M. F. J. (1998). "On Bayesian Modeling of Fat Tails and Skewness." *Journal of the American Statistical Association*, 93(441): 359–371. MR1614601. doi: https://doi.org/10.2307/2669632. 43

Rubio, F. J. and Steel, M. F. J. (2018). "Flexible Linear Mixed Models with Improper Priors for Longitudinal and Survival Data." *Electronic Journal of Statistics*, 12: 572–598. MR3769189. doi: https://doi.org/10.1214/18-EJS1401. 44

Rue, H., Martino, S., and Chopin, N. (2009). "Approximate Bayesian Inference for Latent Gaussian models by using Integrated Nested Laplace Approximations." *Journal of the Royal Statistical Society, Ser. B*, 71: 319–392. MR2649602. doi: https://doi.org/10.1111/j.1467-9868.2008.00700.x. 43

Steel, M. F. J. and Zens, G. (2025). "Model Uncertainty in Latent Gaussian Models with Univariate Link Function." *Bayesian Analysis*, forthcoming. 43

Contributed Discussion

Steven N. MacEachern*

I commend the authors for producing a paper that addresses one of the core problems in our discipline – namely how to create an effective regression modelling strategy when the list of "active" variables is unknown and, specifically, how to effectively handle uncertainty in which variables are active. The paper deals with this problem when some of the data are missing. This partial missingness of observations (item nonresponse in the words of survey sampling) is so common that we might well consider it to be more standard than the setting where the data are said to be complete. The results in the paper are nice. An additional strength of the paper is the extraordinary clarity of the authors' writing, including text, notation and mathematics.

Bayesian model averaging (BMA) has been a huge success for Bayesian statistics. The basic setup for BMA matches the authors setup: there is a finite collection of potential models, indexed by the model indicator, $\boldsymbol{\gamma}$. For each model, there is a prior distribution over the unknown parameters in the model, here a set of regression coefficients and a few additional parameters that might index the scale or shape of the error distribution. These prior distributions consist of a mix of proper and improper portions – improper on common parameters and proper on model-specific parameters. The authors and BMA part ways when describing the goal of the analysis. BMA's typical focus is prediction while the authors' primary interest is in determining which variables are active. Both BMA and the authors rely on Bayes theorem, and both produce results that are convincingly better than many other methods. Simply put, Bayes theorem works!

The success of BMA, is, in my mind, closely tied to results in decision theory. When predictions are evaluated with squared error loss, one quickly passes from prediction error to fitting error, and then to the error in estimating regression coefficients. The old results on Stein estimation transfer to the regression problem. We have both the admissibility of the Bayes estimator and the inadmissibility of such things as the least squares estimator or the least squares estimator after selection of a single model. Averaging over one's uncertainty is effective;[1] ignoring the uncertainty, whether through the device of fitting a model in a very large space or through model selection followed by inference on a given model appears to be less effective. The advantages of averaging are large enough, that, even if one has a relatively poor Bayesian model, one sees gains.

Decision theory provides less for us when it comes to identification of the active variables. The traditional loss functions, of the 0/1 sort, are discontinuous. We know that Bayesian methods (with proper prior distributions) produce admissible estimators for both model selection ($L(\boldsymbol{\gamma}, a) = I(\boldsymbol{\gamma} \neq a)$) and for the selection of individual variables ($L(\boldsymbol{\gamma}, \mathbf{a}) = \sum_j I(\boldsymbol{\gamma}_j \neq \mathbf{a}_j)$). The former loss function leads to the highest probability model while the latter leads to Barbieri and Berger's median probability model.

*Department of Statistics, The Ohio State University, snm@stat.osu.edu

[1] I first saw model averaging for regression in work done in the mid 1980s by Colin Mallows, Lorraine Denby, and Mohan Boodram. They demonstrated the effectiveness of a variety of mechanisms for averaging relative to then-current classical techniques. I have been unable to locate a reference.

Alternative loss functions directly assess the value of a variable in reducing prediction error and also consider the cost of "purchasing" a predictor for future use (e.g., Miyawaki and MacEachern (2023) and the references therein). These alternative loss functions do two things: they move the model/variable selection problem closer to the model averaging problem and they look at the decision problem through the lens of economics. Rephrasing the problem with a loss that is tied to prediction highlights the role of the distribution of the covariates. We might well make a different decision on whether a variable has an important effect based on the distribution of future covariates. Along with this, we might well make a different decision on which covariates to purchase.

The economic rephrasing of data collection raises an important question when combined with the authors' focus on missing data. The decision on whether to purchase a specific covariate for a particular case may depend on the value of the other variables for that case. Such case-specific decisions naturally produce missing data. That is, a fully specified decision problem may tell us to produce a data set with missingness. For some specifications, this will lead to data that are missing at random (MAR).

The authors' approach the model/variable selection problem from an objective Bayesian viewpoint, with modifications needed to handle missing data. My own preference for an analysis runs in the direction of actively modelling the data. To me, this includes consideration of transformations of the variables (both response and potential predictors), creation of new variables, and techniques to handle unusual patterns in the data.

There are two main paths for an active analysis. One is to examine the entire data set, searching for any patterns that might be present, and pursuing those with an apparent effect that is substantially larger than noise. After roughing out the form of the model, or a potential set of models, the model (meta-model) is supplemented with the prior distribution. Here, a part of the effort would include building the model for the missing data. Once all elements are in place, it's on to Bayes theorem and the resulting analysis. Although I often make use of this approach, I do feel the discomfort that comes with using the single data set to form the potential models before proceeding with the Bayesian analysis.

The second approach involves data splitting, where the analyst uses one portion of the data (typically a randomly selected subset) to form the model, processes things internally, and reports their partial posterior distribution. Bayes theorem is then applied to the remainder of the data, taking us to the full-data posterior. For the ozone data, I find substantial departures from linearity and normality and believe that we need to construct a model as is done in Yu et al. (2011). As with beauty, it seems that normality is in the eye of the beholder.

I believe that a well-trained analyst adds value by examining a data set and reacting to patterns in it to create a better model, irrespective of whether the final analysis is classical or Bayesian. Additional benefits follow from a careful description of the loss. The authors' framework appears to be fully compatible with this active-analyst approach. It would be interesting to see whether active modelling, perhaps coupled with objective Bayesian methods for the meta-model, improves predictions and identification of active variables.

Funding

The author was supported by NSF Grant DMS-2413823.

References

Miyawaki, K. and MacEachern, S. N. (2023). "Economic variable selection." *The Canadian Journal of Statistics*, 51: 19–37.
URL https://doi.org/10.1002/cjs.11675 MR4551813. doi: https://doi.org/10.1002/cjs.11675. 47

Yu, Q., MacEachern, S. N., and Peruggia, M. (2011). "Bayesian Synthesis: Combining subjective analyses, with an application to ozone data." *The Annals of Applied Statistics*, 5(2B): 1678–1698.
URL https://doi.org/10.1214/10-AOAS444 MR2849791. doi: https://doi.org/10.1214/10-AOAS444. 47

Contributed Discussion

Leonardo Egidi*

1 Introduction

I would like to congratulate the authors on their impressive contribution to the Bayesian model uncertainty framework in the presence of missing data. Quite surprisingly, the literature still shows a serious gap regarding the use of effective (objective) Bayesian tools to handle random missingness and to obtain reliable posterior distributions. In variable selection problems, missing data significantly complicate the computational steps: indeed, many prior proposals such as Zellner's g-priors (Zellner, 1986) and their extensions are not directly applicable, since they rely on a fixed design matrix. The main novelty of the paper is to allow the covariates of a regression model to vary according to a well-defined mechanism and, in this light, to retrieve valid tools such as Bayes factors and posterior model probabilities.

Several aspects of the paper would be worth further discussion: Γ-open perspective where the true model does not exist or does not belong to the model space; applications to Gaussian mixed-effects models; and connections between the imputed g'-Bayes factor in Section 4.2 and leave-one-out cross-validation scores (Fong and Holmes, 2020). However, in the following section I take up one of the authors' suggestions for future research in the variable selection setting, namely the specification of the hyper-g' imputation prior for g' according to a full Bayes approach. The resulting imputed hyper-g'-Bayes factor is less tractable than the variant in Proposition 2 of the paper; nonetheless, the analysis below sheds light on the difficulties inherent in such a prior specification, and may hopefully serve as a basis for future discussion.

2 Hyper-g prior for g'?

Consider to be framed in a regression setting where the covariates are random and specified through the Equations (5), (8), and (9) in the paper. Dealing with a fixed value for g' in Equation (23) - the authors used $g' = 1$ in their numerical experiments - yields a tractable imputed g'-Bayes factors and allows then to retrieve posterior model probabilities. Many choices for the specification of g are given in the literature in the classical context of deterministic covariates and with no missing data. Among the others, Kass and Wasserman (1995) suggest the use of *unit information priors* by choosing $g = n$, whereas Foster and George (1994) opt for $g = p^2$ from a minimax perspective. George and Foster (2000) show how to calibrate g using criteria such as the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC); empirical Bayes methods are proposed by George and Foster (2000), Clyde and George (2000) and Hansen and Yu (2001). Full Bayesian approaches treating g as a random component

*Department of Economics, Business, Mathematics, and Statistics "*Bruno de Finetti*", University of Trieste, Italy, legidi@units.it

(Bayarri et al., 2012) lead instead to a variety of prior choices. Among these, hyper-g priors were proposed by Liang et al. (2008) as a reliable extension of Zellner's g-prior (Zellner, 1986) and as an alternative to the Zellner–Siow prior (Zellner and Siow, 1980), which can be viewed as a mixture of g-priors. The main advantage of the hyper-g prior is that both the model's marginal posterior distribution of g and the Bayes factor, $B_{\boldsymbol{\gamma}0}\left(\boldsymbol{y}_{(0)}, \boldsymbol{x}_{(0)}\right)$, are available in closed form.

