AStA Wirtschafts- und Sozialstatistisches Archiv
https://doi.org/10.1007/s11943-024-00345-1

ORIGINALVERÖFFENTLICHUNG

# Automated Bayesian variable selection methods for binary regression models with missing covariate data

**Michael Bergrab · Christian Aßmann**

Received: 31 March 2023 / Accepted: 23 July 2024

**Abstract** Data collection and the availability of large data sets has increased over the last decades. In both statistical and machine learning frameworks, two methodological issues typically arise when performing regression analysis on large data sets. First, variable selection is crucial in regression modeling, as it helps to identify an appropriate model with respect to the considered set of conditioning variables. Second, especially in the context of survey data, handling of missing values is important for estimation, which occur even with state-of-the-art data collection and processing methods. Within this paper, we provide an Bayesian approach based on a spike-and-slab prior for the regression coefficients, which allows for simultaneous handling of variable selection and estimation in combination with handling of missing values in covariate data. The paper also discusses the implementation of the approach using Markov chain Monte Carlo techniques and provides results for simulated data sets and an empirical illustration based on data from the German National Educational Panel Study. The suggested Bayesian approach is compared to other statistical and machine learning frameworks such as Lasso, ridge regression, and Elastic net, and is shown to perform well in terms of estimation performance and variable selection accuracy. The simulation results demonstrate that ignoring the handling of missing values in data sets can lead to the generation of biased selection results. Overall, the proposed Bayesian method offers a holistic, flexible, and powerful framework for variable selection in the presence of missing covariate data.

✉ Michael Bergrab · Christian Aßmann
Leibniz Institute for Educational Trajectories, Bamberg, Germany
E-Mail: michael.bergrab@lifbi.de

Christian Aßmann
E-Mail: christian.assmann@lifbi.de

Christian Aßmann
Chair of Survey Statistics and Data Analysis, Otto-Friedrich-Universität, Bamberg, Germany

Published online: 13 September 2024

**Keywords** Bayesian Estimation · Variable Selection · Model Selection · Sparsity · Spike-and-Slab Prior · Elastic Net · Shrinkage · Estimation

**MSC-Classification** JEL classification · C11 · C18 · C55

## 1 Introduction

In regression modeling a long-standing problem is to select an appropriate model in terms of the considered set of conditioning variables. The selection of appropriate variables is always related to the associated selection of models, where various approaches arising in the domain of statistical or machine learning algorithms are discussed in the literature to solve this task. While model selection is in principle straightforward in terms of a decision theoretic approach typically pursued in the context of model averaging, see Hansen (2007), implementation of such a model selection strategy considering all possible model setups is often impossible given available computing capacities. Since the seminal paper of Schwarz (1978) providing a benchmark criterion for model complexity obeying Occam's razor and allowing for model comparison on a common scale, many papers have addressed variable selection in frequentist and Bayesian model setups, see among others Raftery (1995); Tibshirani (1996); O'Hara and Sillanpää (2009); Bottolo and Richardson (2010); Ishwaran and Rao (2005). Clyde and George (2004) depict variable selection problems to become a special part of model selection with every subset of covariates corresponding to a distinct model, and finally every selection problem is a description of uncertainty to the data.[1] To avoid computationally difficulties when considering all possible models, selection methods help to pick up relevant variables and reduce the amount of variables that become part of the modeling process.[2] Further, formalized model selection strategies guard against ad-hoc multiple testing approaches invalidating the use of $p$-values and informal model assessment potentially results in incorrect inference due to strong multicollinearity among the set of variables. This points to the choices with regard to the set of variables considered within model selection. Next to the set of actually observed variables, say $X$, any combination of the variables in terms of higher order moments and cross products, or functional transformation thereof could be considered as well to handle nonlinear relationships. This increases also for moderate $P$ the number of variables to be considered. While Hansen (2007) and Frühwirth-Schnatter (2010), as typical for the applied literature, consider complete data scenarios, model selection strategies at least in the

[1] Model selection is further often related to prediction and estimation performance, which in the sense of model fitness is used as a model selection criterion. In the context of variable selection, the prediction and estimation performance relate to the ability to identify the relevant set of covariates and the implied regression relation correctly with higher prediction and estimating performance indicating higher quality of the applied approach. A side aspect of quality relates to the conceptual stringency and interpretability of the considered approach.

[2] Once a statistical modeling has been agreed upon, model selection problems can also be classified in terms of the number of variables ($P$) and number of observations ($N$). Thereby, the case with $N \gg P$ typically arises for survey data, whereas the case with $P \gg N$ can be found for medical data.

field of surveyed and administrative data should also address potentially incomplete data, i.e. the estimation strategy requires handling of missing data.

Typical formalized approaches to variable selection discussed in the statistical and machine learning literature are shrinkage estimators, with Lasso, ridge regression, and Elastic-net as the prominent variants. Next to established procedures such as stepwise selection, see Marill and Green (1963), also spike-and-slab prior formulations have been suggested, see Ročková and George (2018) and O'Hara and Sillanpää (2009) for an overview. As pointed out by Korobilis and Shimizu (2022) shrinkage estimators can be well aligned to the Bayesian estimation paradigm, where the different penalization terms correspond to assumed prior distributions, whereas the considered loss functions correspond to likelihood functions. However, the Bayesian estimation approach can be readily extended to handle missing values via the device of data augmentation, see Tanner and Wong (1987). Data augmentation in combination with Markov chain Monte Carlo methodology (MCMC) allows for derivation of estimators via sample averages. Further, the more complex the assumed likelihood structures are, the more compelling MCMC approaches to provide estimators may become.[3] When considering binary data or hierarchical model structures, the involved loss and likelihood functions, serving as optimization criteria in the statistical and machine learning context, become more complex although in principle straightforward to handle via MCMC techniques, see among others Aßmann and Boysen-Hogrefe (2011).

In this article, we hence illustrate how model selection for binary regression models can be performed simultaneously to handling missing values in the considered set of covariate data in a Bayesian framework. Comparison is provided regarding alternative statistical and machine learning approaches arising in the context of shrinkage estimation, such as Lasso, ridge regression, and Elastic-net regression for binary dependent variables. We provide the corresponding Bayesian approach based on a MCMC implementation for the handling of missing values accomplished in conjunction with estimation and variable selection and review the close relationships between shrinkage estimators. The described approach uses classification and regression trees to approximate the full conditional distribution of missing values. The holistic Bayesian approach allows for the incorporation of prior uncertainty and the flexibility to consider any function of observed or augmented data within the set of conditioning variables.

We assess the quality of different variable selection approaches when missing values occur, where the considered shrinkage estimation approaches are combined with multiple imputation, via a simulation study and an empirical illustration. As indicators of quality, we use indicators assessing the prediction performance regarding variable selection. The results suggest that for simple setups in terms of numbers of variables, missing mechanism, and dependency structures, all considered approaches perform well with regard to model selection. The more complex the setups becomes, the less reliable standard model selection approaches work, where especially infer-

[3] With more complex model structure also the burden of involved numerical optimization may rise. In combination with efforts to handle missing values via multiple imputation, see Rubin (1984), the computational costs increased considerably.

ence is more accurate for the suggested Bayesian approach based on spike-and-slab priors and simultaneous consideration of missing values. The empirical application illustrates that the considered Bayesian approach is well suited to provide weights that can be used in subsequent analyses. A main finding of the paper is hence that the quality of variable selection approaches remains high in principle even in the context of incomplete data situations. The paper also documents the required computational resources to apply estimation and model selection of this kind. Finally, the Bayesian approach to handling missing values and variable selection is a powerful and flexible framework that offers several advantages over established methods. By providing a coherent and unified framework, the Bayesian approach enables the comparison of different models on a common scale and the incorporation of prior knowledge and expert opinion. Additionally, standardizing the variables and ranking them based on the effect sizes can provide a simple and intuitive way to interpret the results. However, it is important to consider other concepts of variable importance when interpreting the results to gain a more comprehensive understanding of the relationships between the variables and the response variable.

The paper is organized as follows. Sect. 2 reviews the relationship between shrinkage estimators and Bayesian estimation and provides the suggested Bayesian approach towards model selection in the presence of missing covariate data in terms of a spike-and-slab approach for binary regression models. Also, the details with regard to posterior sampling and corresponding inference are provided. Sect. 3 subsumes the variable selection methods in statistical learning and handling of missing values in general and Sect. 4 adds quality assessments of variable selection. Sect. 5 presents results for simulated data sets and Sect. 6 provides the empirical illustration. Sect. 7 concludes.

## 2 Bayesian estimation for binary regression models with variable selection and handling of missing values

A Bayesian estimation approach in general can be motivated in terms of a decision theoretic approach using a loss function to assess the difference between parameter of interest taking value $\theta$ and the corresponding estimator $\theta^*$. A loss function $L(\theta, \theta^*)$ is defined as a mapping of the estimators $\theta^*$ from the set of possible estimators and each of the parameter values $\theta$ within the parameter space in the real line. The optimal estimator $\tilde{\theta}^*$ in terms of minimal expected posterior loss is then defined as

$$\tilde{\theta}^* = \arg\min \int_\theta L(\theta, \theta^*) f(\theta|D) d\theta,$$

where $f(\theta|D)$ denotes the posterior density of all parameters of interest $\theta$. The resulting minimization problem is similar to optimization problems arising in the context of penalized estimation if the involved loss function is defined to imply the mode of the posterior distribution and if the target function in penalized estimation (possibly in log scale) is similar to the structure of the posterior distribution proportional to the product of likelihood and prior distribution.[4]

This framing in terms of a decision theoretic approach points out that also all model selection decisions are an integral part of the estimation process with various model selection approaches being discussed within the literature. Standard estimation procedures are typically conditional on one specific statistical model and the data set under consideration, where properties of the considered data need to be reflected within the statistical model. These properties refer to the scale type of variables within the data, the dimensionality of the data set, and the completeness. Whereas the scale of the variables is reflected in the considered statistical model, strategies to handle the dimension or incompleteness of the data are typically not an integral part of the estimation routine, but considered in a sequential manner. The straightforward strategy to address all issues simultaneously provided by complete enumeration of all possible models (including all possible subsets of variables and incomplete data constellations) is often hindered by the tremendous computational efforts involved and the intractability of the incomplete data likelihood functions. The computational intensity is one of the main reasons why, from a historical viewpoint, strategies such as stepwise variable selection, both forward and backward, have been discussed in the context of complete data early.[5] Thus, the model selection process involves in complete data situations at most the evaluation of $P(P+1)/2$ model specifications instead of $2^P$ model specifications to find a maximum or minimum of the underlying break-up criterion. Thereby, the number of models considered within the selection is dramatically reduced although at the cost of path dependency.[6]

For illustration of the mechanisms involved in complete model enumeration, consider a set of models $\{\mathcal{M}_m\}_{m=1}^M$. Given data $D$, prediction or inference can be based on the asymptotic distribution or the posterior distribution of a parameter estimator for $\theta$ being model specific, i.e., $f(\theta|D, \mathcal{M}_m)$ being specific for the considered models $m = 1, \ldots, M$ with $2^P \leq M$ in case model selection coincide with selecting the appropriate subset of covariate variables, or $M$ even larger than $2^P$ when alternative model frameworks or missing data patterns are considered additionally.[7] A common pitfall is relying on a single model, especially when multiple models are equally likely but provide differing predictions or inferences.

[4] A direct correspondence is given when the prior is directly comparable in structure to the penalization term and the function assessing model fitness in the penalized context reflects the properties of the negative (logarithmic) likelihood function.

[5] For completeness, note that stepwise selection backwards starts with the most general still manageable model specification involving all$P$ variables and selects from $P$ candidate models, where these $P$ models each leave one variable out. If the predefined model selection criterion detects better fit among the candidate models, the candidate model with the largest increase in model fit is chosen, and the process is repeated until no further increase in model fit is detected. Stepwise selection – forwards or backwards – add or removes just one variable per step that changes the criterion most.

[6] Note that a similar strategy is also inherent to other approaches like classification and regression trees, see Breiman et al. (1984), where splitting points are defined in terms of single variables only.

[7] When making prediction or inference, it's essential to recognize that the both are based on a specific model and may be optimal only within the context of the considered models.

Bayesian model averaging provides a formal mechanism to aggregate estimation results from different models arising when performing a complete enumeration. Aggregated prediction or inference can be obtained via

$$f(\theta|D) = \sum_{m=1}^{M} f(\theta|D, \mathcal{M}_m) f(\mathcal{M}_m|D),$$

with

$$f(\mathcal{M}_m|D) = \frac{f(D|\mathcal{M}_m) f(\mathcal{M}_m)}{\sum_{m'=1}^{M} f(D|\mathcal{M}_{m'}) f(\mathcal{M}_{m'})}, \quad m = 1 \ldots, M$$

and $f(\mathcal{M}_m|D)$ denoting the posterior and $f(\mathcal{M}_m)$ the prior model probability, whereas $f(D|\mathcal{M}_m)$ denotes the marginal model specific likelihood implied via

$$\begin{aligned} f(\theta|D, \mathcal{M}_m) &= \frac{\mathcal{L}(D|\theta, \mathcal{M}_m) f(\theta|\mathcal{M}_m)}{f(D|\mathcal{M}_m)} \\ &= \frac{\mathcal{L}(D|\theta, \mathcal{M}_m) f(\theta|\mathcal{M}_m)}{\int \mathcal{L}(D|\theta, \mathcal{M}_m) f(\theta|\mathcal{M}_m) d\theta}, \quad m = 1, \ldots, M. \end{aligned}$$

$\mathcal{L}(D|\theta, \mathcal{M}_m)$ and $f(\theta|\mathcal{M}_m)$ thereby denote the model specific likelihood and prior distributions, respectively. In case the model specification addresses missing values, the likelihood

When a specific model, say $\mathcal{M}_{m'}$, has by far the highest probability aggregated prediction and inference resembles the inference and prediction conditional on model $\mathcal{M}_{m'}$. In case the posterior model probabilities of different models are similar, the aggregated and conditional predictions and inferences will also be similar.

This scheme allows for aggregating inference and prediction from a set of considered models, whether nested or non-nested. However, operationalization and implementation of this scheme require access to the likelihood function as well as tractability of the integration involved to derive the marginal model likelihood. Further, the scheme may be criticized to depend on prior assumption, although the set of considered models may include different prior settings as well for a given likelihood specification. As the computational efforts can become easily prohibitively large, this strategy is applied in the literature in case only relatively small sets of alternative models are considered, see Aßmann and Boysen-Hogrefe (2011), Aßmann (2012), and Frühwirth-Schnatter and Kaufmann (2008), where the computational issues are tractable.

Given the tremendous efforts possibly involved in this general strategy, alternative strategies are discussed in the literature providing model selection based on adaptive strategies. These alternative strategies may consider a restricted class of statistical models only or use alternative model criteria for selection and aggregation purposes which involve tractable computational efforts. In particular, model selection comes down to variable selection, also referred to as feature selection in the literature, when the statistical model is restricted to take the form of a regression model with the conditional expectation of the dependent variable taking the form $X\beta$, where $\beta$

is a $P \times 1$ vector of parameters. In general, looking on the $2^P$ possible regression models in total requires to calculate $2^P$ different comparative measurements, e.g., the Bayes-factor, which leads to model-averaging to get a posterior distribution that takes into account the uncertainty about all $M$ models which requires computing the posterior distribution over the parameters of interest in each model $\mathcal{M}_m$ (Clyde and George 2004). Additional computation is required for the posterior distribution over all such models. For linear regression models with complete covariate data Hansen (2007) discusses model averaging allowing as well as for model selection. Otherwise, in a Bayesian view ignoring the model or parameter by setting the prior to zero violates Cromwells's rule (Jackman 2009). Further, for a restricted model class, the model selection issue can be tackled as well by means of shrinkage estimation often also labeled as penalized estimation approaches. In a model-averaging perspective we are interested in all posterior model probabilities for models in which the regression parameters are unequal to zero which leads to variable selection (George 2000).

In the following, we will consider binary regression models with missing data in covariate variables and discuss shrinkage estimators and Bayesian variable selection approaches to handle the model selection issue arising in form of selecting the most appropriate subset of conditioning variables. The discussion will also point out how these approaches relate to the general strategy. The considered model framework can be described as follows. Let $D = \{y, X\}$ comprise a $N \times 1$ vector $y = (y_1, \ldots, y_N)'$ of binary dependent variables and a $N \times P$ matrix of covariate data $X = (X_1', \ldots, X_N')'$ not including a constant. We will introduce the probit specification for the binary regression model in the following as the involved MCMC sampling scheme is more tractable compared to a corresponding Logit specification. Hence, the binary regression model with probit link is given as

$$y_i = \begin{cases} 1 & \text{if} \quad y_i^* = \alpha + X_i\beta + e_i > 0, \\ 0 & \text{if} \quad y_i^* = \alpha + X_i\beta + e_i \leq 0, \end{cases}$$

where $e = (e_1, \ldots, e_N)'$ is a $N \times 1$ vector of independent standard normally distributed error terms and $y^* = (y_1^*, \ldots, y_N^*)'$ a vector of latent variables. The corresponding likelihood and augmented likelihood functions take the forms

$$\mathcal{L}(y|X, \beta, \alpha) = \prod_{i=1}^{N} \Phi((2y_i - 1)(\alpha + X_i\beta)) \tag{1}$$

and

$$\mathcal{L}(y, y^*|X, \beta, \alpha) = \prod_{i=1}^{N} \phi(y_i^* - (\alpha + X_i\beta))\{y_i\mathcal{I}(y_i^* > 0) + (1 - y_i)\mathcal{I}(y_i^* < 0)\}, \tag{2}$$

where $\Phi(\cdot)$, $\phi(\cdot)$, and $\mathcal{I}(\cdot)$ denote the cumulative density of the standard normal distribution, the density of the standard normal distribution, and the indicator function, respectively.

