Lichter, Jens; Wiemann, Paul F V; Kneib, Thomas

Variational inference: uncertainty quantification in additive models

AStA Advances in Statistical Analysis

Provided in Cooperation with:
Springer Nature

Suggested Citation: Lichter, Jens; Wiemann, Paul F V; Kneib, Thomas (2024) : Variational inference: uncertainty quantification in additive models, AStA Advances in Statistical Analysis, ISSN 1863-818X, Springer, Berlin, Heidelberg, Vol. 108, Iss. 2, pp. 279-331, https://doi.org/10.1007/s10182-024-00492-4

AStA Advances in Statistical Analysis (2024) 108:279–331
https://doi.org/10.1007/s10182-024-00492-4

ORIGINAL PAPER

Variational inference: uncertainty quantification in additive models

Jens Lichter[1] · **Paul F V Wiemann**[1,2] · **Thomas Kneib**[1]

Received: 28 October 2022 / Accepted: 15 September 2023 / Published online: 3 April 2024

Abstract

Markov chain Monte Carlo (MCMC)-based simulation approaches are by far the most common method in Bayesian inference to access the posterior distribution. Recently, motivated by successes in machine learning, variational inference (VI) has gained in interest in statistics since it promises a computationally efficient alternative to MCMC enabling approximate access to the posterior. Classical approaches such as mean-field VI (MFVI), however, are based on the strong mean-field assumption for the approximate posterior where parameters or parameter blocks are assumed to be mutually independent. As a consequence, parameter uncertainties are often underestimated and alternatives such as semi-implicit VI (SIVI) have been suggested to avoid the mean-field assumption and to improve uncertainty estimates. SIVI uses a hierarchical construction of the variational parameters to restore parameter dependencies and relies on a highly flexible implicit mixing distribution whose probability density function is not analytic but samples can be taken via a stochastic procedure. With this paper, we investigate how different forms of VI perform in semiparametric additive regression models as one of the most important fields of application of Bayesian inference in statistics. A particular focus is on the ability of the rivalling approaches to quantify uncertainty, especially with correlated covariates that are likely to aggravate the difficulties of simplifying VI assumptions. Moreover, we propose a method, where we combine both advantages of MFVI and SIVI and compare its performance. The different VI approaches are studied in comparison with MCMC in simulations and an application to tree height models of douglas fir based on a large-scale forestry data set.

Keywords Approximate Bayesian inference · Deep neural networks · Multilayer perceptron · P-splines · Semiparametric regression · Variational Bayes · Simultaneous credible intervals

✉ Jens Lichter
jens.lichter@uni-goettingen.de

1 University of Göttingen, Chair of Statistics, Göttingen, Germany

2 University of Wisconsin-Madison, Department of Statistics,, Madison Wisconsin, United States

1 Introduction

The Bayesian paradigm provides a convenient and attractive framework for performing inference in statistical models, allowing for the incorporation of prior knowledge and, therefore, regularization of the effects of interest. However, the posterior distribution resulting from Bayes' theorem is, beyond simple conjugate cases, in general not analytically tractable. The invention of Markov chain Monte Carlo (MCMC) simulation techniques has revolutionized the applicability of Bayesian inference even in very complex statistical models, providing sampling-based numerical access to the posterior. MCMC provides access to exact posterior even for small samples, including exact uncertainty quantification also for complex functionals of the original model parameters. On the downside, however, MCMC is also known for being notoriously slow due to its sequential construction and it requires careful monitoring of mixing and convergence towards the (unknown) stationary distribution, often including the adaptive choice of tuning parameters. Hence, there is renewed interest in approximate approaches for Bayesian inference that bypass the need for MCMC sampling techniques at the cost of only approximate access to the posterior.

One such approach that has gained considerably in popularity, especially in machine learning, is variational inference (VI) also called variational Bayes. The basic idea is to find the optimal approximation to the posterior distribution within a pre-specified class of variational distributions by searching for the parameters of the approximating distribution with a deterministic optimization scheme (Ormerod and Wand 2010; Blei et al. 2017). In contrast to stochastic optimization techniques such as MCMC, the direct optimization of an objective function promises much faster inference. However, depending on the complexity of the approximating family chosen for VI, the approximate posterior may not capture all aspects of the true posterior distribution and, in particular, it has been reported that simple VI approaches may considerably underestimate uncertainty attached to the parameters of interest (Bishop 2006, Ch. 10). This is particularly the case for the simplest of VI, mean-field VI (MFVI), where the variational family assumes (blocks of) parameters to be mutually independent. This assumption significantly reduces the complexity of the approximation problem and often enables fast optimization steps resembling the structure of Gibbs updates in MCMC. However, the restrictive assumption of posterior independence is often at odds with the true posterior such that MFVI provides sensible point estimates but may severely underestimate parameter uncertainty.

As a consequence, various approaches beyond simple MFVI have been suggested (as reviewed, for example, in Zhang et al. 2018). One obvious remedy is to combine as many parameters as possible in one block such that one multivariate variational distribution is constructed, therefore mitigating the mean-field assumption (see for example Hui et al. 2019; Luts et al. 2014). However, this comes at the price of determining a fully unstructured covariance matrix for all parameters simultaneously, which requires handling of large matrices, especially for a large number of effects. An alternative is the semi-implicit VI (SIVI) approach recently developed by Yin and Zhou (2018). Compared to MFVI, it increases the complexity of the variational distribution allowing for some parameter dependencies. Firstly, SIVI uses a hierarchical construction of the variational parameters to bring back parameter dependencies based on hierarchical

VI (Ranganath et al. 2016). Secondly, the mixing distribution on the higher level of the hierarchy does not need to be an analytic probability density function, meaning a highly flexible implicit distribution can be chosen, i.e. the distribution is not required to have an analytic probability density function but samples can be generated from it (Diggle and Gratton 1984). While this approach brings simulations back into the inferential procedure, the underlying reasoning relies on law of large numbers asymptotics which are much easier to control and monitor than the distributional convergence of a Markov chain towards its limiting stationary distribution.

In this paper, we are focusing on semiparametric additive models as a particularly important special case of statistical modelling where Bayesian inference has gained considerable interest and both Gibbs sampling (Lang and Brezger 2004) and simple MFVI (Luts and Wand 2015; Waldmann and Kneib 2015; Hui et al. 2019) have been developed. More precisely, we

- review different forms of VI, including MFVI and SIVI, in their general form,
- develop their specific forms in semiparametric additive models including an improved MFVI approach where all regression coefficients associated with the additive components are combined in one block following ideas developed in Luts and Wand (2015) and Hui et al. (2019) and a combination of SIVI and MFVI (SIMFVI) that leads to more robust results and speeds up the optimization compared to the SIVI approach,
- investigate the performance of the different forms of VI with a specific focus on quantifying uncertainty in simulations to provide guidance on their reliability and applicability, and
- apply the methods to a data set on tree height of Douglas fir in a large-scale forestry data set.

We find that SIVI and SIMFVI effectively restore parameter uncertainty such that local and simultaneous credible intervals are accurately represented. However, the improved version of MFVI shows comparable performance such that combining all regression parameters in one block and therefore incorporating across effect dependence seems to be the crucial aspect in constructing an appropriate approximating distribution.

The structure of this article is as follows: In Sect. 2, we briefly introduce the necessary background on Bayesian additive regression models. Section 3 describes the methodology of VI and derives the algorithms for the different forms of MFVI and SIVI both in general and in the context of additive models. In Sect. 4, we compare all introduced methods and the Gibbs sampler in a simulation study with a focus on uncertainty quantification. Section 5 describes an application of the presented methods to tree heights of Douglas fir. In the final section, we summarize our results and briefly discuss limitations and potential directions for future research.

2 Bayesian additive models

We consider Bayesian forms of semiparametric additive models for regression data $(y_i, \boldsymbol{x}_i)$, $i = 1, \ldots, n$ where y_i denotes the response variable and $\boldsymbol{x}_i$ is a vector of explanatory variables of different type. More specifically, we assume the model structure

$$y_i = \sum_{j=1}^{p} f_j(\boldsymbol{x}_{ij}) + \epsilon_i,$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ represents the independent Gaussian error term while the effect of the covariates is additively decomposed into p effects $f_j(\cdot)$ that may represent linear, nonlinear, clustered (random), or spatial effects (among others) in a generic form. Each of the effects is then expanded in d_j basis functions as

$$f_j(\boldsymbol{x}_{ij}) = \sum_{l=1}^{d_j} \gamma_{jl} B_l^j(\boldsymbol{x}_{ij})$$

with effect-specific basis functions $B_l^j(\boldsymbol{x}_{ij})$ and corresponding basis coefficients γ_{jl}. In vector–matrix notation, this implies the model

$$\boldsymbol{y} = \boldsymbol{Z}_1 \boldsymbol{\gamma}_1 + \ldots + \boldsymbol{Z}_p \boldsymbol{\gamma}_p + \boldsymbol{\epsilon} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{y}$ and $\boldsymbol{\epsilon}$ are vectors of responses and error terms, the design matrices of basis function evaluations are denoted as $\boldsymbol{Z}_j$ and $\boldsymbol{\gamma}_j$ are the corresponding vectors of basis coefficients. Stacking all design matrices and basis coefficients into the matrix $\boldsymbol{Z}$ and $\boldsymbol{\gamma}$ yields the final representation as a large linear model.

To regularize the estimation of the basis coefficients, we employ multivariate normal priors

$$p(\boldsymbol{\gamma}_j | \tau_j^2) \propto \frac{1}{\left(2\pi\tau_j^2\right)^{\frac{\text{rank}(\boldsymbol{K}_j)}{2}}} \exp\left(-\frac{\boldsymbol{\gamma}_j' \boldsymbol{K}_j \boldsymbol{\gamma}_j}{2\tau_j^2}\right), \tag{2}$$

with zero mean and precision matrix $\boldsymbol{K}_j/\tau_j^2$. The precision matrix is chosen to reflect desirable regularization properties such as smoothness or shrinkage and may contain a non-trivial null space rendering Equation (2) into a partially improper prior specification. The impact of the prior on the posterior is regulated by the prior variance parameter τ_j^2. In the remainder of this paper, we will employ weakly informative inverse gamma priors $\tau_j^2 \sim \text{IG}(a_j, b_j)$, with default values of $a_j = b_j = 0.1$, but other prior distributions are easily conceivable. Similarly, we assign weakly informative inverse Gamma priors to σ^2, $\sigma^2 \sim \text{IG}(a_{\sigma^2}, b_{\sigma^2})$, with the same default values. Analytic forms of the distribution of the likelihood and the priors are shown in Appendix Sect. 7.3.

Each effect type takes a specific form by choosing the basis functions $B_l^j(\boldsymbol{x}_{ij})$ and the penalty matrix $\boldsymbol{K}_j$ (see Fahrmeir et al. 2021, for details):

- For linear effects, the basis functions are the untransformed covariates, $\boldsymbol{Z}_j = \boldsymbol{x}_{\cdot j}$, where $\boldsymbol{x}_{\cdot j}$ is a row vector representation of covariate j and a flat prior is obtained by setting $\boldsymbol{K}_j = \boldsymbol{0}$.
- In the case of clustered "random" effects, the basis functions represent dummy coding for the grouping variables and the penalty matrix equals the identity matrix, i.e. $\boldsymbol{K}_j = \boldsymbol{I}_j$.

- For nonlinear effects of continuous covariates, we use Bayesian P-splines (Lang and Brezger 2004) that are based on B-Spline basis functions in combination with a penalty matrix based on the *kth*-order random walk prior, e.g. a second-order random walk defined as $\gamma_{jl} = 2\gamma_{j,l-1} - \gamma_{j,l-2} + u_j$, with Gaussian errors $u_j \sim \mathcal{N}(0, \tau_j^2)$ and flat priors for γ_{j1} and γ_{j2}. In this way subsequent coefficients are penalized leading to a smoother functional form. The penalty matrix can then be constructed based on a difference matrix $\boldsymbol{D}_j$ such that $\boldsymbol{K}_j = \boldsymbol{D}_j' \boldsymbol{D}_j$.
- The concept of Bayesian P-splines can be extended to bivariate tensor product P-splines for fitting spatial effects or interaction surfaces $f_j(\boldsymbol{x}_{j1}, \boldsymbol{x}_{j2})$. This is achieved by combining the two univariate spline basis matrices $\boldsymbol{Z}_{j1}$ and $\boldsymbol{Z}_{j2}$ in terms of all $d_{j1} \cdot d_{j2}$ pairwise interactions. The penalty matrix is constructed by combining the two univariate spline penalties, $\boldsymbol{K}_{j1}$ and $\boldsymbol{K}_{j2}$ to $\boldsymbol{K}_j = \boldsymbol{K}_{j1} \otimes \boldsymbol{I}_{d_{j2}} + \boldsymbol{I}_{d_{j1}} \otimes \boldsymbol{K}_{j2}$ such that smoothness is enforced in both covariate directions, see also Appendix Sect. 7.1.

For univariate and bivariate effects as described here, the following two points should be considered. Firstly, the penalty matrix $\boldsymbol{K}_j$ is rank deficient and therefore the prior is improper. However, it can be shown that the resulting posterior is still proper (see Appendix Sect. 7.4). Secondly, further restrictions need to be imposed to ensure the identifiability of the model. We use the restriction of a centering constraint in the design matrix (see Appendix Sect. 7.2 for more details).

3 Variational inference in additive model

Variational inference (VI), as used in the Bayesian framework, casts the integration problem associated with obtaining the posterior distribution into an optimization problem. During the optimization, VI searches among a set of candidate distributions for the one approximating the posterior distribution best. If the set of candidate distributions approaches the complexity of the true distribution, VI promises to be computationally faster than MCMC while the quality of the results can be comparable. For instance, You et al. (2014) and Wang and Blei (2019) showed consistency for the VI approach in additive models. However, the procedure requires careful choices to be made which determine the quality of the approximation:

- The variational family $\mathcal{Q}$, i.e. the set of candidate distributions since a misspecification will directly limit the quality of the estimated posterior.
- The measure determining the quality of an element of the variational family relative to the exact posterior distribution. The classical divergence measure is the Kullback–Leibler-divergence (KL-divergence), but also other more general measures as described in' Zhang et al. (2018) are possible.
- The algorithm to searching for the best approximating variational distribution by finding the best combination of variational parameters $\boldsymbol{\psi}$ by optimizing the divergence measure. Again, Zhang et al. (2018) discuss different aspects including algorithms and strategies for variance reduction in the context of stochastic VI.

General overviews of variational inference are given in Bisho (2006, Ch. 10), Ormerod and Wand (2010) and Blei et al. (2017). In this paper, we only address the first point and describe possible extensions to the variational family $\mathcal{Q}$.

In the following, we introduce four different variational families used in this article to approximate the posterior distribution arising in Bayesian additive models. Two of the approaches presented are based on mean field approximations (see Sects. 3.2.1 and 3.2.2) while the remaining families proposed are based on the idea of semi-implicit variational inference (SIVI, Yin and Zhou 2018, see Sects. 3.2.3 and 3.2.4).

We denote the vector of model parameters as $\boldsymbol{\theta}$ and its posterior density with $p(\boldsymbol{\theta}|\mathbf{y})$. The elements of the variational family $\mathcal{Q}$, i.e. the variational distributions, are denoted as $q_{\boldsymbol{\psi}}$ where $\boldsymbol{\psi}$ is the vector of variational parameters. The density of the variational distribution is denoted as $q_{\boldsymbol{\psi}}(\boldsymbol{\theta})$.

To measure the deviation between the variational distribution $q_{\boldsymbol{\psi}}$ and the posterior distribution the Kullback–Leibler (KL) divergence,

$$\mathrm{KL}(q_{\boldsymbol{\psi}}(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathbf{y})) = \mathbb{E}_{\boldsymbol{\theta}\sim q_{\boldsymbol{\psi}}}[\log\ q_{\boldsymbol{\psi}}(\boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{\theta}\sim q_{\boldsymbol{\psi}}}[\log\ p(\boldsymbol{\theta}, \mathbf{y})] + \log\ p(\mathbf{y}),$$

is used (Jordan et al. 1999; Ormerod and Wand 2010). The KL divergence decreases with increasing similarity of the two distributions and is zero for two identical distributions. Hence, we want to find $\boldsymbol{\psi}^\star$ minimizing the KL divergence. Instead of working directly with the KL divergence, the minimization problem is reformulated as an equivalent maximization problem. Precisely, $\boldsymbol{\psi}^\star$ is determined by maximizing the evidence lower bound (ELBO),

$$\mathcal{L}(\boldsymbol{\psi}) = \mathbb{E}_{\boldsymbol{\theta}\sim q_{\boldsymbol{\psi}}}[\log\ p(\boldsymbol{\theta}, \mathbf{y})] - \mathbb{E}_{\boldsymbol{\theta}\sim q_{\boldsymbol{\psi}}}[\log\ q_{\boldsymbol{\psi}}(\boldsymbol{\theta})],$$

not containing the intractable marginal likelihood or model evidence $p(\mathbf{y})$ which does not depend on $\boldsymbol{\psi}$. The ELBO serves as the lower bound to the model evidence.

3.1 Mean-field and semi-implicit VI

3.1.1 Mean-field VI

Mean-field variational inference (MFVI, Parisi 1988; Saul and Jordan 1998) is based on a strong simplification assuming the posterior distribution can be approximated using independent parameter blocks, thus allowing to express the variational density as a product of the independent densities of the parameter blocks. The advantage of this simplification lies in easing the computation (Wainwright and Jordan 2008, p. 127-147) and the resulting speed gains. An iterative optimization scheme can be constructed by iteratively updating the variational parameters associated with one sub-vector such that the update mechanism maximizes the ELBO in each step. For example, the coordinate ascend variational inference algorithm (CAVI, Bishop 2006, Ch. 10) works in this way.

Suppose, the vector of model parameters is divided into p sub-vectors such that $\boldsymbol{\theta} = (\boldsymbol{\theta}'_1, \ldots, \boldsymbol{\theta}'_p)'$. Using the MFVI approach, the variational density can be expressed as

$q_{\boldsymbol{\psi}}(\boldsymbol{\theta}) = \prod_{i=1}^{p} q_{\boldsymbol{\psi}_i}(\boldsymbol{\theta}_i)$, where $\boldsymbol{\psi}_i$ are the variational parameters or a set of variational parameters associated with the variational distribution of the i-th subvector of $\boldsymbol{\theta}$. Now the variational density of the i-th sub-vector maximizing the ELBO is

$$q^{\star}(\boldsymbol{\theta}_i) \propto \exp\left\{\mathbb{E}_{\boldsymbol{\theta}_{-i} \sim q_{\boldsymbol{\psi}_{-i}}}\left[\log(p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}))\right]\right\}, \tag{3}$$

where $\boldsymbol{\theta}_{-i}$ denotes the parameter vector $\boldsymbol{\theta}$ without the i-th sub-vector and $q_{\boldsymbol{\psi}_{-i}}$ the variational distribution with the associated variational parameters in the index of said vector (Bishop 2006, Ch. 10). When selecting $q_{\boldsymbol{\psi}_{-i}}$ suitably and exploiting conditional conjugacy, a closed-form solution to $q^{\star}$ can be constructed by updating the variation parameter $\boldsymbol{\psi}_i$, similar to the parameters describing the sampling distribution in a Gibbs update step. CAVI then repeatedly iterates over i to update $\boldsymbol{\psi}_i$ until convergence of the ELBO.

3.1.2 Semi-implicit VI

SIVI (Yin and Zhou 2018) builds upon the idea of hierarchical variational inference (HVI) proposed by Ranganath et al. (2016) to reintroduce dependencies between the parameter blocks that are assumed independent in MFVI. To illustrate the concept of HVIt, suppose, we have three parameter blocks $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3)$ and for the variational parameters, namely, $\boldsymbol{\psi}_2$ and $\boldsymbol{\psi}_3$, a variational hyper-distribution $q_{\boldsymbol{\phi}}$ is assumed. The variational density of $\boldsymbol{\theta}$ with the variational parameters $\boldsymbol{\psi}_1$ and $\boldsymbol{\phi}$ can then be expressed as

$$q_{\boldsymbol{\psi}_1, \boldsymbol{\phi}}(\boldsymbol{\theta}) = q_{\boldsymbol{\psi}_1}(\boldsymbol{\theta}_1) \int q_{\boldsymbol{\psi}_2}(\boldsymbol{\theta}_2) q_{\boldsymbol{\psi}_3}(\boldsymbol{\theta}_3) q_{\boldsymbol{\phi}}(\boldsymbol{\psi}_2, \boldsymbol{\psi}_3) \mathrm{d}\boldsymbol{\psi}_2 \mathrm{d}\boldsymbol{\psi}_3. \tag{4}$$

Thus, the dependency between $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_3$ in the posterior can be restored in the variational distribution via dependency between $\boldsymbol{\psi}_2$ and $\boldsymbol{\psi}_3$ introduced via $q_{\boldsymbol{\phi}}$. The expansion of the variational family comes at the expense of increasing the computational burden. Hence, there is a trade-off between choosing MFVI as the faster optimization method and HVI which gives better approximations to the posterior in more complex settings but slows down the computational speed.

SIVI takes the idea of HVI a step further by allowing $q_{\boldsymbol{\phi}}$ to be an implicit distribution, meaning a distribution for which the density cannot be evaluated but for which we can sample from. This renders Equation (4) not analytical solvable and we cannot access the ELBO directly. Instead, the authors suggest constructing a lower bound to the ELBO. More precisely, the lower bound $\tilde{\mathcal{L}}_0$ is constructed as,

$$\begin{aligned}\mathcal{L}(\boldsymbol{\psi}_1, \boldsymbol{\phi}) &= -\mathbb{E}_{(\boldsymbol{\psi}_2, \boldsymbol{\psi}_3) \sim q_{\boldsymbol{\phi}}} \mathrm{KL}(q_{\boldsymbol{\psi}}(\boldsymbol{\theta}) || p(\boldsymbol{\theta}|\mathbf{y})) + \log\ p(\mathbf{y}) \\ &\geq -\mathrm{KL}(\mathbb{E}_{(\boldsymbol{\psi}_2, \boldsymbol{\psi}_3) \sim q_{\boldsymbol{\phi}}} q_{\boldsymbol{\psi}}(\boldsymbol{\theta}) || p(\boldsymbol{\theta}|\mathbf{y})) + \log\ p(\mathbf{y}) = \widetilde{\mathcal{L}}_0(\boldsymbol{\psi}_1, \boldsymbol{\phi})\end{aligned}$$

using Jensen's inequality with the observation that the KL-divergence can be viewed as a convex functional (a proof is provided in Yin and Zhou 2018, Appendix A).

$\tilde{\mathcal{L}}_0$ can be used to as a target optimize the variational parameters. In practice, the implicit distribution is implied by the transformation of some noise $\boldsymbol{\varepsilon} \sim \mathcal{D}$ (e.g. $\mathcal{D}$ is the k-dimensional standard normal distribution) with a deep neural net such that $(\boldsymbol{\psi}_2', \boldsymbol{\psi}_3')' = (T_{\boldsymbol{\phi}}(\boldsymbol{\varepsilon})_1', T_{\boldsymbol{\phi}}(\boldsymbol{\varepsilon})_2')' = T_{\boldsymbol{\phi}}(\boldsymbol{\varepsilon})$. However, Yin and Zhou (2018) show in Proposition 1 that optimizing $\tilde{\mathcal{L}}_0$ without early stopping can lead to a degenerated distribution for ψ_1, ψ_2, i.e. a distribution with a single point-mass. To avoid this, the author suggest to add a regularizing term to $\tilde{\mathcal{L}}_0$ yielding

$$\widetilde{\mathcal{L}}(\boldsymbol{\psi}_1, \boldsymbol{\phi}) = \mathbb{E}_{\boldsymbol{\varepsilon} \sim \mathcal{D}} \mathbb{E}_{\boldsymbol{\theta} \sim q_{(\boldsymbol{\psi}_1, T_{\boldsymbol{\phi}}(\boldsymbol{\varepsilon}))}} \mathbb{E}_{\boldsymbol{\varepsilon}^{(1)}, \ldots, \boldsymbol{\varepsilon}^{(K)} \sim \mathcal{D}} \Bigg[\log p(\boldsymbol{\theta}, \mathbf{y}) - \log q_{\boldsymbol{\psi}_1}(\boldsymbol{\theta}_1) - \log \prod_{i=2}^{3} \left(\frac{1}{K+1} \left(q_{T_{\boldsymbol{\phi}}(\boldsymbol{\varepsilon})_i}(\boldsymbol{\theta}_i | T_{\boldsymbol{\phi}}(\boldsymbol{\varepsilon})_i) + \sum_{k=1}^{K} q_{T_{\boldsymbol{\phi}}(\boldsymbol{\varepsilon}^{(k)})_i}(\boldsymbol{\theta}_i | T_{\boldsymbol{\phi}}(\boldsymbol{\varepsilon}^{(k)})_i) \right) \right) \Bigg] \quad (5)$$

as the target for optimization to which refer from here on as the lower bound ELBO (lbELBO).