We extend the prior setup detailed in the paper through Equations (20), (21), and (23), by combining the imputation g'-prior in Equation (23) in the paper with an hyper-g prior for g' as follows - we refer to their Section 2.1 for the whole notation:

$$\begin{aligned} \pi_{\boldsymbol{\gamma}}\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \,\middle|\, \beta_0, \sigma, \boldsymbol{\nu}\right) &= \mathcal{N}_{k_{\boldsymbol{\gamma}}}\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \;\middle|\; \mathbf{0},\, g'\,\sigma^2\,\Sigma_{\boldsymbol{\gamma}\boldsymbol{\gamma}}^{-1}\right) \\ \pi(g') \;&= \frac{a-2}{2}\,(1+g')^{-a/2}, \qquad g' > 0, \end{aligned} \tag{I}$$

with $a > 2$ to ensure to deal with a proper prior distribution - see Liang et al. (2008) for the related discussion on the choice of a. The purpose is to re-evaluate the imputed g'-Bayes factor found in the Proposition 2, $B_{\boldsymbol{\gamma}0}\left(\boldsymbol{y}_{(0)}, \boldsymbol{x}_{(0)}\right)$. According to the proof in the Section B.3 of the supplementary material, the imputed (hyper) g'-Bayes factor is

$$B_{\boldsymbol{\gamma}0}\left(\boldsymbol{y}_{(0)}, \boldsymbol{x}_{(0)}\right) = \frac{m_{\boldsymbol{\gamma}}\left(\boldsymbol{y}_{(0)}, \boldsymbol{x}_{(0)}\right)}{m_0\left(\boldsymbol{y}_{(0)}, \boldsymbol{x}_{(0)}\right)} = \mathbb{E}\left[\frac{m_{\boldsymbol{\gamma}}\left(\boldsymbol{y}_{(0)} \mid \boldsymbol{x}_{(0)}, \boldsymbol{x}_{(1)}, \boldsymbol{\nu}\right)}{m_0\left(\boldsymbol{y}_{(0)} \mid \boldsymbol{x}_{(0)}, \boldsymbol{x}_{(1)}, \boldsymbol{\nu}\right)}\right], \tag{II}$$

where the expectation is with respect to the posterior $x^{(1)}, \boldsymbol{\nu} \mid x^{(0)}$ - the last identity holds because m_0 is a constant inside the expectation. Using Equations (17) and (18) in the paper for the definition of $m_{\boldsymbol{\gamma}}\left(\boldsymbol{y}_{(0)} \mid \boldsymbol{x}_{(0)}, \boldsymbol{x}_{(1)}, \boldsymbol{\nu}\right)$ and $m_0\left(\boldsymbol{y}_{(0)} \mid \boldsymbol{x}_{(0)}, \boldsymbol{x}_{(1)}, \boldsymbol{\nu}\right)$, their ratio $m_{\boldsymbol{\gamma}}/m_0$ in the right side of Equation (II) can be determined as follows after some algebraic manipulations:

$$S_0^{\frac{(n_0-1)}{2}}\,\frac{a-2}{4}\int_0^{\infty} g'^{-p_{\boldsymbol{\gamma}}/2}\,|\Sigma_{\boldsymbol{\gamma}\boldsymbol{\gamma}}|^{1/2}\,|A(g')|^{-1/2}\,S(g')^{-\frac{(n_0-1)}{2}}\,(1+g')^{-a/2}\,dg', \tag{III}$$

where S_0 is $n_0 - 1$ times the sample variance of $\boldsymbol{y}_{(0)}$. The two functions $A(g) = \bar{X}_{\boldsymbol{\gamma}}^t \bar{X}_{\boldsymbol{\gamma}} + \frac{1}{g'}\Sigma_{\boldsymbol{\gamma}\boldsymbol{\gamma}}$ and $S(g) = S_0 - b^t A(g)^{-1} b$, with $b = \bar{X}_{\boldsymbol{\gamma}}^t \boldsymbol{y}_{(0)}$, reduce to $(1 + 1/g')\Sigma_{\boldsymbol{\gamma}\boldsymbol{\gamma}}$ and $S_0 - \frac{g'}{1+g'}SSR$, with $SSR = b^t \Sigma_{\boldsymbol{\gamma}\boldsymbol{\gamma}}^{-1} b$, respectively, in the special Zellner's g-prior case where $\Sigma_{\boldsymbol{\gamma}\boldsymbol{\gamma}} = \bar{X}_{\boldsymbol{\gamma}}^t \bar{X}_{\boldsymbol{\gamma}}$. Then, the *imputed hyper g'-Bayes factor* $B_{\boldsymbol{\gamma}0}\left(\boldsymbol{y}_{(0)}, \boldsymbol{x}_{(0)}\right)$ in (II) is the average of Equation (III) with respect to the posterior $x^{(1)}, \boldsymbol{\nu} \mid x^{(0)}$. This quantity should be computed numerically, since the integral in (III) does not admit a general closed-form expression. One could attempt to approximate the integral using Laplace's method or other numerical integration techniques, such as quadrature rules - Gaussian hypergeometric functions do not easily apply here as in Liang et al. (2008).

This small computation is intended to shed light on the challenges underlying the implementation of the hyper-g prior, and it may encourage further research toward a fully Bayesian (objective) approach to missing data scenarios.

References

Bayarri, M. J., Berger, J. O., Forte, A., and García-Donato, G. (2012). "Criteria for Bayesian model choice with application to variable selection." *The Annals of Statistics*, 40(3): 1550–1577. MR3015035. doi: https://doi.org/10.1214/12-AOS1013. 50

Clyde, M. and George, E. I. (2000). "Flexible empirical Bayes estimation for wavelets." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 62(4): 681–698. MR1796285. doi: https://doi.org/10.1111/1467-9868.00257. 49

Fong, E. and Holmes, C. C. (2020). "On the marginal likelihood and cross-validation." *Biometrika*, 107(2): 489–496. MR4108941. doi: https://doi.org/10.1093/biomet/asz077. 49

Foster, D. P. and George, E. I. (1994). "The risk inflation criterion for multiple regression." *The Annals of Statistics*, 22(4): 1947–1975. MR1329177. doi: https://doi.org/10.1214/aos/1176325766. 49

George, E. and Foster, D. P. (2000). "Calibration and empirical Bayes variable selection." *Biometrika*, 87(4): 731–747. MR1813972. doi: https://doi.org/10.1093/biomet/87.4.731. 49

Hansen, M. H. and Yu, B. (2001). "Model selection and the principle of minimum description length." *Journal of the American Statistical Association*, 96(454): 746–774. MR1939352. doi: https://doi.org/10.1198/016214501753168398. 49

Kass, R. E. and Wasserman, L. (1995). "A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion." *Journal of the American Statistical Association*, 90(431): 928–934. MR1354008. 49

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). "Mixtures of g priors for Bayesian variable selection." *Journal of the American Statistical Association*, 103(481): 410–423. MR2420243. doi: https://doi.org/10.1198/016214507000001337. 50

Zellner, A. (1986). "On assessing prior distributions and Bayesian regression analysis with g-prior distributions." *Bayesian inference and decision techniques*, 233–243. MR0881437. 49, 50

Zellner, A. and Siow, A. (1980). "Posterior odds ratios for selected regression hypotheses." *Trabajos de estadística y de investigación operativa*, 31(1): 585–603. 50

Contributed Discussion

Yichen Ji*, Monica J. Alexander†, and Radu V. Craiu‡

Introduction

We congratulate the authors (henceforth, GDCCQF) on their contribution to the literature on model selection with missing data. Our discussion was inspired by the requirement to estimate the joint distribution of the covariates, $f(X_1, \ldots, X_k|\nu)$, which is at the core of their method. The problem of estimating a joint distribution is central to modern statistics and its positive resolution has been known to impact other influential methods such as the knockoff method (Candes et al., 2018), the study of imputation efficiency in regression models (White and Carlin, 2010), or sample surveys (Gelman et al., 1998).

Estimation of the joint distribution of covariates

Although GDCCQF use, for simplicity, a multivariate Gaussian to generate the covariates and also to model their joint distribution, $f(X_1, \ldots, X_k|\nu)$, real data can depart from such an ideal setup in several ways. First, some covariates may have marginal distributions that are not Gaussian or even continuous, thus falsifying the multivariate Gaussian assumption. Second, the type of dependence captured by the Gaussian distribution may differ substantially from the dependence patterns exhibited by f. For example, it is well known that the tail dependence coefficients are zero for multivariate Gaussian distributions, but not so for other multivariate laws (see Hua and Joe, 2011, and references therein). A general mathematical framework for such comparisons is provided by the copula function which links the marginal and joint distributions of a multivariate vector (Sklar, 1959; Genest and Rivest, 1993). Furthermore, copulas have been increasingly used in the development of statistical methods for multivariate-dependent data (e.g., Genest et al., 2007; Dissmann et al., 2013; Hasler et al., 2018; Zimmerman et al., 2024; Pan et al., 2025). These developments are accompanied by software packages (e.g., Hofert et al., 2020) or programs that make it easier to implement copula-based techniques.

A small numerical study

A copula formulation allows us to study empirically how the performance of GDCCQF's variable selection procedure is impacted when the marginals and the dependence structure of f are misspecified. The analysis produced by GDCCQF's programs assumes

*Department of Statistical Sciences, University of Toronto, yc.ji@mail.utoronto.ca
†Department of Sociology and Department of Statistical Sciences, University of Toronto, monica.alexander@utoronto.ca
‡Department of Statistical Sciences, University of Toronto, radu.craiu@utoronto.ca

that f is a multivariate Gaussian distribution which is equivalent to a Gaussian copula model in which the marginals are all Gaussian. We considered data under three simulation scenarios that vary the copula and the marginals as follows:

CG We used a multivariate Clayton copula in which the Kendall tau between each pair of variables is 0.8, and all marginals are standard Gaussian. This copula choice introduces a strong lower tail dependence, unlike the posited model, which assumes that there is no tail dependence.

CL Same as **CG** but with marginal densities that are generalized Gaussian

$$g(x) \propto \exp(-|x|^8),\ \forall x \in \mathbb{R}$$

and thus have lighter tails than the posited model.

CH Same as **CL** but with marginal densities that are mixtures of a standard Gaussian (weight is 0.2) and an Exponential with parameter 2,

$$g(x) = 0.2\phi(x) + 1.6\exp(-2x)\mathbf{1}_{\{x\geq 0\}}(x),\ \forall x \in \mathbb{R},$$

where ϕ is the density of a standard normal, and $\mathbf{1}_{\{x\geq 0\}}(x)$ is equal to one if $x \geq 0$ and is zero otherwise. This yields marginals with a heavier right tail than in the posited model.

For all scenarios, each generated data set contained $n = 300$ observations, the number of covariates under consideration was $k = 10$, with active covariates X_1, X_2, X_6, X_7 having the corresponding regression coefficients, $\beta_1 = 1$, $\beta_2 = 2$, $\beta_6 = 1$ and $\beta_7 = 2$, and the missing probability was set to $p = 0.1$ under a missing completely at random (MCAR) scheme. We followed GDCCQF to produce $nMC = 500$ imputations. Each scenario was independently replicated $R = 500$ times. Table 1 presents the false negative (FN) and false positive (FP) rates for the three simulation scenarios when the data are analyzed assuming a multivariate Gaussian distribution. We note that the impact on the selection of active covariates seems to vary according to the size of their effect and the missing patterns produced in each replicate. Given that the error rates are not very high, we would need more replicates in order to see similar FP rates for all non-active covariates.

Conclusion

The simulations suggest that misspecification of the dependence can alter the performance of the method. Performance degradation is amplified when marginals are also misspecified. More work is required to understand if the difference in tail dependence between the Clayton copula and the Gaussian copula is responsible for these errors, or other copulas more similar to the Gaussian, such as a t copula, can also wreak havoc. Model misspecification is a well-known issue in Bayesian analysis, but in this case it can be realistically addressed by considering copula models to fit f. Our contribution to this discussion is not meant to be a criticism of GDCCQF's method, but is rather

Scenario	Covariate									
	$\mathbf{X_1}$	$\mathbf{X_2}$	X_3	X_4	X_5	$\mathbf{X_6}$	$\mathbf{X_7}$	X_8	X_9	X_{10}
CG	0	0	0.216	0	0.108	0.002	0	0.002	0.002	0
CL	0.120	0	0.11	0	0.11	0.548	0.01	0	0	0
CH	0.108	0	0.318	0.002	0.106	0.008	0	0.014	0.002	0.002

Table 1: Error rates for each variable in the three scenarios. For the active covariates shown in bold, the error rates represent the fraction of false negatives, while for the remaining inactive covariates the error rates represent the fraction of false positives.

aimed at stimulating the development of flexible estimation methods of multivariate distributions when part of the data are missing. We thank GDCCQF for an inspiring article that opens several directions for further study.