The consideration of the augmented likelihood function is based on Albert and Chib (1993) as it simplifies the implementation of a MCMC sampling scheme for estimation. Contextualizing this for variable selection within binary regression models, a representation for all possible model specifications is required. In general, the model setup is implied via the likelihood and the prior distribution which yields to approximate the posterior distribution as proportion of the likelihood and the prior distribution. To describe differences between Bayesian estimators and shrinkage or Maximum Likelihood estimators in general it may be helpful to recall that Bayesian estimators can be formulated as a decision theoretic problem aiming at minimizing Bayes risk. This risk is associated with an appropriate loss function, see Mood et al. (1974). Depending on the loss function, the posterior mean or median are appropriate Bayesian estimators. In this sense, shrinkage estimators may be interpreted as posterior mode estimators, although as pointed out by Gneiting (2011) it might be hard to reconcile this kind of estimator with the decision theoretic duality of loss functions and estimators in case of interaction between penalization and cross-validation. In this sense, and as the posterior distribution is hardly ever accessible by analytical means, Bayesian estimators are typically derived as sample means, where samples from the assumed posterior distribution are obtained using Markov chain Monte Carlo (MCMC) methods. Different MCMC techniques for Bayesian variable and model selection are developed varying the prior distribution or the mechanism in the MCMC sampler, for details see Yang et al. (2005).

To arrive a general model specification encompassing all possible models, the binary regression setup with $\beta = (\beta_1, \ldots, \beta_P)'$ described above can be extended as follows. Following Lee et al. (2003), we use a $P \times 1$ indicator vector $\gamma$, where each single $\gamma_j$ with $j = 1, \ldots, P$ is defined by

$$\gamma_j = \begin{cases} 1, & \text{if variable } X_j \text{ is considered corresponding to } \beta_j \neq 0, \\ 0, & \text{if variable } X_j \text{ is not considered corresponding to } \beta_j = 0. \end{cases} \tag{3}$$

Taking $\gamma$ as a condition into account, the model described in Eqs. (1) and (2) would become

$$\mathcal{L}(y|X, \beta, \alpha, \gamma) = \prod_{i=1}^{N} \Phi(2y_i - 1)(\alpha + X_i \mathrm{diag}(\gamma)\beta)$$

and

$$\begin{aligned} &\mathcal{L}(y, y^*|X, \beta, \alpha, \gamma) = \\ &= \prod_{i=1}^{N} \phi(y_i^* - (\alpha + X_i \mathrm{diag}(\gamma)\beta))\{y_i \mathcal{I}(y_i^* > 0) + (1 - y_i)\mathcal{I}(y_i^* < 0)\}, \end{aligned} \tag{4}$$

with the diag() operator stacking the indicated vector on the main diagonal of corresponding square matrix. To complete the model setup, priors for $\alpha$, $\beta$, and $\gamma$ need to be specified when variable selection is considered for binary regression models. The implied posterior can be described via

$$p(\beta, \alpha, \gamma | y, X) \propto \mathcal{L}(y | X, \beta, \alpha, \gamma) f(\beta, \alpha, \gamma). \tag{5}$$

Thereby, all quantitative continuous covariate variables in the data set are considered to be standardized via a $z$-transformation.[8] In case an intercept is considered in the model specifications, the intercept is then the common parameter for all possible model and represents the overall mean of the model. Following Lamnisos et al. (2009); George and McCulloch (1993); Lee et al. (2003) we use a normal prior for $\alpha$ with expected value $\alpha_0$ and variance $h$, i.e.

$$f(\beta, \alpha, \gamma) = f(\alpha) f(\beta, \gamma) = \phi(\alpha | \alpha_0, h) f(\beta, \gamma), \tag{6}$$

with $\phi(\cdot|\cdot, \cdot)$ denoting the normal distribution with indicated expectation and variance parameter. Typically, $h$ is set as a large value corresponding to an uninformative prior setting with regard to the intercept (Lamnisos et al. 2009). Table 1 outlines the details with regard to hyperparameters of all prior distributions.

The prior setting for $\beta$ and $\gamma$ is based on George and McCulloch (1993) assuming an independent marginal conditional setup

$$f(\beta, \gamma) = \prod_{j=1}^{P} f(\beta_j, | \gamma_j) f(\gamma_j). \tag{7}$$

The functional forms follow spike-and-slab priors first suggested by Mitchell and Beauchamp (1988) for Bayesian variable selection for normal linear regression models. The according mixture distribution for the coefficients is given as

$$f(\beta_j | \gamma_j) = (1 - \gamma_j) f_1 + \gamma_j f_2, \tag{8}$$

where $f_1$ and $f_2$ are placeholders for any appropriate continuous or discrete probability density function. Given this mixture form, $f_1$ is used to steer coefficients to zero (spike), e.g., $f_1$ assigns a unit point mass at $\beta_j = 0$ and $f_2$ allows for nonzero coefficients (slab), which can be an absolutely continuous density otherwise such as uniform or normal. Hence, this setup directly incorporates selective shrinkage, i.e., the separating effect in the coefficients caused by the spike-and-slab so

[8] The situation with incomplete covariate data requires that each imputed data set is subject to a specific$z$-transformation or that within each iteration of the MCMC algorithm a $z$-transformation is performed as explained below. In addition, while the use of standardized covariate data is an implicit requirement in the context of shrinkage estimation as a single parameter steers the shrinkage, the Bayesian formulation in form of a variable specific prior allows for more flexibility. Most important, while for continuous quantitative variables standardization is possible using a $z$-transformation or other stabilizing transformations such as singular value decomposition, standardization is less straightforward implemented for categorical variables.

**Table 1** Prior specification and MCMC starting values

| Parameter | Functional form | Probability distribution | Starting values |
|---|---|---|---|
| *Intercept* | | | |
| $\alpha$ | $\propto \phi(\alpha_0 = 0, \sigma_\alpha^2 h)$ | Normal | – |
| | Depending on scaling parameter $h$ | | |
| | Wet set to $h = 1$, thus | | |
| $\sigma_\alpha$ | $\propto IG(c_1, c_2)$ | Inverse gamma | – |
| | With $c_1$ and $c_2$ const. | | |
| *Regression vector* | | | |
| $\beta_p$ | $\propto (1 - \gamma_j)\mathcal{N}(\boldsymbol{\beta}_0, \tau_1^2\sigma_\beta^2) + \gamma_j\mathcal{N}(\boldsymbol{\beta}_0, \tau_2^2\sigma_\beta^2)$ | Mixing normal | $\{1\}_{p=1}^P$ |
| | Depending on | | |
| | $\boldsymbol{\beta}_0 = 0$ | | |
| | $\tau_2 \gg \tau_1 > 0$ with $\tau_2 = 1$ and $\tau_2$ | Constant | Set individually |
| $\sigma_\beta$ | $\propto IG(d_1, d_2)$ | Inverse gamma | – |
| | With $d_1$ and $d_2$ const. | | |
| | Mostly $d_1 = 100$, and $d_2 = 100$ | | |
| *Indicator vector, i.e., spike-and-slab* | | | |
| $\gamma$ | $\propto$ Bernoulli$(w)$ | Bernoulli | $\{1\}_{p=1}^P$ |
| | Depending on | | |
| | $w \in (0, 1)$ | Constant | Set individually |
| *Missing values* | | | |
| $X_{\text{mis}}$ | $\propto$ observed sample distribution | Nonparametric | Random draws |

The hyperparameters for the inverse gamma distribution are chosen to provide finite variance and smallest possible prior sample size.

that most coefficients are peaked at zero and significant coefficients are set different from zero. A mixture of two normal distributions with different variances are widely used implying

$$f(\beta_j|\gamma_j) = (1 - \gamma_j)\phi(\beta_j|\beta_0, \tau_1^2\sigma_\beta^2) + \gamma_j\phi(\beta_j|\beta_0, \tau_2^2\sigma_\beta^2), \tag{9}$$

where typically $\tau_2 \gg \tau_1$. Note that different approaches for spike-and-slab prior setups can be found, such as Laplace priors (Tibshirani 1996) or Horseshoe priors (Carvalho et al. 2010) or setting $f_1$ to unit mass at zero (George and McCulloch 1997). Hence, if $\beta_j$ is found to differ substantially from zero, it will be assigned in the model (slab) or otherwise will be skipped out of the model (spike). Note that the different prior setups correspond to different setups of the penalization function in the context of shrinkage estimators.

George and McCulloch (1993) introduced the Stochastic Search Variable Selection (SSVS) method, which is a Bayesian approach for variable selection in linear regression models. SSVS uses a mixture of two normal distributions as a prior for the regression coefficients, where one distribution has a very small variance and the other has a large variance. This prior encourages shrinkage of the coefficients towards zero, and the stochastic search algorithm explores the model space to identify the most important variables. For the prior for $\gamma$ that specifies the model space we follow George and McCulloch (1993, 1997) and Lee et al. (2003) and consider a Bernoulli prior framework given as

$$f(\gamma) = \prod_{j=1}^{P} w_j^{\gamma_j} (1 - w_j)^{1-\gamma_j} \tag{10}$$

with $w_j \in (0, 1)$ governing the probability that the $j$-th column of $X$ is considered within the regression. It is also common to set $w_j = w$ for $j = 1, \ldots, P$, thereby assuming homogeneity of the inclusion probabilities.

In this case, the prior distribution of $\gamma$ is binomial and the a priori expected number of selected variables of $X$ can be modeled in terms of $w$. A fixed value for $w$ can be set if there is consolidated knowledge. If $P \gg N$, small values of $w$ are chosen, to bound the number of variables in the model. Hence, the prior penalizes larger models by setting $w$ to a small percentage.[9] Otherwise, a maximum for the model size $P_{\max}$ can be set as in Dobra (2009).[10] In the following, we use the binomial prior on $\gamma$ with homogeneous inclusion probabilities.

The model setup so far describes the situation with completely observed data and is, as discussed in the literature, accessible to posterior sampling, see also Albert (1992). Data augmentation can also be used to handle missing values. To perform Bayesian inference, an MCMC sampling scheme, see e.g., Geman and Geman (1984); Gelfand and Smith (1990); Aßmann and Preising (2020), is implemented to generate a sample from the posterior distributions of interest. Handling of latent structures and missing values is conceptually straightforward in the Bayesian context via the device of data augmentation since the full conditional distributions of missing values can be added as outlined in Aßmann et al. (2022). The prior is thereby also augmented where we opt for a prior for the missing values proportional to the distribution of observed values, see also below. The parameter vector can be augmented with the missing values, which can then be utilized as conditions for all other full conditional distributions of interest. In the context of a binary probit models, data augmentation involves augmenting the dependent variable $y$ by drawing a new value $y^*$ from a conditional distribution depending on the other current model parameters, thus the observed binary data is augmented with the latent continuous variable. Data augmentation can be operationalized via including a set of appropriately specified full conditional distribution for the missing values within the MCMC sampling scheme. First the latent variable is drawn or the missing variable are handled, then the model parameters are updated by using the current augmented

[9] In the case of a huge more variables than individuals, a small$w$ selects only a few variables. e.g., a data set with 10,000 variables and 1000 individuals a value of $w = \mathbf{0.001}$ means that only 10 variables a expected to be selected into the model.

[10] An alternative way is to specify a hierarchical prior distribution for$w$. Thereby, uncertainty of $w$ can be modeled by implementing a prior for $w$ following a distinct distribution, e.g., a Beta distribution with $\mathcal{B}$ is a Beta function and $w \sim \mathcal{B}(\delta_1, \delta_2)$ (Kohn et al. 2001) so that

$$f(w) = \frac{w^{\delta_1 - 1}(1 - w)^{\delta_2 - 1}}{\mathcal{B}(\delta_1, \delta_2)}$$

with $\mathcal{B}(\cdot, \cdot)$ denoting the Beta function. The prior belief of the model size, i.e., the number if included variables can be parameterized with both $\delta_1 > \mathbf{0}$ and $\delta_2 > \mathbf{0}$.

data. Finally, it allows to estimate the model parameters more accurately and flexibly, and can be used in a variety of applications.[11] This framework has been widely used in various applications, including missing data imputation.

The quantities of interest are hence $y^*, \beta, \alpha, \gamma$, and $X_{\text{mis}}$, where $X_{\text{mis}}$ denotes the missing values of the covariate data $X$ with $X = (X_{\text{obs}}, X_{\text{mis}})$. The corresponding posterior of interest results from Eqs. (4) in combination with the assumed operationalizations of Eq. (6), see also Table 1. The prior for the missing values $X_{\text{mis}}$ is discussed when providing the assumed full conditional distribution. Starting point for sampling and inference is hence the augmented posterior distribution, see also Aßmann et al. (2022), given as

$$p(\beta, \alpha, \gamma, y^*, X_{\text{mis}}|y, X_{\text{obs}}) \propto \mathcal{L}(y, y^*|X, \beta, \alpha, \gamma) f(\alpha) f(\beta, \gamma) f(X_{\text{mis}}|X_{\text{obs}}).$$

Thus, we draw the posterior values iteratively for $m = 1, \ldots, M$ from the respective full conditional distributions of the considered parameter blocks $y^*, \beta, \gamma, X_{\text{mis}}$. After setting appropriate starting values for $\beta, \gamma, X_{\text{mis}}$, the set of full conditional distributions in the Gibbs sampler is set up as follows. Fig. 1 shows the schematic progression of the Gibbs sampler with the setting of the start values and the sequential structure of the full conditional distributions.

$f(y^*|\cdot)$ The full conditional distributions of the latent variables $y^*$ corresponds to a product of truncated normal distributions, since the single elements $y_i^*, i = 1, \ldots, N$, are mutually independent. Sampling for each element is hence performed from a truncated normal distribution with moments $\mu_{y_i^*} = \alpha + X_{i,\gamma}\beta_\gamma$ and variance equal to one with the truncated sphere ranging from $-\infty$ to 0 in case $y_i = 0$ and in case $y_i = 1$ ranging from 0 to $\infty$.

$f(\alpha, \beta|\cdot)$ Following Albert and Chib (1993), the full conditional distribution follows in principle the standard Bayesian linear regression given the latent continuous variable $y^*$. Consideration of the underlying continuous spike-and-slab prior the full conditional distribution has the form of a multivariate normal with variance and expectation given as

$$V_{\alpha,\beta} = (D^{-1} + \tilde{X}'\tilde{X})^{-1} \quad \text{and} \quad m_{\alpha,\beta} = V(D^{-1}(\alpha_0, \beta_0')' + \tilde{X}'y^*)$$

[11] Depending on the scale of the variables under consideration, both parametric and non-parametric models may be appropriate to specify the full conditional distribution of the variables showing missing values. Following Burgette and Reiter (2010) and Doove et al. (2014), we use classification and regression trees (CART) as discussed by Breiman et al. (1984) to approximate the full conditional distributions. This offers a flexible yet computationally feasibly way to model missing values. As previously stated by Aßmann and Preising (2020), data augmentation is employed directly within the MCMC sampler. During the individual draws from the full conditional distributions of an MCMC run, not only are the estimates calculated and a variable selection performed, but also the missing values are imputed. The imputation is conducted using the approach proposed by Tanner and Wong (1987), where a general framework for data augmentation is proposed, which involves introducing latent variables to the model to simplify the estimation of parameters.