Yin and Zhou (2018) show that with increasing K the lbELBO approaches the ELBO reaching equality for $K \to \infty$. The expectations in the lbELBO can be estimated via stochastic approximation. Note that the conditional densities $q_{T_{\boldsymbol{\phi}}(\boldsymbol{\varepsilon})_i}(\boldsymbol{\theta}_i | T_{\boldsymbol{\phi}}(\boldsymbol{\varepsilon})_i)$ can also include non-hierarchical variational parameters, e.g. $q_{T_{\boldsymbol{\phi}}(\boldsymbol{\varepsilon})_i, \boldsymbol{\psi}_{i,2}}(\boldsymbol{\theta}_i | T_{\boldsymbol{\phi}}(\boldsymbol{\varepsilon})_i)$ with additional fixed parameters $\boldsymbol{\psi}_{i,2}$. Finally, updates to the variational parameters are based on the respective gradients. The gradients are available via reverse-mode automatic differentiation exploiting the reparametrization trick (Kingma and Welling 2014). In particular, the updates at iteration ℓ are given by

$$\begin{aligned} \boldsymbol{\phi}^{(\ell)} &= \boldsymbol{\phi}^{(\ell-1)} + \rho_1^{(\ell)} \nabla_{\boldsymbol{\phi}} \widetilde{\mathcal{L}}(\boldsymbol{\psi}_1, \boldsymbol{\phi}), \\ \boldsymbol{\psi_1}^{(\ell)} &= \boldsymbol{\psi_1}^{(\ell-1)} + \rho_2^{(\ell)} \nabla_{\boldsymbol{\psi_1}} \widetilde{\mathcal{L}}(\boldsymbol{\psi}_1, \boldsymbol{\phi}), \end{aligned}$$

with exponential decaying learning rates $\rho_1^{(l)}, \rho_2^{(l)}$. Adding decaying learning rates improved numerical stability and showed better results overall.

The flexibility of the variational family in SIVI is only limited in two ways: First, the implicit variational prior distribution must be reparameterizable. That is, a distribution that can be sampled from using an auxiliary variable $\boldsymbol{\epsilon}$ that is transformed through a differentiable transformation $T(.)$ e.g. $\boldsymbol{\psi} = T_{\boldsymbol{\phi}}(\boldsymbol{\epsilon})$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. Second, the conditional variational distribution of the coefficients must be analytic and reparameterizable or, as demonstrated in Yin and Zhou (2018), the ELBO must be analytic.

3.2 Mean-field and semi-implicit VI for additive models

In this section, we discuss the concrete implementations of two MFVI approaches and two SIVI approaches for the additive model.

3.2.1 Mean-field VI with block-diagonal covariance matrix

In the additive model, the mean field factorization could be blocked as follows $\vec{\theta} = (\vec{\gamma}_1', \ldots, \vec{\gamma}_p', \tau_1, \ldots, \tau_p, \sigma^2)'$ with the variational density given by the factors

$$q_{\boldsymbol{\psi}}(\boldsymbol{\theta}) = q_{\boldsymbol{\psi}_{\sigma^2}}(\sigma^2) \prod_{j=1}^{p} q_{\boldsymbol{\psi}_{\boldsymbol{\gamma}_j}}(\boldsymbol{\gamma}_j)\, q_{\boldsymbol{\psi}_{\tau_j^2}}(\tau_j^2), \quad q_{\boldsymbol{\psi}} \in \mathcal{Q}_{\text{MFb}},$$

(Waldmann and Kneib 2015). Using Equation (3) and exploiting conditional conjugacy results in each $q_{\boldsymbol{\psi}_{\boldsymbol{\gamma}_j}}$ to be multivariate Gaussian and the remaining distributions as inverse gamma parametrized with the parameter vector in the index.

This formulation allows the construction of iterative updates to the variational parameters as follows: For $\boldsymbol{\psi}_{\boldsymbol{\gamma}_j}$ as re-parametrization of the mean vector and covariance matrix, i.e. $\boldsymbol{\psi}_{\boldsymbol{\gamma}_j} = (\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, in the variational distribution of $\boldsymbol{\gamma}_j$, the updates are

$$\begin{aligned}
\boldsymbol{\Sigma}_j &= \left(\frac{\nu_{a_{\sigma^2}}}{\nu_{b_{\sigma^2}}} \boldsymbol{Z}_j' \boldsymbol{Z}_j + \frac{\nu_{a_j}}{\nu_{b_j}} \boldsymbol{K}_j \right)^{-1}, \text{ and} \\
\boldsymbol{\mu}_j &= \frac{\nu_{a_{\sigma^2}}}{\nu_{b_{\sigma^2}}} \boldsymbol{\Sigma}_j \boldsymbol{Z}_j' \left(\boldsymbol{y} - \boldsymbol{Z}_{-j} \boldsymbol{\mu}_{-j} \right] \Big).
\end{aligned}$$

The variational distribution $q^*(\sigma^2)$ of the error variance is $\text{IG}(\nu_{a_{\sigma^2}}, \nu_{b_{\sigma^2}})$ and the updates for the variational parameters are

$$\nu_{a_{\sigma^2}} = a_{\sigma^2} + \frac{N}{2}, \tag{6}$$

$$\nu_{b_{\sigma^2}} = b_{\sigma^2} + \frac{1}{2} \Big((\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\mu})'(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\mu}) + \text{tr}\left(\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Sigma} \right) \Big), \tag{7}$$

where $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_p)$ and $\boldsymbol{\mu} = (\boldsymbol{\mu}_1', \ldots, \boldsymbol{\mu}_p')'$ are the stacked design matrices and mean vectors, respectively. For the covariance matrix $\boldsymbol{\Sigma}$, we can rewrite the component-wise covariance matrices into one block diagonal covariance matrix, i.e. $\boldsymbol{\Sigma} = \text{blockdiag}(\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_p)$. Therefore, we call this approach MFVI with block-diagonal covariance matrix MFVI (block).

The variational distributions $q^*(\tau_j^2)$ of the smoothing parameters are $\text{IG}(\nu_{a_j}, \nu_{b_j})$, $\forall\ j = 1, \ldots, p$ and the updates for the variational parameters are

$$\nu_{a_j} = a_j + \frac{\text{rank}(\boldsymbol{K}_j)}{2}, \tag{8}$$

$$\nu_{b_j} = b_j + \frac{1}{2} \left(\text{tr}\left(\boldsymbol{K}_j \boldsymbol{\Sigma}_j \right) + \boldsymbol{\mu}_j' \boldsymbol{K}_j \boldsymbol{\mu}_j \right). \tag{9}$$

The full derivations for the variational distribution of the coefficients, the error variance and the smoothing parameters are shown in Appendix Sect. 7.4.

3.2.2 Mean-field VI with full covariance matrix

In the special case of a multivariate Gaussian distribution for the coefficients, we can use a single multivariate Gaussian distribution for all coefficients such that the variational density factors to

$$q_{\boldsymbol{\psi}}(\boldsymbol{\theta}) = q_{\boldsymbol{\psi}_{\sigma^2}}(\sigma^2)\, q_{\boldsymbol{\psi}_{\boldsymbol{\gamma}}}(\boldsymbol{\gamma}) \prod_{j=1}^{p} q_{\boldsymbol{\psi}_{\tau_j^2}}(\tau_j^2), \quad q_{\boldsymbol{\psi}} \in \mathcal{Q}_{\text{MFf}},$$

with multivariate Gaussian distribution $q_{\boldsymbol{\psi}_{\boldsymbol{\gamma}}}$. The iterative updates to the variational parameters are,

$$\begin{aligned} \boldsymbol{\Sigma} &= \left(\frac{\nu_{a_{\sigma^2}}}{\nu_{b_{\sigma^2}}} \boldsymbol{Z}'\boldsymbol{Z} + \boldsymbol{K} \right)^{-1}, \\ \boldsymbol{\mu} &= \frac{\nu_{a_{\sigma^2}}}{\nu_{b_{\sigma^2}}} \boldsymbol{\Sigma} \boldsymbol{Z}' \boldsymbol{y}, \\ \text{with } \boldsymbol{K} &= \text{blockdiag}\left(\frac{\nu_{a_1}}{\nu_{b_1}} \boldsymbol{K}_1, \ldots, \frac{\nu_{a_p}}{\nu_{b_p}} \boldsymbol{K}_p \right). \end{aligned}$$

As the covariance matrix $\boldsymbol{\Sigma}$ is a full and unstructured covariance matrix we call the approach MFVI with a full covariance matrix MFVI (full). The updates for the error variance and the smoothing parameters are the same as in MFVI (block).

In MFVI (full), the mean-field assumption plays a crucial part but is not very restrictive. First, the assumption of independence between the error variance and the coefficients is only a mild assumption. For instance, in the case of a diminishing penalty term that is close to zero, we can use the properties of ordinary least squares. That is, all columns of the design matrix are orthogonal to the residuals. For larger influences of the penalty term, however, this assumption is violated. Second, the assumption that smoothing parameters for each component are independent and that they are independent of the error variance and conditionally on the coefficients only lead to a mild restriction as this assumption is only imposed on the hyper-parameters. However, MFVI (full) is computationally very demanding for large variational covariance matrices $\boldsymbol{\Sigma}$ because inverting the $m \times m$ matrix $\boldsymbol{\Sigma}$ involves $O(m^3)$ computations. Furthermore, MFVI (full) as discussed here in the case of conditional conjugate models is not very flexible, as it is limited to the case of a multivariate Gaussian variational distribution for the coefficients of all additive components. For smaller data sets, however, a multivariate Gaussian variational distribution may not capture heavier tails.

3.2.3 Semi-implicit VI

Using SIVI has the advantage of retaining a blocked covariance matrix, but at the same time increasing the flexibility of the variational distribution for the coefficients and thus restoring coefficient dependencies. Additionally, more complex posterior distributions can be captured. The variational density is

$$q_{\boldsymbol{\phi},\boldsymbol{\nu}}(\boldsymbol{\theta}) = q_{\boldsymbol{\psi}_{\sigma^2}}(\sigma^2) \prod_{j=1}^{p} q_{\boldsymbol{\psi}_{\tau_j^2}}(\tau_j^2) \int \left(\prod_{j=1}^{p} q_{\boldsymbol{\psi}_{\boldsymbol{\gamma}_j}}(\boldsymbol{\gamma}_j|\boldsymbol{\psi}_{\boldsymbol{\gamma}_j}) \right) q_{\boldsymbol{\phi}}(\boldsymbol{\psi}_{\boldsymbol{\gamma}}) \mathrm{d}\boldsymbol{\psi}_{\boldsymbol{\gamma}}, \quad q_{\boldsymbol{\phi},\boldsymbol{\nu}} \in \mathcal{Q}_{\text{HVM}}, \tag{10}$$

with $\boldsymbol{\nu} = (\boldsymbol{\psi}_{\sigma^2}, \boldsymbol{\psi}_{\tau_1^2}, \ldots, \boldsymbol{\psi}_{\tau_p^2})$. In this way the variational mean-field parameters for the coefficients, $\boldsymbol{\psi}_{\boldsymbol{\gamma}}$, are marginalized out. The p coefficient blocks are designed to be independent conditioned on $\boldsymbol{\psi}_{\boldsymbol{\gamma}}$, marginally, however, dependencies between the blocks can be captured.

In line with the variational distributions of MFVI, we choose an inverse Gamma distribution for $\sigma^2 \sim \text{IG}(\nu_{a_{\sigma^2}}, \nu_{b_{\sigma^2}})$ with $\boldsymbol{\psi}_{\sigma^2} = (\nu_{a_{\sigma^2}}, \nu_{b_{\sigma^2}})$ and for each $\tau_j^2 \sim \text{IG}(\nu_{a_j}, \nu_{b_j})$ with $\boldsymbol{\psi}_{\tau_j^2} = (\nu_{a_j}, \nu_{b_j})$. For the conditional variational distribution, we use a multivariate Gaussian distribution, i.e. $\boldsymbol{\gamma}_j|\boldsymbol{\psi}_{\boldsymbol{\gamma}_j} \sim \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_{\boldsymbol{\xi}_j})$ with $\boldsymbol{\psi}_{\boldsymbol{\gamma}_j} = (\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_{\boldsymbol{\xi}_j})$. As the optimization was numerically unstable for conditioning on both $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_{\boldsymbol{\xi}_j}$ (also conditioning on variational parameters of smoothing parameters and error variance lead to unreliable results), we apply the hierarchical expansion only on the variational parameters $\boldsymbol{\mu} = (\boldsymbol{\mu}_1', \ldots, \boldsymbol{\mu}_p')'$, i.e. $\boldsymbol{\gamma}_j|\boldsymbol{\mu}_j \sim \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_{\boldsymbol{\xi}_j})$.

The variational parameter of the coefficients are generated as $\boldsymbol{\mu} = T_{\boldsymbol{\phi}}(\boldsymbol{\epsilon})$, with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and where $T_{\boldsymbol{\phi}}$ transforms variables $\boldsymbol{\epsilon}$ through a deep neural network with weight and bias parameters $\boldsymbol{\phi}$.

The covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\xi}_j}$ is parameterized with $\boldsymbol{\xi}_j$. It is the vectorized upper triangular matrix and part of the Cholesky decomposition to build the covariance matrix, i.e. $\boldsymbol{\Sigma}_{\boldsymbol{\xi}_j} = g_U(\boldsymbol{\xi}_j)' g_U(\boldsymbol{\xi}_j)$, where g_U forms a Cholesky factor from its argument vector.

For more details about the structure of SIVI, we provide an illustration in Fig. 6 in Appendix Sect. 9.1 and a pseudo-algorithm in Algorithm 2 in Appendix Sect. 9.

The updates are determined by using a gradient-based approach. For variational hyper-parameters we use the Adam optimizer (Kingma and Ba 2015) and for the parameters $\boldsymbol{\xi}$ and $\boldsymbol{\nu}$ we use a stochastic gradient descend optimizer. In order to estimate the gradients, we rely on the lbELBO for additive models given by

$$\begin{aligned}
&\widetilde{\mathcal{L}}(\boldsymbol{\phi}, \boldsymbol{\xi}, \boldsymbol{\nu}) \\
&\approx \sum_{j=1}^{p} \mathbb{E}_{\tau_j^2 \sim q_{\nu_{a_j}, \nu_{b_j}}} \left[\ln\, p(\tau_j^2) \right] + \mathbb{E}_{\sigma^2 \sim q_{\nu_{a_{\sigma^2}}, \nu_{b_{\sigma^2}}}} \left[\ln\, p(\sigma^2) \right] \\
&\quad - \sum_{j=1}^{p} \mathbb{E}_{\tau_j^2 \sim q_{\nu_{a_j}, \nu_{b_j}}} \left[\ln\, q_{\nu_{a_j}, \nu_{b_j}}(\tau_j^2) \right] - \mathbb{E}_{\sigma^2 \sim q_{\boldsymbol{\psi}_{\sigma^2}}} \left[\ln\, q_{\nu_{a_{\sigma^2}}, \nu_{b_{\sigma^2}}}(\sigma^2) \right] \\
&\quad + \frac{1}{S} \sum_{s=1}^{S} \left\{ \mathbb{E}_{\sigma^2 \sim q_{\nu_{a_{\sigma^2}}, \nu_{b_{\sigma^2}}}} \left[\ln\, p(\mathbf{y}|\boldsymbol{\gamma}_{.,s}, \sigma^2) \right] \right.
\end{aligned}$$

$$+\sum_{j=1}^{p}\mathbb{E}_{\tau_j^2\sim q_{v_{a_j},v_{b_j}}}\left[\ln\, p(\boldsymbol{\gamma}_{j,s}|\tau_j^2)\right]$$
$$-\ln\Bigg(\prod_{j=1}^{p}\frac{1}{K+1}\Big(q_{T_{\boldsymbol{\phi}}(\boldsymbol{\varepsilon}_s)_j,\xi_j}(\boldsymbol{\gamma}_{j,s}|T_{\boldsymbol{\phi}}(\boldsymbol{\varepsilon}_s)_j)$$
$$+\sum_{k=1}^{K}q_{T_{\boldsymbol{\phi}}(\boldsymbol{\varepsilon}_s)_j,\xi_j}(\boldsymbol{\gamma}_j|T_{\boldsymbol{\phi}}(\boldsymbol{\varepsilon}^{(k)})_j)\Big)\Bigg)\Bigg\}, \tag{11}$$

with $\boldsymbol{\gamma}_{.,s}$ as the s-th sample of the stacked coefficient vector. All parts that include the expectation with respect to σ^2 or each τ_j^2 of Formula (11) are available in analytic form. Hence, we can use the same analytic derivations as for MFVI. The expectation with respect to parameters $\boldsymbol{\gamma}_1,\ldots,\boldsymbol{\gamma}_p$ is not tractable and needs to be approximated. We use stochastic approximation by taking S samples of each $\boldsymbol{\gamma}_{j,s}$ and average over them.

The choices of the variational distributions are in accordance with MFVI for reasons of comparison and the choice enables an analytic solution for MFVI. However, different MFVI algorithms have been developed to go beyond conditional conjugate models. In the case of SIVI, any arbitrary variational distribution for the error variance and smoothing parameter(s) can be used. Additionally, the conditional variational distribution is not restricted to be multivariate Gaussian, but other symmetric, as well as asymmetric distributions, can be considered.

3.2.4 Semi-implicit mean field VI

We propose an additional method which we call semi-implicit mean field variational inference (SIMFVI) that can be viewed as a hybrid between SIVI and MFVI. SIMFVI uses the same variational family as SIVI but the variational parameters are updated differently. Wherever possible, we use the analytical updates as in Equation (3) with noisy estimates of the expected values. For the additive model, this algorithm updates the variational hyperparameters with the gradient-based method as in SIVI. $\boldsymbol{\Sigma}_j$ and $\boldsymbol{v}$ are updated similar to the analytic MFVI (block) updates. Indeed, we need to use a stochastic approximation of the expectation with respect to $\boldsymbol{\gamma}_j$. Hence, the updates for the scale parameter of the error variance are,

$$v_{b_{\sigma^2}} = b_{\sigma^2}+\frac{1}{2}\mathbb{E}_{\vec{\gamma}\sim q_{\vec{\psi}_{\vec{\gamma}}}}\left[(\vec{y}-\vec{Z\gamma})'(\vec{y}-\vec{Z\gamma})\right]$$
$$\approx b_{\sigma^2}+\frac{1}{2}\left(\vec{y}'\vec{y}-\frac{2}{S}\sum_{s=1}^{S}\vec{\gamma}_{.,s}\vec{Z}'\vec{y}+\frac{1}{S}\sum_{s=1}^{S}\vec{\gamma}_{.,s}'\vec{Z}'\vec{Z}\vec{\gamma}_{.,s}\right),$$

The updates for the scale parameter of the smoothing parameter for component j are,

$$v_{b_j} = b_j+\frac{1}{2}\mathbb{E}_{\vec{\gamma}\sim q_{\vec{\psi}_{\vec{\gamma}}}}\left[\vec{\gamma}_j'\vec{K}_j\vec{\gamma}_j\right]\approx b_j+\frac{1}{2S}\sum_{s=1}^{S}\vec{\gamma}_{j,s}'\vec{K}_j\vec{\gamma}_{j,s}.$$

We provide a full implementation of the described approaches using the Python programming language. The implementations incorporate the Python libraries `numpy` (Harris et al. 2020) and `pytorch` (Paszke et al. 2019) for generation of random numbers and automated differentiation.

4 Simulation

In the simulation study, we assess the uncertainty estimates across the introduced VI methods based on empirical coverage percentages over repeated simulations. The coverage percentage is the relative frequency of how many times the credible interval (CI) contains the true value. For a 95% CI, we, therefore, expect a coverage percentage that equals about the nominal level of 95%. In Bayesian inference the CI is either based on the highest density interval (Turkkan and Pham-Gia 1993) or on the quantiles of the distribution. In this paper the CI's are based on the quantiles.

In the case of MFVI (block), we suppose the coverage percentage will be below the nominal level due to the strong assumption of independence between coefficient blocks. We also presume that MFVI (full) accurately captures parameter uncertainty, but it comes with the limitation of determining a fully unstructured covariance matrix for all parameters simultaneously which requires handling of large matrices. On the other hand, SIVI and SIMFVI combine the advantage of using a blocked structure with a hierarchical construction to restore parameter dependencies. Therefore, we hypothesize the coverage percentage of SIVI and SIMFVI is close to the nominal level as well. Additionally, we compare the aforementioned methods with the Gibbs sampler, which we use as a reference.

We assess both, the coverage of point-wise and simultaneous CI. For estimating simultaneous CI, we develop an efficient algorithm (see Algorithm 2 in Appendix 8) based on a fully Bayesian quantile-based approach (Krivobokova et al. 2010).

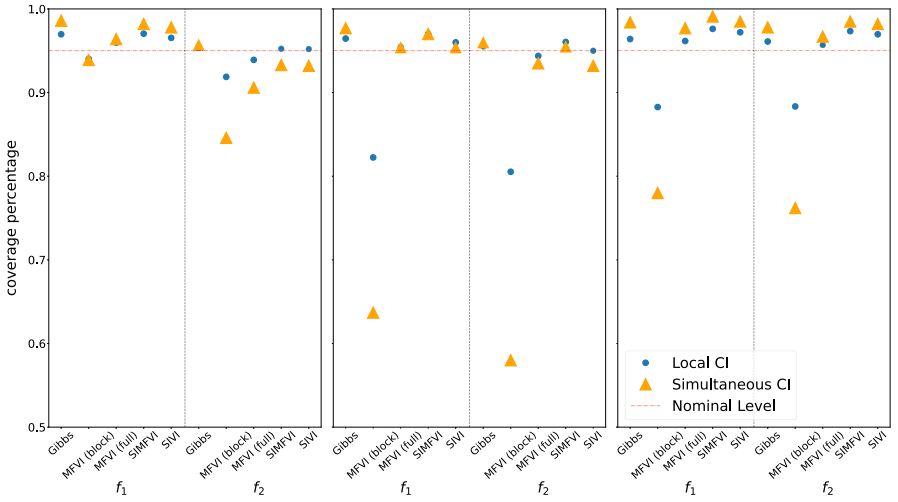
The data generating process (DGP) is based on two covariates that affect the response in a nonlinear way. Hence, for the model, we can use P-splines and can block the covariance matrix accordingly. The DGP has the form

$$\text{(DGP)} \quad y_i = f_1(x_{i1}) + f_2(x_{i2}) + \epsilon_i.$$

The errors are independently generated from a Gaussian distribution with variance 0.5, i.e. $\epsilon_i \sim \mathcal{N}(0, 0.5)$. The values for the covariates x_{i1} and x_{i2} are generated in two steps. In the first step, we generate values from a bivariate normal distribution, $(z_{i1}, z_{i2})' = \boldsymbol{z}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{A})$, with $\boldsymbol{A}$ having ones on the diagonal and the value $\rho \in (-1, 1)$ on the off-diagonals to control for correlations between the variables. We consider correlations of varied intensity, namely, no correlation ($\rho = 0$), medium correlation ($\rho = 0.45$), and strong correlation ($\rho = 0.9$). In the second step, we use a probability integral transform to the variables, such that $x_{i1} = 5 \cdot F(z_{i1})$ and $x_{i2} = 7 \cdot F(z_{i2}) - 1$, with $F(\cdot)$ as univariate standard normal cumulative distribution function.