References

Candes, E., Fan, Y., Janson, L., and Lv, J. (2018). "Panning for gold:'model-X'knockoffs for high dimensional controlled variable selection." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(3): 551–577. 52

Dissmann, J., Brechmann, E. C., Czado, C., and Kurowicka, D. (2013). "Selecting and estimating regular vine copulae and application to financial returns." *Computational Statistics & Data Analysis*, 59: 52–69. 52

Gelman, A., King, G., and Liu, C. (1998). "Not asked and not answered: Multiple imputation for multiple surveys." *Journal of the American Statistical Association*, 93(443): 846–857. 52

Genest, C., Favre, A.-C., Beliveau, J., and Jacques, C. (2007). "Metaelliptical copulas and their use in frequency analysis of multivariate hydrological data." *Water Resources Research*, 43(9). 52

Genest, C. and Rivest, L.-P. (1993). "Statistical inference procedures for bivariate Archimedean copulas." *Journal of the American Statistical Association*, 88: 1034–1043. 52

Hasler, C., Craiu, R. V., and Rivest, L.-P. (2018). "Vine Copulas for Imputation of Monotone Non-response." *International Statistical Review*, 86(3): 488–511. 52

Hofert, M., Kojadinovic, I., Maechler, M., and Yan, J. (2020). *copula: Multivariate Dependence with Copulas*. R package version 1.0-1. 52

Hua, L. and Joe, H. (2011). "Tail order and intermediate tail dependence of multivariate copulas." *Journal of Multivariate Analysis*, 102(10): 1454–1471. 52

Pan, R., Nieto-Barajas, L. E., and Craiu, R. V. (2025). "Bayesian Nonparametric Mixtures of Archimedean Copulas." *Journal of Agricultural, Biological and Environmental Statistics*, 1–25. 52

Sklar, A. (1959). "Fonctions de répartition à n dimensions et leurs marges." *Publications de l'Institut de Statistique de l'Université de Paris*, 8: 229–231. 52

White, I. R. and Carlin, J. B. (2010). "Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values." *Statistics in medicine*, 29(28): 2920–2931. 52

Zimmerman, R., Craiu, R. V., and Leos-Barajas, V. (2024). "Copula modeling of serially correlated multivariate data with hidden structures." *Journal of the American Statistical Association*, 119(548): 2598–2609. 52

Contributed Discussion

Joris Mulder*

García-Donato et al. (2025) present a methodology for handling missing data in a model selection problem using an objective Bayesian approach. The current comment discusses an alternative, existing objective Bayesian method for this problem. First, rather than using the g prior, O'Hagan's fractional Bayes factor (FBF; O'Hagan, 1995) is utilized based on a minimal fraction.[1] Second, and more importantly due to the focus on missing data, Rubin's rules for multiple imputation can directly be used as the fractional Bayes factor can be written as a Savage-Dickey density ratio for a variable selection problem. This attractive property of a Savage-Dickey density ratio was shown by Hoijtink et al. (2018). The use of (adjusted) fractional Bayes factors for a testing problem of a set of predefined equality and/or one-sided constrained hypotheses in the case of missing data was shown in Mulder and Gu (2022) and Mulder et al. (2021). The current comment derives the methodology for a variable selection problem (which is a special case of the above testing problem). Moreover, its implied behavior is illustrated in a numerical experiment, showing competitive results as the method of García-Donato et al. (2025). Throughout this comment, the same notation is used as García-Donato et al. (2025).

The FBF of a model $\boldsymbol{\gamma} \in \Gamma$ against the full model where $\boldsymbol{\gamma}_{full} = \mathbf{1}$ can be written as a Savage-Dickey density ratio (e.g., Mulder and Gu, 2022; Dickey, 1971)

$$B^F_{\boldsymbol{\gamma},\boldsymbol{\gamma}_{full}}(b) \equiv \frac{m_{\boldsymbol{\gamma}}(\mathbf{d})/m_{\boldsymbol{\gamma}}(\mathbf{d}^b)}{m_{\boldsymbol{\gamma}_{full}}(\mathbf{d})/m_{\boldsymbol{\gamma}_{full}}(\mathbf{d}^b)} = \frac{\pi_{\boldsymbol{\gamma}_{full}}(\boldsymbol{\beta}_{-\boldsymbol{\gamma}} = \mathbf{0}|\mathbf{d})}{\pi_{\boldsymbol{\gamma}_{full}}(\boldsymbol{\beta}_{-\boldsymbol{\gamma}} = \mathbf{0}|\mathbf{d}^b)}, \tag{1}$$

where $\boldsymbol{\beta}_{-\boldsymbol{\gamma}}$ denotes the vector of coefficients under the full model which are excluded in model $\boldsymbol{\gamma}$, $m_{\boldsymbol{\gamma}}(\mathbf{d}^b) = \int\int f_{\boldsymbol{\gamma}}(\mathbf{d}|\boldsymbol{\beta}_{\boldsymbol{\gamma}},\boldsymbol{\alpha})^b \pi^N(\boldsymbol{\beta}_{\boldsymbol{\gamma}},\boldsymbol{\alpha}) d\boldsymbol{\beta}_{\boldsymbol{\gamma}} d\boldsymbol{\alpha}$ denotes the marginal likelihood of model $\boldsymbol{\gamma}$ when raising the likelihood to a fraction b, and the numerator (denominator) on the right hand side denotes the marginal posterior (marginal fractional prior (e.g., Gilks, 1995)) under the full model evaluated at the null values of the excluded parameters, i.e., $\boldsymbol{\beta}_{-\boldsymbol{\gamma}} = \mathbf{0}$. Thus, the FBF can be written as a ratio of a posterior and a fractional prior quantity under the full model.

Hence, we can directly apply Rubin's rules for multiple imputation under the full model. For the posterior, this implies

$$\begin{aligned} \pi_{\boldsymbol{\gamma}_{full}}(\boldsymbol{\beta}_{-\boldsymbol{\gamma}} = \mathbf{0}|\mathbf{d}_{(0)}) &= \int \pi_{\boldsymbol{\gamma}_{full}}(\boldsymbol{\beta}_{-\boldsymbol{\gamma}} = \mathbf{0}|\mathbf{d}_{(0)},\mathbf{d}_{(1)}) m_{\boldsymbol{\gamma}_{full}}(\mathbf{d}_{(1)}|\mathbf{d}_{(0)}) d\mathbf{d}_{(1)} \\ &= \text{AVE}[\pi_{\boldsymbol{\gamma}_{full}}(\boldsymbol{\beta}_{-\boldsymbol{\gamma}} = \mathbf{0}|\mathbf{d}_{(0)},\mathbf{d}_{(1)})] \end{aligned} \tag{2}$$

where AVE[·] refers to the average based on repeated imputations drawn from the posterior predictive distribution of missing data given the observed data under the full

*Department of Methodology and Statistics, Tilburg University, j.mulder3@tilburguniversity.edu

[1] Note that the implied fractional prior has a similar covariance structure as the g prior (e.g., see Mulder, 2014).

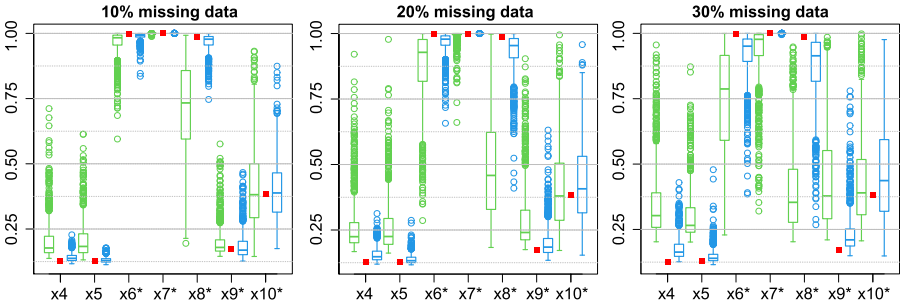

Figure 1: Boxplots of the inclusion probabilities for each variable using FBFs after list-wise deletion (green) and imputed FBFs via Savage-Dickey ratios (blue) when considering a proportion of 0.10 (left), 0.20 (middle), or 0.30 (right) of missing values per variable, for the Ozone dataset. The inclusion probabilities based on the oracle FBF using the full dataset are depicted as red squares. The variable names with '$*$' contained missing observations.

model, $m_{\boldsymbol{\gamma}_{full}}(\mathbf{d}_{(1)}|\mathbf{d}_{(0)})$ (e.g., Rubin, 1996). For the fractional prior, we write

$$\begin{aligned}\pi_{\boldsymbol{\gamma}_{full}}(\boldsymbol{\beta}_{-\boldsymbol{\gamma}} = \mathbf{0}|\mathbf{d}^b_{(0)}) &\equiv \int \pi_{\boldsymbol{\gamma}_{full}}(\boldsymbol{\beta}_{-\boldsymbol{\gamma}} = \mathbf{0}|(\mathbf{d}_{(0)}, \mathbf{d}_{(1)})^b) m_{\boldsymbol{\gamma}_{full}}(\mathbf{d}_{(1)}|\mathbf{d}_{(0)}) d\mathbf{d}_{(1)} \\ &= \text{AVE}[\pi_{\boldsymbol{\gamma}_{full}}(\boldsymbol{\beta}_{-\boldsymbol{\gamma}} = \mathbf{0}|(\mathbf{d}_{(0)}, \mathbf{d}_{(1)})^b)]. \qquad (3)\end{aligned}$$

Hence, similar as for the posterior quantity, the posterior predictive distribution based on all observed data, $\mathbf{d}_{(0)}$, is used to compute the fractional prior quantity. Note that if a minimal fraction of the observed data would have been used in the predictive distribution, the imputed missing data would be unrealistically heterogeneous. Moreover, by taking a fraction of the observed data and the missing data, i.e., $\pi_{\boldsymbol{\gamma}_{full}}(\boldsymbol{\beta}_{-\boldsymbol{\gamma}} = \mathbf{0}|(\mathbf{d}_{(0)}, \mathbf{d}_{(1)})^b)$, the fractional prior is again based on a fraction of the information in the observed data, similar as in the fractional Bayes factor. This construction also allows us to compute the posterior and fractional prior quantities using the same imputed data under the full model using all observed data. Moreover, as the marginal posterior and marginal fractional prior of $\boldsymbol{\beta}_{-\boldsymbol{\gamma}}$ both have multivariate Student t distributions (e.g., Mulder and Gu, 2022), computation is straightforward. The R package `BFpack` (Mulder et al., 2021) computes these quantities for fractional Bayes factors for any equality/one-sided constrained model, which also includes the model $\boldsymbol{\gamma}$. Finally note that by separately computing the numerator and denominator in (1) for the observed data, the fractional Bayes factor is still coherent via this construction (see also O'Hagan, 1997).