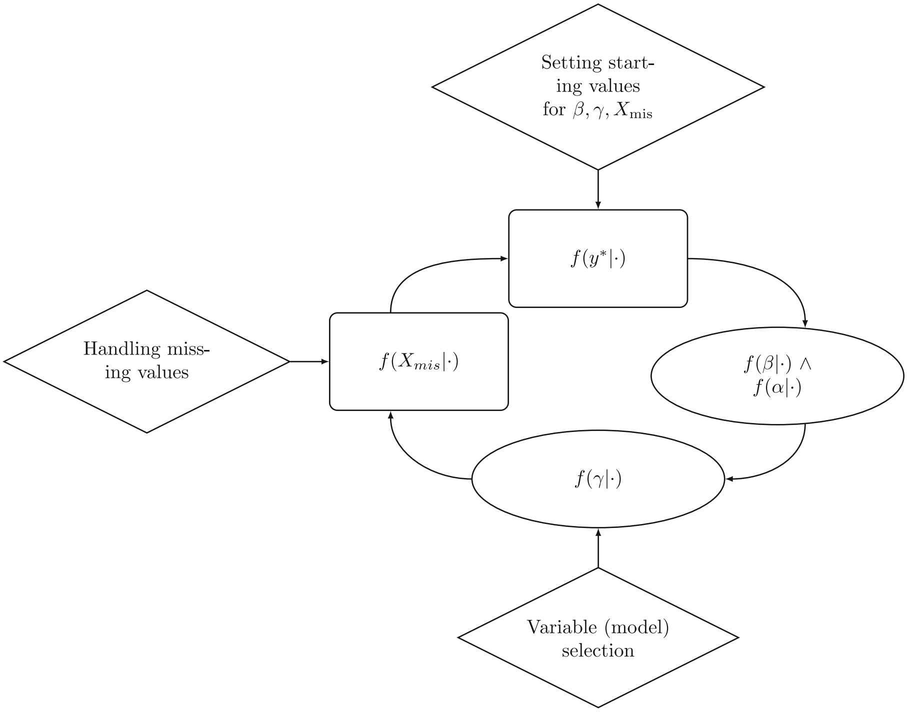
Setting starting values for $\beta, \gamma, X_{\text{mis}}$
$f(y^*|\cdot)$
Handling missing values
$f(X_{mis}|\cdot)$
$f(\beta|\cdot) \wedge f(\alpha|\cdot)$
$f(\gamma|\cdot)$
Variable (model) selection

**Fig. 1** Schematic progress of the sequential structure of the full conditional distributions within the Gibbs sampler. Note that the starting values are set once before starting the $m = 1, \ldots, M$ iterations. The full conditional distributions in the ellipses express extended data, whereas the rectangular blocks represent the full conditional distributions, which provide the output of interest. After subtracting an appropriate burn-in phase, both provide the corresponding estimators based on the median or mean

respectively, where $D$ is a diagonal matrix with $D = \operatorname{diag}(h, (\iota_P - \gamma)\tau_1 + \gamma\tau_2)$ with $\iota_.$ denoting a vector of ones with indicated size and $\tilde{X} = (\iota_N, X)$, see also Biswas et al. (2022) for a discussion of this full conditional distribution.[12]

$f(\gamma|\cdot)$ The full conditional distribution for $\gamma$ is implied via the assumed prior structure and corresponds to $P$ independent Bernoulli distributions as the single elements $\gamma_j$, $j = 1, \ldots, P$, are mutually independent. The corresponding implied probabilities are given as

$$p_j = \frac{w\tau_2^{-1} \exp\left\{-\frac{\beta_j^2}{2\tau_2^2}\right\}}{w\tau_2^{-1} \exp\left\{-\frac{\beta_j^2}{2\tau_2^2}\right\} + (1-w)\tau_1^{-1} \exp\left\{-\frac{\beta_j^2}{2\tau_1^2}\right\}}, \quad j = 1, \ldots, P,$$

[12] Drawing directly from the distribution is very time intensive if $P \gg N$. Following Biswas et al. (2022) the direct drawing routine requires computational cost of $\mathcal{O}(p^3)$ and thus can be modified based on the Woodbury matrix identity summarized by Bhattacharya et al. (2016), which requires still cost of $\mathcal{O}(N^2 p)$ but instead of other approaches converges to the posterior distributions.

with the hyperparameters $0 < w < 1$ and $\tau_2^2 \gg \tau_1^2 > 0$, see Biswas et al. (2022) for discussion and Table 1 for chosen values.

$f(X_{\text{mis}}|\cdot)$ Values of $X_{\text{mis}}$ are sampled sequentially for each column vector of $X$, i.e., $X = (X^{(1)}, \dots, X^{(P)})$, based on the non-parametric approximation suggested in the form of classification and sequential regression trees (CART), see Burgette and Reiter (2010). Let $X_{\text{com}}^{(k)} = (X_{\text{obs}}^{(k)}, X_{\text{mis}}^{(k)})$, $k = 1 \dots, P$, denote the completed variables, and $X_{\text{com}}^{(\backslash k)}$, $k = 1, \dots, P$, denotes the completed matrix of variables except column $k$. It is a major advantage of the data augmentation approach that the latent variables possibly serving as kinds of sufficient statistics can be used for the approximation of the full conditional distribution of missing values. In a first step, a decision tree is built for $X_{\text{com}}^{(k)}$ conditional on the corresponding values of all remaining variables $X_{\text{com}}^{(\backslash k)}$ as well as conditional on $y$ and $y^*$ serving as a kind of sufficient statistic for $y$.[13] In each iteration the covariates are standardized in the spike-and-slab approach as well as in the imputation for the variable selection. To incorporate a prior uncertainty on the hyperparameters of the sequential partitioning regression trees, we build trees with a randomly varying minimum number of elements within nodes. Every missing observation can then be assigned to a node and thus a grouping of observations implied by the binary partition in terms of the conditioning variables. The values within each node provide access to an empirical distribution function serving as an approximation to the full conditional distribution of a missing value and thus as the key element for running the data generating mechanism for missing values. Thereby, this modeling approach is in line with prior distributions of missing values proportional to observed data densities. Draws from the empirical distribution function within a node correspond to draws from the full conditional distributions of missing values, where sampling is performed via the Bayesian bootstrap to account for the estimation uncertainty of the full conditional distribution, see Rubin (1981). The considered approach further offers the flexibility to consider any function of observed or augmented data within the set of conditioning variables as well. We incorporate the implementation available within the rpart package available for R (R Core Team. 2020), see Therneau and Atkinson (2018) for further details. The sampled $X_{\text{mis}}$ values allow to refer to an updated completed matrix of covariates in all other steps of the MCMC algorithm including a renewed standardization of the covariate data.

Given the MCMC output, estimators are readily defined via corresponding sample moments, either arithmetic means or medians. Whether arithmetic means or medians are reported depends on the loss function involved in defining a Bayesian point estimators based on a Risk function, see Mood et al. (1974). Arithmetic means

[13] Note that this specification is also used when performing imputation in the context of alternative approaches serving as comparative benchmarks.

correspond to quadratic Bayesian loss, whereas medians correspond to absolute Bayesian loss. Furthermore, the model structure implies that estimators conditional on $\gamma_j = 1$, $j = 1, \ldots, P$, may be considered as well as unconditional estimators. Whereas unconditional estimators can be obtained via using the complete MCMC output, conditional estimators correspond to discarding those draws for which the sample elements in $\gamma$ are equal to one. In this sense, estimates of the inclusion probabilities reaching at least 50% are necessary to consider a variable as a part of the true underlying data generating process, see also Russu et al. (2012); Bottolo and Richardson (2010); Hans et al. (2007). Finally, note that within the simulation experiments as well as within the empirical illustration, we set $M$ to equal 20,000 after a burn-in phase of 5,000 was found sufficient to discard the effects of initialization both within the simulations study and the empirical illustration. Next, we will discuss variable selection and handling of missing values in the context of shrinkage estimators the relations to Bayesian estimation.

## 3 Shrinkage estimation for binary regression models with missing covariate data

Variable selection as a special case of model selection can be performed in terms of shrinkage estimators. Thereby, the task is to identify a subset of variables that are potential predictors for the dependent variable and is best with respect to predefined optimality criterion. Resulting reduced models give us a higher chance to interpret, visualize and handle the results suitably.

In general, the shrinkage or penalized estimation approach for variable selection is provided in terms of a loss function. Hence, with $\theta = (\alpha, \beta)$, the resulting can be defined as

$$\theta_{\text{shrink}} = \arg\min_{\theta}\{\mathcal{LF}(D, \theta) + p_{\text{shrink}}(\beta, \lambda)\}. \tag{11}$$

Thereby, $\mathcal{LF}(D, \theta)$ measures the ability of the model to fit the data usually taking a form close to a likelihood function, whereas $p(\beta, \lambda)$ penalizes model complexity, i.e., the dimensionality of the parameter vector $\beta$ and $\lambda$ steers the magnitude of penalization. The different shrinkage estimators discussed in the literature differ with respect to alternative specifications of $\mathcal{LF}(D, \theta)$ and $p_{\text{shrink}}(\beta, \lambda)$. In general, the loss function resembles in structure the Mallows criterion, see Mallows (1973).

Prominent choices for measuring model fitness is the residual sum of squares for continuous dependent variables, where the sum of squared residuals is also a constituent part of the log likelihood function given the assumption of multivariate normality or more generally assuming an ellipsoid distribution. The penalization function should be monotonically increasing for larger dimension of $\beta$, where typical functions fulfilling this requirement are quadratic or absolute distance functions. Note that these also play a prominent role in logarithms of densities, for example the normal or Laplace density, respectively. Given this, log likelihood functions and log prior distributions operate in the same way as loss and penalty functions respectively, which in turn makes log likelihood functions and priors natural candidates for defining shrinkage estimators. These relationships will be highlighted in more detail, when discussing specific shrinkage estimators in the following. A final remark relates to the mechanism how the consideration of penalty functions causes the selection of a subset of variables. For illustration consider the case, where the function assessing model fitness takes the form of an ellipsoid with fitness decreasing with larger distance to the center of the ellipsoid. The penalty function contributes minimum loss when parameters take the value zero. The overall loss function is thereby minimized via balancing marginal increase in fitness with marginal loss arising from the penalization in a Lagrangean manner. The point where the marginal fitness and penalization contributions can be balanced depends on the chosen functional forms, where opting for absolute loss may provide shrinkage of single parameters exactly towards zero.

Different specifications of $p(\beta, \lambda)$ imply different shrinkage estimators. A general specification for the penalization function can be stated in form of a linear combination of different distance functions or norms, i.e.,

$$p(\beta, \lambda) = \sum_{j=1}^{J} \lambda_j |\beta' \beta|^{\frac{\kappa_j}{2}},$$

where for $\kappa_j$ taking values $1, 2, \ldots$, i.e., the corresponding $L_\kappa$-norms are involved. The following special cases are prominent in the literature. For $J = 1$ and $\kappa_1 = 2$, the penalization function involves quadratic norms $L_2$ what is referred to as ridge regression, i.e., $p_{\text{ridge}}(\beta, \lambda) = \lambda_1 \beta' \beta$, see Hoerl and Kennard (1970) and Friedman et al. (2010) for a discussion in the context of generalized linear models. $\lambda_1$ controls the impact of the penalization, with higher values pushing more coefficients towards zero. If $\lambda_1 \to \infty$, then $\hat{\beta}_{\text{ridge}} \to 0$, so that the model finally consists only of the intercept. With ridge regression no genuine variable selection is possible because all coefficients stay in the model but are more or less shrunk towards zero. In the context of the binary probit regression model, the loss function is typically chosen as $\mathcal{LF}(D, \theta) = -\ln \mathcal{L}(D|\theta) = -\ln \mathcal{L}(y|X, \beta) = -\sum_{i=1}^{N} \ln \Phi((2y_i - 1)(\alpha + X_i \beta))$. Given this, the ridge regression approach towards binary dependent data resembles a Bayesian estimation approach with multivariate normal prior in the sense that the overall optimality function of the ridge regression has a functional form that is proportional to the logarithm of the implied unnormalized posterior distribution.

A similar situation arises when considering the case with $J = 1$ and $\kappa_1 = 1$, implying the use of absolute values instead of squared ones. Tibshirani (1996) introduced this least absolute shrinkage and selection operator (Lasso) to the linear regression problem, extended to the generalized linear model by Park and Hastie (2007). Given this, the penalization function $p(\beta, \lambda)$ becomes $p_{\text{Lasso}}(\beta, \lambda) = \lambda_1 (\beta' \beta)^{\frac{1}{2}}$. Increasing $\lambda_1$ sets more coefficients to zero, causing the selection of fewer variables with the selected being shrunk, and finally the number of nonzero coefficients decreases. The analytical solution of the Lasso due to the $L_1$-norm penalty is more complicated than with $L_2$-norm.[14] Again, similarity to a Bayesian setup with Laplace

[14] Note that for handling the problem of (high) correlation among the variables, fused Lasso (Tibshirani et al. 2005) or adaptive Lasso (Zou 2006) are discussed in the literature.

prior distribution should be noted when considering the functional forms of the logarithm of the unnormalized posterior density and the overall loss function. Friedman et al. (2010) show that both Lasso and ridge regression have their drawbacks and advantages in the context of correlated variables and over-fitting. Therefore, Zou and Hastie (2005) proposed the Elastic net approach incorporating to include the best of both. This method uses both $L_1$- and $L_2$-penalties and thus is a convex combination of the ridge and Lasso approach towards shrinkage. Friedman et al. (2010) extend the Elastic net to generalized linear models where $J = 2$, $\lambda_1 = 1$, and $\lambda_2 = 2$. After reparametrization the Elastic net manifests as

$$\beta_{\text{ElasticNetMod}} = \arg\min_{\beta} \left\{ \mathcal{L}\mathcal{F}(D, \theta) + \lambda(\varphi|\beta'\beta| + (1-\varphi)|\beta'\beta|)^{\frac{1}{2}} \right\}, \tag{12}$$

thereby using $\varphi$ as control parameter for the weight given to the $L_1$- and $L_2$-norm driven penalty with $0 < \varphi < 1$ and $\lambda > 0$ as weighting parameter given to the penalty. The combination causes the Lasso penalization term to select among the variables thereby putting all weight on the set of selected variables, whereas the ridge term shrinks coefficients towards each other. Hence, Elastic net finds a sparse model with typically high prediction accuracy. Framing this approach from a Bayesian perspective implies that the penalization function is similar to the kernel of the logarithm of a mixture distribution, where the two mixture components follow normal and Laplace distributions, respectively.

As discussed, the Bayesian approach to handling missing values and variable selection, as well as established methods, can set regression coefficients to zero if appropriate. This means that variables are selected by the different approaches, but without ranking the importance of these variables. However, if using standardized variable values, the size of the regression coefficients can be used as a comfortable and simpler way to rank the variables. This ensures that all variables are on the same scale, which facilitates easy comparison of each variable. This approach is recommended by Kyung et al. (2010) for handling different variables with different measurements. Additionally, as discussed by Friedman et al. (2010), standardizing the variables simplifies the analysis. When ranking variables based on the size of the regression coefficients, it is essential to recognize that this ranking reflects the effect sizes. However, it is also crucial to consider other concepts of variable importance, such as significance and permutation feature importance. These alternative concepts can provide additional insights into the relationships between the variables and the response variable and should be taken into account when interpreting the results (Strobl et al. 2007). By considering multiple perspectives on variable importance, it becomes possible to gain a more nuanced understanding of the relationships in the data.

After standardization we resort to various R packages. For Elastic net estimation the glmnet-package (Friedman et al. 2010) provides many setting options for the above-mentioned control and penalty parameters. As hyperparameter, we set both $\varphi = 0.0$ for ridge penalized estimation and $\varphi = 1.0$ for the Lasso penalty and additionally $\varphi = 0.5$. The strength of penalty is controlled by the tuning parameter $\lambda$ which can also be set before estimation or via cross-validation which is most widely used and implemented (Friedman et al. 2010). The $k$-fold cross-validation is used with $k = 10$ folds and a squared loss to use for cross-validation by mean squared error. The shrinkage parameter $\lambda$ is picked up out of the sequence of possible parameters as the within one standard error of the minimum mean cross-validated error value, so that the most regularized model is given. After chosen the shrinkage parameter the Elastic net is finally estimated with the set control and the cross-validated shrinkage parameter for each $m$ data set. Then, the presented results are averaged over the $M$ datasets.

However, the methods are typically discussed and evaluated under the assumption of fully observed covariate data, so that missing values can be handled in a variety of ways beyond the automated Bayesian approach. First, the data augmentation method within Gibbs sampling is particularly well suited to dealing with missing data problems, since the inclusion of missing values in the parameter vector results in the treatment of all other model quantities as if the data were fully observed. In addition, adding the data augmentation step to the Bayesian estimation routine allows for the avoidance of combination rules (Tanner and Wong 1987). Second, to cope with missing values in covariate data in the other above-mentioned approaches, we make use of multiple imputation, see Rubin (1976), where we use the multiple imputation via chained equations (MICE) approach, following Buuren and Groothuis-Oudshoorn (2011). Within the MICE approach, for each variable showing missing values, a full conditional model is specified, where imputations are generated via sampling from these full conditional distributions. Given an appropriate implementation scheme, the MICE algorithm repeatedly iterates over the sequence of assumed full conditional distributions, generates imputations via sampling, and hence updates the data. After an appropriately chosen burn-in phase, obtained draws can be used to build an imputed data set that can be used in subsequent analyses. Repeating the MICE algorithm $M$ times provides $M$ imputed data sets given those the shrinkage estimation routines are performed for each of the imputed data sets.

Furthermore, the treatment of missing values leads to further considerations regarding the evaluation of the quality aspects of the different selection and estimation procedures. The issues and challenges associated with multiple imputation appear widely, in some cases, as convergence issues, imputations model mis-specification, imputation of rare events, large amounts of missing data, challenges while reporting and interpreting or issues while pooling the results. Following Buuren (2018) the general routine associated with MICE of imputing data, analyzing results and pooling results across all $M$ imputed datasets becomes difficult in variable selection because the set of selected variables will differ more or less. In the literature of statistical and machine learning context exists no standard method to pool the results and to combine the information provided by the $M$ different model results. Even for complete datasets, the likelihood-based variable selection methods reach to several limitations (Miller 2019). In the literature different methods are discussed and for a current overview see Du et al. (2022). Brand (1999) presents a two-step solution which pools the results based on the pooled likelihood ratio p-values selecting insignificant variables after applying stepwise regression to each $M$ imputed dataset and exclude variables from a combined supermodel if they have been selected at least less than half of the runs. Otherwise, (Bayesian) model averaging can be resorted to in order to account for the variability of the selected variables across all imputed data sets (Yang et al. 2005). Buuren (2018) distinguishes the literature into three general approaches, referring to Wood et al. (2008) and Vergouwe et al. (2010): the Majority approach counting how often a variable is selected (at least half of the models applied to the imputed datasets), and the Stack approach, so that all imputed datasets are stacked into a single dataset applying variable selection methods with weights, and the Wald approach, especially for stepwise selection, pooling based on the Wald statistics.