The two functional forms of the nonlinear effects are,

$$f_1(x_{i1}) = \sin\left(\frac{\pi}{4} x_{i1} - 1\right) + 2\exp\left(-(x_{i1} - 1)^2\right), \tag{12}$$



(a) $N = 50, \rho = 0.45$ (b) $N = 50, \rho = 0.9$ (c) $N = 250, \rho = 0.9$

Fig. 1 Coverage percentage among different methods for three selected scenarios. The blue dots represent local and the yellow triangulars simultaneous CI coverages

$$f_2(x_{i2}) = \sin\left(\frac{3\pi}{16}\, x_{i2} - \frac{1}{2}\right) + 2\exp\left(-\frac{3}{2}(x_{i2} - \frac{1}{2})^2\right). \tag{13}$$

The functional forms f_1 and f_2 have a similar shape, which additionally increases the difficulty to distinguish between the two effects. Both functions have one sharp peak around 1.2 and 0.6, respectively (see Fig. 7, in Appendix Sect. 9.2).

Moreover, we vary the sample sizes in the simulation study. For each of the three scenarios with varying correlations, we run the simulation study using 50, 250, or 500 observations. This results in total to nine different scenarios. For each scenario, we use 1000 replications.

We argue that the simulation results of using P-splines are transferable to other effect types involving clustered or spatial effects as the coefficient structure in the model remains the same. Only the basis functions and the number of coefficients change. Hence, it appears to be sufficient to limit the extent of the simulation study to two non-linear effects modeled with P-splines.

Simulation results, depicted in Fig. 1, show the coverage percentage for each spline and method for three scenarios with the most significant differences across the methods. These are, not surprisingly, the scenarios with high or medium correlation and a rather small sample size (the results of the other scenarios are shown in Fig. 8 in Appendix Sect. 9.2).

In the scenario with a high correlation between the covariates (middle and right plot), MFVI (block) has a very low coverage for both splines. The simultaneous CI of the estimated function for f_2 in the scenario with 50 observations has a coverage of below 70% that is significantly below the nominal level of 95%. SIVI, SIMFVI and also MFVI (full) show coverages of about the nominal level. For the scenario

with medium correlation and 50 observations, however, the coverage percentage of simultaneous CI for f_2 for MFVI (block) as well as for MFVI (full) is well below the nominal level. This might be due to the fact that for small sample sizes the Gaussian distribution assumption on the coefficients might be too restrictive. The more flexible methods SIVI and SIMFVI show improvements in the coverage, but are also slightly below the nominal level.

For other criteria such as the mean squared error (MSE) for each spline and the overall MSE on the fitted values, no significant differences across the methods are visible. There is only a slight tendency that the different MSEs of SIVI and SIMFVI are larger on average. For instance, in the scenario with 50 observations and a strong correlation the smallest overall MSE value is the one of MFVI (block) with 0.116 and the largest is the one of SIVI with 0.123 (see Table 2 in Appendix Sect. 9.2 for further details about this scenario).

The simulations show very accurate results for the Gibbs sampler across all criteria and scenarios. However, the coverage percentage of the local and, in particular, simultaneous CIs, were above the nominal level by about 0.4 to 4.2 percentage points in all scenarios, indicating that the uncertainty is slightly overestimated.[1] In particular, the simultaneous CI bands of the estimated function for f_1 appear to be too wide. Nevertheless, we use the Gibbs sampler as the reference when comparing the methods in the application, as the MCMC approach is expected to give asymptotically exact results.

Assuming the samples of the MCMC approach to come from the desired posterior distribution, we also evaluate the KL divergence for each VI method based on these samples. This gives a more holistic evaluation over the complete distribution. We approximately evaluate,

$$-\,\mathrm{KL}\left(p(\boldsymbol{\theta}|\boldsymbol{y})||q(\boldsymbol{\theta})\right) \approx \frac{1}{S}\sum_{s=1}^{S}\log\, q(\boldsymbol{\theta}_s) - \frac{1}{S}\sum_{s=1}^{S}\log\, p(\boldsymbol{\theta}_s|\boldsymbol{y})$$

with Gibbs samples $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_S \sim p(\boldsymbol{\theta}|\boldsymbol{y})$. Since we compare the different VI methods based on the same samples we only need to evaluate the term $\frac{1}{S}\sum_{s=1}^{S}\log\, q(\boldsymbol{\theta}_s)$, that is the average logarithmized density given the Gibbs samples (ALDG). Higher values indicate better approximations to the true posterior. For SIVI and SIMFVI, we evaluate the density of the coefficients by averaging them out,

$$\log q(\boldsymbol{\theta}_s) \approx \sum_{j}^{p}\left[\log\frac{1}{\mathrm{M}}\sum_{m=1}^{\mathrm{M}}\left(q_{\boldsymbol{\mu}_{jm},\boldsymbol{\Sigma}_j}(\boldsymbol{\gamma}_{js}|\boldsymbol{\mu}_{jm})\right) + \log\, q_{a_j,b_j}(\tau_{js}^2)\right] + \log q_{a,b}(\sigma_s^2),$$

with M samples out of the neural network. For MFVI a full factorization applies.

The results confirm our previous findings. In Fig. 2 we show the ALDG of the coefficients for each VI method (we show figures of the total ALDG and for the different model parameters for all scenarios in Figs. 9, 10, 11, 12, 13, 14, 15, 16 and

[1] A tendency to overestimate the uncertainty is also found in other studies related to MCMC approximations (Fahrmeir et al. 2004)

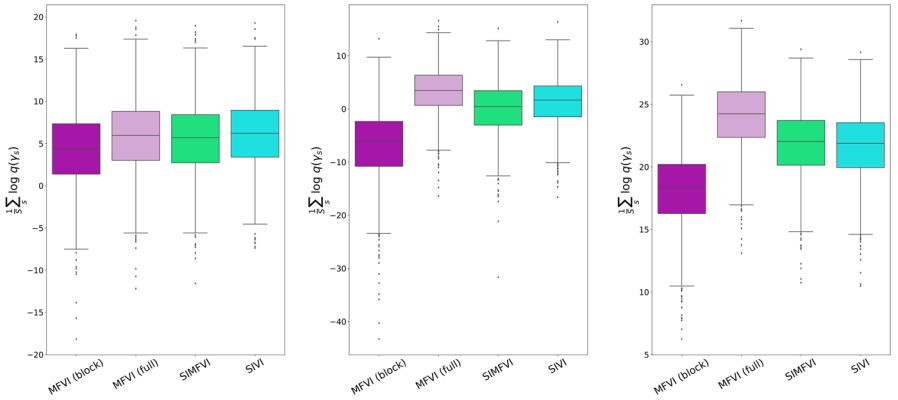


Fig. 2 The coefficient ALDG across all VI methods for three selected scenarios: 50 observations and medium correlation (left), 50 observations and high correlation (middle) and 250 observations and high correlation (right). The boxplots are based on 1000 simulations

17) in Appendix Sect. 9). MFVI (block) does not accurately capture the complete distribution, whereas SIVI and SIMFVI show significant improvements. However, evaluating the ALDG for high correlation between the covariates reveal that MFVI (full) performs slightly better. Hence, the hierarchical approach with flexible mixing distribution restores parameters dependencies to a large extend, but may fails to restore all dependencies.

The extend to how good SIVI and SIMFVI restores parameter dependencies is highly sensitive to the specifications of the neural network. Important considerations are the neural net structure and the activation function. We see a deterioration of the performance, if the neural net structure has more than 3 hidden layers and if the activation function is Sigmoid, instead of ReLU or Tanh. The specification for the input dimension, however, does not effect the results significantly (one example about the sensitivity is shown in Fig. 18 Appendix Sect. 9.2). Most importantly, however, is the choice of the learning rates for the parameter updates. In the simulation study it appears that in general higher learning rates (up to 0.1) improve the results, but the algorithm becomes less numerically stable. We show further details about the model set up and the choice of hyper-parameters in Appendix Sect. 7.7.

5 Application to tree height models of douglas fir

Douglas fir is a non-native conifer species to Germany. It is expected to be resilient to drought events and higher temperatures and, thus, with changing climatic conditions may serve as an important addition to the tree species portfolio of German climate-smart forestry. Modeling tree heights of Douglas fir is of high value for economic and climate consideration concerning, e.g. returns from investment and carbon storage potentials.

We use data from the national forestry inventory (NFI) of Germany and a climate data set provided by the Nordwestdeutsche Forstliche Versuchsanstalt.

We model heights of Douglas fir in Germany using two types of covariates, namely, tree- and climate-specific covariates. Some of those covariates have strong correlations. Accordingly, a method should be used that can reflect the increased uncertainty of the estimates. Therefore, we use the novel SIVI and SIMFVI methods and compare the results to the standard MFVI. Additionally, we use the Gibbs sampler as the benchmark method.

5.1 Statistical additive model

For modeling the mean tree height of Douglas fir, we employ an additive model (similar to Pya and Schmidt 2016). With the combined tree and climate data, we fit the following model for the observed tree height,

$$\begin{aligned} h_i &= \beta_0 + \beta_1 dbh_i + \beta_2 dbh_i^2 + f_1(age_i) \\ &\quad + f_2(prec_i) + f_3(t_i) + f_4(alt_i) + f_{\text{geo}}(long_i, lat_i) + \epsilon_i, \end{aligned} \tag{14}$$

where we assume a Gaussian distributed error ϵ_i.

The tree-specific data are the tree height in meters (h) that we use as response variable, the DBH in decimeters (dbh), and the age (age).

For all climate-specific variables, i.e. the accumulated precipitation per tree over its lifespan ($prec$) and the accumulated temperature per tree over its lifespan (t), and also for the adjusted altitude per location (alt) we use nonlinear effects as well.

Finally, to account for the spatial effect, we include a tensor product spline with the approximated coordinates for each tract, i.e. the longitude and latitude ($long, lat$). We provide more details about the model set up in Appendix Sect. 7.7.

We further split the data into 70% training and 30% test data to evaluate the predictive performance of each method. In total, we have 7,082 in the training and 3,035 observations in the test data set and overall 826 coefficients in the model.

5.2 Results

The results, as shown in Fig. 3, reveal the importance of considering climate variables to understand the tree heights of Douglas fir. Both increasing accumulated precipitation and temperature increase the expected tree height when looking at moderate values of the covariates. In case of more extreme values for accumulated temperature, the effect on tree height is less clear. For high values of accumulated precipitation, the estimated effect starts to decrease.

The estimated effect of age shows an expected functional form, basically following the typical height growth pattern over age, i.e. large height increment in younger ages and levelling off in older ages. The altitude of the tree location does not seem to play an important role.

When comparing the different proposed methods, both similarities and distinct differences are visible. The estimates for the mean effects are similar across all methods. Only SIVI deviates in some parameters. For higher values of the estimated mean effect of altitude, SIVI tends to be slightly below the other methods. Additionally, the mean

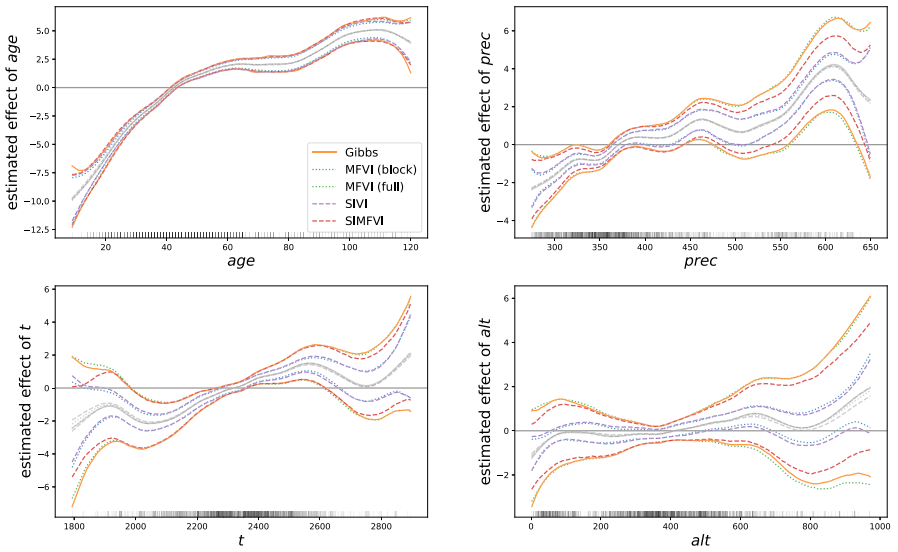


Fig. 3 Estimated effects with 95% simultaneous CIs colored by method

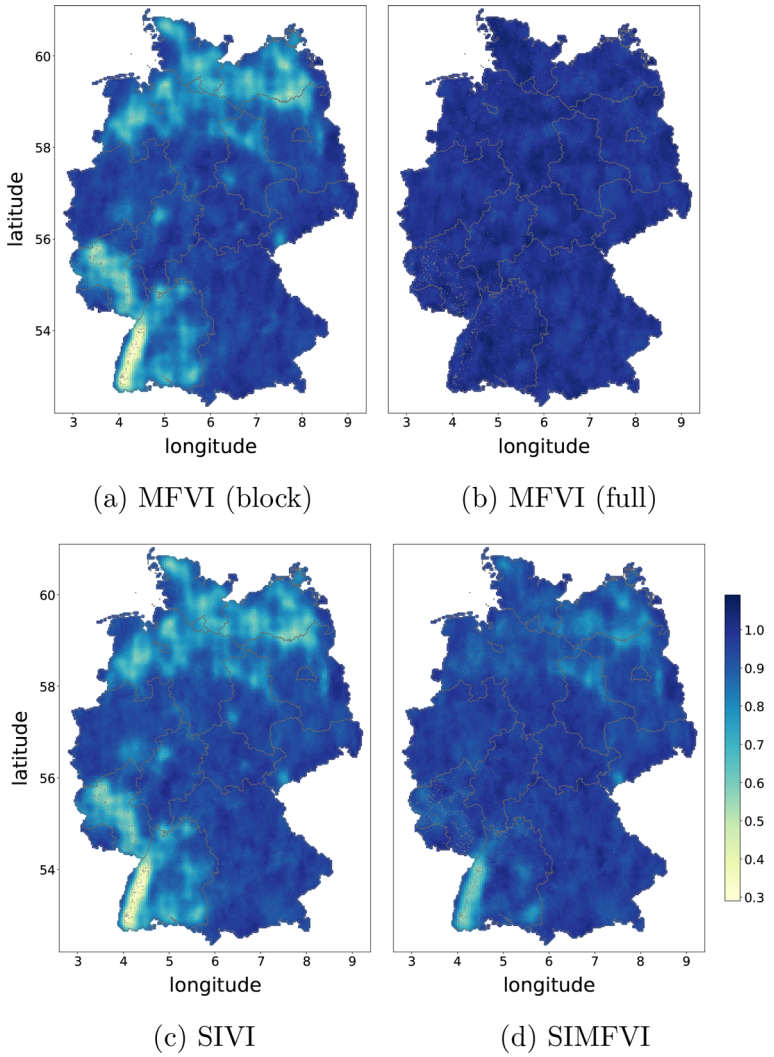
of the error variance term for SIVI is with about 7.97 estimated higher compared to all other methods which estimates are in the range between 7.59 and 7.61 (see Appendix Sect. 9.3, Table 4 for more information).

The differences between the methods become more apparent when considering the CIs, in particular for the estimated effect of accumulated precipitation, accumulated temperature, and altitude. Here, we deal with the additional problem, that the covariates are correlated. The results show a similar pattern for correlated effects as discovered in the simulation study: the widths of the 95% simultaneous CI bands of MFVI (block) are narrower compared to the other methods. However, as opposed to the simulation, also SIVI is not able to match the CI widths of the Gibbs sampler. SIMFVI and MFVI (full) show results much closer to the one from Gibbs sampler. In regions with only a few observations, SIMFVI tends to have narrower CI bands compared to Gibbs and MFVI (full).

The biggest differences are in the CI bands of altitude. For MFVI (block) and SIVI, the CI bands are at one occasion above and at one below the zero line. Whereas for Gibbs, MFVI (full), and SIMFVI, the CI bands cover the zero line across all values of altitude.

Similarly, the spatial effects for the different methods show differences in uncertainty levels (see Fig. 4). To highlight the differences in CI width, we visualize the CI width of each VI method as share to the width of the Gibbs sampler. Darkblue areas mean the CI is on par with Gibbs and yellow means the CI width is about 30% of the one from Gibbs, which is the lowest measured share on a location.

The biggest differences are in south-west Germany. MFVI (block) and also SIVI show much narrower CI widths at these locations compared to the other methods. Similarly, in north Germany, the areas of MFVI (block) and SIVI are lighter shaded. In areas without observations (without grey dots), there appears to be no difference between the methods.



(a) MFVI (block) (b) MFVI (full)

(c) SIVI (d) SIMFVI

Fig. 4 Width of 95% simultaneous CI of the two-dimensional spline as share to the CI width of Gibbs sampler across different VI methods. 100% (darkblue) stands for the same width as Gibbs sampler. Tracts with at least one Douglas fir are marked as grey dots

The accuracy of parameter uncertainty can still be improved in SIMFVI,[2] when altering some of the parameters for the algorithm. In particular, the number of samples K out of the neural net that is used to approximate the ELBO from below. Higher values further tighten the lower bound but to the expense of computational time. We opt to draw $K = 100$ samples out of the neural net for SIVI and SIMFVI, but increasing the number to $K = 300$, brings the CI width slightly closer to the one of Gibbs (see an example of the spatial effect with SIMFVI in Fig. 20 in Appendix Sect. 9.3). However, it significantly comes to the expense of computational time.

[2] For this application, the accuracy of parameter uncertainty is not improved with SIVI.

Table 1 Estimated ALDG across VI methods

	MFVI (block)	MFVI (full)	SIMF K=100	SIMF K=300	SIVI MFVI init. K=100
$\frac{1}{S}\sum_{s=1}^{S} \log\, q(\boldsymbol{\theta}_s)$	−2020.33	−1796.92	−1827.44	−1822.59	−1968.47
$\frac{1}{S}\sum_{s=1}^{S} \log\, q(\boldsymbol{\gamma}_s)$	−2003.32	−1779.98	−1809.08	−1803.98	−1951.39
$\frac{1}{S}\sum_{s=1}^{S} \log\, q(\boldsymbol{\tau}_s^2)$	−17.56	−17.53	−18.92	−19.18	−17.56
$\frac{1}{S}\sum_{s=1}^{S} \log\, q(\sigma_s^2)$	0.56	0.58	0.57	0.57	0.48

Numbers highlighted in green and yellow are the best and second best approximations to the Gibbs posterior, respectively. Additionally, the model parameter contributions to the ALDG are listed

The improvement of an increase in K is only marginal as can be seen when evaluating the whole distribution based on the Gibbs samples (see Table 1). SIMFVI with $K = 300$ is marginally closer to but still slightly worse than MFVI (full). In general, the ALDG confirms the findings of improved approximations in SIMFVI over MFVI (block). There is a substantial gap between the ALDG of those 2 methods, whereas SIVI shows only minor improvements.

Finally, we compare the predictive performance of each method. The predictive power is similar across all methods. For the MSE and the predictive coverage on the test data, we do not find significant distinctions between the methods. The predictive coverage is about 93 to 96% for all methods just as expected from the nominal level (see Table 5 in Appendix Sect. 9.3 for more details).

6 Conclusion

As there is growing access to ever more data resources and with it a growing interest in fast approximate methods, variational inference has gained considerably in popularity. In our analyses of additive models, variational inference performs well in terms of point estimates and parameter uncertainty, even despite making use of the strong mean-field assumption. However, the performance might degrade, and in particular, the parameter uncertainty is underestimated, if the mean-field assumption is placed on critical parameters, such as different coefficient blocks. This, might nevertheless be of interest, as treating all coefficients simultaneously, might need handling and estimation of large matrices, in particular when using combinations of spatial and cluster effects in a model, that requires numerous coefficients.

The SIVI and SIMFVI algorithms proposed are capable of using a blocked structure on the coefficients, but they still give accurate results on parameter uncertainty. In cases when a variational Gaussian distribution on the coefficients is too restrictive, SIVI and SIMFVI can even outperform MFVI with a full covariance structure, due to allowing more flexibility on the coefficients posterior distribution. Yet, the performance of SIVI seems to deteriorate, if dealing with large matrices from e.g. spatial effects or a large number of observations. In these cases, the gradient based approach to estimate parameters of the covariance matrices appears to be rather inefficient. The SIMFVI algorithm solves this issue and additionally, needs less computational time.

There is, however, still much room for improvement when considering the computational time of SIVI and SIMFVI. More efficient implementations could make SIVI and SIMFVI much faster.

We only investigate the use case of blocked versus fully unstructured covariance matrices for the coefficients. Future research can address more complex scenarios including complex hierarchical models and extensions to generalized additive models.

Acknowledgements The authors gratefully acknowledge funding via the Deutsche Forschungsgemeinschaft for the Research Training Group 2300 "Enrichment of European beech forests with conifers". Furthermore, we thank Nordwestdeutsche Forstliche Versuchsanstalt and in particular, Jan Schick, who is also part of the Research Training Group 2300 (subproject 9), for the provision of the climate data set adjusted for each stand.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability The tree specific data is available online, https://bwi.info/Download/de/BWI-Basisdaten/. The climate specific data is not freely available.

Code availability All code is available upon request from the authors. The code will be publicly available with the publication of the manuscript.

Declarations

Conflict of interest The authors declare no conflict of interest.

7 Design matrices, model and derivations

7.1 Bivariate tensor product P-spline

A two-dimensional surface is fitted by allowing the smooth of one continuous covariate i.e. $\boldsymbol{x}_{j1}$ with basis function $B(\boldsymbol{x}_{j1}) = \boldsymbol{Z}_{j1}$ to vary smoothly with another continuous covariate $\boldsymbol{x}_{j2}$ with basis function $B(\boldsymbol{x}_{j2}) = \boldsymbol{Z}_{j2}$. This is achieved by combining the two univariate smooth design matrices $\boldsymbol{Z}_{j1}$ and $\boldsymbol{Z}_{j2}$ with the Kronecker product ($\otimes$) for each row. The ith row of the constructed bivariate design matrix $\boldsymbol{Z}_j$ becomes,

$$\boldsymbol{z}_{ij} = \boldsymbol{z}_{ij1} \otimes \boldsymbol{z}_{ij2}.$$

The penalty is constructed by combining the univariate smooth penalties. The resulting penalty matrix that accounts for row-wise and column-wise differences is given by:

$$\boldsymbol{K}_j = \boldsymbol{I}_{d_{j2}} \otimes \boldsymbol{K}_{j1} + \boldsymbol{K}_{j2} \otimes \boldsymbol{I}_{d_{j1}}$$

7.2 Centering constraint in design matrix

To ensure identifiability in a model with more then one nonlinear predictor component we set the constraint (Wood 2017, chapter 1.8.1 & 4.2):

$$\mathbf{1}'\boldsymbol{Z}_u\,\boldsymbol{\gamma}_u = \mathbf{0},$$

where $\boldsymbol{Z}_u$ is the unconstraint $N \times p$ design matrix (so $\mathbf{1}'\boldsymbol{Z}_u$ is a $1 \times p$ vector) and unconstraint coefficients $\boldsymbol{\gamma}_u$. One way of imposing the constraint is by using $p-1$ unconstraint parameters with the QR decomposition. The column sums of the design matrix can be factored to:

$$\boldsymbol{Z}_u'\mathbf{1} = \boldsymbol{U}\begin{pmatrix}\boldsymbol{a}\\ \mathbf{0}\end{pmatrix},$$

where $\boldsymbol{U}$ is a $p \times p$ orthogonal matrix and $\boldsymbol{a}$ is a scalar. As a next step, $\boldsymbol{U}$ can be partitioned to $\boldsymbol{U} = (\boldsymbol{D} : \boldsymbol{C})$ where $\boldsymbol{C}$ is a $p \times (p-1)$ matrix. Then

$$\boldsymbol{C}\,\boldsymbol{\gamma}_u = \boldsymbol{\gamma},$$

will meet the constraints for any value of the $p-1$ dimensional vector $\boldsymbol{\gamma}_u$.
The new design matrix is:

$$\boldsymbol{Z} = \boldsymbol{Z}_u\boldsymbol{C},$$

such that,

$$\mathbf{1}'\boldsymbol{Z}\boldsymbol{\gamma} = \mathbf{0}$$

7.3 Model likelihood and priors

1. Likelihood for $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{Z}\boldsymbol{\gamma}, \sigma^2\boldsymbol{I}_n)$:

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{\gamma}, \sigma^2) &= \prod_{i=1}^{\mathrm{n}} \mathcal{N}(y_i|\boldsymbol{\gamma}, \sigma^2) \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{\mathrm{n}}{2}}} \exp\left(-\frac{(\boldsymbol{y}-\boldsymbol{Z}\boldsymbol{\gamma})'(\boldsymbol{y}-\boldsymbol{Z}\boldsymbol{\gamma})}{2\sigma^2}\right) \end{aligned}$$

2. Prior distribution for the error variance $\sigma^2 \sim \mathrm{IG}(a_{\sigma^2}, b_{\sigma^2})$:

$$p(\sigma^2) = \frac{b_{\sigma^2}^{a_{\sigma^2}}}{\Gamma(a_{\sigma^2})}\left(\frac{1}{\sigma^2}\right)^{a_{\sigma^2}+1} \exp\left(-\frac{b_{\sigma^2}}{\sigma^2}\right)$$

3. Prior distribution for coefficients $\boldsymbol{\gamma}_j$:

$$p(\boldsymbol{\gamma}_j|\tau_j^2) \propto \frac{1}{\left(2\pi\tau_j^2\right)^{\frac{\mathrm{rank}(\boldsymbol{K}_j)}{2}}} \exp\left(-\frac{\boldsymbol{\gamma}_j'\boldsymbol{K}_j\boldsymbol{\gamma}_j}{2\tau_j^2}\right)$$

4. Prior distribution for smoothing parameter of the $\tau_j^2 \sim \text{IG}(a_j, b_j)$:

$$p(\tau_j^2) = \frac{b_j^{a_j}}{\Gamma(a_j)} \left(\frac{1}{\tau_j^2}\right)^{a_j+1} \exp\left(-\frac{b_j}{\tau_j^2}\right)$$

7.4 Derivation of variational densities for MFVI

For notational convenience, we further do not add the distribution to the model parameter, when using the expectation.. Instead of e.g. $\mathbb{E}_{\boldsymbol{\gamma} \sim q_{\boldsymbol{\psi}_{\boldsymbol{\gamma}}}}$, we write $\mathbb{E}_{\boldsymbol{\gamma}}$.