I end this comment by illustrating the implied selection behavior of this method for Experiment 2 of García-Donato et al. (2025) using the `Ozone35` dataset with 7 potential predictors and the same missing data mechanisms. Inclusion probabilities were obtained using FBFs after list-wise deletion and when using FBFs using the above methodology for handling missing data. The R package `BFpack` was used for computing the posterior probabilities for all $7^2 = 128$ possible models from which the inclusion probabilities can

be computed.[2] Figure 1 shows the boxplots of the inclusion probabilities for all predictors based on proportions of missing values for the variables x_6 to x_{10} of 10%, 20%, or 30% when using list-wise deletion (green plots) and the method discussed above (blue plots). Similar as García-Donato et al.'s method, the proposed method is clearly superior over list-wise deletion by better preserving the evidence. Moreover, the resulting inclusion probabilities using this method are very similar as compared to García-Donato et al.'s method (comparing Figure 1 here with Figure 1 of García-Donato et al.).

References

Dickey, J. (1971). "The Weighted Likelihood Ratio, Linear Hypotheses on Normal Location Parameters." *The Annals of Statistics*, 42(1): 204–223. MR0309225. doi: https://doi.org/10.1214/aoms/1177693507. 56

García-Donato, G., Castellanos, M. E., Cabras, S., Quirós, A., and Forte, A. (2025). "Model Uncertainty and Missing Data: An Objective Bayesian Perspective." *Bayesian Analysis*, 1(1): 1–26. 56, 57

Gilks, W. R. (1995). "Discussion to fractional Bayes factors for model comparison (by O'Hagan)." *Journal of the Royal Statistical Society Series B*, 56: 118–120. MR1325379. 56

Hoijtink, H., Gu, X., Mulder, J., and Rosseel, Y. (2018). "Computing Bayes Factors From Data With Missing Values." *Psychological Methods*, 24(2): 253–268. 56

Mulder, J. (2014). "Prior Adjusted Default Bayes Factors for Testing (In)equality Constrained Hypotheses." *Computational Statistics and Data Analysis*, 71: 448–463. MR3131982. doi: https://doi.org/10.1016/j.csda.2013.07.017. 56

Mulder, J. and Gu, X. (2022). "Bayesian testing of scientific expectations under multivariate normal linear models." *Multivariate Behavioral Research*, 57(5): 767–783. 56, 57

Mulder, J., Williams, D. R., Gu, X., Tomarken, A., Böing-Messing, F., Olsson-Collentine, A., Meijerink-Bosman, M., Menke, J., van Aert, R., Fox, J.-P., et al. (2021). "BFpack: Flexible Bayes factor testing of scientific theories in R." *Journal of Statistical Software*, 100: 1–63. 56, 57

O'Hagan, A. (1995). "Fractional Bayes Factors for Model Comparison (with discussion)." *Journal of the Royal Statistical Society B*, 57(1): 99–138. MR1325379. 56

O'Hagan, A. (1997). "Properties of intrinsic and fractional Bayes factors." *Test*, 6(1): 101–118. MR1466435. doi: https://doi.org/10.1007/BF02564428. 57

Rubin, D. B. (1996). "Multiple Imputation After 18+ Years." *Journal of the American statistical Association*, 91(434): 473–489. MR1294072. 57

[2]The R code can be found here: http://github.com/jomulder/missing-data-BFpack-FBFs

Contributed Discussion

Guido Consonni* and Dimitris Fouskakis†

We commend the Authors on their valuable contribution on the important topic of missing data under model uncertainty. Their objective Bayes perspective is also welcome. We found the paper very clear, highly readable and truly enjoyable!

Gaussian Graphical Imputation Model In their Conclusions the Authors briefly touch the issue of model determination when the number of covariates k is very high. We also address scalability to k but focus on the imputation model.

In their experiments the Authors adopt the imputation model $(x_{i1}, x_{i2}, \ldots, x_{ik}) \mid \boldsymbol{\nu} \overset{iid}{\sim} \mathrm{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is symmetric and p.d. but otherwise unrestricted. When k is large, it is reasonable to assume that the dependence structure among the k base covariates $(x_1, \ldots, x_k)$ will exhibit some degree of sparsity. This fact can be leveraged through an undirected graph G whose vertices are the $\{x_j\}$'s. The corresponding graphical model assumes that the joint distribution is such that x_j is independent of x_h, conditionally on all other variables, if and only if there is no edge between them under G. In a Gaussian Graphical Model this translates to $\boldsymbol{\Omega}_{jh} = 0$ where $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ is the precision matrix (Lauritzen, 1996).

Let M_G denote the space of $\boldsymbol{\Sigma}$ constrained by $\boldsymbol{\Omega}_{jh} = 0$ whenever there is no edge between x_j and x_h in G. When G is decomposable, vertices can be arranged recursively into a sequence of cliques $C \in \mathcal{C}$ and separators $S \in \mathcal{S}$. A convenient conjugate class of priors for $\boldsymbol{\Sigma} \in M_G$ is the hyper-inverse Wishart (Dawid and Lauritzen, 1993), written $\boldsymbol{\Sigma} \mid G \sim \mathrm{HIW}_G(b, \boldsymbol{D})$, where b is a degrees-of-freedom hyper-parameter and $\boldsymbol{D}$ a s.p.d.matrix. The corresponding density is given by $p(\boldsymbol{\Sigma} \mid G) = \frac{\prod_{C \in \mathcal{C}} p(\boldsymbol{\Sigma}_C \mid b, D_C)}{\prod_{S \in \mathcal{S}} p(\boldsymbol{\Sigma}_S \mid b, D_S)}$, where each marginal density is a standard Inverse-Wishart, so that in particular $\boldsymbol{\Sigma}_C \sim \mathrm{IW}(b, D_C)$ and similarly for $\boldsymbol{\Sigma}_S$.

For given G, we can sample from the target posterior $p(\mathbf{x}_{(1)}, \boldsymbol{\Sigma}, \boldsymbol{\mu} \mid \mathbf{x}_{(0)})$ using the required full conditionals. Regarding $\boldsymbol{\mu}$, updating is standard under a normal prior with a variance matrix proportional to $\boldsymbol{\Sigma}$ (or under a limiting flat prior). To update $\boldsymbol{\Sigma}$, notice first that $\boldsymbol{\Sigma} \mid \boldsymbol{\mu}, X \sim \mathrm{HIW}_G\big(b+n,\ \mathbf{D} + \sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top\big)$. To update $\boldsymbol{\Sigma}$, at each iteration compute and cache clique scatter matrices, then draw $\boldsymbol{\Sigma}^{(j)}$ using a standard junction-tree sampler (Højsgaard et al., 2012).

To update the missing covariates, for each row i in the data matrix X, let $\mathbf{x}_{i,0}$ be the observed variables and $\mathbf{x}_{i,1}$ the missing ones. Then, using the precision matrix $\boldsymbol{\Omega}$, $\mathbf{x}_{i,1} \mid \mathbf{x}_{i,0}, \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim \mathrm{N}(\boldsymbol{\mu}_1 - \boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{10}(\mathbf{x}_{i,0} - \boldsymbol{\mu}_0), \boldsymbol{\Omega}_{11}^{-1})$ so that for each missing variable only its neighbors in the graph G will appear in the expectation because of the zero-entries in $\boldsymbol{\Omega}_{10}$.

*Department of Statistical Sciences, Università Cattolica del Sacro Cuore, guido.consonni@unicatt.it
†Department of Mathematics, National Technical University of Athens, fouskakis@math.ntua.gr

The previous structure highlights two forms of locality which arise under decomposable graphical models with an *Hyper Inverse Wishart* (HIW) prior: clique-wise updates of Σ and neighbor-based imputations of $\mathbf{x}_{i,1}$. These properties deliver scalability advantages especially for sparse graphs. However, to evaluate the posterior expectation of functionals such as the imputed g' Bayes factor of formula (24) -which only depends on a subset of the data matrix and parameter- no automatic decoupling of such subset is possible.

The above discussion is predicated on a specific decomposable graph G which is typically unknown. Useful Bayesian references for graph-structure learning are (Carvalho and Scott, 2009) (which makes use of an objective version of the HIW) and (Rajaratnam et al., 2008) (large k). Their implementation with missing data would require suitable adaptations, and would represent an interesting line of research.

Use Imaginary Data to Construct the Prior Distribution for Regression Parameters
In the context of model uncertainty, one principled strategy for defining objective priors is the use of *imaginary training samples* (Consonni et al., 2018). A prominent construction is the *Expected Posterior Prior* (EPP; (Pérez and Berger, 2002)). The idea is to begin with default priors (often improper) and then generate imaginary data from a common predictive distribution. For each competing model, the (proper) posterior distribution of the parameters is computed using the default prior and the generated data, and the EPP is defined as the expectation of this posterior with respect to the predictive distribution. In this way, one obtains compatible priors across models. Typically, EPP methods rely on training samples of minimal size. An extension is given by the *Power Expected Posterior* (PEP) prior (Fouskakis and Ntzoufras, 2022), in which a power parameter regulates the effective size and influence of the imaginary data. Exploring EPP and PEP priors in the context of missing data could be fruitful, for example by generating random imaginary design matrices from a baseline imputation model and using them to define the prior distribution of the regression coefficients.

Within the normal linear model, it was proved in (Fouskakis and Ntzoufras, 2022) that the PEP prior (and the EPP), can be represented as a mixture of g-priors, like a wide range of prior distributions (see for example (Fouskakis and Ntzoufras, 2022)). This enhances computational tractability because posterior distributions and Bayes factors are derived in closed form. In particular, when the reference model is the null (intercept and no explanatory variables) and the imaginary covariates are the actual ones, centred at their means, the PEP prior is a mixture of g-priors with g following a shifted generalized beta prime distribution. In principle this methodology could be applied under the setup of missing data, deriving the imputation PEP prior, using arguments similar to those in the paper, by replacing the sample variance-covariance matrix with the corresponding parametric version of the imputation model. In this context, an interesting topic to explore would be to find the corresponding PEP prior or EPP after replacing in the mixture of g-priors the $X^T X$ by the variance covariance matrix of the covariates under the imputation model, as defined in the paper.

References

Carvalho, C. M. and Scott, J. G. (2009). "Objective Bayesian model selection in Gaussian graphical models." *Biometrika*, 96(2): 497–512. MR2538753. doi: https://doi.org/10.1093/biomet/asp017. 60

Consonni, G., Fouskakis, D., Liseo, B., and Ntzoufras, I. (2018). "Prior Distributions for Objective Bayesian Analysis." *Bayesian Analysis*, 13(2): 627–679. MR3807861. doi: https://doi.org/10.1214/18-BA1103. 60

Dawid, A. and Lauritzen, S. (1993). "Hyper-Markov laws in the statistical analysis of decomposable graphical models." *Annals of Statistics*, 21: 1272–11317. MR1241267. doi: https://doi.org/10.1214/aos/1176349260. 59

Fouskakis, D. and Ntzoufras, I. (2022). "Power-Expected-Posterior Priors as Mixtures of g-Priors in Normal Linear Models." *Bayesian Analysis*, 17(4): 1073–1099. MR4506022. doi: https://doi.org/10.1214/21-ba1288. 60

Højsgaard, S., Edwards, D., and Lauritzen, S. L. (2012). *Graphical Models with R*. Use R! New York: Springer. MR2905395. doi: https://doi.org/10.1007/978-1-4614-2299-0. 59

Lauritzen, S. (1996). *Graphical Models*. Oxford: Clarendon Press. MR1419991. 59

Pérez, J. M. and Berger, J. O. (2002). "Expected-Posterior Prior Distributions for Model Selection." *Biometrika*, 89(3): 491–512. MR1929158. doi: https://doi.org/10.1093/biomet/89.3.491. 60

Rajaratnam, B., Massam, H., and Carvalho, C. M. (2008). "Flexible covariance estimation in high dimensions." *Annals of Statistics*, 36(6): 2818–2849. MR2485014. doi: https://doi.org/10.1214/08-AOS619. 60

Invited Discussion

Merlise A. Clyde*

1 Introduction

The article by García-Donato and co-authors addresses the dual challenges of accounting for model uncertainty and missing data within the Gaussian regression frameworks from an objective Bayesian perspective. Through the use of an imputation g-prior that replaces $\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{\boldsymbol{\gamma}}$ in the covariance of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ with $\boldsymbol{\Sigma}_{\mathbf{X}_{\boldsymbol{\gamma}}}$, the authors develop a coherent approach to addressing the missing data problem and model uncertainty simultaneously with random $\mathbf{X}_{\boldsymbol{\gamma}}$ in the missing at random (MAR) or missing completely at random (MCAR) settings, while still being computationally tractable.