Considering the above-mentioned variable selection approaches while handling missing data, we present the following routines. First, we present an average-based approach (Average), where the final shrinkage estimator is obtained as the arithmetic mean of the $M$ estimators obtained for each imputed data set and the combined variance estimator is given as the sum of within and between variance of the $M$ estimators obtained from the imputed data sets.[15] Hence, this procedure is a rough way of pooling estimates, but common in daily practice. As a second approach (Majority), we set up an imputation based on the above-mentioned MICE settings, where we perform variable selection techniques on each imputed dataset $m$ resulting to $M$ different selection models with partly different selected variables. For pooling, we extract the selected variables from each model and sum across the imputations to identify variables that were selected in at least half of the imputed datasets. Then, we estimate a probit model as supermodel with the mostly selected variables and pool results according to the Rubin's rules for generalized linear models. However, this procedure implicitly involves a loss of information. Finally, following Buuren (2018) we implement a pooling approach based on the Wald test (Wald) and expand the Majority-approach by extracting the redundant variable by testing. After counting who often a variable is selected in the $M$ imputed datasets, we compare the variable appearing in more than 50% of the $M$ models and apply a Wald test to determine which variable of the sorted counts.

## 4 Quality assessments of variable selection while handling missing values

Value of data is in general linked to the ability to form informed decisions based on the available data information. This value depends on the quality of statistical or machine learning algorithms to use the available data information thoroughly. The proposed Bayesian approach illustrates in the context of a binary regression model the possibilities to integrate all available information into the analysis of factors influencing the binary dependent variable. The proposed method covers the data constellation with many although incompletely observed variables and relatively few observations. The proposed algorithm learns in a machine learning like manner the best combination of covariate variables explaining the dependent variable thereby addressing the entailed problem to decide for a subset of variables and their corresponding influence. To assess the quality in terms of the statistical efficiency of the proposed method, we present a couple of statistical measures typically used

[15] Typically rules for asymptotically normally distributed estimators are considered.

to compare different statistical approaches. In the context of variable selection and missing value imputation, the quality aspects refer to the indicators used to assess the performance of different approaches in selecting the correct variables and handling missing values. These quality aspects are typically related to the prediction performance and accuracy of the model. This means that the quality of the approaches is evaluated based on how well they can identify the correct variables while imputing missing values. First, a common quality aspect is to evaluate the performance of variable selection approaches including the accuracy of selection in both model and variable, i.e. the ability of the approach to select the correct variables and exclude irrelevant ones. The following criteria are used to assess the performance of the different strategies within the different scenarios. For evaluation diagnostics of the different approaches, we use the precision rate ($PR$), the recall rate ($RR$), and the $F$-measure assessing the number of covariate variables (in-)correctly identified as (false) true, i.e., whether the decision to incorporate them in the model is in line with the DGP or not. A true positive ($TP$) is a correctly selected positive, in our case correctly selecting a variable which is indeed part of the assumed DGP. A true negative ($TN$) is a correctly non-selected non-important one. A false positive ($FP$) is an incorrect selection that a variable is important, when in fact, was non-important. Finally, false negative ($FN$) is an incorrect selection that a variable is non-important, when in fact is important. Based on these definitions the above-mentioned performance measurements are defined as

$$\begin{aligned} PR &= \frac{TP}{TP+FP}, \quad RR = \frac{TP}{TP+FN}, \quad \text{and} \\ F &= \frac{2 \cdot PR \cdot RR}{PR + RR} = \frac{2TP}{2TP + FP + FN}. \end{aligned} \tag{13}$$

Note that the $F$-measure is given as the weighted harmonic mean of the precision and sensitivity. The $F$-measure balances hence both and is useful if the recall $RR$ has large values, but the precision $PR$ has small ones.

In addition to the aforementioned considerations, the consistency of variable selection across different iterations in the case of $M$ imputed datasets represents a further quality aspect in shrinkage estimation approaches. But as above mentioned the calculus of counting the selected variables seems not to be a suitable quality aspect. Due to space limitations only the average estimates over the $M = 100$ datasets are reported, thn the biases are straightforward. Note that the number of estimators per shrinkage approaches varies. For instance, the ridge regression provides $M$ estimates even for redundant variables because estimates are shrunk to zero. In contrast, the Elastic net and the Lasso set the impact of variables to zero. The Bayesian spike-and-slab approach, on the other hand, provides $M$ estimates for all variables.

Generally, the root mean square error (RMSE) over the results from the $M$ datasets is an quality indicator for the robustness not only in the complete cases, but also in the handling of missing values. For a parameter $\hat{\theta}_p$ the RMSE is calculated

$$\mathrm{RMSE}(\hat{\theta}_p) = \sqrt{\mathrm{MSE}(\hat{\theta}_p)} = \sqrt{E\left[\left(\beta_p^{\text{true}} - \hat{\theta}_p\right)^2\right]} = \sqrt{\frac{1}{D}\sum_{d=1}^{D}\left(\beta_p^{\text{true}} - \hat{\theta}_p^{(d)}\right)^2}, \tag{14}$$

where MSE denotes the mean square error. In the context of ridge regression and the Bayesian approach, the value of $D$ is equal to $M$. However, in the case of the Elastic net, Lasso, and stepwise regression, the value of $D$ differs from $M$ for each variable. In these approaches, redundant variables are typically excluded from the model, and in a few datasets, important variables are attenuated, thus $D$ varies.

## 5 Simulation study

The following simulation experiments aim at a comparison of different strategies to achieve variable selection within a binary probit regression framework. The simulations were implemented in R (R Core Team. 2020) and Julia (Bezanson et al. 2017) and compare the performance of the Bayesian variable selection approach – hereafter referred to as Bayesian Spike-and-Slab (SnS) – with the stepwise regression (SR) strategy as implemented in the MASS R package (Venables and Ripley 2002)) and different variations of Elastic net (EN) like Lasso and ridge regression as available within the R package glmnet (Friedman et al. 2010). For the shrinkage estimators, we consider for the Elastic net setup different control parameter $\varphi = 0.0$ for ridge regression (EN.0), $\varphi = 0.5$ for a variation of Elastic net (EN.5), and $\varphi = 1.0$ for Lasso (EN1.). Via cross-validation $\lambda$ is chosen such that the error is within one standard error of the minimum shrinkage parameter (Friedman et al. 2010). Thereby, stepwise regression strategy depends on the choice of the selection criterion, where we opt for the Bayesian information criterion (BIC). For the Bayesian estimation approach, estimates are each based on MCMC chains of length 40,000. After discarding the first 10,000 iterations as burn-in, inference is based on the remaining 30,000 simulated draws from the joint posterior distribution. Convergence is monitored via Geweke statistics, the Gelman-Rubin statistics, and the effective sample size, see Geweke (1991); Gelman et al. (2023). The convergence diagnostics indicate overall convergence.

The simulated data is generated to follow a probit regression model with $N = 1000$ and $P = 10$. The considered data generating process satisfies the following conditions. Next to a constant, the covariate data $X_{(p)}$ with $1, \ldots, 9$ is generated from standard normal distributions in each variable. Only the first five out of the total ten parameters are set to a non-zero value by setting the indicator vector, accordingly, including the intercept. The signs of the 10 parameters are chosen to alternate. For the sake of variation, $X_1$ to $X_4$ are drawn from a multivariate normal distribution with expectation $\mu = (0, 0, 0, 0)'$ and covariance $\mathrm{vech}(\Sigma) =$

**Table 2** Overview of the missing design of the experimental studies

| Design | Missing mechanism | Total missing rate (%) | Results |
|---|---|---|---|
| Ex 1 | $Pr(X_{2,i} = \text{missing}) = 0.2$ | $35.99^1$ | Table 3 |
| | $Pr(X_{3,i} = \text{missing}) = 0.2$ | | |
| Ex 2 | $Pr(X_{2,i} = \text{missing}) = 0.2$ | $49.51^1$ | Table 4 |
| | $Pr(X_{3,i} = \text{missing}) = 0.3$ | | |
| | $Pr(X_{4,i} = \text{missing}) = 0.1$ | | |
| Ex 3 | $X_{2,i} = \text{missing if } 1/(1 + \exp(-\omega_{2,i}))$ | 25.00 | Table 5 |
| | $w_{2,i} = 0.2X_{2,i} + \rho_i$ and $\rho_i \overset{\text{i.i.d.}}{\sim} N(0, 1)$ | | |

The experimental studies Ex 1 and Ex 2 can be characterized as missing completely at random (MCAR) and experimental study Ex 3, where the missing probability depends on the variable itself as missing at random (MAR). All simulation runs have been performed with 40,000 Gibbs iterations with the first 10,000 iterations as burn-in. [1] Average over $M = 100$ datasets.

$(1, 0.85, 0.65, 0.45, 1, 0.45, 0.35, 1, 0.25, 1)'$ mimicking a situation with correlated covariate data.[16] Finally, a total of 100 simulated data sets are generated through replication.

Based on this data generating process, we consider two different variations for missingness in the covariate data: missing completely at random (MCAR) and missing at random (MAR). Simulating missing values as MCAR, we randomly set 20% of the values in the covariates $X_2$ and $X_3$ to missing later named as experimental study 1 (Ex 1) and we randomly set 20% in $X_2$, 30% in $X_3$ and 10% in $X_3$ to missing as experimental study 2 (Ex 2). For the MAR variation (Ex 3), we consider a missing generating mechanism for $X_{2,i}$ where $X_{2,i}$ is missing if $F_U(U_i) > 0.75$, where $F_U(U_i)$ denotes the empirical distribution function of the random variable $U_i$ which is

$$U_i = \frac{1}{1 + \exp\{\omega_{2,i}\}} \quad i = 1, \ldots, N,$$

with $\omega_{2,i} = 0.2X_{2,i} + \rho_i$ and $\rho_i$ being standard normally distributed. Thus, a missing rate of 25% for $X_2$ is generated. For further details on the described missing designs, see Table 2.

To assess the different estimation strategies in case of missing values, we designed a comprehensive simulation study which is split in different scenarios. First, we consider a benchmark estimation without missing values labeled as before deletion (BD), followed by estimation of complete cases (CC) only, and finally a scenario with missing values (MIS), where missing values are handled either via multiple imputation before estimation as for the shrinkage estimators or embedded within the MCMC algorithm as for the Bayesian approach.

Regarding estimation results, we provide the true parameter values used in the DGP, mean posterior medians and averaged estimates, respectively over the 100 replications obtained for the BD, CC, MIS scenarios. We report on both the regression coefficients and conditional variance parameters. Beside the averaged estimates,

[16] Note that vech(·) denotes the half-vectorization operator as defined in Lütkepohl (1996).

**Table 3** Experimental study 1: Results of MCAR with $20\%$ missing rate in $X_1$ and $X_2$, and with correlation between $X_1, \ldots, X_4$, and with true parameter values, mean posterior medians and root mean squared errors (RMSEs) of important $(\alpha, \beta_1, \ldots, \beta_4)$ and non-important $(\beta_5, \ldots, \beta_9)$ regression coefficients over 100 replications obtained by Elastic net with control parameter $\varphi = 0.0$ for ridge regression (EN.0), $\varphi = 0.5$ for a variation of Elastic net (EN.5), and $\varphi = 1.0$ for Lasso (EN1.) and stepwise regression (SR) and Bayesian spike-and-slab (SnS), where the first column presents the estimates $(\hat{\beta})$ and the second one the inclusion probabilities $(\hat{\gamma})$. The prior setting for the spike-and-slab were set to $\tau_2 = 2 \gg \tau_1 = 0.2$ and controlling the number of selecting variables with $w = 0.5$

| | Average estimates | | | | | | | RMSEs | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| True values | Probit | EN.0 | EN.5 | EN1. | SR | SnS | | Probit | EN.0 | EN.5 | EN1. | SR | SnS |
| Before Deletion | | | | | | | | | | | | | |
| $\alpha = 0.5$ | 0.513 | 0.501 | 0.509 | 0.508 | 0.515 | 0.519 | 1.000 | 0.046 | 0.043 | 0.045 | 0.045 | 0.047 | 0.048 |
| $\beta_1 = -0.5$ | –0.514 | –0.362 | –0.434 | –0.430 | –0.516 | –0.511 | 0.993 | 0.108 | 0.156 | 0.130 | 0.129 | 0.108 | 0.106 |
| $\beta_2 = 0.5$ | 0.518 | 0.384 | 0.449 | 0.447 | 0.520 | 0.518 | 0.999 | 0.096 | 0.135 | 0.109 | 0.106 | 0.096 | 0.096 |
| $\beta_3 = -0.5$ | –0.505 | –0.496 | –0.507 | –0.508 | –0.507 | –0.511 | 1.000 | 0.058 | 0.047 | 0.056 | 0.057 | 0.058 | 0.059 |
| $\beta_4 = 0.5$ | 0.513 | 0.459 | 0.489 | 0.488 | 0.514 | 0.517 | 1.000 | 0.055 | 0.062 | 0.056 | 0.057 | 0.056 | 0.058 |
| $\beta_5 = 0.0$ | – | –0.001 | 0.000 | 0.000 | 0.000 | 0.009 | 0.060 | – | 0.041 | 0.040 | 0.038 | 0.088 | 0.045 |
| $\beta_6 = 0.0$ | – | –0.003 | –0.003 | –0.003 | –0.012 | 0.007 | 0.058 | – | 0.041 | 0.040 | 0.038 | 0.085 | 0.045 |
| $\beta_7 = 0.0$ | – | 0.000 | –0.001 | 0.000 | 0.015 | –0.010 | 0.070 | – | 0.045 | 0.044 | 0.042 | 0.091 | 0.049 |
| $\beta_8 = 0.0$ | – | 0.008 | 0.007 | 0.007 | 0.041 | 0.019 | 0.066 | – | 0.043 | 0.041 | 0.039 | 0.089 | 0.048 |
| $\beta_9 = 0.0$ | – | –0.002 | –0.001 | –0.001 | 0.002 | 0.008 | 0.070 | – | 0.046 | 0.044 | 0.043 | 0.083 | 0.050 |
| Complete case | | | | | | | | | | | | | |
| $\alpha = 0.5$ | 0.513 | 0.499 | 0.505 | 0.505 | 0.515 | 0.522 | 1.000 | 0.063 | 0.060 | 0.062 | 0.062 | 0.047 | 0.064 |
| $\beta_1 = -0.5$ | –0.514 | –0.359 | –0.399 | –0.399 | –0.516 | –0.514 | 0.975 | 0.137 | 0.169 | 0.173 | 0.179 | 0.108 | 0.142 |
| $\beta_2 = 0.5$ | 0.523 | 0.387 | 0.425 | 0.425 | 0.520 | 0.526 | 0.992 | 0.128 | 0.145 | 0.148 | 0.153 | 0.096 | 0.131 |
| $\beta_3 = -0.5$ | –0.512 | –0.503 | –0.514 | –0.515 | –0.507 | –0.521 | 1.000 | 0.079 | 0.064 | 0.075 | 0.077 | 0.058 | 0.082 |
| $\beta_4 = 0.5$ | 0.512 | 0.458 | 0.477 | 0.479 | 0.514 | 0.519 | 1.000 | 0.066 | 0.070 | 0.051 | 0.066 | 0.056 | 0.069 |
| $\beta_5 = 0.0$ | – | –0.002 | –0.002 | –0.001 | 0.000 | 0.012 | 0.162 | – | 0.054 | 0.051 | 0.049 | 0.088 | 0.059 |
| $\beta_6 = 0.0$ | – | 0.002 | 0.002 | 0.003 | –0.012 | 0.015 | 0.168 | – | 0.053 | 0.050 | 0.048 | 0.085 | 0.059 |
| $\beta_7 = 0.0$ | – | –0.003 | –0.001 | –0.001 | 0.015 | 0.011 | 0.164 | – | 0.053 | 0.050 | 0.049 | 0.091 | 0.058 |
| $\beta_8 = 0.0$ | – | 0.006 | 0.005 | 0.006 | 0.041 | 0.021 | 0.160 | – | 0.050 | 0.047 | 0.045 | 0.089 | 0.056 |
| $\beta_9 = 0.0$ | – | –0.004 | –0.004 | –0.004 | 0.002 | –0.008 | 0.175 | – | 0.056 | 0.052 | 0.050 | 0.083 | 0.060 |