1. Variational density for $\boldsymbol{\gamma}$ coefficients with $\boldsymbol{Z}\boldsymbol{\gamma}_{-j} = \sum_{r \neq j} \boldsymbol{Z}_r \boldsymbol{\gamma}_r$. Therefore, we have $\mathbb{E}_{-\boldsymbol{\gamma}_j}\left[\boldsymbol{Z}\boldsymbol{\gamma}_{-j}\right] = \sum_{r \neq j} \boldsymbol{Z}_r \mathbb{E}_{\boldsymbol{\gamma}_r}\left[\boldsymbol{\gamma}_r\right]$:

$$\begin{aligned}
q(\boldsymbol{\gamma_j}) &\propto \exp\Bigg\{\mathbb{E}_{-\boldsymbol{\gamma}_j}\Bigg[\ln p(\boldsymbol{y}|\boldsymbol{\gamma}, \sigma^2)\, p(\boldsymbol{\gamma}_j|\tau_j^2) \underbrace{p(\boldsymbol{\gamma}_{-j}|\tau_{-j}^2)\, p(\tau_j^2)\, p(\sigma^2)}_{\text{const}}\Bigg]\Bigg\} \\
&\propto \exp\Bigg\{\mathbb{E}_{-\boldsymbol{\gamma}_j}\Bigg[\ln\Bigg(\frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \\
&\quad \exp\left(-\frac{(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\gamma}_{-j} - \boldsymbol{Z}_j\boldsymbol{\gamma}_j)'(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\gamma}_{-j} - \boldsymbol{Z}_j\boldsymbol{\gamma}_j)}{2\sigma^2}\right) \\
&\quad \frac{1}{\left(2\pi\tau^2\right)^{\frac{d_j-1}{2}}} \exp\left(-\frac{1}{2\tau_j^2}\boldsymbol{\gamma}_j'\boldsymbol{K}_j\boldsymbol{\gamma}_j\right)\Bigg)\Bigg]\Bigg\} \\
&= \exp\Bigg\{\mathbb{E}_{-\boldsymbol{\gamma}_j}\Bigg[-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\gamma}_{-j} - \boldsymbol{Z}_j\boldsymbol{\gamma}_j)'(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\gamma}_{-j} - \boldsymbol{Z}_j\boldsymbol{\gamma}_j) \\
&\quad -\frac{1}{2\tau_j^2}\boldsymbol{\gamma}_j'\boldsymbol{K}_j\boldsymbol{\gamma}_j + \text{const}\Bigg]\Bigg\} \\
&\propto \exp\Bigg\{\mathbb{E}_{-\boldsymbol{\gamma}_j}\Bigg[-\frac{1}{2\sigma^2}\Big(\underbrace{\boldsymbol{y}'\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\gamma}'_{-j}\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{Z}\boldsymbol{\gamma}_{-j} + \boldsymbol{\gamma}'_{-j}\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\gamma}_{-j}}_{\text{does not contain } \boldsymbol{\gamma}_j} \\
&\quad -2\boldsymbol{\gamma}_j'\boldsymbol{Z}_j'\boldsymbol{y} + 2\boldsymbol{\gamma}_j'\boldsymbol{Z}_j'\boldsymbol{Z}\boldsymbol{\gamma}_{-j} + \boldsymbol{\gamma}_j'\boldsymbol{Z}_j'\boldsymbol{Z}_j\boldsymbol{\gamma}_j\Big) \\
&\quad -\frac{1}{2\tau_j^2}\boldsymbol{\gamma}_j'\boldsymbol{K}_j\boldsymbol{\gamma}_j\Bigg]\Bigg\} \\
&\propto \exp\Bigg\{\mathbb{E}_{-\boldsymbol{\gamma}_j}\Bigg[\frac{1}{\sigma^2}\boldsymbol{\gamma}_j'\boldsymbol{Z}_j'(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\gamma}_{-j}) - \frac{1}{2\sigma^2}\boldsymbol{\gamma}_j'\boldsymbol{Z}_j'\boldsymbol{Z}_j\boldsymbol{\gamma}_j - \frac{1}{2\tau_j^2}\boldsymbol{\gamma}_j'\boldsymbol{K}_j\boldsymbol{\gamma}_j\Bigg]\Bigg\} \\
&= \exp\Bigg\{\mathbb{E}_{\sigma^2}\left[\frac{1}{\sigma^2}\right]\boldsymbol{\gamma}_j'\boldsymbol{Z}_j'\left(\boldsymbol{y} - \mathbb{E}_{\boldsymbol{\gamma}_{-j}}\left[\boldsymbol{Z}\boldsymbol{\gamma}_{-j}\right]\right) - \mathbb{E}_{\sigma^2}\left[\frac{1}{\sigma^2}\right]\frac{1}{2}\boldsymbol{\gamma}_j'\boldsymbol{Z}_j'\boldsymbol{Z}_j\boldsymbol{\gamma}_j \\
&\quad -\mathbb{E}_{\tau_j^2}\left[\frac{1}{\tau_j^2}\right]\frac{1}{2}\boldsymbol{\gamma}_j'\boldsymbol{K}_j\boldsymbol{\gamma}_j\Bigg\}
\end{aligned}$$

$$
\begin{aligned}
&= \exp\Bigg\{\mathbb{E}_{\sigma^2}\left[\frac{1}{\sigma^2}\right]\boldsymbol{\gamma}_j'\boldsymbol{Z}_j'\left(\boldsymbol{y} - \mathbb{E}_{\boldsymbol{\gamma}_{-j}}\left[\boldsymbol{Z}\boldsymbol{\gamma}_{-j}\right]\right)\\
&\quad - \frac{1}{2}\boldsymbol{\gamma}_j'\left(\mathbb{E}_{\sigma^2}\left[\frac{1}{\sigma^2}\right]\boldsymbol{Z}_j'\boldsymbol{Z}_j + \mathbb{E}_{\tau_j^2}\left[\frac{1}{\tau_j^2}\right]\boldsymbol{K}_j\right)\boldsymbol{\gamma}_j\Bigg\}\\
&= \exp\Bigg\{-\frac{1}{2}\boldsymbol{\gamma}_j'\underbrace{\left(\mathbb{E}_{\sigma^2}\left[\frac{1}{\sigma^2}\right]\boldsymbol{Z}_j'\boldsymbol{Z}_j + \mathbb{E}_{\tau_j^2}\left[\frac{1}{\tau_j^2}\right]\boldsymbol{K}_j\right)}_{\boldsymbol{\Sigma}_j^{-1}}\boldsymbol{\gamma}_j\\
&\quad + \boldsymbol{\gamma}_j'\boldsymbol{\Sigma}_j^{-1}\underbrace{\boldsymbol{\Sigma}_j\mathbb{E}_{\sigma^2}\left[\frac{1}{\sigma^2}\right]\boldsymbol{Z}_j'\left(\boldsymbol{y} - \mathbb{E}_{\boldsymbol{\gamma}_{-j}}\left[\boldsymbol{Z}\boldsymbol{\gamma}_{-j}\right]\right)}_{\boldsymbol{\mu}_j}\Bigg\}
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}(\boldsymbol{\gamma}_j) = \boldsymbol{\mu}_j &= \mathbb{E}_{\sigma^2}\left[\frac{1}{\sigma^2}\right]\boldsymbol{\Sigma}_j\boldsymbol{Z}_j'\left(\boldsymbol{y} - \mathbb{E}_{\boldsymbol{\gamma}_{-j}}\left[\boldsymbol{Z}\boldsymbol{\gamma}_{-j}\right]\right)\\
\mathrm{Var}(\boldsymbol{\gamma}_j) = \boldsymbol{\Sigma}_j &= \left(\mathbb{E}_{\sigma^2}\left[\frac{1}{\sigma^2}\right]\boldsymbol{Z}_j'\boldsymbol{Z}_j + \mathbb{E}_{\tau_j^2}\left[\frac{1}{\tau_j^2}\right]\boldsymbol{K}_j\right)^{-1}
\end{aligned}
$$

For $\boldsymbol{\gamma}_j = \left(\gamma_{j1}, \gamma_{j2}, ..., \gamma_{jd_j}\right)'$ and $\boldsymbol{\mu}_j = \left(\mu_{j1}, \mu_{j2}, ..., \mu_{jd_j}\right)'$
Full Covariance update:

$$
\begin{aligned}
\mathbb{E}(\boldsymbol{\gamma}) = \boldsymbol{\mu} &= \mathbb{E}_{\sigma^2}\left[\frac{1}{\sigma^2}\right]\boldsymbol{\Sigma}\boldsymbol{Z}'\boldsymbol{y}\right)\\
\mathrm{Var}(\boldsymbol{\gamma}) = \boldsymbol{\Sigma} &= \left(\mathbb{E}_{\sigma^2}\left[\frac{1}{\sigma^2}\right]\boldsymbol{Z}'\boldsymbol{Z} + \mathbb{E}_{\boldsymbol{\tau}^2}\left[\boldsymbol{K}\right]\right)^{-1}
\end{aligned}
$$

With $\boldsymbol{\mu} = (\boldsymbol{\mu}_1', ..., \boldsymbol{\mu}_p')'$, $\boldsymbol{Z} = (\boldsymbol{Z}_1, ..., \boldsymbol{Z}_p)$ and
$\mathbb{E}_{\boldsymbol{\tau}^2}\left[\boldsymbol{K}\right] = \mathrm{diag}\big(\mathbb{E}_{\tau_1^2}\left[\frac{1}{\tau_1^2}\right]\boldsymbol{K}_1, ..., \mathbb{E}_{\tau_p^2}\left[\frac{1}{\tau_p^2}\right]\boldsymbol{K}_p\big)$

2. Variational density for $\boldsymbol{\gamma}$ coefficients variance $q(\tau_j^2)$:

$$
\begin{aligned}
q(\tau_j^2) &\propto \exp\left\{\mathbb{E}_{-\tau^2}\left[\ln\left(p(\tau_j^2)\,p(\boldsymbol{\gamma}_j|\tau_j^2)\right)\right]\right\}\\
&= \exp\Bigg\{\mathbb{E}_{-\tau_j^2}\Bigg[\ln\Bigg(\frac{b_j^{a_j}}{\Gamma(a_j)}\left(\frac{1}{\tau^2}\right)^{a_j+1}\exp\left(-\frac{b_j}{\tau_j^2}\right)\\
&\quad \frac{1}{(2\pi\tau_j^2)^{\frac{\mathrm{rank}(\boldsymbol{K}_j)}{2}}}\exp\left(-\frac{1}{2\tau_j^2}\boldsymbol{\gamma}_j'\boldsymbol{K}_j\boldsymbol{\gamma}_j\right)\Bigg)\Bigg]\Bigg\}\\
&\propto \exp\Bigg\{\mathbb{E}_{-\tau_j^2}\Bigg[(a_j+1)\ln\left(\frac{1}{\tau_j^2}\right) - \frac{b_j}{\tau_j^2}
\end{aligned}
$$

$$
\begin{aligned}
&+ \frac{\text{rank}(\boldsymbol{K}_j)}{2} \ln\left(\frac{1}{2\pi\tau_j^2}\right) - \frac{1}{2\tau_j^2}\boldsymbol{\gamma}_j'\boldsymbol{K}_j\boldsymbol{\gamma}_j\Big]\Big\} \\
&\propto \exp\Big\{(a_j+1)\ln\left(\frac{1}{\tau_j^2}\right) + \frac{\text{rank}(\boldsymbol{K}_j)}{2}\ln\left(\frac{1}{\tau_j^2}\right) \\
&\quad - \frac{b_j}{\tau_j^2} - \frac{1}{2\tau_j^2}\mathbb{E}_{\boldsymbol{\gamma}_j}\left[\boldsymbol{\gamma}_j'\boldsymbol{K}_j\boldsymbol{\gamma}_j\right]\Big\} \\
&= \exp\Big\{\ln\left(\frac{1}{\tau_j^2}\right)\left(a_j + 1 + \frac{\text{rank}(\boldsymbol{K}_j)}{2}\right) - \frac{1}{\tau_j^2}\left(b_j + \frac{1}{2}\mathbb{E}_{\boldsymbol{\gamma}_j}\left[\boldsymbol{\gamma}_j'\boldsymbol{K}_j\boldsymbol{\gamma}_j\right]\right)\Big\} \\
&= \left(\frac{1}{\tau_j^2}\right)^{\underbrace{a_j + \frac{\text{rank}(\boldsymbol{K}_j)}{2}}_{\nu_{a_j}}+1} \exp\Big\{-\frac{1}{\tau_j^2}\underbrace{\left(b_j + \frac{1}{2}\mathbb{E}_{\boldsymbol{\gamma}_j}\left[\boldsymbol{\gamma}_j'\boldsymbol{K}_j\boldsymbol{\gamma}_j\right]\right)}_{\nu_{b_j}}\Big\}
\end{aligned}
$$

$$
\begin{aligned}
\nu_{a_j} &= a_j + \frac{\text{rank}(\boldsymbol{K}_j)}{2} \\
\nu_{b_j} &= b_j + \frac{1}{2}\mathbb{E}_{\boldsymbol{\gamma}}\left[\boldsymbol{\gamma}_j'\boldsymbol{K}_j\boldsymbol{\gamma}_j\right] \\
&= b_j + \frac{1}{2}\left(\text{tr}\left(\boldsymbol{K}_j\boldsymbol{\Sigma}_j\right) + \boldsymbol{\mu}_j'\boldsymbol{K}_j\boldsymbol{\mu}_j\right) \\
\mathbb{E}(\tau_j^2) &= \frac{\nu_{b_j}}{\nu_{a_j}-1} \\
\mathbb{E}\left(\frac{1}{\tau_j^2}\right) &= \frac{\nu_{a_j}}{\nu_{b_j}}
\end{aligned}
$$

3. Variational density for the variance of the error term:

$$
\begin{aligned}
q(\sigma^2) &\propto \exp\Big\{\mathbb{E}_{-\sigma^2}\Big[\ln\Big(p(\boldsymbol{y}|\boldsymbol{\gamma},\sigma)\,p(\sigma^2)\Big)\Big]\Big\} \\
&= \exp\Big\{\mathbb{E}_{-\sigma^2}\Big[\ln\Big(\frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}\exp\left(-\frac{(\boldsymbol{y}-\boldsymbol{Z}\boldsymbol{\gamma})'(\boldsymbol{y}-\boldsymbol{Z}\boldsymbol{\gamma})}{2\sigma^2}\right) \\
&\quad \frac{b_{\sigma^2}^{a_{\sigma^2}}}{\Gamma(a)}\left(\frac{1}{\sigma^2}\right)^{a_{\sigma^2}+1}\exp\left(-\frac{b_{\sigma^2}}{\sigma^2}\right)\Big)\Big]\Big\} \\
&\propto \exp\Big\{\mathbb{E}_{-\sigma^2}\Big[\ln\Big(\frac{1}{\sigma^2}\Big)\frac{n}{2} + \ln\left(\frac{1}{\sigma^2}\right)(a_{\sigma^2}+1) \\
&\quad - \frac{(\boldsymbol{y}-\boldsymbol{Z}\boldsymbol{\gamma})'(\boldsymbol{y}-\boldsymbol{Z}\boldsymbol{\gamma})}{2\sigma^2} - \frac{b_{\sigma^2}}{\sigma^2}\Big]\Big\} \\
&= \exp\Big\{\ln\Big(\frac{1}{\sigma^2}\Big)\left(\frac{n}{2} + a_{\sigma^2} + 1\right)
\end{aligned}
$$

$$
\begin{aligned}
&- \frac{1}{\sigma^2}\left(b_{\sigma^2} + \frac{1}{2}\mathbb{E}_{\boldsymbol{\gamma}}\left[(\boldsymbol{y} - \boldsymbol{Z\gamma})'(\boldsymbol{y} - \boldsymbol{Z\gamma})\right]\right)\Bigg\} \\
&= \left(\frac{1}{\sigma^2}\right)^{\underbrace{a_{\sigma^2} + \frac{n}{2}}_{\nu_{a_{\sigma^2}}} + 1} \exp\Bigg\{ - \frac{1}{\sigma^2}\underbrace{\left(b_{\sigma^2} + \frac{1}{2}\mathbb{E}_{\boldsymbol{\gamma}}\left[(\boldsymbol{y} - \boldsymbol{Z\gamma})'(\boldsymbol{y} - \boldsymbol{Z\gamma})\right]\right)}_{\nu_{b_{\sigma^2}}}\Bigg\}
\end{aligned}
$$

$$
\begin{aligned}
\nu_{a_{\sigma^2}} &= a_{\sigma^2} + \frac{n}{2} \\
\nu_{b_{\sigma^2}} &= b_{\sigma^2} + \frac{1}{2}\mathbb{E}_{\boldsymbol{\gamma}}\left[(\boldsymbol{y} - \boldsymbol{Z\gamma})'(\boldsymbol{y} - \boldsymbol{Z\gamma})\right] \\
&= b_{\sigma^2} + \frac{1}{2}\left(\boldsymbol{y}'\boldsymbol{y} - 2\mathbb{E}_{\boldsymbol{\gamma}}\left[\boldsymbol{\gamma}'\right]\boldsymbol{Z}'\boldsymbol{y} + \mathbb{E}_{\boldsymbol{\gamma}}\left[\boldsymbol{\gamma}'\boldsymbol{Z}'\boldsymbol{Z\gamma}\right]\right) \\
&= b_{\sigma^2} + \frac{1}{2}\left(\boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{\mu}'\boldsymbol{Z}'\boldsymbol{y} + \operatorname{tr}\left(\boldsymbol{Z}'\boldsymbol{Z\Sigma}\right) + \boldsymbol{\mu}'\boldsymbol{Z}'\boldsymbol{Z\mu}\right) \\
&= b_{\sigma^2} + \frac{1}{2}\left((\boldsymbol{y} - \boldsymbol{Z\mu})'(\boldsymbol{y} - \boldsymbol{Z\mu}) + \operatorname{tr}\left(\boldsymbol{Z}'\boldsymbol{Z\Sigma}\right)\right) \\
\mathbb{E}(\sigma^2) &= \frac{\nu_{b_{\sigma^2}}}{\nu_{a_{\sigma^2}} - 1} \\
\mathbb{E}\left(\frac{1}{\sigma^2}\right) &= \frac{\nu_{a_{\sigma^2}}}{\nu_{b_{\sigma^2}}}
\end{aligned}
$$

Note that $\boldsymbol{\Sigma}$ is either a blockdiagonal matrix with $\boldsymbol{\Sigma} = \operatorname{diag}(\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_p)$ or fully unstructured.