Under the MCAR/MAR framework and assumptions on the prior distributions, one of the key results that permits tractable computation is the expression for the marginal distribution of $\mathbf{Y}_{\mathsf{obs}}$ given $\mathbf{X}_{\mathsf{obs}}$ and $\boldsymbol{\gamma}$

$$m_{\boldsymbol{\gamma}}(\mathbf{Y}_{\mathsf{obs}} \mid \mathbf{X}_{\mathsf{obs}}) = \int \left[\int f(\mathbf{Y}_{\mathsf{obs}} \mid \mathbf{X}_{\mathsf{obs}}, \mathbf{X}_{\mathsf{miss}}, \alpha, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2, \boldsymbol{\gamma}) \pi(\alpha, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2 \mid \boldsymbol{\nu}, \boldsymbol{\gamma}) d[\alpha, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2] \right] \times p(\mathbf{X}_{\mathsf{miss}} \mid \mathbf{X}_{\mathsf{obs}}, \boldsymbol{\nu}) \pi(\boldsymbol{\nu} \mid \mathbf{X}_{\mathsf{obs}}) d[\mathbf{X}_{\mathsf{miss}}, \boldsymbol{\nu}]$$

where $\boldsymbol{\nu} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are the hyper-parameters of the distribution of $\mathbf{X}$. By replacing the usual $\frac{1}{n}(\mathbf{X}_{\boldsymbol{\gamma}} - \mathbf{1}_n \bar{\mathbf{x}}_{\boldsymbol{\gamma}})^T(\mathbf{X}_{\boldsymbol{\gamma}} - \mathbf{1}_n \bar{\mathbf{x}}_{\boldsymbol{\gamma}})$ in the g-prior with its expectation, $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}$, the inner integral (the marginal likelihood under the complete data $(\mathbf{Y}_{\mathsf{obs}}, \mathbf{X}_{\mathsf{obs}}, \mathbf{X}_{\mathsf{miss}})$ conditional on $\boldsymbol{\nu}$) continues to be available in closed form due to conjugacy. The outer integral over $\mathbf{X}_{\mathsf{miss}}$ and $\boldsymbol{\nu}$ is not tractable, but through the factorization of the posteriors, can be approximated via Monte Carlo integration leading to a "Rao-Blackwellized" estimator.

$$\begin{aligned}
\boldsymbol{\nu}^{(j)} &\sim p(\boldsymbol{\nu} \mid \mathbf{X}_{\mathsf{obs}}) \\
\mathbf{X}_{\mathsf{miss}}^{(j)} &\sim p(\mathbf{X}_{\mathsf{miss}} \mid \mathbf{X}_{\mathsf{obs}}, \boldsymbol{\nu}^{(j)}) \\
m(\mathbf{Y}_{\mathsf{obs}} \mid \mathbf{X}_{\mathsf{obs}}, \mathbf{X}_{\mathsf{miss}}^{(j)}, \boldsymbol{\gamma}, \boldsymbol{\nu}^{(j)}) &= \int f(\mathbf{Y}_{\mathsf{obs}} \mid \mathbf{X}_{\mathsf{obs}}, \mathbf{X}_{\mathsf{miss}}^{(j)}, \alpha, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2, \boldsymbol{\gamma}) \\
&\quad \times \pi(\alpha, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2 \mid \boldsymbol{\nu}^{(j)}, \boldsymbol{\gamma}) d[\alpha, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2] \\
\hat{m}_{\boldsymbol{\gamma}}(\mathbf{Y}_{\mathsf{obs}} \mid \mathbf{X}_{\mathsf{obs}}) &= \frac{1}{J} \sum_{j=1}^{J} m(\mathbf{Y}_{\mathsf{obs}} \mid \mathbf{X}_{\mathsf{obs}}, \mathbf{X}_{\mathsf{miss}}^{(j)}, \boldsymbol{\gamma}, \boldsymbol{\nu}^{(j)}).
\end{aligned}$$

This is similar in spirit to the Impute Then Select (ITS) approach of Yang et al. (2005) although rather than using the same set of imputed $\mathbf{X}_{\mathsf{miss}}, \boldsymbol{\nu}$ across all models, it appears that here new imputations are generated for each model – I would expect

arXiv: 2025.00000
*Department of Statistical Science, Duke University, clyde@duke.edu

that using the same set of imputations across all models would reduce the Monte Carlo error for estimating posterior model probabilities. Beyond the substantial difference in choice of prior distributions, the implementation of ITS in Yang et al. (2005) does not take advantage of collapsing the Gibbs sampler over α and $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, which would lead to less efficient estimators Liu et al. (1995); Ghosh and Clyde (2011).

2 Imputation g-Prior

For those who oppose g-priors on philosophical grounds because of the dependence on the design matrix $\mathbf{X}_{\boldsymbol{\gamma}}$ and potentially sample size n, the imputation g-prior may be more palatable as it replaces the sample covariance of $\mathbf{X}_{\boldsymbol{\gamma}}$ with the population parameter $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}$. Under random sampling this is the limiting version of the g-prior with $g = n$ as $n \to \infty$, satisfying the fourth criterion "Intrinsic Prior Consistency" of Bayarri et al. (2012).

The imputation g-prior of course could be used as the "real prior" with complete data. The expression for the Bayes factor in (24) is still valid with no missing observations based on the hierarchical prior, however, there is no need to integrate over $\mathbf{X}_{\mathsf{miss}}$ in this case. Because of the non-linearity in $\boldsymbol{\Sigma}$, it was not clear to me that with complete data, the imputation prior and usual g-prior would lead to the equivalent results in finite samples, as there is uncertainty about $\boldsymbol{\Sigma}$ that should be propagated. There are computational advantages, however, of the usual g-prior over the imputation g-prior with complete data, which may limit its adoption in general even with random sampling. An alternative approach is to continue using the usual g-prior based on the observed $\mathbf{X}_{\mathsf{obs}}$ and imputed $\mathbf{X}_{\mathsf{miss}}$ instead of the imputation g-prior with or without missing data. This would provide an avenue to handle transformations of $\mathbf{X}$ such as quadratic terms and interactions. For example, in many analyses of the ozone data, linear, quadratic, and two-way interactions are considered, e.g. (Casella and Moreno, 2006; Liang et al., 2008). While joint normality may be appropriate for the linear terms, joint normality is not appropriate for all variables in the design matrix. While this might preclude use of the imputation g-prior, one could proceed with the imputation step as described for the missing linear terms, and then use the completed $\mathbf{X}$ to construct the usual g-prior for each model. A related approach is described in the context of multiple imputation of missing covariates with non-linear and interactions by Seaman et al. (2012) who compare 'passive imputation' of the non-linear and interaction terms with passive imputation with predictive matching (PMM), to 'just another variable' (JAV) imputation where the non-linear and interaction terms are treated as additional variables in the multivariate imputation model. When $\mathbf{X}$ is missing at random, JAV may be biased, but this bias was generally less than for passive imputation and PMM. When quadratic effects were pronounced, they found that JAV sometimes led to large bias and poor coverage. For logistic regression, JAV's performance was sometimes very poor, with PMM generally improving on passive imputation, in terms of bias and coverage, but did not eliminate the bias. Clearly more work is needed in this area to extend Bayesian Variable Selection (BVS) and Bayesian Model Averaging methods to handle non-linearity and interactions with missing data.

3 Scalability to Large Model Spaces

Estimates of posterior model probabilities

$$\hat{p}(\boldsymbol{\gamma} \mid \mathbf{Y}_{\mathsf{obs}}, \mathbf{X}_{\mathsf{obs}}) = \frac{\hat{m}_{\boldsymbol{\gamma}}(\mathbf{Y}_{\mathsf{obs}} \mid \mathbf{X}_{\mathsf{obs}})\pi(\boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma} \in \mathbf{G}} \hat{m}_{\boldsymbol{\gamma}}(\mathbf{Y}_{\mathsf{obs}} \mid \mathbf{X}_{\mathsf{obs}})\pi(\boldsymbol{\gamma})} \tag{1}$$

and marginal posterior inclusion probabilities are feasible for model spaces where enumeration is possible. However, for larger model spaces that preclude enumeration, Markov chain Monte Carlo (MCMC) of some form will be necessary to sample models with estimates of marginal posterior probabilities and inclusion probabilities based on ergodic averages of $\boldsymbol{\gamma}$. While one could iterate over J imputations of $\mathbf{X}_{mis}$ and $\boldsymbol{\nu}$ for each model to obtain more accurate estimates of the marginal likelihoods and Bayes factors used to accept proposals of γ, using the Monte Carlo averaged marginal likelihoods and summing over sampled models will lead to biased estimates as the normalizing constant in (1) is underestimated (Clyde and Ghosh, 2012). As with ITS, one could use a two stage approach, by imputing J sets of $\mathbf{X}_{\mathsf{miss}}$ and $\boldsymbol{\nu}$ and then running a standard MCMC for M iterations to sample models given each imputed dataset, which has the advantage of being parallelizable. Alternatively, one could embed the imputation step within the MCMC similar to the Simultaneous Impute and Select (SIAS) algorithm of Yang et al. (2005), but instead draw new imputations of $\mathbf{X}_{\mathsf{miss}}$ and $\boldsymbol{\nu}$ given $\mathbf{X}_{\mathsf{obs}}$ at each iteration and accepting or rejecting the proposed $\boldsymbol{\gamma}, \mathbf{X}_{\mathsf{miss}}, \boldsymbol{\nu}$, where the imputation step requires sampling $\boldsymbol{\nu} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, given $\mathbf{X}_{\mathsf{obs}}$ and generating $\mathbf{X}_{\mathsf{miss}}$ conditional on $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as before. If one is estimating posterior probabilities based on ergodic averages from a MCMC, a question that arises is whether this is more efficient than proposing $\mathbf{X}_{\mathsf{miss}}$ given $\mathbf{Y}_{\mathsf{obs}}, \mathbf{X}_{\mathsf{obs}}$ and $\boldsymbol{\nu}$ in a Gibbs step and then accepting/rejecting a proposed $\boldsymbol{\gamma}$ given the completed $\mathbf{X}$ and $\mathbf{Y}_{\mathsf{obs}}$ in a Metropolis Hastings step. Using the information in $\mathbf{Y}_{\mathsf{obs}}$ may lead to more informative proposals for the missing data, at the cost of potentially slower mixing.