**Table 3** (Continued)

| | Average estimates | | | | | | | RMSEs | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| True values | Probit | EN.0 | EN.5 | EN1. | SR | SnS | | Probit | EN.0 | EN.5 | EN1. | SR | SnS |
| Imputation – average method[a] | | | | | | | | | | | | | |
| $\alpha = 0.5$ | 0.514 | 0.501 | 0.509 | 0.509 | 0.515 | 0.511 | 1.000 | 0.048 | 0.044 | 0.046 | 0.046 | 0.048 | 0.045 |
| $\beta_1 = -0.5$ | –0.512 | –0.361 | –0.432 | –0.431 | –0.516 | –0.344 | 0.999 | 0.135 | 0.166 | 0.155 | 0.154 | 0.135 | 0.187 |
| $\beta_2 = 0.5$ | 0.515 | 0.384 | 0.447 | 0.447 | 0.518 | 0.378 | 1.000 | 0.124 | 0.146 | 0.133 | 0.132 | 0.124 | 0.153 |
| $\beta_3 = -0.5$ | –0.505 | –0.496 | –0.506 | –0.508 | –0.507 | –0.546 | 1.000 | 0.061 | 0.049 | 0.058 | 0.060 | 0.061 | 0.078 |
| $\beta_4 = 0.5$ | 0.512 | 0.459 | 0.488 | 0.488 | 0.513 | 0.491 | 1.000 | 0.058 | 0.064 | 0.059 | 0.058 | 0.059 | 0.053 |
| $\beta_5 = 0.0$ | – | –0.001 | –0.002 | –0.002 | 0.004 | 0.002 | 0.175 | – | 0.042 | 0.040 | 0.039 | 0.076 | 0.044 |
| $\beta_6 = 0.0$ | – | –0.003 | –0.004 | –0.004 | –0.027 | 0.000 | 0.174 | – | 0.041 | 0.040 | 0.039 | 0.085 | 0.044 |
| $\beta_7 = 0.0$ | – | 0.000 | –0.001 | 0.000 | 0.032 | 0.003 | 0.170 | – | 0.045 | 0.043 | 0.042 | 0.085 | 0.048 |
| $\beta_8 = 0.0$ | – | 0.008 | 0.007 | 0.007 | 0.014 | 0.011 | 0.170 | – | 0.043 | 0.042 | 0.040 | 0.076 | 0.047 |
| $\beta_9 = 0.0$ | – | –0.002 | –0.001 | –0.001 | –0.009 | 0.001 | 0.184 | – | 0.046 | 0.044 | 0.042 | 0.067 | 0.050 |
| Imputation – majority method | | | | | | | | | | | | | |
| $\alpha = 0.5$ | – | 0.517 | 0.517 | 0.517 | 0.515 | – | – | – | 0.049 | 0.049 | 0.049 | 0.048 | – |
| $\beta_1 = -0.5$ | – | –0.516 | –0.522 | –0.526 | –0.519 | – | – | – | 0.138 | 0.134 | 0.133 | 0.132 | – |
| $\beta_2 = 0.5$ | – | 0.520 | 0.517 | 0.513 | 0.514 | – | – | – | 0.125 | 0.130 | 0.139 | 0.128 | – |
| $\beta_3 = -0.5$ | – | –0.508 | –0.509 | –0.511 | –0.508 | – | – | – | 0.061 | 0.065 | 0.068 | 0.064 | – |
| $\beta_4 = 0.5$ | – | 0.515 | 0.514 | 0.513 | 0.512 | – | – | – | 0.059 | 0.060 | 0.060 | 0.059 | – |
| $\beta_5 = 0.0$ | – | –0.001 | –0.001 | –0.002 | –0.023 | – | – | – | 0.045 | 0.051 | 0.055 | 0.091 | – |
| $\beta_6 = 0.0$ | – | –0.004 | –0.006 | –0.008 | –0.023 | – | – | – | 0.045 | 0.052 | 0.055 | 0.089 | – |
| $\beta_7 = 0.0$ | – | 0.000 | –0.001 | –0.001 | 0.021 | – | – | – | 0.048 | 0.054 | 0.057 | 0.096 | – |
| $\beta_8 = 0.0$ | – | 0.009 | 0.011 | 0.011 | 0.050 | – | – | – | 0.047 | 0.054 | 0.055 | 0.092 | – |
| $\beta_9 = 0.0$ | – | –0.002 | –0.002 | –0.002 | 0.003 | – | – | – | 0.050 | 0.054 | 0.056 | 0.092 | – |

**Table 3** (Continued)

| | Average estimates | | | | | | | RMSEs | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| True values | Probit | EN.0 | EN.5 | EN1. | SR | SnS | | Probit | EN.0 | EN.5 | EN1. | SR | SnS |
| Imputation – Wald method | | | | | | | | | | | | | |
| $\alpha = 0.5$ | – | 0.504 | 0.505 | 0.505 | 0.515 | – | – | – | 0.050 | 0.051 | 0.051 | 0.048 | – |
| $\beta_1 = -0.5$ | – | –0.359 | –0.412 | –0.412 | –0.516 | – | – | – | 0.168 | 0.168 | 0.168 | 0.130 | – |
| $\beta_2 = 0.5$ | – | 0.388 | 0.480 | 0.480 | 0.515 | – | – | – | 0.145 | 0.143 | 0.142 | 0.128 | – |
| $\beta_3 = -0.5$ | – | –0.503 | –0.538 | –0.543 | –0.509 | – | – | – | 0.060 | 0.055 | 0.055 | 0.065 | – |
| $\beta_4 = 0.5$ | – | 0.458 | 0.458 | 0.460 | 0.513 | – | – | – | 0.069 | 0.065 | 0.065 | 0.059 | – |
| $\beta_5 = 0.0$ | – | –0.002 | –0.001 | –0.001 | –0.015 | – | – | – | 0.043 | 0.043 | 0.043 | 0.091 | – |
| $\beta_6 = 0.0$ | – | –0.002 | –0.004 | –0.004 | –0.027 | – | – | – | 0.042 | 0.042 | 0.042 | 0.089 | – |
| $\beta_7 = 0.0$ | – | –0.003 | –0.001 | –0.001 | 0.016 | – | – | – | 0.045 | 0.044 | 0.043 | 0.090 | – |
| $\beta_8 = 0.0$ | – | 0.006 | 0.007 | 0.007 | 0.043 | – | – | – | 0.043 | 0.044 | 0.044 | 0.090 | – |
| $\beta_9 = 0.0$ | – | –0.004 | 0.000 | 0.000 | –0.007 | – | – | – | 0.045 | 0.042 | 0.038 | 0.080 | – |

[a] The imputation results are presented for the spike-and-slab approach from the Gibbs sampler and for the probit from multiple imputation only once to avoid redundancy.

**Table 4** Experimental study 2: Results of MCAR with missings in $X_1$ (20%), $X_2$ (30%), and $X_3$ (10%), and with correlation between $X_1, \ldots, X_4$, and with true parameter values, mean posterior medians and root mean squared errors (RMSEs) of important $(\alpha, \beta_1, \ldots, \beta_4)$ and non-important $(\beta_5, \ldots, \beta_9)$ regression coefficients over 100 replications obtained by Elastic net with control parameter $\varphi = \mathbf{0.0}$ for ridge regression (EN.0), $\varphi = \mathbf{0.5}$ for a variation of Elastic net (EN.5), and $\varphi = \mathbf{1.0}$ for Lasso (EN1.) and stepwise regression (SR) and Bayesian spike-and-slab (SnS), where the first column presents the estimates $(\hat{\beta})$ and the second one the inclusion probabilities $(\hat{\gamma})$. The prior setting for the spike-and-slab were set to $\tau_2 = 2 \gg \tau_1 = \mathbf{0.2}$ and controlling the number of selecting variables with $w = \mathbf{0.5}$

| | Average estimates | | | | | | | RMSEs | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| True values | Probit | EN.0 | EN.5 | EN1. | SR | SnS | | Probit | EN.0 | EN.5 | EN1. | SR | SnS |
| Before Deletion | | | | | | | | | | | | | |
| $\alpha = \mathbf{0.5}$ | 0.507 | 0.495 | 0.503 | 0.503 | 0.509 | 0.513 | 1.000 | 0.046 | 0.043 | 0.044 | 0.044 | 0.046 | 0.047 |
| $\beta_1 = \mathbf{-0.5}$ | –0.484 | –0.365 | –0.440 | –0.436 | –0.520 | –0.516 | 0.992 | 0.119 | 0.152 | 0.130 | 0.131 | 0.107 | 0.106 |
| $\beta_2 = \mathbf{0.5}$ | 0.494 | 0.378 | 0.446 | 0.518 | 0.513 | 0.511 | 0.997 | 0.095 | 0.141 | 0.119 | 0.120 | 0.100 | 0.100 |
| $\beta_3 = \mathbf{-0.5}$ | –0.511 | –0.493 | –0.504 | –0.470 | –0.504 | –0.507 | 1.000 | 0.063 | 0.056 | 0.066 | 0.067 | 0.069 | 0.069 |
| $\beta_4 = \mathbf{0.5}$ | 0.499 | 0.450 | 0.480 | 0.437 | 0.505 | 0.508 | 1.000 | 0.051 | 0.069 | 0.058 | 0.058 | 0.055 | 0.056 |
| $\beta_5 = \mathbf{0.0}$ | – | 0.002 | 0.002 | 0.002 | 0.012 | 0.012 | 0.069 | – | 0.044 | 0.043 | 0.041 | 0.096 | 0.049 |
| $\beta_6 = \mathbf{0.0}$ | – | –0.001 | –0.002 | –0.002 | –0.015 | 0.008 | 0.065 | – | 0.044 | 0.042 | 0.040 | 0.088 | 0.047 |
| $\beta_7 = \mathbf{0.0}$ | – | 0.006 | 0.006 | 0.005 | 0.036 | –0.016 | 0.059 | – | 0.040 | 0.038 | 0.037 | 0.081 | 0.045 |
| $\beta_8 = \mathbf{0.0}$ | – | 0.000 | 0.000 | 0.000 | 0.016 | –0.010 | 0.062 | – | 0.042 | 0.040 | 0.039 | 0.087 | 0.046 |
| $\beta_9 = \mathbf{0.0}$ | – | 0.003 | 0.003 | 0.004 | 0.003 | 0.013 | 0.084 | – | 0.050 | 0.049 | 0.047 | 0.091 | 0.055 |
| Complete case | | | | | | | | | | | | | |
| $\alpha = \mathbf{0.5}$ | 0.501 | 0.490 | 0.493 | 0.492 | 0.503 | 0.511 | 1.000 | 0.062 | 0.062 | 0.063 | 0.064 | 0.064 | 0.066 |
| $\beta_1 = \mathbf{-0.5}$ | –0.472 | –0.329 | –0.338 | –0.329 | –0.495 | –0.470 | 0.929 | 0.172 | 0.208 | 0.238 | 0.260 | 0.148 | 0.179 |
| $\beta_2 = \mathbf{0.5}$ | 0.477 | 0.350 | 0.361 | 0.352 | 0.477 | 0.479 | 0.971 | 0.140 | 0.183 | 0.202 | 0.226 | 0.143 | 0.141 |
| $\beta_3 = \mathbf{-0.5}$ | –0.503 | –0.491 | –0.502 | –0.503 | –0.509 | –0.517 | 0.998 | 0.086 | 0.071 | 0.080 | 0.084 | 0.094 | 0.093 |
| $\beta_4 = \mathbf{0.5}$ | 0.493 | 0.440 | 0.449 | 0.446 | 0.493 | 0.500 | 1.000 | 0.069 | 0.086 | 0.087 | 0.093 | 0.071 | 0.073 |
| $\beta_5 = \mathbf{0.0}$ | – | –0.004 | –0.004 | –0.003 | 0.011 | 0.012 | 0.204 | – | 0.057 | 0.052 | 0.050 | 0.121 | 0.063 |
| $\beta_6 = \mathbf{0.0}$ | – | 0.001 | 0.003 | 0.003 | 0.015 | 0.017 | 0.242 | – | 0.067 | 0.061 | 0.059 | 0.123 | 0.073 |
| $\beta_7 = \mathbf{0.0}$ | – | 0.009 | 0.009 | 0.008 | 0.056 | –0.027 | 0.215 | – | 0.057 | 0.051 | 0.049 | 0.111 | 0.065 |
| $\beta_8 = \mathbf{0.0}$ | – | 0.004 | 0.004 | 0.004 | 0.032 | –0.021 | 0.226 | – | 0.063 | 0.058 | 0.057 | 0.126 | 0.070 |
| $\beta_9 = \mathbf{0.0}$ | – | –0.010 | –0.010 | –0.010 | –0.049 | 0.006 | 0.211 | – | 0.059 | 0.053 | 0.050 | 0.111 | 0.064 |

**Table 4** (Continued)

| | Average estimates | | | | | | | RMSEs | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| True values | Probit | EN.0 | EN.5 | EN1. | SR | SnS | | Probit | EN.0 | EN.5 | EN1. | SR | SnS |
| Imputation – average method[a] | | | | | | | | | | | | | |
| $\alpha = 0.5$ | 0.500 | 0.488 | 0.495 | 0.495 | 0.502 | 0.492 | 1.000 | 0.047 | 0.047 | 0.047 | 0.047 | 0.048 | 0.055 |
| $\beta_1 = -0.5$ | –0.468 | –0.333 | –0.389 | –0.388 | –0.485 | –0.391 | 0.964 | 0.156 | 0.198 | 0.194 | 0.196 | 0.136 | 0.134 |
| $\beta_2 = 0.5$ | 0.480 | 0.360 | 0.412 | 0.411 | 0.483 | 0.431 | 0.995 | 0.128 | 0.167 | 0.160 | 0.160 | 0.129 | 0.089 |
| $\beta_3 = -0.5$ | –0.514 | –0.502 | –0.514 | –0.515 | –0.516 | –0.522 | 1.000 | 0.077 | 0.063 | 0.074 | 0.076 | 0.080 | 0.064 |
| $\beta_4 = 0.5$ | 0.497 | 0.447 | 0.471 | 0.472 | 0.498 | 0.471 | 1.000 | 0.054 | 0.072 | 0.063 | 0.062 | 0.055 | 0.050 |
| $\beta_5 = 0.0$ | – | –0.005 | –0.005 | –0.004 | –0.002 | 0.011 | 0.089 | – | 0.044 | 0.041 | 0.040 | 0.081 | 0.060 |
| $\beta_6 = 0.0$ | – | –0.001 | 0.000 | 0.000 | 0.040 | 0.013 | 0.083 | – | 0.041 | 0.038 | 0.037 | 0.074 | 0.046 |
| $\beta_7 = 0.0$ | – | 0.007 | 0.006 | 0.005 | 0.046 | 0.029 | 0.080 | – | 0.039 | 0.036 | 0.034 | 0.073 | 0.072 |
| $\beta_8 = 0.0$ | – | 0.007 | 0.007 | 0.007 | 0.020 | 0.024 | 0.101 | – | 0.045 | 0.044 | 0.042 | 0.083 | 0.060 |
| $\beta_9 = 0.0$ | – | –0.005 | –0.005 | –0.004 | 0.003 | 0.001 | 0.054 | – | 0.039 | 0.037 | 0.035 | 0.064 | 0.028 |
| Imputation – majority method | | | | | | | | | | | | | |
| $\alpha = 0.5$ | – | 0.503 | 0.502 | 0.502 | 0.501 | – | – | – | 0.048 | 0.048 | 0.048 | 0.048 | – |
| $\beta_1 = -0.5$ | – | –0.474 | –0.505 | –0.513 | –0.496 | – | – | – | 0.157 | 0.127 | 0.123 | 0.130 | – |
| $\beta_2 = 0.5$ | – | 0.485 | 0.473 | 0.467 | 0.476 | – | – | – | 0.128 | 0.151 | 0.160 | 0.227 | – |
| $\beta_3 = -0.5$ | – | –0.517 | –0.522 | –0.525 | –0.519 | – | – | – | 0.080 | 0.089 | 0.092 | 0.060 | – |
| $\beta_4 = 0.5$ | – | 0.500 | 0.497 | 0.496 | 0.496 | – | – | – | 0.055 | 0.058 | 0.059 | 0.065 | – |
| $\beta_5 = 0.0$ | – | –0.005 | –0.008 | –0.009 | –0.017 | – | – | – | 0.048 | 0.055 | 0.056 | 0.082 | – |
| $\beta_6 = 0.0$ | – | 0.000 | 0.000 | 0.001 | 0.004 | – | – | – | 0.044 | 0.051 | 0.054 | 0.078 | – |
| $\beta_7 = 0.0$ | – | 0.008 | 0.009 | 0.010 | 0.062 | – | – | – | 0.042 | 0.048 | 0.051 | 0.080 | – |
| $\beta_8 = 0.0$ | – | 0.008 | 0.010 | 0.011 | 0.035 | – | – | – | 0.050 | 0.057 | 0.060 | 0.086 | – |
| $\beta_9 = 0.0$ | – | –0.006 | –0.007 | –0.008 | –0.004 | – | – | – | 0.043 | 0.051 | 0.053 | 0.074 | – |

**Table 4** (Continued)

| | Average estimates | | | | | | | RMSEs | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| True values | Probit | EN.0 | EN.5 | EN1. | SR | SnS | | Probit | EN.0 | EN.5 | EN1. | SR | SnS |
| Imputation – Wald method | | | | | | | | | | | | | |
| $\alpha = 0.5$ | – | 0.504 | 0.505 | 0.505 | 0.515 | – | – | – | 0.049 | 0.051 | 0.052 | 0.050 | – |
| $\beta_1 = -0.5$ | – | –0.340 | –0.385 | –0.386 | –0.496 | – | – | – | 0.195 | 0.193 | 0.193 | 0.130 | – |
| $\beta_2 = 0.5$ | – | 0.370 | 0.420 | 0.415 | 0.481 | – | – | – | 0.168 | 0.165 | 0.165 | 0.128 | – |
| $\beta_3 = -0.5$ | – | –0.503 | –0.512 | –0.512 | –0.510 | – | – | – | 0.060 | 0.060 | 0.059 | 0.065 | – |
| $\beta_4 = 0.5$ | – | 0.458 | 0.466 | 0.466 | 0.499 | – | – | – | 0.071 | 0.068 | 0.068 | 0.059 | – |
| $\beta_5 = 0.0$ | – | –0.003 | –0.002 | –0.002 | –0.015 | – | – | – | 0.043 | 0.043 | 0.043 | 0.091 | – |
| $\beta_6 = 0.0$ | – | –0.001 | –0.002 | –0.002 | –0.007 | – | – | – | 0.040 | 0.039 | 0.038 | 0.089 | – |
| $\beta_7 = 0.0$ | – | –0.003 | –0.001 | –0.001 | 0.006 | – | – | – | 0.044 | 0.045 | 0.043 | 0.090 | – |
| $\beta_8 = 0.0$ | – | 0.006 | 0.007 | 0.007 | 0.023 | – | – | – | 0.042 | 0.042 | 0.041 | 0.090 | – |
| $\beta_9 = 0.0$ | – | –0.004 | –0.003 | –0.003 | 0.009 | – | – | – | 0.045 | 0.042 | 0.038 | 0.080 | – |

[a] The imputation results are presented for the spike-and-slab approach from the Gibbs sampler and for the probit from multiple imputation only once to avoid redundancy.