7.5 Derivation of ELBO for MFVI

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\nu}) =& \mathbb{E}_{(\boldsymbol{\gamma}, \boldsymbol{\tau}^2, \sigma^2) \sim q_{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\nu}}}\left[\ln\, p(\boldsymbol{\gamma}, \boldsymbol{\tau}^2, \sigma^2, \boldsymbol{y})\right] \\
&- \mathbb{E}_{(\boldsymbol{\gamma}, \boldsymbol{\tau}^2, \sigma^2) \sim q_{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\nu}}}\left[\ln\, q(\boldsymbol{\gamma}, \boldsymbol{\tau}^2, \sigma^2)\right] \\
=& \underbrace{\mathbb{E}_{\boldsymbol{\gamma} \sim q_{\boldsymbol{\psi}_{\boldsymbol{\gamma}}}, \sigma^2 \sim q_{\boldsymbol{\psi}_{\sigma^2}}}\left[\ln\, p(\boldsymbol{y}|\boldsymbol{\gamma}, \sigma^2)\right]}_{①} \\
&+ \sum_{j=1}^{p} \underbrace{\mathbb{E}_{\boldsymbol{\gamma}_j \sim q_{\boldsymbol{\psi}_{\boldsymbol{\gamma}_j}}, \tau_j^2 \sim q_{\boldsymbol{\psi}_{\tau_j^2}}}\left[\ln\, p(\boldsymbol{\gamma}_j|\tau_j^2)\right]}_{②}
\end{aligned}
$$

$$+\sum_{j=1}^{p}\underbrace{\mathbb{E}_{\tau_j^2\sim q_{\boldsymbol{\psi}_{\tau_j^2}}}\left[\ln\, p(\tau_j^2)\right]}_{(3)}+\underbrace{\mathbb{E}_{\sigma^2\sim q_{\boldsymbol{\psi}_{\sigma^2}}}\left[\ln\, p(\sigma^2)\right]}_{(4)}$$

$$-\sum_{j=1}^{p}\underbrace{\mathbb{E}_{\boldsymbol{\gamma}_j\sim q_{\boldsymbol{\psi}_{\boldsymbol{\gamma}_j}}}\left[\ln\, q(\boldsymbol{\gamma}_j)\right]}_{(5)}-\sum_{j=1}^{p}\underbrace{\mathbb{E}_{\tau_j^2\sim q_{\boldsymbol{\psi}_{\tau_j^2}}}\left[\ln\, q(\tau_j^2)\right]}_{(6)}$$

$$-\underbrace{\mathbb{E}_{\sigma^2\sim q_{\boldsymbol{\psi}_{\sigma^2}}}\left[\ln\, q(\sigma^2)\right]}_{(7)}$$

For notational convenience, we further do not add the distribution to the model parameter, when using the expectation. Instead of e.g. $\mathbb{E}_{\boldsymbol{\gamma}\sim q_{\boldsymbol{\psi}_{\boldsymbol{\gamma}}}}$, we write $\mathbb{E}_{\boldsymbol{\gamma}}$.

$$(1)\ \mathbb{E}_{\boldsymbol{\gamma},\sigma^2}\left[\ln\, p(\boldsymbol{y}|\boldsymbol{\gamma},\sigma^2)\right]$$

$$\begin{aligned}
&=\mathbb{E}_{\boldsymbol{\gamma},\sigma^2}\left[\ln\left(\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}}\exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{Z}\boldsymbol{\gamma})'(\boldsymbol{y}-\boldsymbol{Z}\boldsymbol{\gamma})\right)\right)\right]\\
&\propto\mathbb{E}_{\boldsymbol{\gamma},\sigma^2}\left[\frac{n}{2}\ln\left(\frac{1}{\sigma^2}\right)-\frac{1}{2\sigma^2}\left((\boldsymbol{y}-\boldsymbol{Z}\boldsymbol{\gamma})'(\boldsymbol{y}-\boldsymbol{Z}\boldsymbol{\gamma})\right)\right]\\
&=\frac{n}{2}\mathbb{E}_{\sigma^2}\left[\ln\left(\frac{1}{\sigma^2}\right)\right]-\frac{1}{2}\mathbb{E}_{\sigma^2}\left[\frac{1}{\sigma^2}\right]\left(\boldsymbol{y}'\boldsymbol{y}-2\mathbb{E}_{\boldsymbol{\gamma}}\left[\boldsymbol{\gamma}'\right]\boldsymbol{Z}'\boldsymbol{y}+\mathbb{E}_{\boldsymbol{\gamma}}\left[\boldsymbol{\gamma}'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\gamma}\right]\right)\\
&=\frac{n}{2}\mathbb{E}_{\sigma^2}\left[\ln\left(\frac{1}{\sigma^2}\right)\right]-\frac{1}{2}\frac{a^*}{b^*}\left((\boldsymbol{y}-\boldsymbol{Z}\boldsymbol{\mu}_{\boldsymbol{\gamma}})'(\boldsymbol{y}-\boldsymbol{Z}\boldsymbol{\mu}_{\boldsymbol{\gamma}})+\mathrm{tr}\left(\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\right)\right)\\
&=\frac{n}{2}\left(\digamma(a^*)-\ln(b^*)\right)-\frac{1}{2}\frac{a^*}{b^*}\left((\boldsymbol{y}-\boldsymbol{Z}\boldsymbol{\mu}_{\boldsymbol{\gamma}})'(\boldsymbol{y}-\boldsymbol{Z}\boldsymbol{\mu}_{\boldsymbol{\gamma}})+\mathrm{tr}\left(\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\right)\right)
\end{aligned}$$

Note if $x\sim\mathrm{InvGa}(a,b)\Rightarrow\frac{1}{x}\sim\mathrm{Ga}(a,b)$ then
$\mathbb{E}(\ln(\frac{1}{x}))=\digamma(a)-\ln(b)$, where $\digamma(x)$ is the digamma function.

$$(2)\ \mathbb{E}_{\boldsymbol{\gamma}_j,\tau^2}\left[\ln\, p(\boldsymbol{\gamma}_j|\tau_j^2)\right]$$

$$\begin{aligned}
&=\mathbb{E}_{\boldsymbol{\gamma}_j,\tau_j^2}\left[\ln\left(\frac{1}{(2\pi\tau^2)^{\frac{\mathrm{rank}(\boldsymbol{K}_j)}{2}}}\exp\left(-\frac{\boldsymbol{\gamma}_j'\boldsymbol{K}_j\boldsymbol{\gamma}_j}{2\tau_j^2}\right)\right)\right]\\
&\propto\mathbb{E}_{\boldsymbol{\gamma}_j,\tau_j^2}\left[\frac{\mathrm{rank}(\boldsymbol{K}_j)}{2}\ln\left(\frac{1}{\tau_j^2}\right)-\frac{1}{2\tau_j^2}\boldsymbol{\gamma}_j'\boldsymbol{K}_j\boldsymbol{\gamma}_j\right]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\operatorname{rank}(\boldsymbol{K}_j)}{2} \mathbb{E}_{\tau_j^2}\left[\ln\left(\frac{1}{\tau_j^2}\right)\right] - \frac{1}{2}\mathbb{E}_{\tau_j^2}\left[\frac{1}{\tau_j^2}\right]\mathbb{E}_{\boldsymbol{\gamma}_j}\left[\boldsymbol{\gamma}_j' \boldsymbol{K}_j \boldsymbol{\gamma}_j\right] \\
&= \frac{\operatorname{rank}(\boldsymbol{K}_j)}{2}\left(F(\nu_{a_j}) - \ln(\nu_{b_j})\right) - \frac{1}{2}\frac{\nu_{a_j}}{\nu_{b_j}}\left(\operatorname{tr}\left(\boldsymbol{K}_j \boldsymbol{\Sigma}_j\right) + \boldsymbol{\mu}_j' \boldsymbol{K}_j \boldsymbol{\mu}_j\right)
\end{aligned}
$$

③ $\mathbb{E}_{\tau_j^2}\left[\ln p(\tau_j^2)\right]$

$$
\begin{aligned}
&= \mathbb{E}_{\tau_j^2}\left[\ln\left(\frac{b_j^{a_j}}{\Gamma(a_j)}\left(\frac{1}{\tau_j^2}\right)^{a_j+1}\exp\left(-\frac{b_j}{\tau_j^2}\right)\right)\right] \\
&\propto \mathbb{E}_{\tau_j^2}\left[(a_j+1)\ln\left(\frac{1}{\tau_j^2}\right) - \frac{b_j}{\tau_j^2}\right] \\
&= (a_j+1)\mathbb{E}_{\tau_j^2}\left[\ln\left(\frac{1}{\tau_j^2}\right)\right] - b_j \mathbb{E}_{\tau_j^2}\left[\frac{1}{\tau_j^2}\right] \\
&= (a_j+1)\left(F(\nu_{a_j}) - \ln(\nu_{b_j})\right) - b_j \frac{\nu_{a_j}}{\nu_{b_j}}
\end{aligned}
$$

④ $\mathbb{E}_{\sigma^2}\left[\ln p(\sigma^2)\right]$

$$
\begin{aligned}
&= \mathbb{E}_{\sigma^2}\left[\ln\left(\frac{b_{\sigma^2}^{a_{\sigma^2}}}{\Gamma(a_{\sigma^2})}\left(\frac{1}{\sigma^2}\right)^{a_{\sigma^2}+1}\exp\left(-\frac{b_{\sigma^2}}{\sigma^2}\right)\right)\right] \\
&\propto \mathbb{E}_{\sigma^2}\left[(a_{\sigma^2}+1)\ln\left(\frac{1}{\sigma^2}\right) - \frac{b_{\sigma^2}}{\sigma^2}\right] \\
&= (a_{\sigma^2}+1)\mathbb{E}_{\sigma^2}\left[\ln\left(\frac{1}{\sigma^2}\right)\right] - b_{\sigma^2}\mathbb{E}_{\sigma^2}\left[\frac{1}{\sigma^2}\right] \\
&= (a_{\sigma^2}+1)\left(F(\nu_{a_{\sigma^2}}) - \ln(\nu_{b_{\sigma^2}})\right) - b_{\sigma^2}\frac{\nu_{a_{\sigma^2}}}{\nu_{b_{\sigma^2}}}
\end{aligned}
$$

⑤ $\mathbb{E}_{\boldsymbol{\gamma}_j}\left[\ln q(\boldsymbol{\gamma}_j)\right]$

$$
\begin{aligned}
&= \mathbb{E}_{\boldsymbol{\gamma}_j}\left[\ln\left((2\pi)^{-\frac{d_j}{2}}\det(\boldsymbol{\Sigma}_j)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\boldsymbol{\gamma}_j - \boldsymbol{\mu}_j)'\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{\gamma}_j - \boldsymbol{\mu}_j)\right)\right)\right] \\
&\propto \mathbb{E}_{\boldsymbol{\gamma}_j}\left[-\frac{1}{2}\ln\big(\det(\boldsymbol{\Sigma}_j)\big) - \frac{1}{2}(\boldsymbol{\gamma}_j - \boldsymbol{\mu}_j)'\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{\gamma}_j - \boldsymbol{\mu}_j)\right] \\
&= -\frac{1}{2}\ln\big(\det(\boldsymbol{\Sigma}_j)\big) - \frac{1}{2}\mathbb{E}_{\boldsymbol{\gamma}_j}\left[(\boldsymbol{\gamma}_j - \boldsymbol{\mu}_j)'\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{\gamma}_j - \boldsymbol{\mu}_j)\right] \\
&= -\frac{1}{2}\ln\big(\det(\boldsymbol{\Sigma}_j)\big) - \frac{1}{2}\left(\mathbb{E}_{\boldsymbol{\gamma}_j}\left[\boldsymbol{\gamma}_j'\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\gamma}_j\right] - 2\mathbb{E}_{\boldsymbol{\gamma}_j}\left[\boldsymbol{\gamma}_j'\right]\boldsymbol{\Sigma}_j^{-1} + \boldsymbol{\mu}_j'\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\mu}_j\right)
\end{aligned}
$$

$$
\begin{aligned}
&= -\frac{1}{2}\ln\big(\det(\boldsymbol{\Sigma}_j)\big) - \frac{1}{2}\left(\operatorname{tr}\left(\boldsymbol{\Sigma}_j\boldsymbol{\Sigma}_j^{-1}\right) + \boldsymbol{\mu}_j'\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\mu}_j - 2\boldsymbol{\mu}_j'\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\mu}_j + \boldsymbol{\mu}_j'\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\mu}_j\right)\\
&= -\frac{1}{2}\ln\big(\det(\boldsymbol{\Sigma}_j)\big) - \underbrace{\frac{1}{2}\operatorname{tr}\left(\boldsymbol{I}_j\right)}_{\text{const}}\\
&\propto -\frac{1}{2}\ln\big(\det(\boldsymbol{\Sigma}_j)\big)
\end{aligned}
$$

For full covariance:

$$
\mathbb{E}_{\boldsymbol{\gamma}_j}\left[\ln q(\boldsymbol{\gamma})\right] \propto -\frac{1}{2}\ln\big(\det(\boldsymbol{\Sigma})\big)
$$

(6) $\mathbb{E}_{\tau_j^2}\left[\ln q(\tau_j^2)\right]$

$$
\begin{aligned}
&= \mathbb{E}_{\tau_j^2}\left[\ln \frac{{\nu_{b_j}}^{\nu_{a_j}}}{\Gamma(\nu_{a_j})}\left(\frac{1}{\tau_j^2}\right)^{\nu_{a_j}+1}\exp\left(-\frac{\nu_{b_j}}{\tau_j^2}\right)\right]\\
&= \mathbb{E}_{\tau_j^2}\left[\nu_{a_j}\ln(\nu_{b_j}) - \ln(\Gamma(\nu_{a_j})) + (\nu_{a_j}+1)\ln\left(\frac{1}{\tau_j^2}\right) - \frac{\nu_{b_j}}{\tau_j^2}\right]\\
&= \nu_{a_j}\ln(\nu_{b_j}) - \ln(\Gamma(\nu_{a_j})) + (\nu_{a_j}+1)\mathbb{E}_{\tau_j^2}\left[\ln\left(\frac{1}{\tau_j^2}\right)\right] - \nu_{b_j}\mathbb{E}_{\tau_j^2}\left[\frac{1}{\tau_j^2}\right]\\
&= \nu_{a_j}\ln(\nu_{b_j}) - \ln(\Gamma(\nu_{a_j})) + (\nu_{a_j}+1)\left(F(\nu_{a_j}) - \ln(\nu_{b_j})\right) - \nu_{b_j}\frac{\nu_{a_j}}{\nu_{b_j}}\\
&= \nu_{a_j}\ln(\nu_{b_j}) - \ln(\Gamma(\nu_{a_j})) + (\nu_{a_j}+1)F(\nu_{a_j}) - \nu_{a_j}\ln(\nu_{b_j}) - \ln(\nu_{b_j}) - \nu_{a_j} +\\
&= (\nu_{a_j}+1)F(\nu_{a_j}) - \ln(\nu_{b_j}) - \nu_{a_j} - \ln(\Gamma(\nu_{a_j}))
\end{aligned}
$$

(7) $\mathbb{E}_{\sigma^2}\left[\ln q(\sigma^2)\right]$

$$
\begin{aligned}
&= \mathbb{E}_{\sigma^2}\left[\ln \frac{{\nu_{b_{\sigma^2}}}^{\nu_{a_{\sigma^2}}}}{\Gamma(\nu_{a_{\sigma^2}})}\left(\frac{1}{\sigma^2}\right)^{\nu_{a_{\sigma^2}}+1}\exp\left(-\frac{\nu_{b_{\sigma^2}}}{\sigma^2}\right)\right]\\
&= \mathbb{E}_{\sigma^2}\left[\nu_{a_{\sigma^2}}\ln(\nu_{b_{\sigma^2}}) - \ln(\Gamma(\nu_{a_{\sigma^2}})) + (\nu_{a_{\sigma^2}}+1)\ln\left(\frac{1}{\sigma^2}\right) - \frac{\nu_{b_{\sigma^2}}}{\sigma^2}\right]\\
&= \nu_{a_{\sigma^2}}\ln(\nu_{b_{\sigma^2}}) - \ln(\Gamma(\nu_{a_{\sigma^2}})) + (\nu_{a_{\sigma^2}}+1)\mathbb{E}_{\sigma^2}\left[\ln\left(\frac{1}{\sigma^2}\right)\right] - \nu_{b_{\sigma^2}}\mathbb{E}_{\sigma^2}\left[\frac{1}{\sigma^2}\right]\\
&= \nu_{a_{\sigma^2}}\ln(\nu_{b_{\sigma^2}}) - \ln(\Gamma(\nu_{a_{\sigma^2}})) + (\nu_{a_{\sigma^2}}+1)\left(F(\nu_{a_{\sigma^2}}) - \ln(\nu_{b_{\sigma^2}})\right) - \nu_{b_{\sigma^2}}\frac{\nu_{a_{\sigma^2}}}{\nu_{b_{\sigma^2}}}\\
&= \nu_{a_{\sigma^2}}\ln(\nu_{b_{\sigma^2}}) - \ln(\Gamma(\nu_{a_{\sigma^2}})) + (\nu_{a_{\sigma^2}}+1)F(\nu_{a_{\sigma^2}}) - \nu_{a_{\sigma^2}}\ln(\nu_{b_{\sigma^2}}) - \ln(\nu_{b_{\sigma^2}}) - \nu_{a_{\sigma^2}}\\
&= (\nu_{a_{\sigma^2}}+1)F(\nu_{a_{\sigma^2}}) - \ln(\nu_{b_{\sigma^2}}) - \nu_{a_{\sigma^2}} - \ln(\Gamma(\nu_{a_{\sigma^2}}))
\end{aligned}
$$

7.6 Derivation of ELBO for SIVI

$$
\begin{aligned}
\widetilde{\mathcal{L}}(\boldsymbol{\phi}, \boldsymbol{\xi}, \boldsymbol{\nu}) \approx \sum_{j=1}^{p} \underbrace{\mathbb{E}_{\tau_j^2 \sim q_{\boldsymbol{\psi}_{\tau_j^2}}}\left[\ln\, p(\tau_j^2)\right]}_{①} + \underbrace{\mathbb{E}_{\sigma^2 \sim q_{\boldsymbol{\psi}_{\sigma^2}}}\left[\ln\, p(\sigma^2)\right]}_{②} \\
- \sum_{j=1}^{p} \underbrace{\mathbb{E}_{\tau_j^2 \sim q_{\boldsymbol{\psi}_{\tau_j^2}}}\left[\ln\, q(\tau_j^2)\right]}_{③} - \underbrace{\mathbb{E}_{\sigma^2 \sim q_{\boldsymbol{\psi}_{\sigma^2}}}\left[\ln\, q(\sigma^2)\right]}_{④} \\
+ \frac{1}{S} \sum_{s=1}^{S} \Bigg\{ \underbrace{\mathbb{E}_{\sigma^2 \sim q_{\boldsymbol{\psi}_{\sigma^2}}}\left[\ln\, p(\boldsymbol{y}|\boldsymbol{\gamma}_s, \sigma^2)\right]}_{⑤} + \sum_{j=1}^{p} \underbrace{\mathbb{E}_{\tau_j^2 \sim q_{\boldsymbol{\psi}_{\tau_j^2}}}\left[\ln\, p(\boldsymbol{\gamma}_{j,s}|\tau_j^2)\right]}_{⑥} \\
- \sum_{j=1}^{p} \Bigg(\underbrace{\ln\left(q(\boldsymbol{\gamma}_{j,s}|\boldsymbol{\mu}_{j,s}) + \sum_{k=1}^{K} q(\boldsymbol{\gamma}_j|\boldsymbol{\mu}_j^{(k)})\right)}_{⑦} - \ln\,(K+1) \Bigg) \Bigg\}
\end{aligned}
$$

① $\mathbb{E}_{\tau_j^2 \sim q_{\boldsymbol{\psi}_{\tau_j^2}}}\left[\ln\, p(\tau_j^2)\right]$

$$\propto (a_j + 1)\left(F(\nu_{a_j}) - \ln(\nu_{b_j})\right) - b_j \frac{\nu_{a_j}}{\nu_{b_j}}$$

② $\mathbb{E}_{\sigma^2 \sim q_{\boldsymbol{\psi}_{\sigma^2}}}\left[\ln\, p(\sigma_s^2)\right]$

$$\propto (a_{\sigma^2} + 1)\left(F(\nu_{a_{\sigma^2}}) - \ln(\nu_{b_{\sigma^2}})\right) - b_{\sigma^2} \frac{\nu_{a_{\sigma^2}}}{\nu_{b_{\sigma^2}}}$$

③ $\mathbb{E}_{\tau_j^2 \sim q_{\boldsymbol{\psi}_{\tau_j^2}}}\left[\ln\, q(\tau_j^2)\right]$

$$\propto (\nu_{a_j} + 1)F(\nu_{a_j}) - \ln(\nu_{b_j}) - \nu_{a_j} - \ln(\Gamma(\nu_{a_j}))$$

④ $\mathbb{E}_{\sigma^2 \sim q_{\boldsymbol{\psi}_{\sigma^2}}}\left[\ln\, q(\sigma^2)\right]$

$$\propto (\nu_{a_{\sigma^2}} + 1)F(\nu_{a_{\sigma^2}}) - \ln(\nu_{b_{\sigma^2}}) - \nu_{a_{\sigma^2}} - \ln(\Gamma(\nu_{a_{\sigma^2}}))$$

$$\textcircled{5}\ \mathbb{E}_{\sigma^2\sim q_{\psi_{\sigma^2}}}\left[\ln\, p(\boldsymbol{y}|\boldsymbol{\gamma}_s,\sigma^2)\right]$$

$$\propto \frac{n}{2}\left(\digamma(a^*)-\ln(b^*)\right)-\frac{1}{2}\frac{a^*}{b^*}\Big((\boldsymbol{y}-\boldsymbol{Z}\boldsymbol{\gamma}_s)'(\boldsymbol{y}-\boldsymbol{Z}\boldsymbol{\gamma}_s)+\operatorname{tr}\left(\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Sigma}\right)\Big)$$

$$\textcircled{6}\ \mathbb{E}_{\tau_j^2\sim q_{\psi_{\tau_j^2}}}\left[\ln\, p(\boldsymbol{\gamma}_{j,s}|\tau_j^2)\right]$$

$$\propto \frac{\operatorname{rank}(\boldsymbol{K}_j)}{2}\left(\digamma(\nu_{a_j})-\ln(\nu_{b_j})\right)-\frac{1}{2}\frac{\nu_{a_j}}{\nu_{b_j}}\left(\operatorname{tr}\left(\boldsymbol{K}_j\boldsymbol{\Sigma}_j\right)+\boldsymbol{\gamma}'_{j,s}\boldsymbol{K}_j\boldsymbol{\gamma}_{j,s}\right)$$

$$\textcircled{7}\ \ln\left[q(\boldsymbol{\gamma}_{j,s}|\boldsymbol{\mu}_{j,s})+\textstyle\sum_{k=1}^{K} q(\boldsymbol{\gamma}_j|\boldsymbol{\mu}_j^{(k)})\right]$$

$$\begin{aligned}
&\propto \ln\Bigg[\sum_{k=1}^{K}\det(\boldsymbol{\Sigma}_j)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\boldsymbol{\gamma}_{j,s}|\boldsymbol{\mu}_j^{(k)}-\boldsymbol{\mu}_j^{(k)})'(\boldsymbol{\Sigma}_j)^{-1}(\boldsymbol{\gamma}_{j,s}|\boldsymbol{\mu}_j^{(k)}-\boldsymbol{\mu}_j^{(k)})\right)\\
&\quad+\det(\boldsymbol{\Sigma}_j)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\boldsymbol{\gamma}_{j,s}|\boldsymbol{\mu}_{j,s}-\boldsymbol{\mu}_{j,s})'(\boldsymbol{\Sigma}_j)^{-1}(\boldsymbol{\gamma}_{j,s}|\boldsymbol{\mu}_{j,s}-\boldsymbol{\mu}_{j,s})\right)\Bigg]\\
&=-\frac{1}{2}\ln\det(\boldsymbol{\Sigma}_j)\\
&\quad+\ln\Bigg[\sum_{t=1}^{T}\exp\left(-\frac{1}{2}(\boldsymbol{\gamma}_{j,s}|\boldsymbol{\mu}_j^{(t)}-\boldsymbol{\mu}_j^{(t)})'(\boldsymbol{\Sigma}_j)^{-1}(\boldsymbol{\gamma}_{j,s}|\boldsymbol{\mu}_j^{(t)}-\boldsymbol{\mu}_j^{(t)})\right)\\
&\quad+\exp\left(-\frac{1}{2}(\boldsymbol{\gamma}_{j,s}|\boldsymbol{\mu}_{j,s}-\boldsymbol{\mu}_{j,s})'(\boldsymbol{\Sigma}_j)^{-1}(\boldsymbol{\gamma}_{j,s}|\boldsymbol{\mu}_{j,s}-\boldsymbol{\mu}_{j,s})\right)\Bigg]
\end{aligned}$$

7.7 Model specifications

Settings for simulation

We use P-Splines with 25 knots, third degree polynomials and 2nd order penalty.

For the Gibbs sampler, we choose a burnin of 100 and a thinning of 2. The length of the chain is then between 5000 and 12,500 depending on the scenario. Due to the very high auto-correlation in the chains for some scenarios and in particular for the scenario with 50 observations and high correlation between the covariates, we increased the the sample size.

For MFVI we use a tolerance level of 0.0001 and a maximum of iterations of 5000 to define convergence.

For SIVI and SIMFVI we employ a fully connected multilayer perceptron for $T_{\boldsymbol{\phi}}$ and set the default input dimension to 30 as in Yin and Zhou (2018). As activation function, we use the ReLu activation function. We employ 3 hidden layers with 50, 100 and 50 neurons, respectively. The learning rates for $\boldsymbol{\phi}$ are 0.01, for $\boldsymbol{\xi}$ are 0.01, 0.005 and 0.0025 with increasing observations and for $\boldsymbol{\nu}$ are 0.01 with 50 and 250

observations and 0.005 with 500 observations. We initialize the neural net according to the method described in He et al. (2015). We initialize $\boldsymbol{\xi}$ with 0.4 and $\boldsymbol{\nu}_{\sigma^2}$ with $(\frac{N}{2}, \frac{N}{2})$ and $\boldsymbol{\nu}_{\tau_j^2} = (e, e)$ (Euler's number). As decay rate we choose 0.9 (and decay starts after 300 iterations). The Adam optimizer takes 2 more hyper-parameters, that the coefficients to control for the exponential decay rate that we set to $\beta = (0.9, 0.999)$ and a term to improve numerical stability that we set to $\epsilon = 1e - 08$. For the early stopping criterion, we take the the average slope of the lower bound ELBO over the last 200 to 250 iterations (based on OLS) and stop optimizing after the slope is less then the tolerance level of 0.001 for SIVI and 0.00001 for SIMFVI. For comparison, we check the results after 5,000 iteration (without any early stopping). The presented results are the one based on early stopping.

Settings for application

The prior and spline settings are as follows: We use non-informative priors on the linear coefficients. For all nonlinear effects, we use Bayesian P-splines with 25 interior knots and a second-order random walk prior. For the spatial effect, we vary the number of knots in each direction. As Germany has a north–south distance of about 830 km and a west-east distance of about 650 km, using more knots in the latitude direction results in a distance-wise equal spread of knots in each direction. We, therefore, chose to place 22 knots in the longitude direction and 28 knots in the latitude direction. This results in a knot placement of about every 30 km in each direction. Choosing an equal number of knots in both directions leads to strong biases for location-specific covariates. For the random walk order, we chose the order 2 in each direction.