3.1 Objective Graphical Model Selection and g-Priors

For large p, sampling the full $p \times p$ covariance matrix $\boldsymbol{\Sigma}$ may be memory intensive and computationally expensive depending on the rejection rate for proposing $\boldsymbol{\Sigma}$; a concern if $\boldsymbol{\Sigma}$ is nearly singular (Sun and Berger, 2007). One potential approach to reduce the dimension of $\boldsymbol{\Sigma}$ is to consider selection in the covariance structure of $\mathbf{X}$ in addition to the variables in the regression of $\mathbf{Y}$ on $\mathbf{X}$. We define the vector

$$\mathbf{Z} = \begin{pmatrix} Y \\ \mathbf{x}_{\boldsymbol{\gamma}} \\ \mathbf{x}_{-\boldsymbol{\gamma}} \end{pmatrix} \in \mathbb{R}^{p+1}$$

where Y is a scalar, $\mathbf{x}_{\boldsymbol{\gamma}}$ are the elements of $\mathbf{x}$ where $\gamma_j = 1$ and $\mathbf{x}_{-\boldsymbol{\gamma}}$ are the elements of $\mathbf{x}$ where $\gamma_j = 0$. Then $\mathbf{Z}$ has a multivariate normal distribution implied by the conditional distribution of $Y \mid \mathbf{x}_{\boldsymbol{\gamma}}$ and the marginal distribution of $\mathbf{x}$,

$$\mathbf{Z} =\mid \boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim \mathsf{N}\left(\begin{pmatrix} \mu_y \\ \mu_{\mathbf{x}_{\boldsymbol{\gamma}}} \\ \mu_{\mathbf{x}_{-\boldsymbol{\gamma}}} \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \Sigma_{yy} & \boldsymbol{\Sigma}_{y\mathbf{x}_{\boldsymbol{\gamma}}} & \boldsymbol{\Sigma}_{y\mathbf{x}_{-\boldsymbol{\gamma}}} \\ \boldsymbol{\Sigma}_{\mathbf{x}_{\boldsymbol{\gamma}}y} & \boldsymbol{\Sigma}_{\mathbf{x}_{\boldsymbol{\gamma}}\mathbf{x}_{\boldsymbol{\gamma}}} & \boldsymbol{\Sigma}_{\mathbf{x}_{\boldsymbol{\gamma}},\mathbf{x}_{-\boldsymbol{\gamma}}} \\ \boldsymbol{\Sigma}_{\mathbf{x}_{-\boldsymbol{\gamma}}y} & \boldsymbol{\Sigma}_{\mathbf{x}_{-\boldsymbol{\gamma}}\mathbf{x}_{\boldsymbol{\gamma}}} & \boldsymbol{\Sigma}_{\mathbf{x}_{-\boldsymbol{\gamma}}\mathbf{x}_{-\boldsymbol{\gamma}}} \end{pmatrix} \right)$$

with conditional distribution of $Y \mid \mathbf{x}_{\boldsymbol{\gamma}}$ given by

$$\begin{aligned} Y \mid \mathbf{x}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma} &\sim \mathsf{N}(\mu_{Y|\mathbf{x}_{\boldsymbol{\gamma}}}, \sigma^2_{y|\mathbf{x}_{\boldsymbol{\gamma}}}) \\ \mu_{Y|\mathbf{x}_{\boldsymbol{\gamma}}} &= \mu_y + \boldsymbol{\Sigma}_{y\mathbf{x}_{\boldsymbol{\gamma}}} \boldsymbol{\Sigma}^{-1}_{\mathbf{x}_{\boldsymbol{\gamma}}\mathbf{x}_{\boldsymbol{\gamma}}} (\mathbf{x}_{\boldsymbol{\gamma}} - \mu_{\mathbf{x}_{\boldsymbol{\gamma}}}) \\ \sigma^2_{y|\mathbf{x}_{\boldsymbol{\gamma}}} &= \Sigma_{yy} - \boldsymbol{\Sigma}_{y\mathbf{x}_{\boldsymbol{\gamma}}} \boldsymbol{\Sigma}^{-1}_{\mathbf{x}_{\boldsymbol{\gamma}}\mathbf{x}_{\boldsymbol{\gamma}}} \boldsymbol{\Sigma}_{\mathbf{x}_{\boldsymbol{\gamma}}y} \end{aligned}$$

and marginal distribution of $\mathbf{x}$ given by

$$\begin{pmatrix} \mathbf{x}_{\boldsymbol{\gamma}} \\ \mathbf{x}_{-\boldsymbol{\gamma}} \end{pmatrix} \sim \mathsf{N}\left(\begin{pmatrix} \boldsymbol{\mu}_{\mathbf{x}_{\boldsymbol{\gamma}}} \\ \boldsymbol{\mu}_{xng} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{x}_{\boldsymbol{\gamma}}\mathbf{x}_{\boldsymbol{\gamma}}} & \boldsymbol{\Sigma}_{\mathbf{x}_{\boldsymbol{\gamma}}\mathbf{x}_{-\boldsymbol{\gamma}}} \\ \boldsymbol{\Sigma}_{\mathbf{x}_{-\boldsymbol{\gamma}}\mathbf{x}_{\boldsymbol{\gamma}}} & \boldsymbol{\Sigma}_{\mathbf{x}_{-\boldsymbol{\gamma}}\mathbf{x}_{-\boldsymbol{\gamma}}} \end{pmatrix} \right).$$

Under model $\boldsymbol{\gamma}$, $\Sigma_{Y\mathbf{x}_{-\boldsymbol{\gamma}}} = \mathbf{0}$. While $\sigma^2_{y|\mathbf{x}_{\boldsymbol{\gamma}}}$ is generally treated as constant across models, it does appear to depend on the model through $\boldsymbol{\Sigma}_{\mathbf{x}_{\boldsymbol{\gamma}}\mathbf{x}_{\boldsymbol{\gamma}}}$ and $\Sigma_{y\mathbf{x}_{\boldsymbol{\gamma}}}$.

This may be viewed as a special case of model selection in Gaussian graphical models, where the graph $\mathbf{G}$ is defined by edges between all pairs of variables in $\mathbf{x}$ (a fully connected graph) and edges between Y and variables in $\mathbf{x}_{\boldsymbol{\gamma}}$. By relaxing, the assumption that $\mathbf{x}$ is fully connected, one could consider more general graphical models for $\mathbf{Z}$ leading to more parsimonious models for large p (Jones et al., 2005). As with model specific parameters in linear regression, neither improper or vague priors on $\boldsymbol{\Sigma}$ are generally allowable. In all but the smallest of problems, $\pi(\boldsymbol{\Sigma} \mid \mathbf{G})$ must be a conjugate hyper-inverse Wishart prior (HIW) (Dawid and Lauritzen, 1993; Giudici and Green, 1999) for tractable computation of marginal likelihoods. Seeking well behaved objective priors for graphical model selection, Carvalho and Scott (2009) developed a HIW g-prior, $\boldsymbol{\Sigma} \sim \mathsf{HIW}_{\mathbf{G}}(gn, g\mathbf{Z}^T\mathbf{Z})$ constructed using fractional Bayes factors (O'Hagan, 1995). Under this approach, some fraction, $0 < g < 1$ of the likelihood is used for training an improper prior on $\boldsymbol{\Sigma}$, where the resulting fractional prior is proportional to the improper prior times the fractional likelihood. Integrating the remaining $1 - g$ fraction of the likelihood with respect to the fractional prior leads to a fractional marginal likelihood. They recommend the choice $g = 1/n$ based on minimal training samples leading to the vector $\mathbf{z}$ having a marginal Cauchy distribution and prove that the resulting Bayes Factors avoid the information paradox (Liang et al., 2008; Bayarri et al., 2012).

Focusing on the implied conditional regression for $\mathbf{Y} \mid \mathbf{X}_{\boldsymbol{\gamma}}$ implied by the graph Carvalho and Scott (2009) prove the fractional prior for $\boldsymbol{\Sigma}$ induces a g-prior for $\boldsymbol{\beta}_{\boldsymbol{\gamma}} = \Sigma_{y\mathbf{x}_{\boldsymbol{\gamma}}} \boldsymbol{\Sigma}^{-1}_{\mathbf{x}_{\boldsymbol{\gamma}}\mathbf{x}_{\boldsymbol{\gamma}}}$,

$$\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \sigma^2_{y|\mathbf{x}_{\boldsymbol{\gamma}}} \sim \mathsf{N}\left(\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}, \frac{\sigma^2_{y|\mathbf{x}_{\boldsymbol{\gamma}}}}{g} (\mathbf{X}'_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}})^{-1} \right) \tag{2}$$

$$1/\sigma^2_{y|\mathbf{x}_{\boldsymbol{\gamma}}} \sim \mathsf{Ga}\left(\frac{gn + p_{\boldsymbol{\gamma}}}{2}, \frac{g\mathsf{RSS}_{\mathbf{x}_{\boldsymbol{\gamma}}}}{2} \right) \tag{3}$$

where $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ is the usual least squares estimate and $\mathsf{RSS}_{\mathbf{x}_{\boldsymbol{\gamma}}}$ is the residual sum of squares for regression of $\mathbf{Y}$ on $\mathbf{X}_{\boldsymbol{\gamma}}$ and $p_{\boldsymbol{\gamma}} = \sum \gamma_j$. In the case where no selection in the covariance structure of $\mathbf{X}$ is done, this provides an alternative justification of a g-prior using the completed $\mathbf{X}$.

The general problem of objective graphical model selection with the HIW g-prior is with missing data is an interesting direction to pursue. One potential approach under the HIW g-prior is to initiate the chain at some draw of $\mathbf{Z}_{\mathsf{miss}}^{(0)}$ and iterate for $i = 1, \ldots, M$:

$$\begin{aligned}
&\text{Propose } \mathbf{G}^* \sim q(\mathbf{G}* \mid \mathbf{G}^{(i)}) \\
&\text{Propose } \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^* \mid \mathbf{G}^*, \mathbf{Z}_{\mathsf{miss}}^{(i)}, \mathbf{Z}_{\mathsf{obs}} \text{ from the joint posterior conditional} \\
&\text{Propose } \mathbf{Z}_{\mathsf{miss}}^* \mid \mathbf{Z}_{\mathsf{obs}}, \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*, \mathbf{G}^* \text{ from the posterior conditional} \\
&\text{Accept/Reject } \mathbf{G}^*, \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*, \mathbf{Z}_{\mathsf{miss}}^* \text{ based on the standard Metropolis Hastings ratio.} \\
&\text{Update } \mathbf{G}^{(i+1)}, \boldsymbol{\mu}^{(i+1)}, \boldsymbol{\Sigma}^{(i+1)}, \mathbf{Z}_{\mathsf{miss}}^{(i+1)}
\end{aligned}$$

at iteration $i+1$ to either the proposed values (Accept) or the previous values (Reject). For complete data, we can collapse by integrating out $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Other factorizations, approaches to collapse the conditional distributions or approximations for proposing $\mathbf{Z}_{\mathsf{miss}}$ given $\mathbf{G}$ could lead to more efficient sampling approaches that improve mixing. The choice of improper prior in Carvalho and Scott (2009) leads to a computationally convenient fractional prior as a g-prior for graphical models. Further improvements, could be achieved by adapting recent recommendations of Berger et al. (2020) for objective priors on $\boldsymbol{\Sigma}$.