**Table 5** Experimental study 3: Average method results of MAR with missings in $X_1$ (20%), and with correlation between $X_1, \ldots, X_4$, and with true parameter values, mean posterior medians and root mean squared errors (RMSEs) of important $(\alpha, \beta_1, \ldots, \beta_4)$ and non-important $(\beta_5, \ldots, \beta_9)$ regression coefficients over 100 replications obtained by Elastic net with control parameter $\varphi = \mathbf{0.0}$ for ridge regression (EN.0), $\varphi = \mathbf{0.5}$ for a variation of Elastic net (EN.5), and $\varphi = \mathbf{1.0}$ for Lasso (EN1.) and stepwise regression (SR) and Bayesian spike-and-slab (SnS), where the first column presents the estimates $(\hat{\beta})$ and the second one the inclusion probabilities $(\hat{\gamma})$. The prior setting for the spike-and-slab were set to $\tau_2 = \mathbf{2} \gg \tau_1 = \mathbf{0.2}$ and controlling the number of selecting variables with $w = \mathbf{0.5}$

| | Average estimates | | | | | | | RMSEs | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| True values | Probit | EN.0 | EN.5 | EN1. | SR | SnS | | Probit | EN.0 | EN.5 | EN1. | SR | SnS |
| Before Deletion | | | | | | | | | | | | | |
| $\alpha = \mathbf{0.5}$ | 0.507 | 0.495 | 0.502 | 0.503 | 0.509 | 0.506 | 1.000 | 0.048 | 0.047 | 0.047 | 0.048 | 0.048 | 0.046 |
| $\beta_1 = \mathbf{-0.5}$ | –0.493 | –0.349 | –0.412 | –0.412 | –0.495 | –0.481 | 0.981 | 0.110 | 0.168 | 0.139 | 0.140 | 0.110 | 0.121 |
| $\beta_2 = \mathbf{0.5}$ | 0.488 | 0.363 | 0.419 | 0.420 | 0.490 | 0.494 | 0.997 | 0.086 | 0.150 | 0.116 | 0.116 | 0.086 | 0.096 |
| $\beta_3 = \mathbf{-0.5}$ | –0.504 | –0.495 | –0.506 | –0.508 | –0.506 | –0.516 | 1.000 | 0.058 | 0.047 | 0.056 | 0.057 | 0.057 | 0.065 |
| $\beta_4 = \mathbf{0.5}$ | 0.501 | 0.450 | 0.477 | 0.478 | 0.503 | 0.504 | 1.000 | 0.052 | 0.068 | 0.057 | 0.058 | 0.053 | 0.053 |
| $\beta_5 = \mathbf{0.0}$ | – | –0.001 | 0.003 | –0.001 | –0.032 | 0.003 | 0.062 | – | 0.040 | 0.038 | 0.037 | 0.082 | 0.046 |
| $\beta_6 = \mathbf{0.0}$ | – | 0.000 | –0.003 | 0.000 | 0.004 | 0.009 | 0.057 | – | 0.045 | 0.043 | 0.042 | 0.089 | 0.045 |
| $\beta_7 = \mathbf{0.0}$ | – | 0.001 | –0.009 | 0.002 | –0.001 | 0.016 | 0.055 | – | 0.041 | 0.039 | 0.038 | 0.080 | 0.044 |
| $\beta_8 = \mathbf{0.0}$ | – | –0.002 | –0.002 | –0.002 | –0.003 | 0.017 | 0.074 | – | 0.043 | 0.041 | 0.040 | 0.081 | 0.051 |
| $\beta_9 = \mathbf{0.0}$ | – | 0.002 | –0.005 | 0.001 | 0.010 | 0.003 | 0.050 | – | 0.041 | 0.039 | 0.038 | 0.080 | 0.041 |
| Complete case | | | | | | | | | | | | | |
| $\alpha = \mathbf{0.5}$ | 0.512 | 0.499 | 0.506 | 0.506 | 0.514 | 0. 520 | 1.000 | 0.055 | 0.053 | 0.055 | 0.055 | 0.056 | 0.058 |
| $\beta_1 = \mathbf{-0.5}$ | –0.501 | –0.354 | –0.406 | –0.407 | –0.504 | –0.503 | 0.984 | 0.134 | 0.171 | 0.164 | 0.161 | 0.133 | 0.134 |
| $\beta_2 = \mathbf{0.5}$ | 0.495 | 0.367 | 0.415 | 0.415 | 0.498 | 0.499 | 0.995 | 0.112 | 0.154 | 0.145 | 0.139 | 0.112 | 0.111 |
| $\beta_3 = \mathbf{-0.5}$ | –0.509 | –0.499 | –0.512 | –0.513 | –0.512 | –0.517 | 1.000 | 0.075 | 0.061 | 0.072 | 0.074 | 0.076 | 0.077 |
| $\beta_4 = \mathbf{0.5}$ | 0.505 | 0.453 | 0.476 | 0.478 | 0.507 | 0.513 | 1.000 | 0.059 | 0.070 | 0.064 | 0.063 | 0.060 | 0.061 |
| $\beta_5 = \mathbf{-0.5}$ | – | 0.003 | 0.003 | 0.002 | 0.004 | 0.014 | 0.126 | – | 0.045 | 0.043 | 0.041 | 0.094 | 0.050 |
| $\beta_6 = \mathbf{0.5}$ | – | 0.003 | 0.002 | 0.002 | 0.014 | 0.014 | 0.145 | – | 0.051 | 0.049 | 0.049 | 0.101 | 0.057 |
| $\beta_7 = \mathbf{0.5}$ | – | –0.001 | –0.001 | –0.000 | 0.014 | 0.010 | 0.127 | – | 0.047 | 0.044 | 0.043 | 0.089 | 0.051 |
| $\beta_8 = \mathbf{0.5}$ | – | –0.006 | –0.007 | –0.006 | –0.017 | 0.006 | 0.137 | – | 0.050 | 0.048 | 0.047 | 0.094 | 0.054 |
| $\beta_9 = \mathbf{0.5}$ | – | 0.002 | 0.003 | 0.002 | 0.013 | 0.014 | 0.148 | – | 0.051 | 0.049 | 0.047 | 0.099 | 0.057 |

**Table 5** (Continued)

| | Average estimates | | | | | | | RMSEs | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| True values | Probit | EN.0 | EN.5 | EN1. | SR | SnS | | Probit | EN.0 | EN.5 | EN1. | SR | SnS |
| Imputation – average method[a] | | | | | | | | | | | | | |
| $\alpha = 0.5$ | 0.515 | 0.501 | 0.510 | 0.510 | 0.517 | 0.520 | 1.000 | 0.050 | 0.046 | 0.048 | 0.048 | 0.050 | 0.052 |
| $\beta_1 = -0.5$ | –0.498 | –0.354 | –0.419 | –0.417 | –0.501 | –0.408 | 0.958 | 0.137 | 0.173 | 0.156 | 0.156 | 0.138 | 0.153 |
| $\beta_2 = 0.5$ | 0.490 | 0.364 | 0.422 | 0.421 | 0.492 | 0.435 | 0.991 | 0.103 | 0.153 | 0.127 | 0.127 | 0.103 | 0.116 |
| $\beta_3 = -0.5$ | –0.504 | –0.494 | –0.505 | –0.507 | –0.505 | –0.545 | 1.000 | 0.062 | 0.049 | 0.059 | 0.060 | 0.061 | 0.073 |
| $\beta_4 = 0.5$ | 0.502 | 0.450 | 0.478 | 0.478 | 0.503 | 0.477 | 1.000 | 0.054 | 0.069 | 0.058 | 0.058 | 0.055 | 0.057 |
| $\beta_5 = -0.5$ | – | –0.001 | –0.001 | –0.001 | –0.039 | 0.008 | 0.085 | – | 0.040 | 0.038 | 0.037 | 0.082 | 0.043 |
| $\beta_6 = 0.5$ | – | 0.000 | 0.000 | 0.001 | –0.017 | 0.009 | 0.101 | – | 0.045 | 0.044 | 0.042 | 0.084 | 0.049 |
| $\beta_7 = 0.5$ | – | 0.001 | 0.001 | 0.001 | 0.010 | 0.010 | 0.089 | – | 0.042 | 0.040 | 0.038 | 0.069 | 0.046 |
| $\beta_8 = 0.5$ | – | –0.002 | –0.003 | –0.002 | –0.030 | –0.006 | 0.092 | – | 0.043 | 0.041 | 0.040 | 0.084 | 0.047 |
| $\beta_9 = 0.5$ | – | 0.002 | 0.001 | 0.001 | 0.011 | 0.011 | 0.088 | – | 0.042 | 0.039 | 0.038 | 0.069 | 0.046 |
| Imputation – majority method | | | | | | | | | | | | | |
| $\alpha = 0.5$ | – | 0.518 | 0.518 | 0.518 | 0.517 | – | - | – | 0.051 | 0.051 | 0.051 | 0.050 | - |
| $\beta_1 = -0.5$ | – | –0.503 | –0.503 | –0.505 | –0.501 | – | - | – | 0.137 | 0.137 | 0.137 | 0.137 | - |
| $\beta_2 = 0.5$ | – | 0.493 | 0.493 | 0.488 | 0.492 | – | - | – | 0.103 | 0.103 | 0.116 | 0.103 | - |
| $\beta_3 = -0.5$ | – | –0.507 | –0.507 | –0.509 | –0.505 | – | - | – | 0.061 | 0.061 | 0.063 | 0.061 | - |
| $\beta_4 = 0.5$ | – | 0.505 | 0.505 | 0.503 | 0.503 | – | - | – | 0.055 | 0.055 | 0.056 | 0.055 | - |
| $\beta_5 = -0.5$ | – | –0.001 | 0.000 | 0.000 | –0.043 | – | - | – | 0.044 | 0.049 | 0.050 | 0.085 | - |
| $\beta_6 = 0.5$ | – | 0.000 | 0.000 | 0.000 | 0.002 | – | - | – | 0.049 | 0.054 | 0.056 | 0.092 | - |
| $\beta_7 = 0.5$ | – | 0.001 | –0.002 | 0.003 | 0.004 | – | - | – | 0.044 | 0.050 | 0.053 | 0.089 | - |
| $\beta_8 = 0.5$ | – | –0.003 | –0.003 | –0.004 | –0.020 | – | - | – | 0.047 | 0.055 | 0.057 | 0.087 | - |
| $\beta_9 = 0.5$ | – | 0.002 | 0.001 | 0.001 | 0.024 | – | - | – | 0.045 | 0.050 | 0.051 | 0.083 | - |

**Table 5** (Continued)

| | Average estimates | | | | | | | RMSEs | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| True values | Probit | EN.0 | EN.5 | EN1. | SR | SnS | | Probit | EN.0 | EN.5 | EN1. | SR | SnS |
| Imputation – Wald method | | | | | | | | | | | | | |
| $\alpha = 0.5$ | – | 0.505 | 0.510 | 0.510 | 0.584 | – | – | – | 0.046 | 0.048 | 0.048 | 0.070 | - |
| $\beta_1 = -0.5$ | – | –0.237 | –0.406 | –0.405 | –0.490 | – | – | – | 0.274 | 0.167 | 0.167 | 0.125 | - |
| $\beta_2 = 0.5$ | – | 0.361 | 0.355 | 0.358 | 0.349 | – | – | – | 0.149 | 0.158 | 0.159 | 0.164 | - |
| $\beta_3 = -0.5$ | – | –0.421 | –0.472 | –0.472 | –0.494 | – | – | – | 0.090 | 0.061 | 0.063 | 0.053 | - |
| $\beta_4 = 0.5$ | – | 0.352 | 0.439 | 0.438 | 0.478 | – | – | – | 0.158 | 0.099 | 0.099 | 0.077 | - |
| $\beta_5 = 0.5$ | – | 0.002 | 0.002 | 0.001 | 0.028 | – | – | – | 0.039 | 0.039 | 0.039 | 0.079 | - |
| $\beta_6 = 0.5$ | – | –0.002 | –0.002 | –0.002 | –0.031 | – | – | – | 0.038 | 0.037 | 0.037 | 0.083 | - |
| $\beta_7 = 0.5$ | – | 0.009 | 0.010 | –0.009 | –0.031 | – | – | – | 0.040 | 0.040 | 0.039 | 0.076 | - |
| $\beta_8 = 0.5$ | – | 0.002 | 0.002 | –0.002 | –0.022 | – | – | – | 0.037 | 0.037 | 0.036 | 0.083 | - |
| $\beta_9 = 0.5$ | – | –0.005 | –0.005 | –0.005 | –0.018 | – | – | – | 0.037 | 0.037 | 0.036 | 0.083 | - |

[a] The imputation results are presented for the spike-and-slab approach from the Gibbs sampler and for the probit from multiple imputation only once to avoid redundancy.

**Table 6** For experimental study 1 (Ex1), experimental study 2 (Ex2), and experimental study 3 (Ex3) comparison of precision, recall, and F-measure obtained by Elastic net with control parameter $\varphi = \mathbf{0.0}$ for ridge regression (EN.0), $\varphi = \mathbf{0.5}$ for a variation of Elastic net (EN.5), and $\varphi = \mathbf{1.0}$ for Lasso (EN1.) and stepwise regression (SR) and Bayesian spike-and-slab (SnS) over 100 replications

| | Average precision | | | | | Average recall | | | | | Average F-measure | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | EN0.0 | EN0.5 | EN1.0 | SR | SnS | EN0.0 | EN0.5 | EN1.0 | SR | SnS | EN0.0 | EN0.5 | EN1.0 | SR | SnS |
| *Experimental study 1* | | | | | | | | | | | | | | | |
| BD | 1.00 | 0.99 | 0.98 | 1.00 | 1.00 | 0.50 | 0.53 | 0.54 | 0.89 | 0.98 | 0.67 | 0.68 | 0.69 | 0.93 | 0.99 |
| CC | 1.00 | 0.95 | 0.93 | 0.98 | 1.00 | 0.50 | 0.54 | 0.57 | 0.87 | 0.95 | 0.67 | 0.68 | 0.69 | 0.91 | 0.97 |
| IMP-Average | 1.00 | 0.99 | 0.99 | 0.95 | 1.00 | 0.50 | 0.54 | 0.55 | 0.88 | 0.99 | 0.67 | 0.70 | 0.70 | 0.93 | 1.00 |
| IMP-Majority | 1.00 | 0.98 | 0.99 | 1.00 | 1.00 | 0.50 | 0.52 | 0.52 | 0.89 | 0.99 | 0.67 | 0.68 | 0.69 | 0.94 | 1.00 |
| IMP-Wald | 1.00 | 0.97 | 0.98 | 0.98 | 1.00 | 0.50 | 0.55 | 0.54 | 0.89 | 0.99 | 0.67 | 0.70 | 0.70 | 0.93 | 1.00 |
| *Experimental study 2* | | | | | | | | | | | | | | | |
| BD | 1.00 | 1.00 | 0.99 | 1.00 | 1.00 | 0.50 | 0.52 | 0.54 | 0.88 | 0.99 | 0.67 | 0.68 | 0.69 | 0.93 | 1.00 |
| CC | 1.00 | 0.93 | 0.91 | 0.95 | 0.99 | 0.50 | 0.53 | 0.57 | 0.85 | 0.95 | 0.67 | 0.67 | 0.69 | 0.89 | 0.97 |
| IMP-Average | 1.00 | 0.99 | 0.99 | 1.00 | 1.00 | 0.50 | 0.52 | 0.53 | 0.88 | 0.99 | 0.67 | 0.68 | 0.69 | 0.93 | 1.00 |
| IMP-Majority | 1.00 | 0.98 | 0.99 | 1.00 | 1.00 | 0.50 | 0.52 | 0.52 | 0.89 | 0.99 | 0.67 | 0.68 | 0.69 | 0.94 | 1.00 |
| IMP-Wald | 1.00 | 0.97 | 0.98 | 0.98 | 1.00 | 0.50 | 0.55 | 0.54 | 0.90 | 0.99 | 0.67 | 0.70 | 0.70 | 0.95 | 1.00 |
| *Experimental study 3* | | | | | | | | | | | | | | | |
| BD | 1.00 | 0.99 | 0.99 | 1.00 | 1.00 | 0.50 | 0.52 | 0.53 | 0.90 | 0.99 | 0.67 | 0.69 | 0.70 | 0.94 | 1.00 |
| CC | 1.00 | 0.95 | 0.92 | 0.98 | 0.99 | 0.50 | 0.55 | 0.57 | 0.89 | 0.96 | 0.67 | 0.68 | 0.69 | 0.93 | 0.98 |
| IMP-Average | 1.00 | 0.99 | 0.99 | 1.00 | 0.99 | 0.50 | 0.52 | 0.53 | 0.88 | 0.98 | 0.67 | 0.68 | 0.69 | 0.93 | 0.98 |
| IMP-Majority | 1.00 | 0.98 | 0.99 | 0.99 | 0.99 | 0.50 | 0.52 | 0.52 | 0.89 | 0.98 | 0.67 | 0.68 | 0.69 | 0.94 | 0.98 |
| IMP-Wald | 1.00 | 0.97 | 0.97 | 0.98 | 0.99 | 0.50 | 0.55 | 0.54 | 0.88 | 0.98 | 0.67 | 0.70 | 0.70 | 0.94 | 0.98 |