The settings for the algorithms that change compared to the simulation are the following: As the SIVI algorithm was very unstable, we initialize the parameters $\boldsymbol{\nu}$ and $\boldsymbol{\Sigma}_1, \ldots \boldsymbol{\Sigma}_p$ with the results obtained in MFVI (block). Additionally, we reduce the learning rates for $\boldsymbol{\nu}$ and for $\boldsymbol{\xi}$ to 1×10^{-9} and 1×10^{-6}, respectively, such that no large changes for the optimized parameters occur in the optimization. For SIMFVI, we change the neural net structure. Instead of three hidden layers, we choose only two with 100 and 1000 neurons, respectively. Choosing a less deeper neural net structure for SIMFVI accelerates learning and improves the result. For MFVI, we keep the same settings except for the tolerance level[3] that we reduce by one decimal point. Hence, the optimization runs longer.

We also vary the length of the posterior sample size. For the Gibbs sampler, we run a chain of length 62,000. The first 2000 samples are considered burn-in and the chains are thinned to every 10th sample. One of the coefficient chains of the spline for the altitude has an effective sample size of 319.9 that is the minimum across all chains. The 1% quantile of effective sample sizes across all chains is 385.5 and the 5% quantile is 688.3. We deem this sufficient to evaluate 95% CIs of the marginal posterior distributions.

For all VI methods, we draw 3000 samples from the estimated posterior distribution for each parameter.

[3] The tolerance level is the minimum difference between the ELBO's of 2 consecutive iterations. If this minimum is reached the algorithm stops.

Using this setup, the SIVI and SIMFVI algorithm had problems fitting some of the coefficients due to occurrences of extreme covariate values. Therefore, we classified values as outliers based on the boundaries considering quartiles and the interquartile range for each covariate. Classified outliers are 2.5 to 3 times the interquartile range below or above the 25% or 75% quartile, respectively. In total, we exclude about 1.2% of the data from the analysis.

7.8 Tree and climate data sets

We use data from the national forestry inventory (NFI) of Germany. The inventory is a large-scale survey conducted in 2012. To determine the locations of the observation points for the survey, a grid is spanned all over Germany. The grid has nodes every 2, 2.83 or 4 km, depending on the region. Around each node, four measurement points are determined. The distance of the four measurement points is based on a square with a side length of 150 m, whereby each measurement point is located at the corner of the square and the node is at the center of the square. A square with measurement points is also called a tract. From each measurement point, an inventory of the surrounding trees is taken if the point is within a forest. Trees must have a minimum diameter at breast height[4] (DBH) of 7 cm and must belong to the circular inclusion zone around the measurement point. As the sampling method is based on angle count sampling, the inclusion zone is proportional to the basal area of the tree, whereby the basal area is the cross-sectional area at breast height (for more details see Gregoire and Valentine 2007, Ch. 8). In simplified words, the idea is to draw a virtual circle around a tree that is proportional to its DBH. If the measurement point is within the virtual circle of a surrounding tree, the tree is included in the sample. Otherwise, the tree is not part of the sample. Hence, a greater variety of tree sizes are included in the sampling without increasing the sampling effort. But also trees with larger DBH are more likely included in the sampling procedure. The study design is further illustrated in Fig. 5.

Douglas firs are observed all over Germany, but particularly frequently in north-east and west to south-west of Germany. In some cases, Douglas firs are observed in neighboring tracts, and in other cases, as shown in the middle picture of Fig. 5, tracts with Douglas fir observations are randomly dispersed over the area.

The rightmost picture in Fig. 5 exemplarily shows one tract with Douglas firs. Three of the four measurement points have Douglas fir observed (red points). For the picture, however, we use simulated data, because the exact coordinates of each tree are not provided due to data confidentiality. Publicly available are only the approximated coordinates of each tract. The data can be found at the website of the Thünen-Institute.[5] We only consider Douglas firs that belong to the primary stand of the observed stand around the measurement point, that is trees with similar light availability. Hence, small trees covered by larger trees and therefore, with limited light availability are excluded.

We additionally use a climate data set provided by the Nordwestdeutsche Forstliche Versuchsanstalt. The data set contains the accumulated temperature and precipitation

[4] The diameter is always taken at a height of about 1.30 m.

[5] Thuenen-Institut, Dritte Bundeswaldinventur—Basisdaten (Stand 20.03.2015)—https://bwi.info/Download/de/BWI-Basisdaten/.

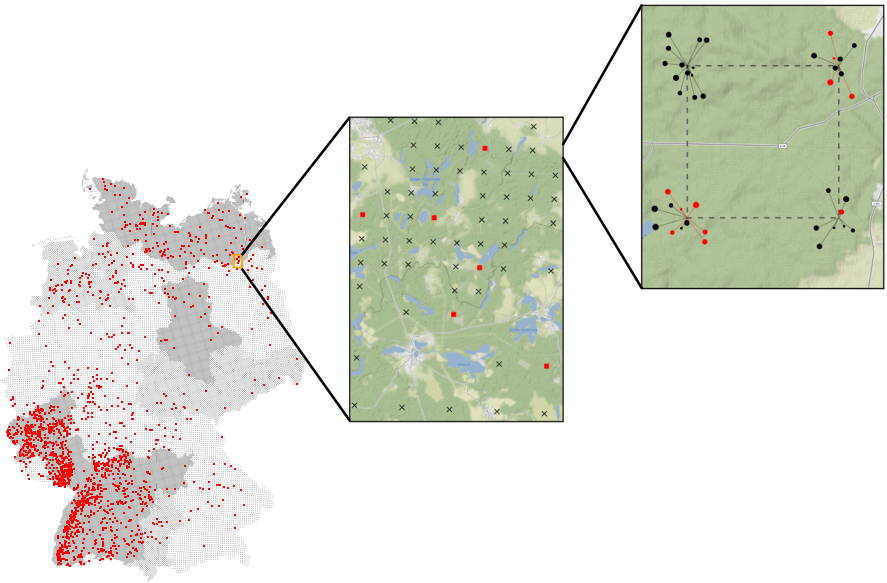


Fig. 5 Study design: A grid is placed all over Germany with 2 km, 2.83 km, and 4 km distance (dark to light regions (left picture)) between each node (the shape file is online available: https://bwi.info/Download/de/BWI-Basisdaten/ACCESS2003/). All red squares mark occurrences of at least one Douglas fir in a tract. The picture in the middle illustrates the change of the grid density in an extract from north Germany with 2 km density on the upper half and 4 km density on the lower half (crosses mark a tract without Douglas fir). The third picture on the right is a simulated example of how a tract might look like with Douglas firs in red and other species in black (exact coordinates of each tree are not provided due to confidentiality). The map data are obtained from the `osmdata` package in R (Padgham et al. 2017)

measurements in the vegetation period over the lifespan of each Douglas fir[6]. The measurements are taken in various German weather stations from year 1900 until the considered NFI. As no measurements are taken between weather stations, the data is interpolated up to a 50 m scale for temperature and up to a 100 m scale for precipitation. For the interpolation, the elevation of each location and a spatial effect are considered. The model is an additive model,

$$y_i = \beta_0 + f_1(alt_i) + f_{\text{geo}}(long_i, lat_i) + \epsilon_i,$$

with Gaussian distributed error ϵ_i. The response y_i is either the summed temperature or the summed precipitation over the vegetation period. The covariate alt, the altitude, is modeled nonlinearly and the other effect type is a spatial effect on the coordinates.

The elevation data come from the Copernicus GLO-90 Digital Elevation Model.[7]

[6] The vegetation periods are determined with the R-package `vegperiod`—https://cran.r-project.org/web/packages/vegperiod/index.html.

[7] Link to Copernicus Space Component Data Access Portal: https://spacedata.copernicus.eu/web/cscda/dataset-details?articleId=394198.

8 Algorithms

Algorithm 1: SIVI/SIMFVI algorithm

Input : Observations $\mathbf{X}$ and $\mathbf{y}$, joint distribution $p(\mathbf{y}|\mathbf{X}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \boldsymbol{\tau^2}, \sigma^2)$, explicit variational distributions $\boldsymbol{\gamma} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and variational distributions $\sigma^2 \sim \text{IG}(\nu_{a_{\sigma^2}}, \nu_{b_{\sigma^2}})$ and $\tau^2 \sim \text{IG}(\nu_a, \nu_b)$, neural network structure $T(.)$

Output : Variational parameters $\boldsymbol{\phi}$, $\boldsymbol{\xi}$ and $\boldsymbol{\nu}$

Initialize $\boldsymbol{\varphi}$, $\boldsymbol{\xi}$ and $\boldsymbol{\nu}$

Set iteration $i = 1$ and initial learning rates $\rho_{1,0}$, $\rho_{2,0}$ and $\rho_{3,0}$

while *not converged or not maximum number of iterations* **do**

$$\widetilde{\mathcal{L}}_i = \sum_{j=1}^{p} \mathbb{E}_{\tau_j^2 \sim q_{\nu_{a_j}, \nu_{b_j}}} \left[\log\, p(\tau_j^2) \right] + \mathbb{E}_{\sigma^2 \sim q_{\nu_{a_{\sigma^2}}, \nu_{b_{\sigma^2}}}} \left[\log\, p(\sigma^2) \right] - \sum_{j=1}^{p} \mathbb{E}_{\tau_j^2 \sim q_{\nu_{a_j}, \nu_{b_j}}} \left[\log\, q_{\nu_{a_j}, \nu_{b_j}}(\tau_j^2) \right] - \mathbb{E}_{\sigma^2 \sim q_{\boldsymbol{\psi}_{\sigma^2}}} \left[\log\, q_{\nu_{a_{\sigma^2}}, \nu_{b_{\sigma^2}}}(\sigma^2) \right]$$

sample $\boldsymbol{\epsilon}^{(\boldsymbol{\mu})^{(k)}} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ for $k = 1, ..., K$

for *s=1* **to** *S* **do**

Sample $\boldsymbol{\mu}_s = g_{\boldsymbol{\varphi}}(\boldsymbol{\epsilon}_s^{(\boldsymbol{\mu})}), \qquad \boldsymbol{\epsilon}_s^{(\boldsymbol{\mu})} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$

Sample $\boldsymbol{\gamma}_{j,s} = \boldsymbol{\mu}_{j,s} + \boldsymbol{\xi}_j \boldsymbol{\epsilon}_{j,s}^{(\boldsymbol{\gamma})}, \quad \boldsymbol{\epsilon}_{j,s}^{(\boldsymbol{\gamma})} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ for $j = 1, ..., p$

$$\widetilde{\mathcal{L}}_i = \widetilde{\mathcal{L}}_i + \frac{1}{S} \sum_{s=1}^{S} \Bigg\{ \mathbb{E}_{\sigma^2 \sim q_{\nu_{a_{\sigma^2}}, \nu_{b_{\sigma^2}}}} \left[\log\, p(\mathbf{y}|\boldsymbol{\gamma}_{.,s}, \sigma^2) \right] + \sum_{j=1}^{p} \mathbb{E}_{\tau_j^2 \sim q_{\nu_{a_j}, \nu_{b_j}}} \left[\log\, p(\boldsymbol{\gamma}_{j,s}|\tau_j^2) \right] - \log \Bigg(\prod_{j=1}^{p} \frac{1}{K+1} \Big(q_{T_{\boldsymbol{\phi}}(\boldsymbol{\varepsilon}_s)_j, \boldsymbol{\xi}_j}(\boldsymbol{\gamma}_{j,s}|T_{\boldsymbol{\phi}}(\boldsymbol{\varepsilon}_s)_j) + \sum_{k=1}^{K} q_{T_{\boldsymbol{\phi}}(\boldsymbol{\varepsilon}_s)_j, \boldsymbol{\xi}_j}(\boldsymbol{\gamma}_j|T_{\boldsymbol{\phi}}(\boldsymbol{\varepsilon}^{(k)})_j) \Big) \Bigg) \Bigg\}.$$

end

Compute $\rho_{1,i} = f_{\rho_1}(\rho_{1,i-1})$, $\rho_{2,i} = f_{\rho_2}(\rho_{2,i-1})$ and $\rho_{3,i} = f_{\rho_3}(\rho_{3,i-1})$

$\boldsymbol{\phi} = \boldsymbol{\phi} + \rho_{1,i}\, \Delta_{\boldsymbol{\phi}}\, \widetilde{\mathcal{L}}_i$

$\boldsymbol{\xi} = \boldsymbol{\xi} + \rho_{2,i}\, \Delta_{\boldsymbol{\xi}}\, \widetilde{\mathcal{L}}_i$ (for SIMFVI take MFVI updates)

$\boldsymbol{\nu} = \boldsymbol{\nu} + \rho_{3,i}\, \Delta_{\boldsymbol{\nu}}\, \widetilde{\mathcal{L}}_i$ (for SIMFVI take MFVI updates)

$i = i + 1$

end

Algorithm 2: General Simultaneous Credible Intervals.

Input : Samples of posterior splines $\boldsymbol{S}$ with N observations and M samples
Output : Lower and upper bound band of size $N \rightarrow \boldsymbol{b}_l, \boldsymbol{b}_u$

1. Compute expected number of splines within CI:
   $k = (1 - \alpha)\, M$
2. Compute mean and quantiles:
   $\boldsymbol{m} = (m_1, ..., m_N), \quad \boldsymbol{q}_l = (lq_1, ..., lq_N), \quad \boldsymbol{q}_u = (uq_1, ..., uq_N)$

   where $m_i = \frac{1}{M} \sum_{j=1}^{M} \boldsymbol{S}_{ij}, \quad lq_i = Q_{\alpha/2}(\boldsymbol{S}_{i\cdot})$ and $uq_i = Q_{(1-\alpha)/2}(\boldsymbol{S}_{i\cdot})$
3. Compute $\boldsymbol{b}_l$ and $\boldsymbol{b}_u$ and number of splines within bounds with init $c = 1$:
   $\boldsymbol{b}_l = m - c\,(\boldsymbol{q}_l - \boldsymbol{m}), \quad \boldsymbol{b}_u = m - c\,(\boldsymbol{q}_u - \boldsymbol{m})$

   $\tilde{k} = \sum_{j=1}^{M} \left(\boldsymbol{b}_l \leq \boldsymbol{S}_{\cdot j} \leq \boldsymbol{b}_u\right)$
4. Set $c_l = 0, c_u = 0, d_l = 0, d_u = 0$ and run loop:

while $\tilde{k} - k \neq 0$ **do**
  if $c_l = 0$ ***or*** $c_u = 0$ **then**
    $\rho = (1 + |k - \tilde{k}|/n))$
    if $\tilde{k} < k$ **then**
      $c = c\,\rho$
    else
      $c = c/\rho$
  else
    $w_1 = 1/d_l, \quad w_2 = 1/d_u, \quad w_{12} = w_1 + w_2$
    $c = (c_l\, w_1 + c_u\, w_2)/w_{12}$
  repeat Step 3. with updated c to get new bounds and new $\tilde{k}$

  if $\tilde{k} < k$ **then**
    Set $c_l = c$ and $d_l = k - \tilde{k}$
  else
    Set $c_u = c$ and $d_u = \tilde{k} - k$

end

9 Additional tables and figures

9.1 SIVI for additive models

See Fig. 6.

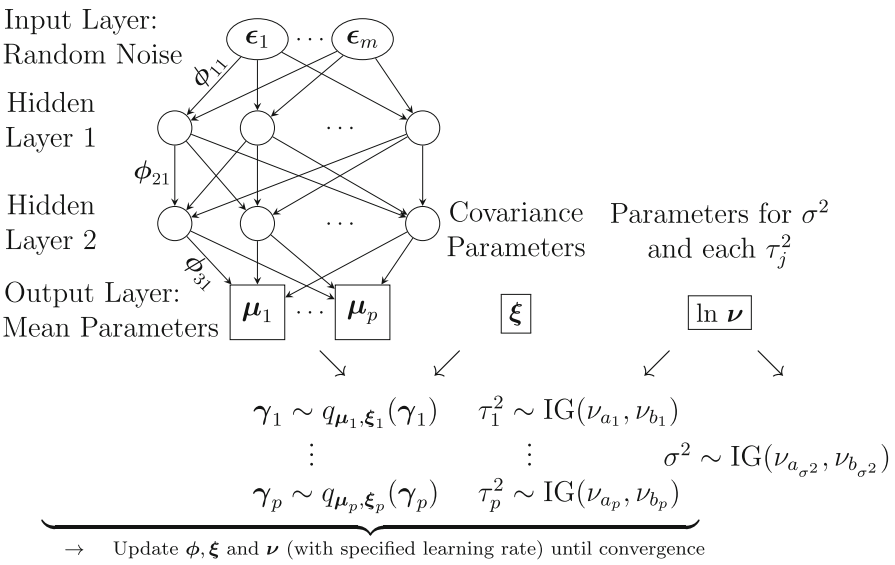


Fig. 6 Illustration of SIVI design for additive models. The neural network takes as input random noise variables $\epsilon_1, \ldots, \epsilon_m$, with each e.g. $\epsilon_1 = (\epsilon_{11}, \ldots, \epsilon_{1K})$, that are transformed through one or more hidden layers with specified activation functions and parameters ϕ to parameters $\mu_1, \ldots, \mu_m$, with each $\mu_j = (\mu_{j1}, \ldots, \mu_{jK})$. We obtain samples of the coefficients γ_j from $q_{\mu_j,\xi_j}(\gamma_j)$ for which we marginalize out μ_j using Monte Carlo integration with K samples. Apart from the neural network parameters ϕ, parameters $\xi = (\xi_1, \ldots, \xi_p)$ and $\nu = ((\nu_{a_1}, \nu_{b_1}), \ldots, (\nu_{a_p}, \nu_{b_1 p}), (\nu_{a_{\sigma^2}}, \nu_{b_{\sigma^2}}))$ are optimized utilizing the Mean-Field assumption

9.2 Simulation results

See Table 2, 3 and Figs. 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17.

Table 2 M2 results for 1000 simulations each 50 observations and strong correlation between covariates

	Gibbs	MFVI (block)	MFVI (full)	SIMFVI early $K = 100$	SIMFVI $K = 100$	SIVI early $K = 100$	SIVI $K = 100$
Intercept							
∅ of lb 95% CI's	1.604	1.608	1.610	1.594	1.595	1.615	1.613
∅ of means	1.800	1.800	1.800	1.800	1.800	1.800	1.800
∅ of ub 95% CI's	1.997	1.993	1.990	2.010	2.009	1.994	1.992
∅ of variances	0.010	0.010	0.010	0.011	0.011	0.009	0.009
Within CI	0.951	0.946	0.945	0.960	0.962	0.932	0.940
Spline 1							
95% CI coverage							
∅ local	0.965	0.822	0.955	0.964	0.97	0.957	0.96
∅ simultaneous	0.977	0.637	0.954	0.955	0.97	0.946	0.954
∅ of MSE's	0.142	0.138	0.142	0.148	0.149	0.154	0.154
Spline 2							
95% CI coverage							
∅ local	0.955	0.805	0.944	0.953	0.961	0.947	0.95
∅ simultaneous	0.96	0.58	0.935	0.942	0.955	0.927	0.932
∅ of MSE's	0.157	0.154	0.158	0.164	0.165	0.168	0.168
Variance σ^2							
∅ of lb 95% CI's	0.310	0.342	0.335	0.356	0.364	0.336	0.337
∅ of means	0.504	0.509	0.498	0.529	0.541	0.499	0.500
∅ of ub 95% CI's	0.804	0.754	0.738	0.784	0.802	0.738	0.741
Within CI	0.958	0.917	0.907	0.917	0.911	0.901	0.906
∅ of MSE's	0.118	0.116	0.117	0.118	0.118	0.123	0.121
n_{eff}							
0	342.0						
0.01	849.1						
0.5	3807.7						

For SIVI and SIMFVI K is set to 100 and S to 50 with results after early stopping was triggered and after 5000 iterations. Shown are the average values over all 1000 simulations, e.g. the lower bound (lb), the upper bound (ub), percentage coverage of local and simultaneous CI's in case of splines, otherwise depicted in row "Within CI", the MSE per spline and overall and also different percentiles for the effective sample size of Gibbs sampler. The columns of SIMFVI and SIVI that do not include the phrase "early" are the results after 5,000 iterations without early stopping. The algorithm for early stopping usually stops around 2000 iterations

Table 3 M2 results for 1000 simulations each 50 observations and strong correlation between covariates

	Gibbs	MFVI (block)	MFVI (full)	SIMFVI early $K = 100$	SIMFVI $K = 100$	SIVI early $K = 100$	SIVI $K = 100$
Spline 1							
95% CI's							
∅ 0–2	0.963	0.792	0.952	0.961	0.969	0.955	0.957
∅ 2–4	0.969	0.89	0.96	0.968	0.971	0.962	0.965
∅ 4–6	0.961	0.749	0.953	0.962	0.971	0.951	0.957
Spline 2							
95.0 % CI's							
∅ −1.5–0.5	0.944	0.705	0.926	0.939	0.951	0.93	0.934
∅ 0.5–2.5	0.951	0.848	0.939	0.944	0.952	0.944	0.946
∅ 2.5–4.5	0.963	0.883	0.955	0.96	0.965	0.957	0.959
∅ 4.5–6.5	0.961	0.753	0.953	0.964	0.972	0.953	0.958

For SIVI/SIMFVI K is set to 100 and S to 50 with results after early stopping was triggered and after 5000 iterations. Depicted are the local CI coverage per predefined interval (e.g. 0–2 shows the average coverage within the interval [0, 2])

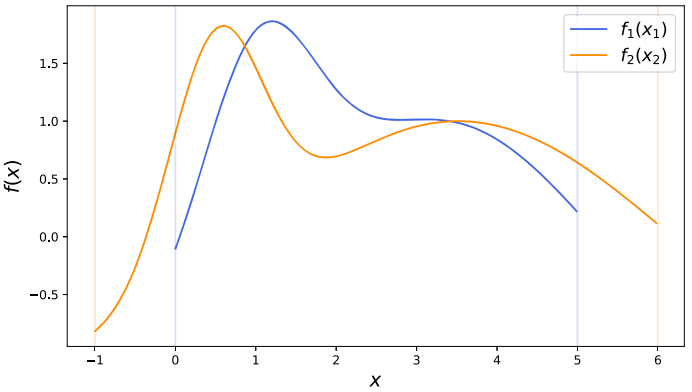


Fig. 7 The true nonlinear marginal effects of the DGP

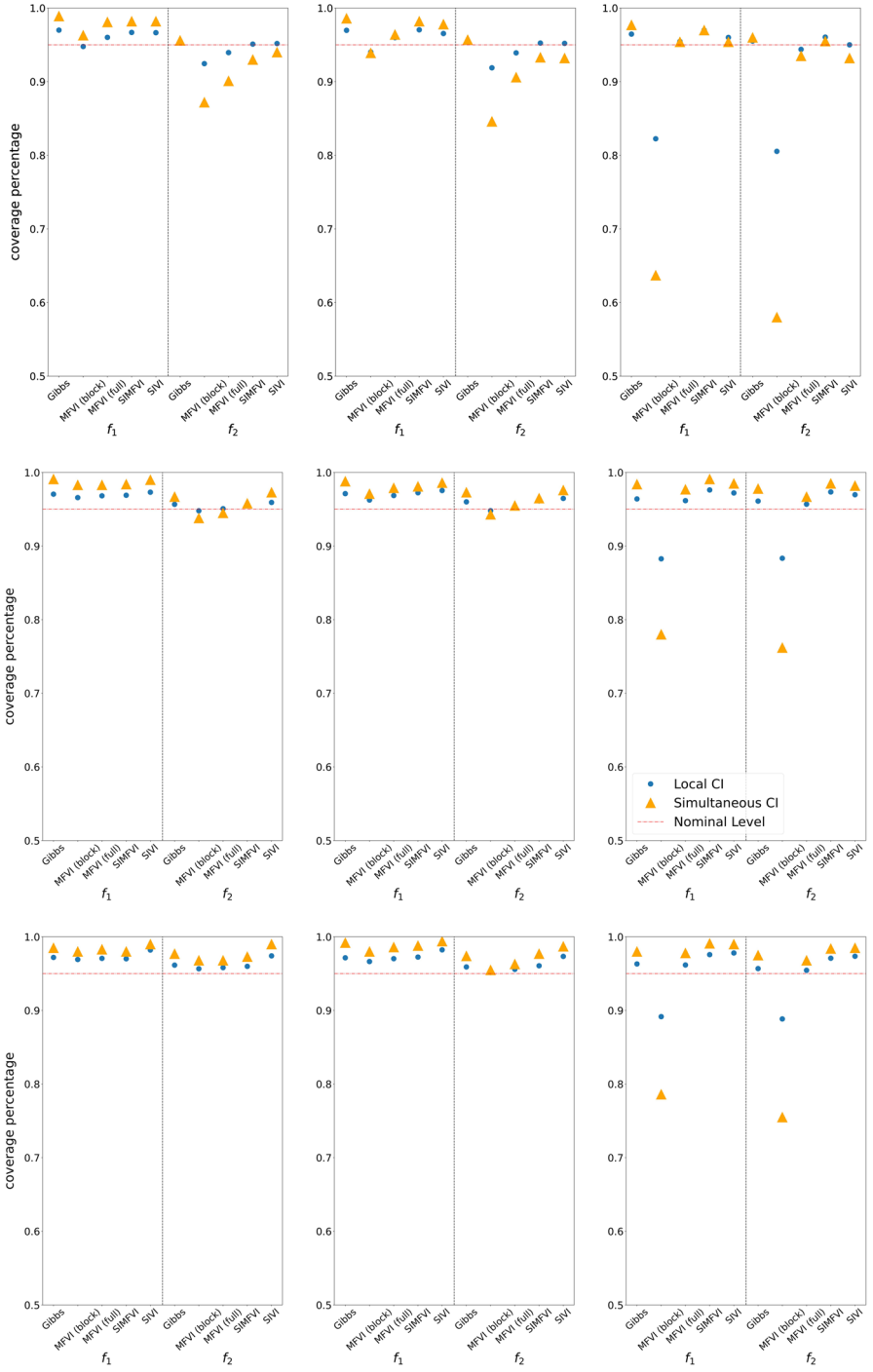


Fig. 8 Coverage percentage among different methods for scenario with 50, 250 and 500 observation (from top to bottom row) and low to high correlation (left to right) between covariates