References

Bayarri, M. J., Berger, J. O., Forte, A., and García-Donato, G. (2012). "Criteria for Bayesian Model Choice with Application to Variable Selection." *The Annals of Statistics*, 40(3): 1550–1577. MR3015035. doi: https://doi.org/10.1214/12-AOS1013. 63, 65

Berger, J. O., Sun, D., and Song, C. (2020). "Bayesian analysis of the covariance matrix of a multivariate normal distribution with a new class of priors." *Annals of Statistics*, 48(4): 2381–2403. URL https://doi.org/10.1214/19-AOS1891 MR4134799. doi: https://doi.org/10.1214/19-AOS1891. 66

Carvalho, C. M. and Scott, J. G. (2009). "Objective Bayesian model selection in Gaussian graphical models." *Biometrika*, 96(3): 497–512. URL https://doi.org/10.1093/biomet/asp017 MR2538753. doi: https://doi.org/10.1093/biomet/asp017. 65, 66

Casella, G. and Moreno, E. (2006). "Objective Bayesian Variable Selection." *Journal of the American Statistical Association*, 101(473): 157–167. MR2268035. doi: https://doi.org/10.1198/016214505000000646. 63

Clyde, M. A. and Ghosh, J. (2012). "Finite population estimators in stochastic search variable selection." *Biometrika*, 99: 981–988. URL http://biomet.oxfordjournals.org/content/early/2012/09/30/biomet.ass040.abstract MR2999173. doi: https://doi.org/10.1093/biomet/ass040. 64

Dawid, A. P. and Lauritzen, S. L. (1993). "Hyper Markov laws in the statistical analysis of decomposable graphical models." *Annals of Statistics*, 21(3): 1272–1317. URL

https://doi.org/10.1214/aos/1176349260 MR1241267. doi: https://doi.org/10.1214/aos/1176349260. 65

Ghosh, J. and Clyde, M. A. (2011). "Rao-Blackwellization for Bayesian Variable Selection and Model Averaging in Linear and Binary Regression: A Novel Data Augmentation Approach." *Journal of the American Statistical Association*, 106(495): 1041–1052. URL http://dx.doi.org/10.1198/jasa.2011.tm10518 MR2894762. doi: https://doi.org/10.1198/jasa.2011.tm10518. 63

Giudici, P. and Green, P. J. (1999). "Decomposable graphical Gaussian model determination." *Biometrika*, 86(4): 785–801. URL https://doi.org/10.1093/biomet/86.4.785 MR1741977. doi: https://doi.org/10.1093/biomet/86.4.785. 65

Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C., and West, M. (2005). "Experiments in Stochastic Computation for High-Dimensional Graphical Models." *Statistical Science*, 20(4): 388–400. URL https://doi.org/10.1214/088342305000000304 MR2210226. doi: https://doi.org/10.1214/088342305000000304. 65

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). "Mixtures of g-priors for Bayesian Variable Selection." *Journal of the American Statistical Association*, 103(481): 410–423. URL http://ideas.repec.org/a/bes/jnlasa/v103y2008mmarchp410-423.html MR2420243. doi: https://doi.org/10.1198/016214507000001337. 63, 65

Liu, J. S., Wong, W. H., and Kong, A. (1995). "Covariance structure and convergence rate of the Gibbs sampler with various scans." *Journal of the Royal Statistical Society – Series B*, 57: 157–169. MR1325382. 63

O'Hagan, A. (1995). "Fractional Bayes factors for model comparison." *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1): 99–138. URL https://www.jstor.org/stable/2346088 MR1325379. 65

Seaman, S. R., Bartlett, J. W., and White, I. R. (2012). "Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods." *BMC Medical Research Methodology*, 12(1): 46. URL https://doi.org/10.1186/1471-2288-12-46 63

Sun, D. and Berger, J. O. (2007). "Objective Bayesian analysis for the multivariate normal model." In Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M. (eds.), *Bayesian Statistics 8*, Oxford Science Publications, 525–562. Oxford: Oxford University Press. MR2433206. 64

Yang, X., Belin, T. R., and Boscardin, W. J. (2005). "Imputation and Variable Selection in Linear Regression Models with Missing Covariates." *Biometrics*, 61(2): 498–506. URL http://www.jstor.org/stable/3695970 MR2140922. doi: https://doi.org/10.1111/j.1541-0420.2005.00317.x. 62, 63, 64

Invited Discussion

Sebastian Arnold* and Alexander Ly†

Congratulations to García-Donato, Castellanos, Cabras, Quirós, and Forte (2025; henceforth, GCCQF) for their valuable work extending the Bayesian variable selection methodology to handle missing data, which significantly broadens its applicability to real-world scenarios. The paper offers a valuable basis for further discussion and we are pleased by the invitation to write this comment.

Our comment explores a further extension of the proposed methodology. Specifically, we consider the sequential setting where (potentially missing) data accumulate over time, with the goal of continuously monitoring statistical evidence, as opposed to assessing it only once data collection terminates. To this end, we replicated the first setting of their Experiment 1 using data generated under the linear regression model

$$y_i = x_{i1} + 2x_{i2} + x_{i6} + 2x_{i7} + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \text{ where } \sigma^2 = 2.5, \tag{1}$$

where the four active variables $x_{i1}, x_{i2}, x_{i6}, x_{i7}$ are correlated with 6 other covariates, all jointly normally distributed. In each replication, 100 outcomes $y \in \mathbb{R}$ and potential covariates $\boldsymbol{x} \in \mathbb{R}^p$, where $p = 10$, were generated. As in the main text, 40% of the covariates were made missing at random. Instead of running their `R` procedures only at $n = 100$, we did so repeatedly at $n = 19, 20, \ldots, 100$, resulting in $t = 1, \ldots, 82$ imputed covariate sets and posterior inclusion probabilities (the used imputation method needs a minimum number of data points to run properly, and we could, thus, not track the posterior probabilities for $n < 19$). As opposed to the original setting, we did decrease the number of mice imputations from 500 to 50, but this did not qualitatively change the reported results at $n = 100$.

1 Exploring sequential Bayesian variable selection

In the simulation study we applied the sequential procedure to 100 data sets, and the top row of Figure 1 shows the result of the GCCQF procedure for simulation 19 and 76, respectively. The posterior inclusion probabilities of the active and inactive covariates are depicted as green and brown curves, respectively. The top left panel shows that the inclusion probabilities of the active covariates remained above 0.5 throughout, but that the inactive covariates are 17 times misidentified as being active during data collection. The top right panel shows 4 and 46 misidentifications of the active and inactive covariates, respectively. To stabilise inference, we combined an initial approach based on sequential model confidence sets (bottom row of Figure 1) with the GCCQF procedure, yielding the results shown in the second row of Figure 1.

arXiv: 2509.22901
*Machine Learning, Centrum Wiskunde & Informatica, Sebastian.Arnold@cwi.nl
†Machine Learning, Centrum Wiskunde & Informatica, Alexander.Ly@cwi.nl

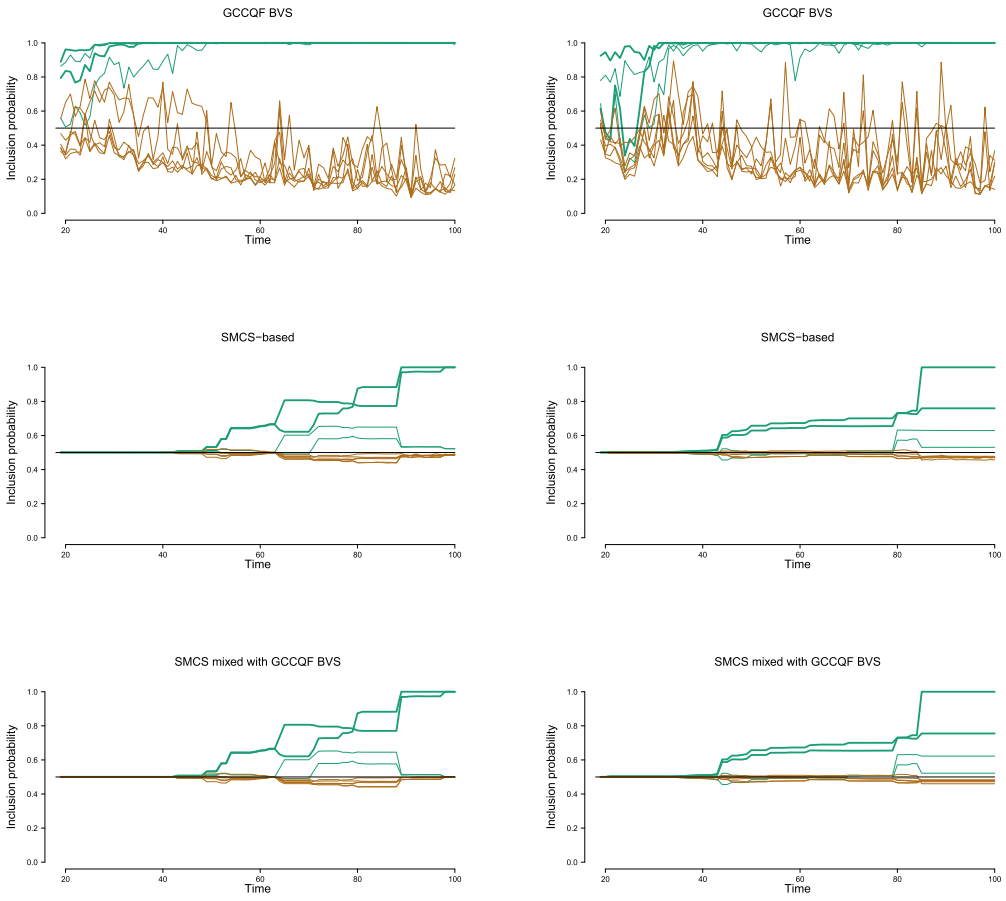

Figure 1: Posterior inclusion probabilities of the active (green) and inactive (brown) covariates over time for two runs of the simulation with respect to the sequential application of GCCQF (top), the sequential model confidence sets, hence, SMCS-based approach (bottom), and the mixture approach (middle), where $\alpha = 0.1$ and $\lambda = 1/(8\varsigma^2) \approx 0.3$, for $\varsigma = 0.65 \leq \sigma$. Thick lines indicate the active covariates x_2 and x_7 with larger regression coefficients.