**Table 7** Model estimating the individual participation propensity for students in Wave 12 of SC 4 used to derive adjustment factors for adjusted wave-specific cross-sectional and longitudinal weights. From left-to-right the estimates for stepwise regression backwards, Elastic net with control mixing parameter $\alpha = 1.0$, i.e., Lasso penalty, and Bayesian Variable selection (BVS) with spike-and-slab prior. Additionally, the BVS estimates $\hat{\beta}$ are completed by the corresponding inclusion probabilities $\hat{\gamma}$

| Variables | Stepwise regression | Elastic net | Bayesian SnS | |
|---|---|---|---|---|
| | | | $\hat{\beta}$ | $\hat{\gamma}$ |
| Intercept | –1.452*** | –0.638 | **− 1.571** | **1.000** |
| | (0.096) | | | |
| Migration Background | –0.013*** | –0.017 | –0.037 | 0.227 |
| Yes | (0.035) | | | |
| Student participated in | –0.121*** | –0.004 | **− 0.129** | **0.693** |
| wave 6 | (0.031) | | | |
| Student participated in | 0.218*** | – | **0.129** | **0.560** |
| wave 8 | (0.056) | | | |
| Student participated in | 0.371*** | 0.069 | **0.343** | **1.000** |
| wave 9 | (0.053) | | | |
| Student participated in | 0.356*** | 0.001 | **0.251** | **0.973** |
| wave 10 | (0.056) | | | |
| Student participated in | 1.278*** | 1.159 | **1.214** | **1.000** |
| wave 11 | (0.036) | | | |

Reference categories are: Migration background *no*. To model individual participation, for the stepwise regression the glm-function with a probit link provided in R (R Core Team. 2020) was used. ***, **, and * denote significance at the 0.1%, 1%, and 5% level, respectively. Standard errors are given in parentheses. BIC based backward selection was used, and only significant coefficients are reported. BIC for the final model with selected variables: BIC $= 9{,}454.741$. For Elastic net only non-negligible variables are reported. The shrinkage parameter $\lambda$ is set to the largest value such that the error is 1 standard error of the minimum: $\lambda_{\min} = 0.015$ obtained with the cv.glmnet-function in R. The bold results in the last two columns show variables with an inclusion probability higher than 50%. The other two variables are not selected by the Bayesian approach, but are reported for the sake of completeness. The prior setting for the spike-and-slab were set to $\tau_2 = 1 \gg \tau_1 = 0.015$ and controlling the number of selecting variables with $w = 0.1$.

simulation results are also evaluated in terms of the root mean square error (RMSE) and the inclusion probabilities where possible. Therefore, it is important to bear in mind that we assess the models' performance on two levels. Firstly, the model selection, that is, regardless of the classification rule, how well do the different routines predict the initial model parameter based on the DGP. And secondly, how exactly the parameters underlying the data generating process are estimated.

The results of the different variations are presented in Tables 3 for experimental study 1 (Ex 1), Table 4 for experimental study 2 (Ex 2), and Table 5 for experimental study 3 (Ex 3) summarizing the three scenarios BD, CC, and MIS with results as average estimates and root mean square errors (RMSEs). Looking on Ex 1, the tables differ only in the results of estimation and pooling of the treatment of missing values, which are examined under quality aspects. As benchmark we report results for a simple probit model covering all important variables $(\alpha, \beta_1, \ldots, \beta_4)$ for the three different experimental studies. In the BD scenario we find overall unbiased results for all parameters, therefore we can assume that the considered routines

**Table 8** Model estimating the individual participation propensity for students in Wave 12 of SC 4 used to derive adjustment factors for adjusted wave-specific cross-sectional and longitudinal weights. Sensitivity analysis – Bayesian Variable selection (BVS) with spike-and-slab prior: results of the estimates of regression coefficients $\hat{\beta}$ for different $\tau_1$. Bold printed values have an inclusion probability $\hat{\gamma} > \mathbf{0.50}$. The prior setting for the spike-and-slab were set to $\tau_2 = 1$ and the spike prior $tau_1$ is set gradually in 0.025 steps from 0.0050 to 0.0375 controlling the number of selecting variables with $w = \mathbf{0.1}$

| | $\tau_0 = \mathbf{0.050}$ | | $\tau_0 = \mathbf{0.075}$ | | $\tau_0 = \mathbf{0.100}$ | | $\tau_0 = \mathbf{0.125}$ | | $\tau_0 = \mathbf{0.150}$ | | $\tau_0 = \mathbf{0.175}$ | | $\tau_0 = \mathbf{0.200}$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\hat{\beta}$ | $\hat{\gamma}$ | $\hat{\beta}$ | $\hat{\gamma}$ | $\hat{\beta}$ | $\hat{\gamma}$ | $\hat{\beta}$ | $\hat{\gamma}$ | $\hat{\beta}$ | $\hat{\gamma}$ | $\hat{\beta}$ | $\hat{\gamma}$ | $\hat{\beta}$ | $\hat{\gamma}$ |
| Intercept | –1.920 | 1.000 | –1.970 | 1.000 | –1.916 | 1.000 | –1.917 | 1.000 | –1.919 | 1.000 | –1.921 | 1.000 | –1.927 | 1.000 |
| Age: younger half | 0.051 | 0.014 | 0.051 | 0.014 | 0.049 | 0.015 | 0.049 | 0.014 | 0.050 | 0.017 | 0.050 | 0.019 | 0.050 | 0.022 |
| Gender: female | 0.023 | 0.009 | 0.023 | 0.009 | 0.023 | 0.011 | 0.023 | 0.014 | 0.023 | 0.017 | 0.023 | 0.020 | 0.023 | 0.023 |
| Nationality: German | 0.064 | 0.083 | 0.064 | 0.083 | 0.063 | 0.019 | 0.064 | 0.018 | 0.063 | 0.021 | 0.063 | 0.024 | 0.063 | 0.024 |
| Primary language: German | 0.078 | 0.148 | 0.079 | 0.045 | 0.079 | 0.024 | 0.078 | 0.022 | 0.081 | 0.021 | 0.078 | 0.023 | 0.080 | 0.024 |
| Migration Background: yes | –0.095 | 0.106 | –0.095 | 0.029 | –0.095 | 0.019 | –0.095 | 0.020 | –0.096 | 0.018 | –0.094 | 0.022 | –0.094 | 0.026 |
| Participation in wave 1 | –0.061 | 0.548 | –0.057 | 0.398 | –0.064 | 0.283 | –0.065 | 0.197 | –0.068 | 0.142 | –0.063 | 0.105 | –0.060 | 0.087 |
| Participation in wave 2 | 0.200 | 0.680 | 0.196 | 0.353 | 0.196 | 0.151 | 0.198 | 0.081 | 0.197 | 0.049 | 0.196 | 0.045 | 0.198 | 0.035 |
| Participation in wave 3 | 0.057 | 0.056 | 0.057 | 0.021 | 0.056 | 0.014 | 0.058 | 0.018 | 0.057 | 0.021 | 0.056 | 0.022 | 0.058 | 0.020 |
| Participation in wave 4 | –0.001 | 0.028 | –0.003 | 0.013 | –0.001 | 0.012 | –0.001 | 0.015 | –0.003 | 0.016 | –0.000 | 0.019 | –0.002 | 0.020 |
| Participation in wave 5 | 0.155 | 0.481 | 0.156 | 0.182 | 0.157 | 0.072 | 0.157 | 0.040 | 0.155 | 0.030 | 0.157 | 0.032 | 0.157 | 0.035 |
| Participation in wave 6 | –0.130 | 0.241 | –0.130 | 0.052 | –0.131 | 0.027 | –0.132 | 0.025 | –0.130 | 0.025 | –0.131 | 0.026 | –0.130 | 0.023 |
| Participation in wave 7 | 0.030 | 0.030 | 0.030 | 0.014 | 0.031 | 0.014 | 0.028 | 0.019 | 0.032 | 0.017 | 0.031 | 0.020 | 0.030 | 0.024 |
| Participation in wave 8 | 0.196 | 0.719 | 0.197 | 0.324 | 0.197 | 0.123 | 0.197 | 0.059 | 0.197 | 0.044 | 0.195 | 0.038 | 0.197 | 0.037 |
| Participation in wave 9 | 0.356 | 0.999 | 0.358 | 0.971 | 0.356 | 0.771 | 0.358 | 0.464 | 0.357 | 0.239 | 0.359 | 0.154 | 0.357 | 0.103 |
| Participation in wave 10 | 0.337 | 0.999 | 0.337 | 0.940 | 0.337 | 0.684 | 0.337 | 0.374 | 0.339 | 0.206 | 0.339 | 0.126 | 0.339 | 0.098 |
| Participation in wave 11 | 1.269 | 1.000 | 1.270 | 1.000 | 1.269 | 1.000 | 1.267 | 1.000 | 1.268 | 1.000 | 1.269 | 1.000 | 1.269 | 1.000 |

Reference categories are: Age *older half*, Gender *male* Nationality *other than German*, primary language *other than German*, migration background *no*.

are implemented correctly. The different pooling approaches produce very different results in terms of accuracy and variation in the estimation results. It is surprising that the results of simple averaging come to significant improvements in the estimation results compared to the before deletion values, which is set as a benchmark for the three pooling strategies. It seems that the majority method tends to over-fit, as the estimation results are clearly too unbiased. This can be seen in all three experimental studies. Compared to our Bayesian spike-and-slab approach, which also treats missing values, there are sometimes large differences in the estimation results compared to the respective pooling approaches. However, the average and Wald methods show comparable patterns in that the RMSEs are better compared to the complete cases, whereas they are in part more distorted compared to the before deletion results. When data sets with many of missing values in several variables are available, the Bayesian spike-and-slab method shows its strengths. The RMSE increases slightly for all methods in the CC scenario, whereas the absolute estimation bias decrease for the three Elastic net methods. The bias of the estimators increases for the stepwise regression method and for the Spike-and-Slab approach. In contrast, imputation shows that performance decreases for all five approaches measured with the RSME. With the Bayesian Spike-and-Slab, the inclusion probability for $X_2$ increases again, so that the variables are selected with a very high probability, but the bias increases so that the RSME is high compared to the stepwise regression. Interestingly, all three Elastic net approaches do not show altogether large deviations in the imputation scenario due to the combining rules. Thus, the results of stepwise regression and Bayesian Spike-and-Slab are quite similar. The experimental study 2 shows similar results, but due to the higher missing dropout, more biased results are to be expected. This shows the strength of our Bayesian approach, which does not only treat the missing values in the run-up to selection and estimation, making the estimation and selection results more precise and the selection of variables more correct in terms of precision, recall and $F$-measure. Experimental Study 3 shows less precise results in estimation and selection due to the MAR failure of data for all presented selection strategies, however, the Bayesian spike-and-slab approach can score here by having precision ahead of the other methods.

However, it must be noted that the estimation accuracy is more accurate with stepwise regression (SR) and Bayesian Spike-and-Slab (SnS), and thus the bias is small and the RMSE is half towards the Elastic net (EN) approaches. Table 6 shows the results of the calculations of Accuracy: precision, recall, and F-measure for before deletion, complete case and imputation obtained with Elastic net, stepwise regression, and Bayesian Spike-and-Slab. Looking on the important measurement, e.g., $F$-measure, shows in principle the same result. Accuracy, as measured for ridge regression (EN.00), always yields the same results because all variables are always included in the model and no selection is made, only shrinkage. The spike-and-slab is almost among the top performers in terms of $F$-measure. This shows that the ridge approach does not perform selection in the strict sense, but merely shrinks the values close to zero. Thus, the values for the ridge results never reach the accuracy of the other estimation methods. On the other hand, it is striking that the accuracy values of the Lasso estimates do not approach those of stepwise regression or our Bayesian approach. However, stepwise regression also shows its strengths in correctly selecting appropriate variables that were involved in the data generating process. In general, the accuracy decreases when estimating the complete cases compared to the before deletion values and then increases when dealing with missing values, as intended. Here, it is also shown that the average and the Wald method of pooling yields comparable results to the Bayesian approach and provides the intended accuracy gain of handling missing values compared to the complete cases.

## 6 Empirical illustration

In order to illustrate the usefulness of the suggested Bayesian spike and slab approach in empirical analysis, we provide exemplary applications using the scientific data use file of the German National Educational Panel Study (NEPS), Starting Cohort Grade 9, see NEPS (2021) and Blossfeld and Roßbach (2019). For this purpose, a random sample of schools, stratified by school type, was drawn throughout Germany. Within the schools, all students of two randomly selected ninth grades were invited to participate in the survey. The technical details on weighting are reported for each wave, see e.g., Bergrab (2020). Here, variable selection finds its application in selecting the appropriate variables that describe a student's probability of participation in a specific wave, where we analyze wave twelve here. From the set of available covariate variables ($P = 16$), only a few have to be selected in order to determine the participation probabilities and subsequently prepare suitable weights for further analyses. In the starting cohort considered here all students, regardless of whether in vocational education or academic track, willing to participate in the NEPS are followed up over time. The students entering the academic track usually remain within their school context. In contrast, students entering the vocational education leave school for a vocational training. In wave twelve all students left their school context and are surveyed individually. To account for the wave-specific participation decision of students' response propensity re-weighting is used to provide corresponding weights. To model binary participation decisions a model with probit link function is used for all three variable selection methods: backward selection, Elastic net, Bayesian spike-and-slab. By wave twelve, the panel cohort has reduced to $n = 7,911$ students in the age of mean 26.76 (standard deviation 0.73). For our analysis, we included all students and $p = 16$ variables. In contrast to the weighting report, all variables were standardized before selection to show comparability with the above-mentioned experimental studies.

Table 7 summarizes the results for stepwise regression, Elastic net, and Bayesian spike-and-slab (Bayesian SnS) approach. To model individual participation, for the stepwise regression the `glm`-function with a probit link provided in R (R Core Team. 2020) was used. BIC based backward selection was used and only significant coefficients are reported. For Elastic net only non-negligible variables are reported. The dash indicates that this variable was not included in the estimation model. The shrinkage parameter of Elastic net $\lambda$ is set to the largest value such that the error is 1 standard error of the minimum: $\lambda_{\text{min}} = 0.015$ obtained with the `cv.glmnet`-function in R. The Elastic net represents a Lasso selection with a control parameter of 1.0.

For the results of our Bayesian approach both, the estimates $\hat{\beta}$ as median of posterior and the corresponding marginal posterior inclusion probabilities $\hat{\gamma}$, are presented. The bold results in the last two columns show variables with an inclusion probability higher than 50%, where the other results are listed for the sake of completeness. The prior setting for the spike-and-slab were set to $\tau_2 = 1 \gg \tau_1 = 0.015$ and the shrinkage $w = 0.015$. The posterior estimates are based on MCMC chains of length 20,000. After discarding the first 5,000 iterations as burn-in, inference is based on the remaining 15,000 simulated draws from the joint posterior distribution. The convergence diagnostics indicate overall convergence.

Table 7 show less differences among the three approaches. Participation in the last waves is selected according to all three approaches, except for participation in wave 6, which is not selected by Elastic net. The total estimate and the selection of migration background vary across the three models. The most parsimonious model is determined by Elastic net and our spike-and-slab approach, which includes a total of six variables. In contrast, stepwise regression only includes one additional variable in the model. The selection of stepwise regression shows overall significant results, whereas the Bayesian spike-and-slab approach extracts migration background as a redundant variable with an inclusion probability of $\hat{\gamma} = 0.227$. The remaining inclusion probabilities indicate a markedly distinct decision to include. The model based on elastic net does not consider participation in Wave 6. However, it includes all other variables that are also selected by stepwise regression and the Bayesian approach. An analysis of the inclusion probabilities reveals that, in addition to the intercept, positive participation in the surveys in waves 9, 10, and 11 also strongly influences the probability of participating in wave 12. The inclusion probabilities are 1 or nearly 1 for all four variables. This demonstrates that stepwise regression and the Bayesian approach yield nearly identical outcomes, as the most crucial variables are selected in a comparable sequence. The advantage of the Bayesian approach is that, in addition to the assessment based on the significance level in stepwise regression, the inclusion probability offers a more direct approach to the evaluation.

It is important to recognize that the Bayesian approach is susceptible to the influence of prior beliefs, which requires further investigation. The impact of the variance priors on the values of $\beta$ and $\gamma$ is negligible, as well as the starting values. However, when holding $\tau_2 = 1$, the exploration of the gradual adjustment of the hyperparameter $\tau_1$ from 0.050 to 0.200 in steps of 0.0.025 reveals sensitive changes. The results of this can be found in Table 8. The Bayesian spike-and-slab algorithm is sensitive to the specific choice of spike, i.e., $\tau_2$; in detail, climbing up the increments on $\tau_2$ keeps the variables participation in the last wave and migration background in the model with constant high inclusion probabilities. Furthermore, the estimates of the latter two variables alternate, but the estimate for participation in the previous wave is extremely constant. The other selected variables vanish out gradually. It is noteworthy that a similar outcome is achieved when the shrinkage parameter of Elastic net, defined as $\lambda_{\min} = 0.015$, is employed in conjunction with the Bayesian spike-and-slab approach.