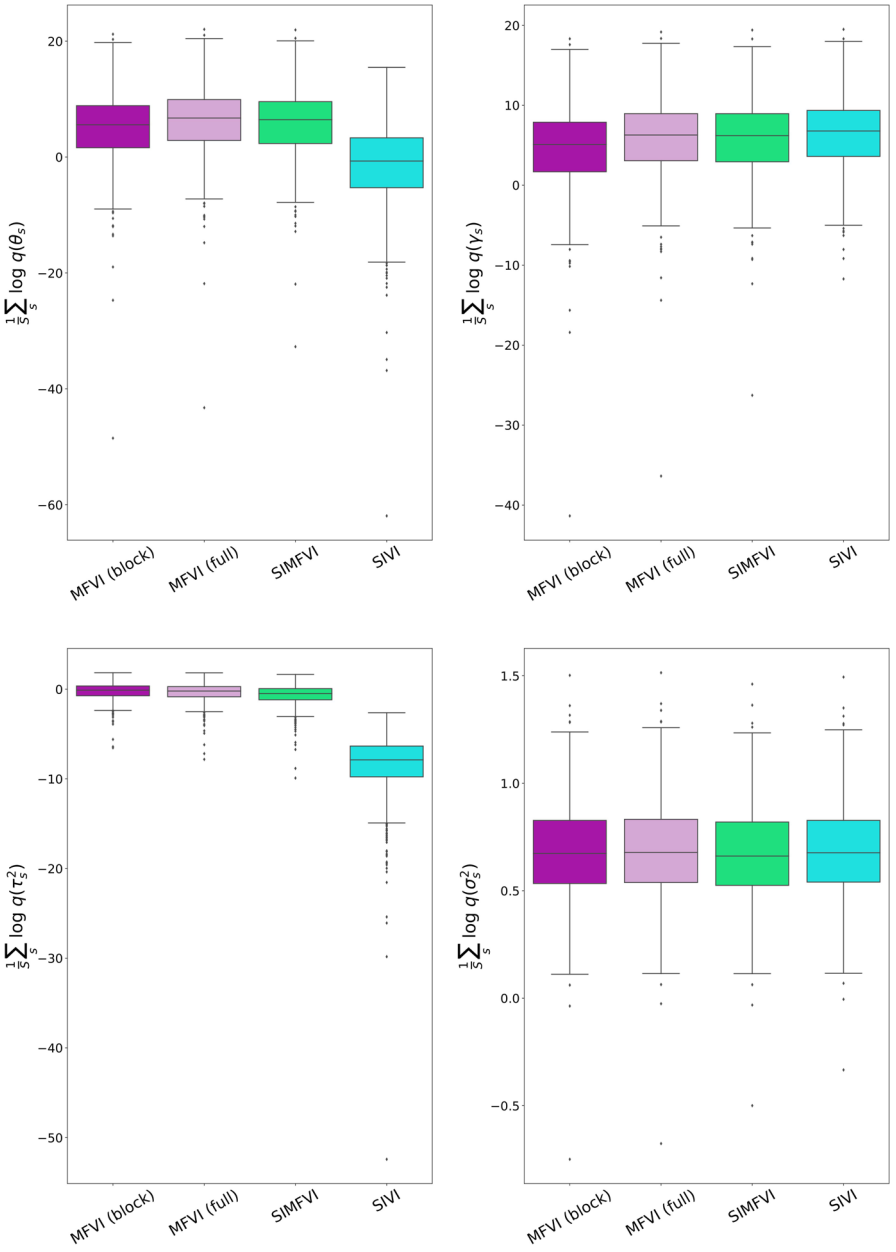


Fig. 9 Average log densities of VI methods given Gibbs samples for scenario with 50 observations and no correlation based on 1000 simulations. The total ALDG (with parameter vector $\boldsymbol{\theta}$) is decomposed into its model parameter components (plots for coefficients $\boldsymbol{\gamma}$, smoothing parameters $\boldsymbol{\tau}^2$ and error variance σ^2)

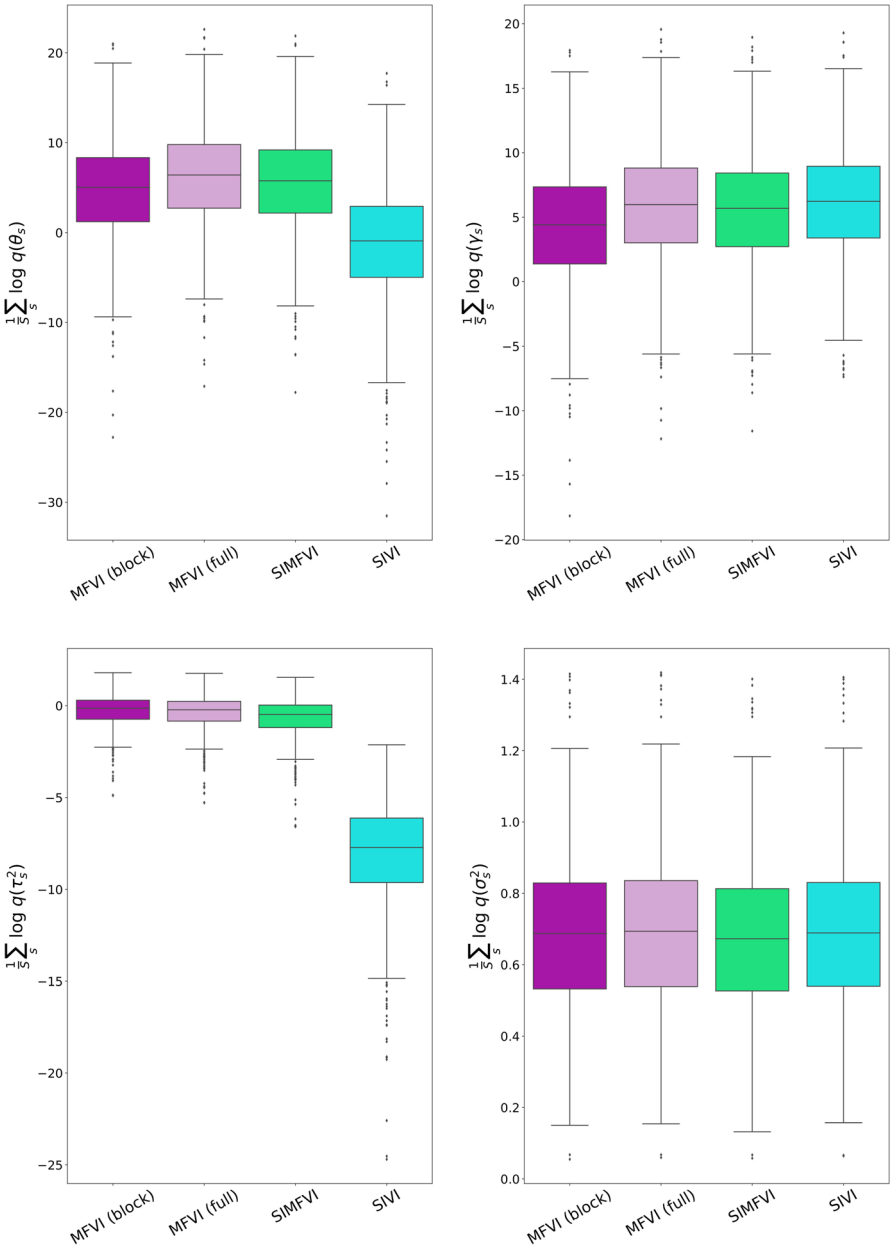


Fig. 10 Average log densities of VI methods given Gibbs samples for scenario with 50 observation and medium correlation based on 1000 simulations. The total ALDG (with parameter vector $\boldsymbol{\theta}$) is decomposed into its model parameter components (plots for coefficients $\boldsymbol{\gamma}$, smoothing parameters $\boldsymbol{\tau}^2$ and error variance σ^2)

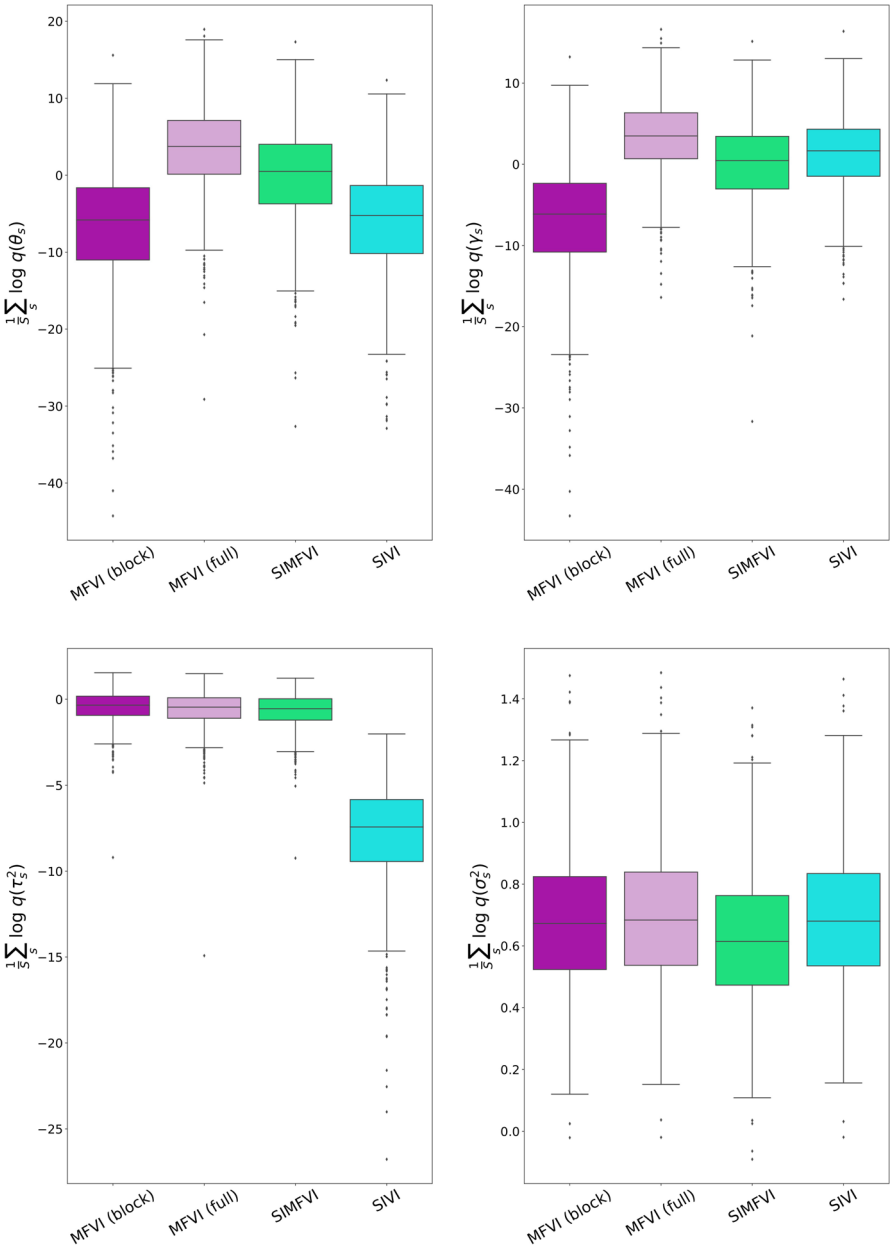


Fig. 11 Average log densities of VI methods given Gibbs samples for scenario 50 observations and high corrlation based on 1000 simulations. The total ALDG (with parameter vector $\boldsymbol{\theta}$) is decomposed into its model parameter components (plots for coefficients $\boldsymbol{\gamma}$, smoothing parameters $\boldsymbol{\tau}^2$ and error variance σ^2)

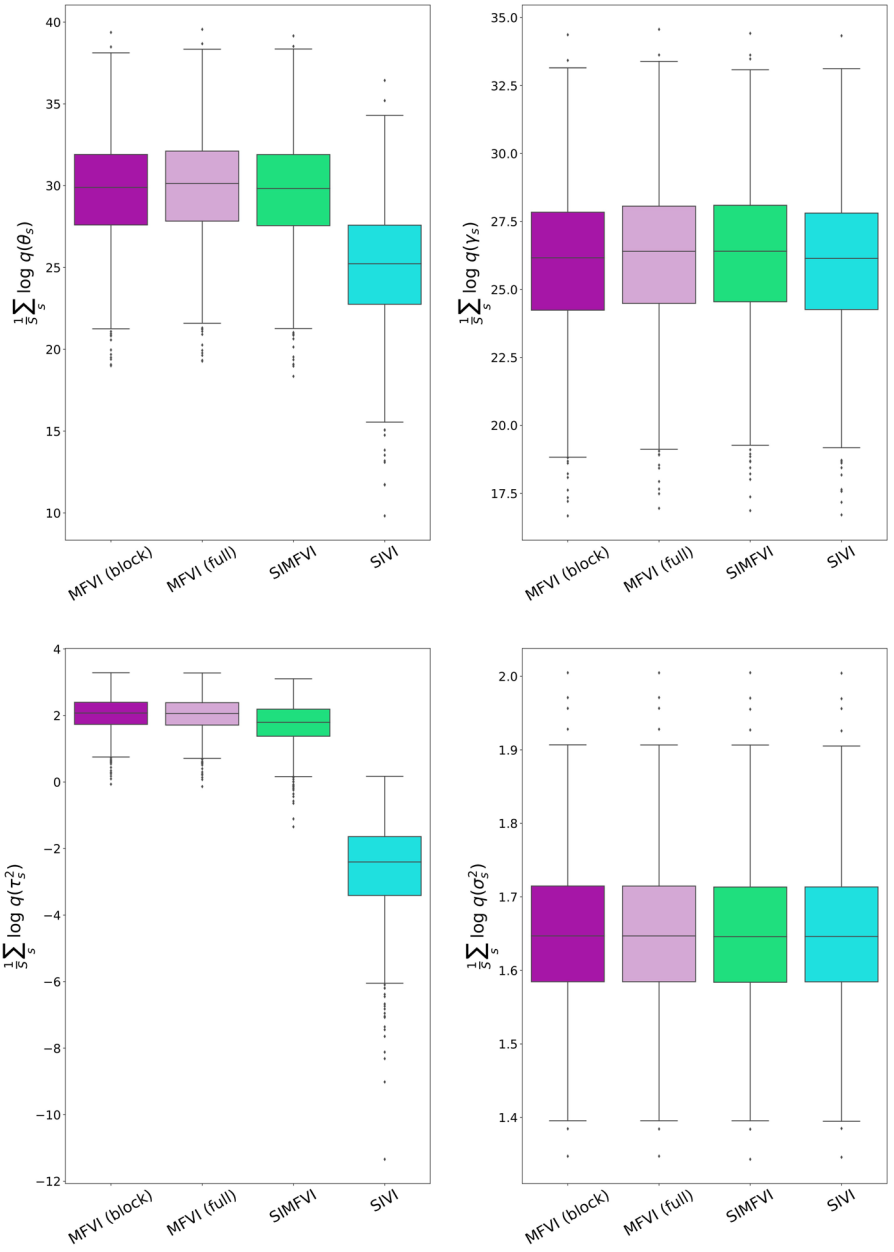


Fig. 12 Average log densities of VI methods given Gibbs samples for scenario 250 observations and no correlation based on 1000 simulations. The total ALDG (with parameter vector $\boldsymbol{\theta}$) is decomposed into its model parameter components (plots for coefficients $\boldsymbol{\gamma}$, smoothing parameters $\boldsymbol{\tau}^2$ and error variance σ^2)

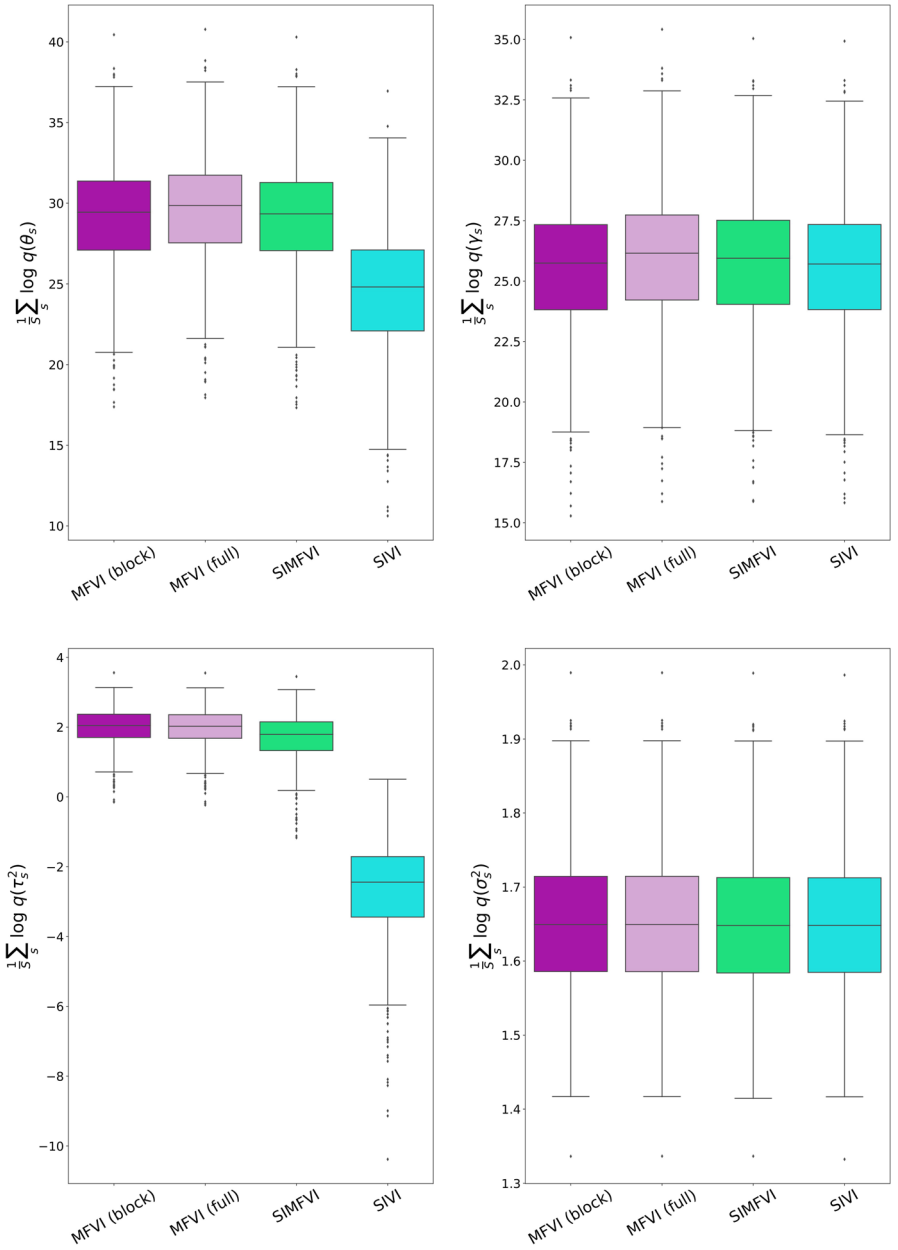


Fig. 13 Average log densities of VI methods given Gibbs samples for scenario with 250 observations and medium correlation based on 1000 simulations. The total ALDG (with parameter vector $\boldsymbol{\theta}$) is decomposed into its model parameter components (plots for coefficients $\boldsymbol{\gamma}$, smoothing parameters $\boldsymbol{\tau}^2$ and error variance σ^2)

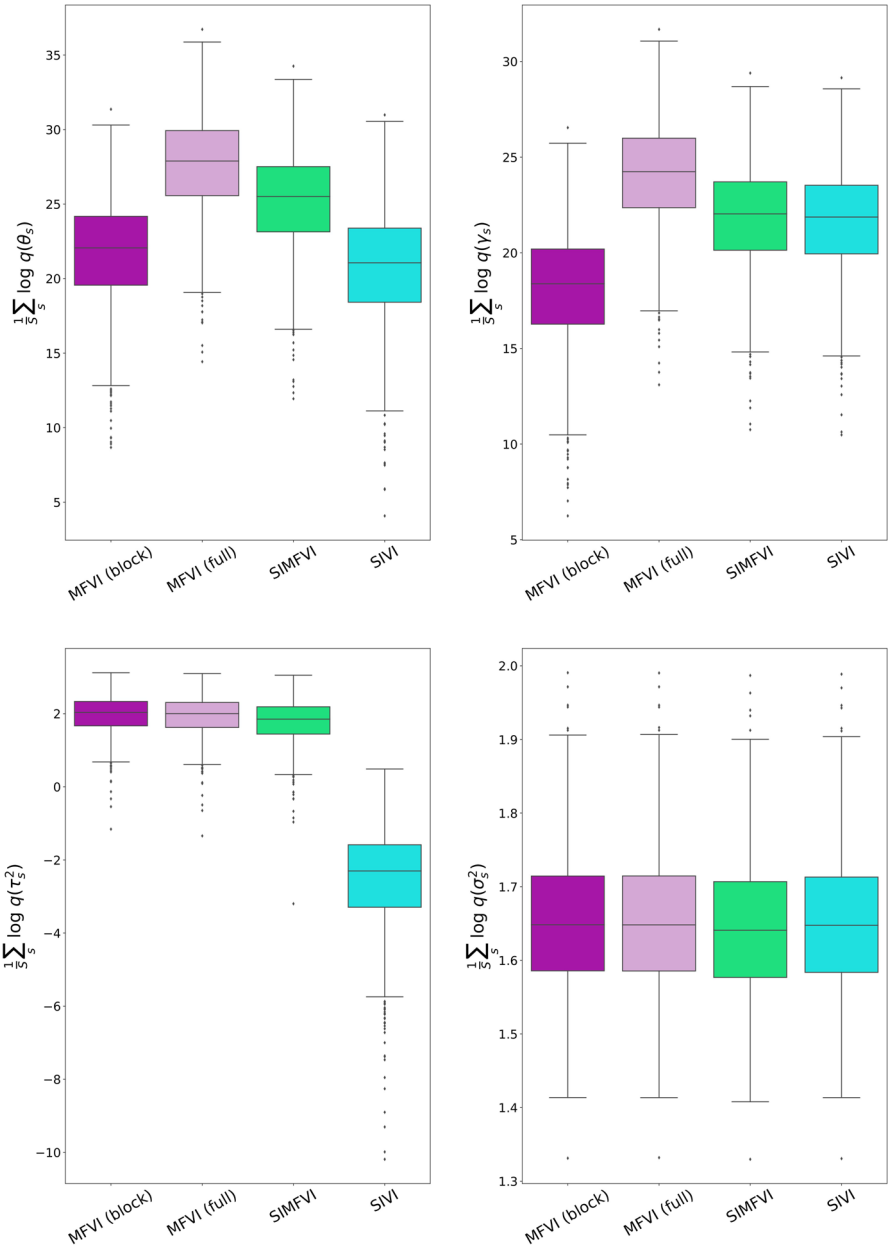


Fig. 14 Average log densities of VI methods given Gibbs samples for scenario with 250 observations and high correlation based on 1000 simulations. The total ALDG (with parameter vector $\boldsymbol{\theta}$) is decomposed into its model parameter components (plots for coefficients $\boldsymbol{\gamma}$, smoothing parameters $\boldsymbol{\tau}^2$ and error variance σ^2)