2 Sequential model confidence sets

In our study, we use sequential model confidence sets, as proposed by Arnold et al. (2024), as a meta-algorithm for sequential model selection. This method compares the $m = 2^p$ candidate models represented by binary vectors $\boldsymbol{\gamma}^i = (\gamma_1^i, \ldots, \gamma_p^i) \in \{0,1\}^p$, $i \in [m] := \{1, \ldots, m\}$, based on their predictive performance. At time $t \in \mathbb{N}$, we observe y_t and compute the loss of model i, denoted by $L_{i,t} = \ell(\hat{y}_{i,t}, y_t)$, where $\hat{y}_{i,t}$ is the model's prediction and ℓ some negatively oriented loss function. The loss differences $d_{ij,t} = L_{i,t} - L_{j,t} \leq 0$, iff model i outperforms model j at time t. Proposition 3.3 in

Arnold et al. (2024) implies that, if (i) the data are generated by some model $i^\star \in [m]$ under consideration, and (ii) the loss differences are conditionally sub-exponential with tuning parameter $\lambda \geq 0$, then the collection $\widehat{\mathcal{M}}_t$ consisting of all models $i \in [m]$ for which the following holds

$$E_{i,t} = \sup_{r \leq t} \frac{1}{m-1} \sum_{j \neq i} \exp\left\{ \lambda \left(\sum_{s=1}^{r} d_{ij,s} \right) - r/8 \right\} \leq 1/\alpha, \quad t \in \mathbb{N}, \tag{2}$$

forms a so-called *sequence of model confidence sets* (SMCS) or simply *sequential model confidence sets* (SMCSs) at level $\alpha \in (0,1)$, which means that

$$\mathbb{P}(\forall t \geq 1 : i^\star \in \widehat{\mathcal{M}}_t) \geq 1 - \alpha. \tag{3}$$

That is, the SMCS guarantees simultaneous coverage of the true data-generating model $i^\star$ for all times t with high probability. Property (3) is equivalent to the statement that the probability of the true model ever being excluded from the SMCS is at most α:

$$\mathbb{P}(\exists t \geq 1 : i^\star \notin \widehat{\mathcal{M}}_t) \leq \alpha. \tag{4}$$

This bound is known as Ville's inequality and generally holds for exceedance probabilities of E-processes, which are the key objects in safe anytime-valid inference, see for instance Grünwald et al. (2024), Howard et al. (2021), and in simplified form Ly et al. (2025). Inspired by the Bayesian variable selection approach advocated by the authors, we can derive the SMCS-based inclusion probability of covariate $k = 1, \ldots, p$ as

$$\hat{p}_t^{\text{SMCS}}(\gamma_k = 1 \mid \boldsymbol{x}_1, y_1, \ldots, \boldsymbol{x}_t, y_t) = \frac{|\{i \in \widehat{\mathcal{M}}_t \mid \gamma_k^i = 1\}|}{|\{i \in \widehat{\mathcal{M}}_t\}|}, \quad t \in \mathbb{N}. \tag{5}$$

Raw counting results from $\widehat{\mathcal{M}}_t$ being a frequentist construct that either does, or does not, contain the true data generating model. Under repeated use, this procedure is expected to – at all times – cover the truth in $1 - \alpha\%$ of the cases, unlike Bayesian posterior model probabilities, which represent gradually updated beliefs.

3 Exploring SMCS inclusion probabilities

Our initial attempt to use the predictions of each model assessed by the ℓ_2 loss fits the general theory, but did unfortunately not yield the results we were searching for, perhaps caused by sub-optimally chosen parameters λ and α. Eventually, we used the logarithm of the (imputed) GCCQF Bayes factors to score each model. In particular, the "loss" at time t of model $i \in [m]$ was computed as

$$L_{i,t} := \frac{1}{m-1} \sum_{j \neq i} \log \text{BF}_{ji,t}, \text{ where } \text{BF}_{ji,t} := \frac{m_{\boldsymbol{\gamma}^j}(y_1, \ldots, y_t \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_t)}{m_{\boldsymbol{\gamma}^i}(y_1, \ldots, y_t \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_t)}. \tag{6}$$

The inclusion probabilities for the two simulation runs are depicted in the bottom row of Figure 1. Compared to the sequential application of the GCCQF procedure, there

is a significant reduction in both the fluctuations and number of crossings through the critical threshold 0.5 (from 17 to 5 in the left, and from 50 to 9 in the right column). For the simulation run on the right, this increase in stability came at no cost of accuracy at $n = 100$, thus, $t = 82$. However, the SMCS based inclusion probabilities do misclassify x_1 and x_6 as inactive from about $n = 98$, thus, $t = 80$, onward in the left column. At this time point about 256 models remain in $\widehat{\mathcal{M}}_t$, and the results show that these weak active covariates appear in exactly half of them.

To borrow strength from the Bayesian variable selection approach, we explored two methods. The first simply sets the GCCQF posterior model probabilities for model $i \in [m]$ at time t to zero, whenever $i \notin \widehat{\mathcal{M}}_t$. After renormalisation, we then qualitatively recovered the standard Bayesian variable selection approach, which suggests close alignment of the posterior model probabilities and sequential model confidence sets. The second method mixes the GCCQF and SMCS inclusions probabilities, where the last one is proportionally weighted by $|\widehat{\mathcal{M}}_t|/m$, the relative number of surviving models at time t. This implies that the posterior model probabilities will dominate once a significant number of candidate models are removed from $\widehat{\mathcal{M}}_t$. The latter typically occurs for large t, when the Bayesian inclusions probabilities are hopefully stabilised. The middle row of Figure 1 shows that this mixed approach identifies the active coefficients more accurately, while at the same time leading to more stable inference.

Method	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
GCCQF BVS	1.08	0.09	3.82	2.65	2.96	2.46	0.16	3.79	4.35	3.96
Mixed	0.47	0.05	0.88	0.74	0.63	0.82	0.15	0.98	1.29	0.93
SMCS based	0.46	0.05	0.8	0.66	0.7	0.75	0.15	0.87	1.02	0.94

Table 1: The average number of crossing of each covariate.

Method	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
GCCQF BVS	1	1	0.03	0.01	0.04	0.99	1	0.02	0.09	0.05
Mixed	0.98	1	0.04	0.10	0.02	0.97	1	0.25	0.19	0.10
SMCS based	0.86	1	0.08	0.13	0.02	0.84	1	0.26	0.20	0.11

Table 2: The relative frequency of each covariate being included at time $n = 100$ ($t = 82$).

The stability improvements hold consistently across all 100 replicated data sets (Table 1). While this gain in stability incurs minimal cost in identifying active covariates (Table 2), it does increase the misclassification rate of inactive covariates as active. Figure 2 visualises the decrease in variability of the total number of crossings through the critical value 0.5.

4 Concluding comments and further discussion

With this note we do not want to critique the contribution of GCCQF, but rather highlight the many interesting challenges involved with sequential decision making. The

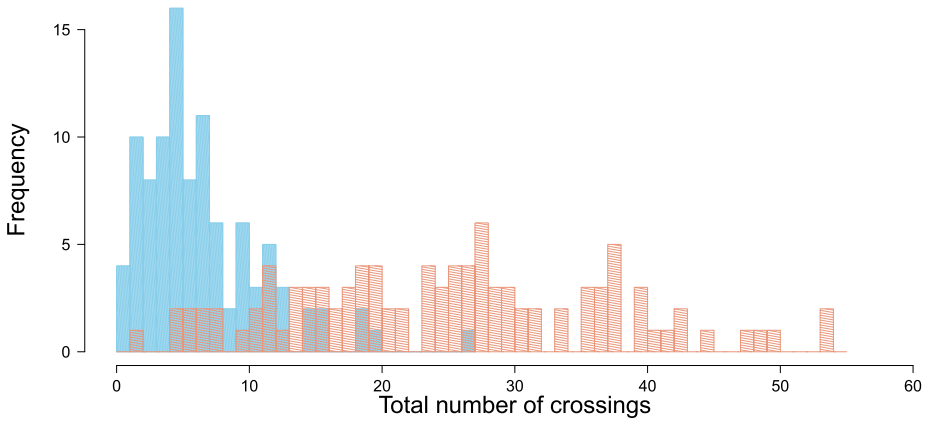

Figure 2: The averaged total number of crossings through the critical value 0.5 for the posterior inclusion probabilities (red) and the mixture approach (blue).

approach we took here was fully exploratory, as we honestly do not know what can be reasonably expected from a sequential variable selection method. It is unrealistic to demand such method to perfectly identify the (in)active covariates as (in)active as soon as possible, and retain those classifications throughout. But what is realistic? The increase in stability based on mixing the SMCS and GCCQF inclusion probabilities was significant, but we do not know whether it is an optimal procedure. In fact, we do not even know if our choice of λ and α is optimal for our mixed procedure. The chosen $\lambda \approx 0.3$ works well for the underlying true β, and we can make it adaptive to other βs. For instance, with a prior distribution, or using a prequential plugin approach. The role of α requires more explorations and thought. It is also unclear whether the cost of increased misclassification of inactive covariates, as shown in Table 2, is necessary. Furthermore, it remains unclear how the theoretical guarantees of SMCS, derived in Arnold et al. (2024), translate from the level of models to the derived inclusion probabilities (5). The use of SMCSs, however, is not totally arbitrary as admissible anytime-valid inference should rely on E-processes (Ramdas et al., 2020), and SMCSs provide a natural candidate for sequential model selection. A central result in the theory of safe sequential inference also suggests that (log) optimal procedures should be Bayesian in nature (Grünwald et al., 2024; Larsson et al., 2025). Our meta-approach moves out of this realm, and we wonder whether the authors can see avenues in making the approach Bayesian again. Perhaps they suggest a different approach to sequential Bayesian variable selection over our naive approach. And would an increase in the rate of misclassification of an inactive covariate be worth the cost of more stable inference?

Acknowledgments

The authors would like to thank Stefano Cabras and Gonzalo García-Donato for their code, upon which this work builds. They also acknowledge Udo Boehm for valuable discussions and inputs.

Funding

This research was funded by the Dutch Research Council (NWO) through the VENI fellowship grant "Increasing Scientific Efficiency with Sequential Methods" (VI.Veni.211G.040) awarded to AL. SA acknowledges funding by the ERC advanced grant (101142168) awarded to Peter Grünwald.

References

Arnold, S., Gavrilopoulos, G., Schulz, B., and Ziegel, J. (2024). "Sequential model confidence sets." *arXiv preprint arXiv:2404.18678*. 69, 70, 72

García-Donato, G., Castellanos, M. E., Cabras, S., Quirós, A., and Forte, A. (2025). "Model Uncertainty and Missing Data: An Objective Bayesian Perspective." *Bayesian Analysis*, 1–51. URL https://doi.org/10.1214/25-BA1531 doi: https://doi.org/10.1214/25-BA1531. 68

Grünwald, P., de Heide, R., and Koolen, W. (2024). "Safe testing." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(5): 1091–1128. MR4825000. doi: https://doi.org/10.1093/jrsssb/qkae011. 70, 72

Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. (2021). "Time-uniform, nonparametric, nonasymptotic confidence sequences." *The Annals of Statistics*, 49(2): 1055–1080. MR4255119. doi: https://doi.org/10.1214/20-aos1991. 70

Larsson, M., Ramdas, A., and Ruf, J. (2025). "The numeraire e-variable and reverse information projection." *The Annals of Statistics*, 53(3): 1015 – 1043. URL https://doi.org/10.1214/24-AOS2487 MR4925114. doi: https://doi.org/10.1214/24-aos2487. 72

Ly, A., Boehm, U., Grünwald, P. D., Ramdas, A., and van Ravenzwaaij, D. (2025). "A Tutorial on Safe Anytime-Valid Inference: Practical Maximally Flexible Sampling Designs for Experiments Based on e-Values." *PsyArxiv preprint*. URL https://doi.org/10.31234/osf.io/h5vae_v2 70

Ramdas, A., Ruf, J., Larsson, M., and Koolen, W. (2020). "Admissible anytime-valid sequential inference must rely on nonnegative martingales." *arXiv preprint arXiv:2009.03167*. MR4897884. doi: https://doi.org/10.1093/jrsssb/qkae061. 72