As previously discussed, these methods will set regression coefficients to zero, if necessary. The three approaches compared provide results that are essentially similar, although differences in depth are apparent. The results offer an intriguing perspective on the Bayesian approach to handling missing values, as well as variable selection and its performance relative to established methods for variable selection. Further investigations regarding the weights calculated with the models were not carried out.

## 7 Conclusion

This paper provides assessment of different strategies, i.e., applying statistical as well as machine learning algorithms, for variable selection in the context of binary regression models with missing values in the covariate variables. We show how our algorithmic strategies can be combined and how they can accommodate inference over the prior inclusion probability and which prior settings affect the posterior estimates. To handle missing values within the Bayesian estimation paradigm, the device of data augmentation can be used. The discussion of the various strategies highlights similarities and differences between shrinkage estimators and Bayesian estimation approaches. The choice of hyperparameters is in all methods a sensitive issue. The tuning parameter for using shrinkage estimators is typically estimated by cross-validation, whereas the hyperparameters in Bayesian estimation are fixed a priori.

From a methodological view, the conceptual strengths of the Bayesian spike-and-slab model and the ridge regression are revealed in the sense that they do not exhaust predictive information in trying to determine which variables have exactly zero effect. Any attempt to select variables after the fact, as is done in Lasso or Elastic net, does not lead to the loss of information to worse predictions. Therefore, in the Bayesian setting it is important to understand that the set of selected variables that have no predictive effect has a probability of (approximately) zero. Whereas all variables that have an inclusion probability that is above a certain threshold can be considered truly predictive. The evidence provided in the empirical illustration suggests that for the purposes of weighting in the $N > P$ setting, all strategies work well, but with small advantages of the Bayesian approach, especially when missing values occur in the covariate data. As discussed in Du et al. (2022) our Bayesian approach, simultaneously imputing the missing values, selecting and estimating jointly the model parameters, is time-consuming and computationally intensive not only for a large amount of missingness. Therefore, we show that dealing with missing values in the context of statistics and machine learning requires a convincing strategy for imputing the missing values. Not considering the missing data pattern as well as the averaging of imputed estimation results without taking pooling rules into account leads to a loss of quality, which can be seen e.g., in biased estimation results or lower variances of the estimators. Hence, case-deletion or complete-case strategies are frequently used when individuals are excluded from the analysis, if they are missing any of the variables or items. whereas the quality of the data analysis suffers as a result. As shown, non-Bayesian variable selection approaches cannot be easily applied within the framework of multiple imputation. The difficulty lies in how to combine the selection results across the multiply imputed data sets, because non-Bayesian variable selection approaches would commonly identify different redundant coefficients for the various imputed data sets and thus will lead to different numbers of coefficients to compare. In conclusion, we present a strategy clearly combining estimation, shrinkage and handling of missing values. The Bayesian holistic method for combining variable selection and imputation of missing values in covariates offers several advantages over traditional methods. By treating missing values as parameters and assigning priors to them, this approach provides a more accurate and reliable estimation of regression coefficients. While it may appear to users that the Elastic net and stepwise regression approaches are assumption-free, there are nevertheless assumptions involved regarding the setting of the control and shrinkage parameters, which are set via the priors in the Bayesian approach. Likewise, the pooling step based on chained equation is non-trivial, while in the Bayesian approach this can be implemented as an additional step in the iterative Gibbs sampler.

There are several approaches in the literature for imputing missing values and selecting variables. Note that Heymans et al. (2007) suggest a boot-strapped variable selection under multiple imputation to overcome the pitfalls of applying the combining rules to stepwise regression and Panken and Heymans (2022) provide similar frameworks for the logistic setup based on the majority based approach dividing the imputed datasets in test and train data. Chen and Wang (2013) extend the Lasso to multiple imputation with grouping the imputed data, combining multiple imputation and random Lasso see Liu et al. (2016), and combining Lasso with the Expectation-Maximization algorithm (Sabbe et al. 2013). The presented Bayesian holistic method highlights the potential of combining imputation with advanced techniques with accuracy and reliability in statistical results and offers promising avenues for future research.

**Electronic supplementary material** The online version of this article (https://doi.org/10.1007/s11943-024-00345-1) contains supplementary material, which is available to authorized users.

**Acknowledgements** This paper utilizes data from the German National Educational Panel Study (NEPS), as detailed in Blossfeld and Roßbach (2019). The NEPS is conducted by the Leibniz Institute for Educational Trajectories (LIfBi, Germany) in collaboration with a nationwide network. Additionally, we would like to express our gratitude to the Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities for providing the necessary system resources for the simulation studies.

**Funding** Open Access funding enabled and organized by Projekt DEAL.

**Conflicts of interest** The authors declare to have no conflict of interest.

## References

Albert JH (1992) Bayesian estimation of normal ogive item response curves using Gibbs sampling. J Educ Stat 17(3):251–269. https://doi.org/10.2307/1165149

Albert JH, Chib S (1993) Bayesian analysis of binary and polychotomous response data. J Am Stat Assoc 88(422):669–679. https://doi.org/10.1080/01621459.1993.10476321

Aßmann C (2012) Determinants and costs of current account reversals under heterogeneity and serial correlation. Appl Econ 44(13):1685–1700. https://doi.org/10.1080/00036846.2011.554370

Aßmann C, Boysen-Hogrefe J (2011) A Bayesian approach to model-based clustering for binary panel probit models. Comput Stat Data Anal 55(1):261–279. https://doi.org/10.1016/j.csda.2010.04.016

Aßmann C, Gaasch JC, Stingl D (2023) A Bayesian approach towards missing covariate data in multilevel latent regression models. Psychometrika 88:1495–1528. https://doi.org/10.1007/s11336-022-09888-0

Aßmann C, Preising M (2020) Bayesian estimation and model comparison for linear dynamic panel models with missing values. Aust N Z J Stat 62(4):536–557. https://doi.org/10.1111/anzs.12316

Bergrab M (2020) *Samples, weights, and nonresponse: The sample of starting cohort 4 of the national educational panel study (wave 12)* (tech. rep.). Leibniz Institute for Eduational Trajectories, National Educational Panel Study, Bamberg (https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC4/12-0-0/SC4_12-0-0_W.pdf)

Bezanson J, Edelman A, Karpinski S, Shah VB (2017) Julia: a fresh approach to numerical computing. SIAM Rev 59(1):65–98. https://doi.org/10.1137/141000671

Bhattacharya A, Chakraborty A, Mallick BK (2016) Fast sampling with gaussian scale mixture priors in high-dimensional regression. Biometrika 103(4):985–991. https://doi.org/10.1093/biomet/asw042

Biswas N, Mackey L, Meng XL (2022) Scalable spike-and-slab. In: Chaudhuri K, Jegelka S, Song L, Szepesvari C, Niu G, Sabato S (eds) Proceedings of the 39th international conference on machine learning. PMLR, pp 2021–2040 (https://proceedings.mlr.press/v162/biswas22a.html)

Blossfeld HP, Roßbach HG (eds) (2019) Education as a lifelong process. Springer https://doi.org/10.1007/978-3-658-23162-0

Bottolo L, Richardson S (2010) Evolutionary stochastic search for Bayesian model exploration. Bayesian Anal https://doi.org/10.1214/10-BA523

Brand J (1999) Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets. TNO Prevention; Health (Erasmus University Rotterdam) (ph.d. thesis)

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Chapman & Hall CRC

Burgette LF, Reiter JP (2010) Multiple imputation for missing data via sequential regression trees. Epidemiol Rev 172(9):1070–1076. https://doi.org/10.1093/aje/kwq260

van Buuren S (2018) Flexible imputation of missing data, second edition. Chapman Hall CRC https://doi.org/10.1201/9780429492259

van Buuren S, Groothuis-Oudshoorn K (2011) Mice: Multivariate imputation by chained equations in R. J Stat Soft 45(3):1–67. https://doi.org/10.18637/jss.v045.i03

Carvalho CM, Polson NG, Scott JG (2010) The horseshoe estimator for sparse signals. Biometrika 97(2):465–480. https://doi.org/10.1093/biomet/asq017

Chen Q, Wang S (2013) Variable selection for multiply-imputed data with application to dioxin exposure study. Statist Med 32(21):3646–3659. https://doi.org/10.1002/sim.5783

Clyde M, George EI (2004) Model uncertainty. Stat Sci 19(1):81–94. https://doi.org/10.1214/088342304000000035

Dobra A (2009) Variable selection and dependency networks for genomewide data. Biostatistics 10(4):621–639.https://doi.org/10.1093/biostatistics/kxp018

Doove LL, van Buuren S, Dusseldorp E (2014) Recursive partitioning for missing data imputation in the presence of interaction effects. Comput Stat Data Anal 72:92–104. https://doi.org/10.1016/j.csda.2013.10.025

Du J, Boss J, Han P, Beesley LJ, Kleinsasser M, Goutman SA, Batterman S, Feldman EL, Mukherjee B (2022) Variable selection with multiply-imputed datasets: Choosing between stacked and grouped methods. J Comput Graph Stat 31(4):1063–1075. https://doi.org/10.1080/10618600.2022.2035739

Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. J Stat Soft 33(1):1–22. https://doi.org/10.18637/jss.v033.i01

Frühwirth-Schnatter S (2010) Finite mixture and markov switching models. Springer

Frühwirth-Schnatter S, Kaufmann S (2008) Model-based clustering of multiple time series. J Bus Econ Stat 26(1):78–89. https://doi.org/10.1198/073500107000000106

Gelfand AE, Smith AFM (1990) Sampling-based approaches to calculating marginal densities. J Am Stat Assoc 85(410):398–409. https://doi.org/10.1080/01621459.1990.10476213

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2023) Bayesian data analysis. Chapman Hall, CRC https://doi.org/10.1201/b16018

Geman S, Geman D (1984) Stochastic relaxation, gibbs distributions, and the Bayesian restoration of images. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-6, vol 6, pp 721–741. https://doi.org/10.1109/tpami.1984.4767596

George EI (2000) The variable selection problem. J Am Stat Assoc 95(452):1304–1308. https://doi.org/10.1080/01621459.2000.10474336

George EI, McCulloch RE (1993) Variable selection via Gibbs sampling. J Am Stat Assoc 88(423):881–889. https://doi.org/10.1080/01621459.1993.10476353

George EI, McCulloch RE (1997) Approaches to Bayesian variable selection. Stat Sinica 7(2):339–373

Geweke J (1991) Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. Staff report (Federal Reserve Bank of Minneapolis. Research Department) 148. https://doi.org/10.21034/sr.148

Gneiting T (2011) Making and evaluating point forecasts. J Am Stat Assoc 106(494):746–762. https://doi.org/10.1198/jasa.2011.r10138

Hans C, Dobra A, West M (2007) Shotgun stochastic search for ”large p” regression. J Am Stat Assoc 102(478):507–516. https://doi.org/10.1198/016214507000000121

Hansen BE (2007) Least squares model averaging. Econometrica 75(4):1175–1189. https://doi.org/10.1111/j.1468-0262.2007.00785.x

Heymans MW, van Buuren S, Knol DL, van Mechelen W, de Vet HC (2007) Variable selection under multiple imputation using the bootstrap in a prognostic study. BMC Med Res Methodol. https://doi.org/10.1186/1471-2288-7-33

Hoerl AE, Kennard RW (1970) Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12(1):55–67. https://doi.org/10.1080/00401706.1970.10488634

Ishwaran H, Rao JS (2005) Spike and slab variable selection: Frequentist and Bayesian strategies. Ann Stat 33(2):730–773. https://doi.org/10.1214/009053604000001147

Jackman S (2009) Bayesian analysis for the social sciences. Wiley

Kohn R, Smith M, Chan D (2001) Nonparametric regression using linear combinations of basis functions. Stat Comput 11(4):313–322. https://doi.org/10.1023/a:1011916902934

Korobilis D, Shimizu K (2022) Bayesian approaches to shrinkage and sparse estimation. Found Trends Econom 11(4):230–354. https://doi.org/10.1561/0800000041

Kyung M, Gill J, Ghosh M, Casella G (2010) Penalized regression, standard errors, and bayesian lassos. Bayesian Anal 5(2):369–411. https://doi.org/10.1214/10-ba607

Lamnisos D, Griffin JE, Steel MFJ (2009) Transdimensional sampling algorithms for Bayesian variable selection in classification problems with many more variables than observations. J Comput Graph Stat 18(3):592–612. https://doi.org/10.1198/jcgs.2009.08027

Lee KE, Sha N, Dougherty ER, Vannucci M, Mallick BK (2003) Gene selection: a Bayesian variable selection approach. Bioinformatics 19(1):90–97. https://doi.org/10.1093/bioinformatics/19.1.90

Liu Y, Wang Y, Feng Y, Wall MM (2016) Variable selection and prediction with incomplete high-dimensional data. Ann Appl Stat 10(1):418–450. https://doi.org/10.1214/15-AOAS899

Lütkepohl H (1996) Handbook of matrices. Wiley

Mallows CL (1973) Some comments on cp. Technometrics 15(4):661–675. https://doi.org/10.1080/00401706.1973.10489103

Marill T, Green D (1963) On the effectiveness of receptors in recognition systems. IEEE Trans Inf Theory 9(1):11–17. https://doi.org/10.1109/tit.1963.1057810

Miller A (2019) Subset selection in regression. Taylor & Francis

Mitchell TJ, Beauchamp JJ (1988) Bayesian variable selection in linear regression. J Am Stat Assoc 83(404):1023–1032. https://doi.org/10.1080/01621459.1988.10478694

Mood AM, Graybill FA, Boes DC (1974) Introduction to the theory of statistics. McGraw-Hill

NEPS, National Educational Panel Study (2021) Neps-startkohorte 4: Klasse 9 (sc4 12.0.0) https://doi.org/10.5157/NEPS:SC4:12.0.0

O’Hara RB, Sillanpää MJ (2009) A review of Bayesian variable selection methods: What, how and which. Bayesian Anal 4(1):85–117. https://doi.org/10.1214/09-BA403

Panken AM, Heymans MW (2022) A simple pooling method for variable selection in multiply imputed datasets outperformed complex methods. BMC Med Res Methodol. https://doi.org/10.1186/s12874-022-01693-8

Park MY, Hastie T (2007) L1-regularization path algorithm for generalized linear models. J Royal Stat Soc Ser B (statistical Methodol 69(4):659–677. https://doi.org/10.1111/j.1467-9868.2007.00607.x

R Core Team (2020) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna (https://www.R-project.org/)

Raftery AE (1995) Bayesian model selection in social research. Sociol Methodol 25:111–163. https://doi.org/10.2307/271063

Ročková V, George EI (2018) The spike-and-slab LASSO. J Am Stat Assoc 113(521):431–444. https://doi.org/10.1080/01621459.2016.1260469

Rubin D (1976) Inference and missing data. Biometrika 63(3):581–592. https://doi.org/10.1093/biomet/63.3.581

Rubin D (1981) The Bayesian bootstrap. Ann Stat 9(1):130–134. https://doi.org/10.1214/aos/1176345338

Rubin D (1984) Bayesianly justifiable and relevant frequency calculations for the applied statistician. Ann Stat 12(4):1151–1172. https://doi.org/10.1214/aos/1176346785

Russu A, Malovini A, Puca A, Bellazzi R (2012) Stochastic model search with binary outcomes for genome-wide association studies. J Am Med Inform Assoc 19(e1):e13–e20. https://doi.org/10.1136/amiajnl-2011-000741

Sabbe N, Thas O, Ottoy JP (2013) EMlasso: logistic lasso with missing data. Statist Med 32(18):3143–3157. https://doi.org/10.1002/sim.5760

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6(2):461–464. https://doi.org/10.1214/aos/1176344136

Strobl C, Boulesteix AL, Zeileis A, Hothorn T (2007) Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinform. https://doi.org/10.1186/1471-2105-8-25

Tanner M, Wong W (1987) The calculation of posterior distributions by data augmentation. J Am Stat Assoc 82(398):528–540. https://doi.org/10.1080/01621459.1987.10478458

Therneau T, Atkinson B (2018) *Rpart: Recursive partitioning and regression trees, [computer software manual].*. Version (R package version 4.1-13). https://CRAN.R-project.org/package=rpart

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J Royal Stat Soc Ser B (methodological) 58(1):267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K (2005) Sparsity and smoothness via the fused lasso. J Royal Stat Soc Ser B Stat Methodol 67(1):91–108. https://doi.org/10.1111/j.1467-9868.2005.00490.x

Venables WN, Ripley B (2002) Modern applied statistics with s. Springer

Vergouwe Y, Royston P, Moons KG, Altman DG (2010) Development and validation of a prediction model with missing predictor data: a practical approach. J Clin Epidemiol 63(2):205–214. https://doi.org/10.1016/j.jclinepi.2009.03.017

Wood AM, White IR, Royston P (2008) How should variable selection be performed with multiply imputed data? Statist Med 27(17):3227–3246. https://doi.org/10.1002/sim.3177

Yang X, Belin TR, Boscardin WJ (2005) Imputation and variable selection in linear regression models with missing covariates. Biometrics 61(2):498–506. https://doi.org/10.1111/j.1541-0420.2005.00317.x

Zou H (2006) The adaptive lasso and its oracle properties. J Am Stat Assoc 101(476):1418–1429. https://doi.org/10.1198/016214506000000735

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J Royal Stat Soc Ser B Stat Methodol 67(2):301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x