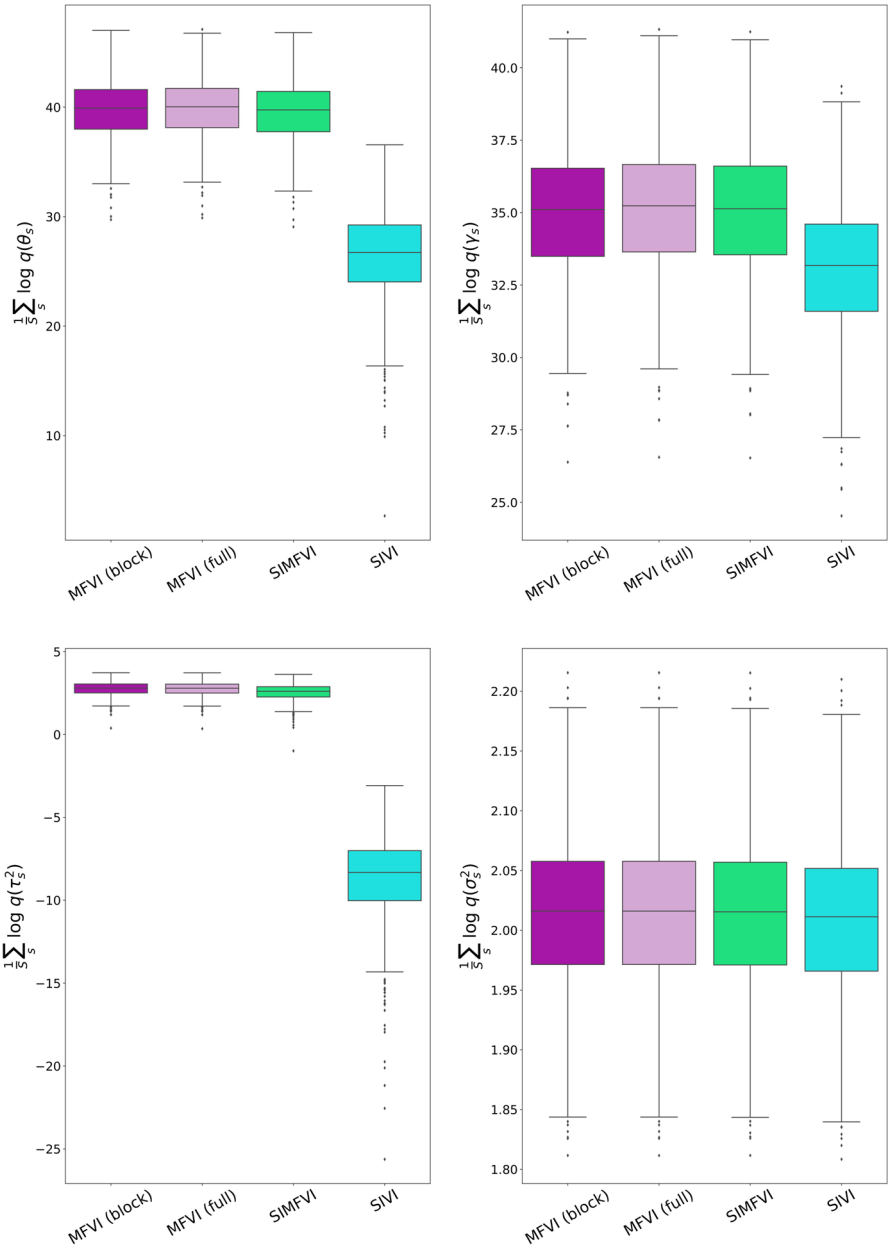


Fig. 15 Average log densities of VI methods given Gibbs samples for scenario with 500 observations and no correlation based on 1000 simulations. The total ALDG (with parameter vector $\boldsymbol{\theta}$) is decomposed into its model parameter components (plots for coefficients $\boldsymbol{\gamma}$, smoothing parameters $\boldsymbol{\tau}^2$ and error variance σ^2)

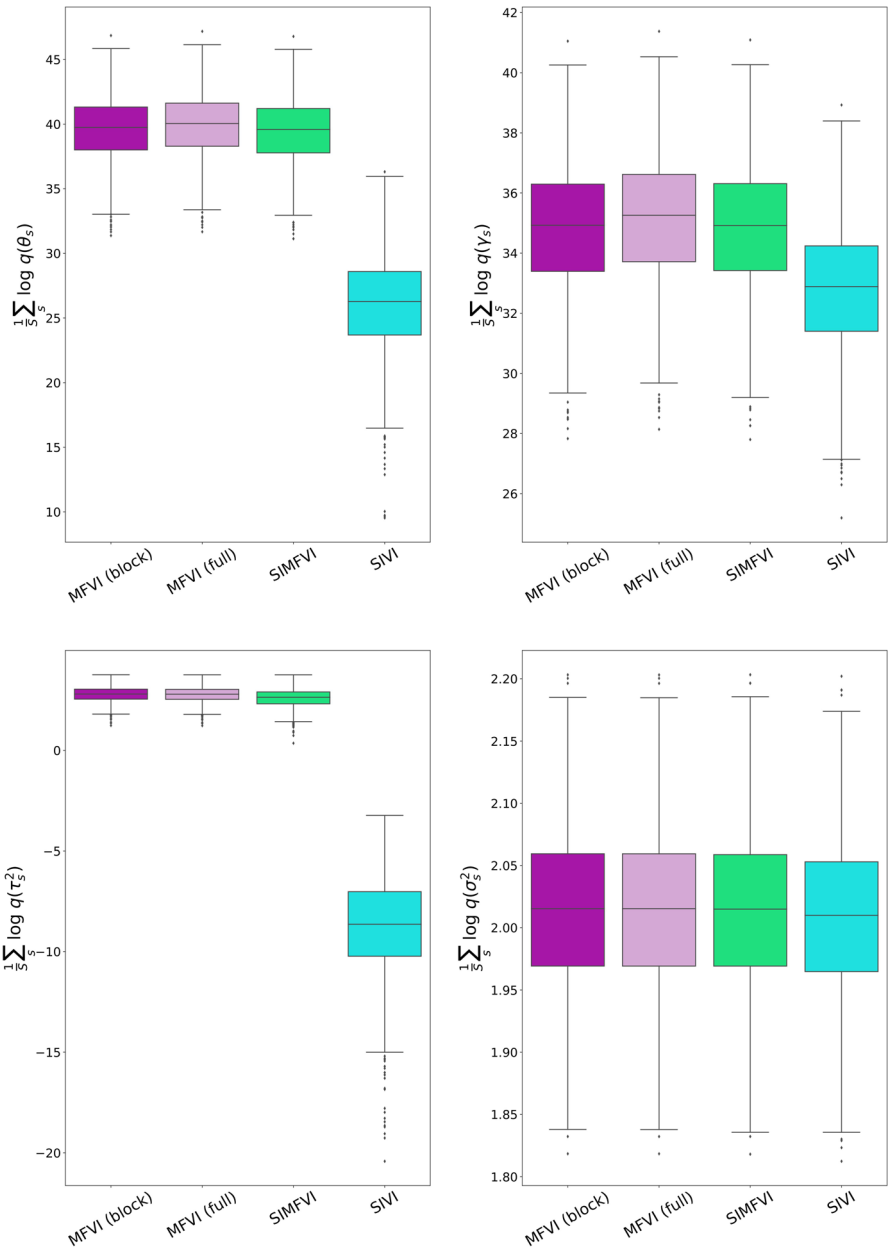


Fig. 16 Average log densities of VI methods given Gibbs samples for scenario with 500 observations and medium correlation based on 1000 simulations. The total ALDG (with parameter vector $\boldsymbol{\theta}$) is decomposed into its model parameter components (plots for coefficients $\boldsymbol{\gamma}$, smoothing parameters $\boldsymbol{\tau}^2$ and error variance σ^2)

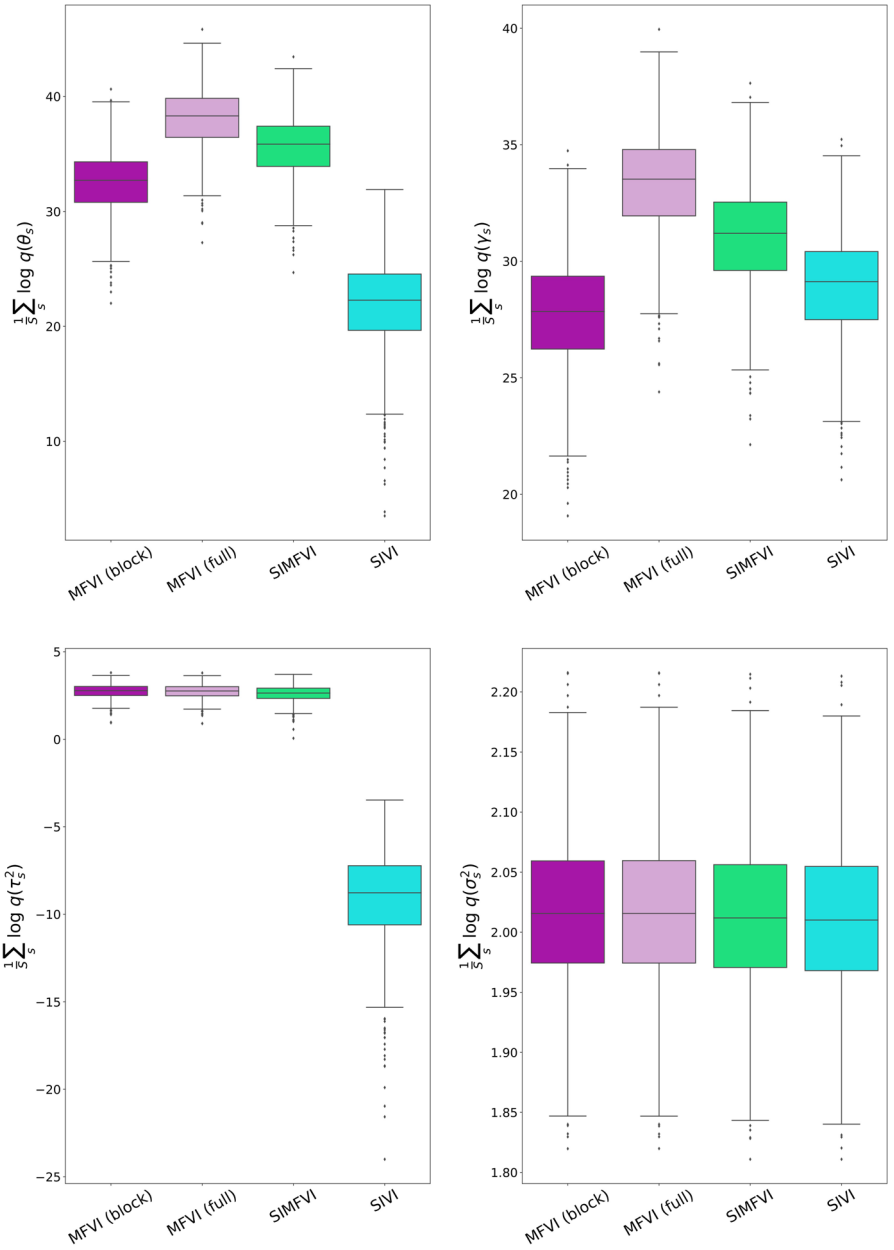


Fig. 17 Average log densities of VI methods given Gibbs samples for scenario with 500 observations and high correlation based on 1000 simulations. The total ALDG (with parameter vector $\boldsymbol{\theta}$) is decomposed into its model parameter components (plots for coefficients $\boldsymbol{\gamma}$, smoothing parameters $\boldsymbol{\tau}^2$ and error variance σ^2)

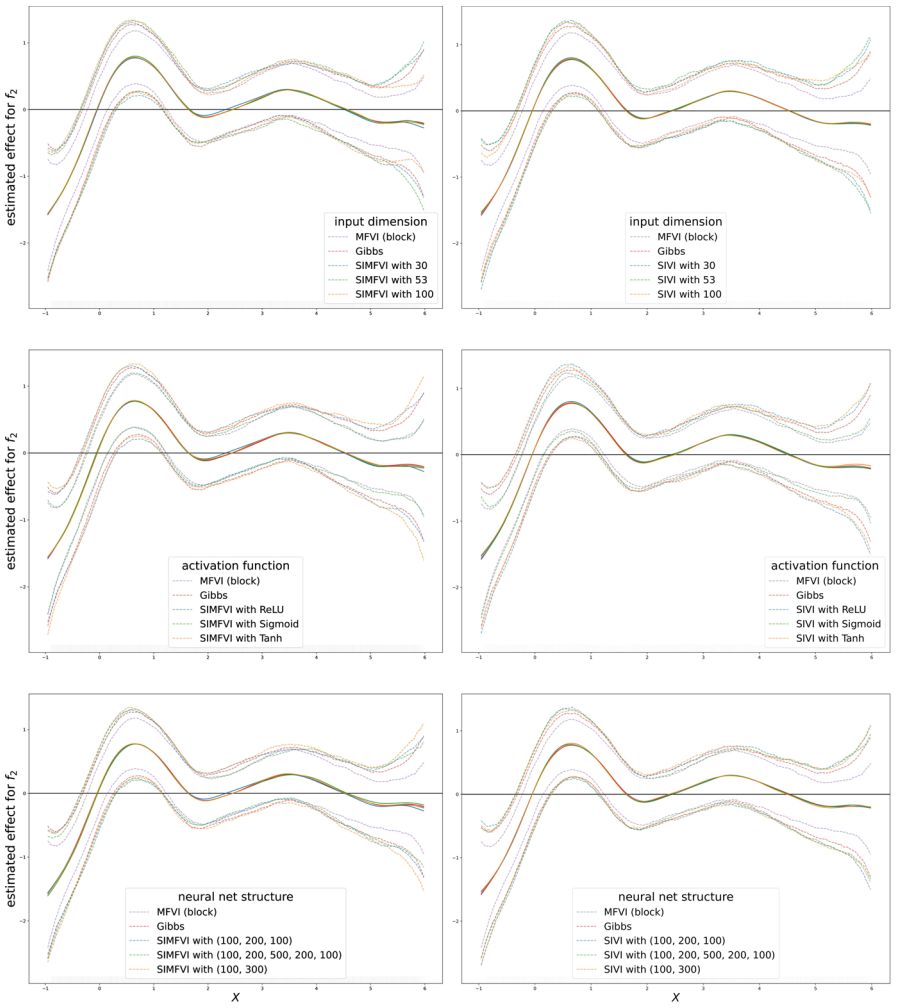


Fig. 18 Sensitivity analysis for SIMFVI (left column) and SIVI (right column) for different neural net specifications. Results for simulation 613 in scenario with 50 observations and high correlation between the covariates

9.3 Tree height model

See Tables 4, 5 and Figs. 19, 20.

Table 4 Results for linear coefficients and error variance for douglas fir height model

	Gibbs	MFVI (block)	MFVI (full)	SIMF K=100	SIVI $K = 100$
Intercept					
2.5% CI	10.2429	10.3604	10.2297	10.218	10.1865
Mean	10.7051	10.6942	10.6985	10.6906	10.6608
97.5% CI	11.1771	11.0279	11.1673	11.1561	11.1875
Coefficient bhd					
2.5% CI	5.3089	5.3688	5.3114	5.3119	5.3561
Mean	5.5239	5.5274	5.5263	5.5238	5.5516
97.5% CI	5.7379	5.686	5.7412	5.7396	5.7718
Coefficient bhd^2					
2.5% CI	−0.2869	−0.2817	−0.2874	−0.2889	−0.3102
Mean	−0.2641	−0.2643	−0.2643	−0.2644	−0.2674
97.5% CI	−0.241	−0.2469	−0.2412	−0.2395	−0.2303
Error Variance σ^2					
2.5% CI	7.3251	7.3675	7.3457	7.3602	7.7121
Mean	7.5866	7.6143	7.5918	7.6068	7.9735
97.5% CI	7.8505	7.8692	7.8459	7.8614	8.2436

Table 5 Results for predictions for douglas fir height model on 30% test data set

	Gibbs	MFVI (block)	MFVI (full)	SIMF K=100	SIVI $K = 100$
MSE					
Known locations	6.8203	6.8426	6.8271	6.8300	6.8500
Unknown locations	11.5759	11.5575	11.5668	11.5496	11.5591
Prediction Interval Coverage					
Known locations	0.9584	0.9591	0.9577	0.9598	0.9635
Unknown locations	0.9318	0.9356	0.9394	0.9356	0.9432

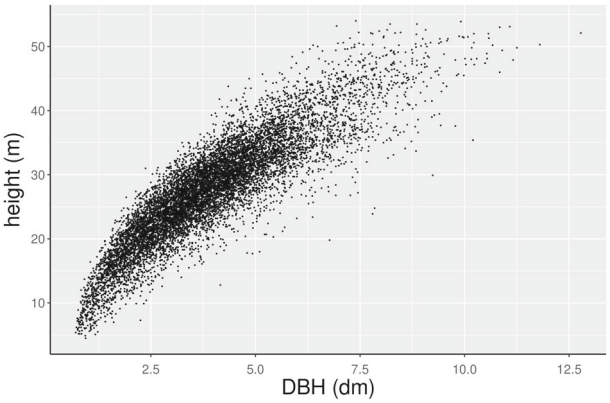


Fig. 19 Descriptive plot on height-DBH relation

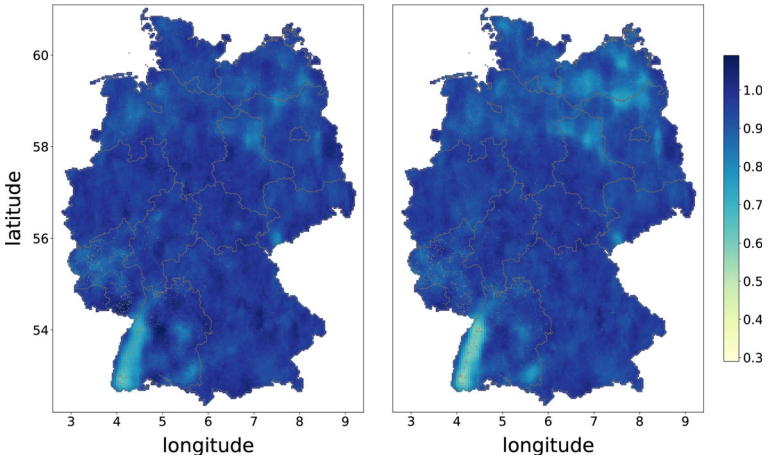


Fig. 20 Width of 95% simultaneous CI of 2-dimensional spline as share to the CI width of Gibbs sampler for SIMFVI with $K = 300$ (left) and $K = 100$ (right). 100% (darkblue) stands for same width as Gibbs sampler

References

Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, New York (2006)

Blei, D.M., Kucukelbir, A., McAuliffe, J.D.: Variational inference: a review for statisticians. J. Am. Stat. Assoc. **112**(518), 859–877 (2017)

Diggle, P.J., Gratton, R.J.: Monte Carlo methods of inference for implicit statistical models. J. R. Stat. Soc. B **46**(2), 193–212 (1984)

Fahrmeir, L., Kneib, T., Lang, S.: Penalized structured additive regression for space-time data: a Bayesian perspective. Stat. Sin. **14**, 731–761 (2004)

Fahrmeir, L., Kneib, T., Lang, S., Marx, B.: Regression—Models, Methods and Applications. Springer, Berlin (2021)

Gregoire, T.G., Valentine, H.T.: Sampling Strategies for Natural Resources and the Environment. Chapman and Hall/CRC, London (2007)

Harris, C.R., Millman, K.J., Walt, S.J.V., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., Kerkwijk, M.H., Brett, M., Haldane, A., Río,

J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E.: Array programming with NumPy. Nature **585**(7825), 357–362 (2020)
He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2015)
Hui, F.K.C., You, C., Shang, H.L., Müller, S.: Semiparametric regression using variational approximations. J. Am. Stat. Assoc. **114**(528), 1765–1777 (2019)
Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., Saul, L.K.: An introduction to variational methods for graphical models. Mach. Learn. **37**(2), 183–233 (1999)
Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. CoRR, arXiv:1412.6980 (2015)
Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings. arXiv:1312.6114v10 (2014)
Krivobokova, T., Kneib, T., Claeskens, G.: Simultaneous confidence bands for penalized spline estimators. J. Am. Stat. Assoc. **105**(490), 852–863 (2010)
Lang, S., Brezger, A.: Bayesian P-Splines. J. Comput. Graph. Stat. **13**(1), 183–212 (2004)
Luts, J., Broderick, T., Wand, M.P.: Real-time semiparametric regression. J. Comput. Graph. Stat. **23**(3), 589–615 (2014)
Luts, J., Wand, M.P.: Variational inference for count response semiparametric regression. Bayesian Anal. **10**(4), 991–1023 (2015)
Ormerod, J.T., Wand, M.P.: Explaining variational approximations. Am. Stat. **64**(2), 140–153 (2010)
Padgham, M., Lovelace, R., Salmon, M., Rudis, B.: Osmdata. J. Open Source Softw. (2017). https://doi.org/10.21105/joss.00305
Parisi, G.: Statistical Field Theory. Addison-Wesley, Redwood City (1988)
Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: PyTorch: an imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems, vol. 32. Curran Associates, Inc, pp. 8024–8035 (2019)
Pya, N., Schmidt, M.: Incorporating shape constraints in generalized additive modelling of the height-diameter relationship for Norway spruce. For Ecosyst **3**, 1–14 (2016)
Ranganath, R., Tran, D., Blei, D.M.: Hierarchical variational models. In: International Conference on Machine Learning, pp. 324–333 (2016)
Saul, L.K., Jordan, M.I.: Exploiting tractable substructures in intractable networks. In: Advances in Neural Information Processing Systems, Cambridge, Mass. MIT Press, pp. 486–492 (1998)
Turkkan, N., Pham-Gia, T.: Computation of the highest posterior density interval in Bayesian analysis. J. Stat. Comput. Simul. **44**(3–4), 243–250 (1993). (**Publisher: Taylor & Francis**)
Wainwright, M.J., Jordan, M.I.: Graphical models, exponential families, and variational inference. Found. Trends® Mach. Learn. **1**(1–2), 1–305 (2008)
Waldmann, E., Kneib, T.: Variational approximations in geoadditive latent Gaussian regression: mean and quantile regression. Stat. Comput. **25**(6), 1247–1263 (2015)
Wang, Y., Blei, D.M.: Frequentist consistency of variational Bayes. J. Am. Stat. Assoc. **114**(527), 1147–1161 (2019)
Wood, S.N.: Generalized additive models: an introduction with R, 2 edition Chapman and Hall/CRC, London (2017)
Yin, M., Zhou, M.: Semi-implicit variational inference. In: Proceedings of the 35th International Conference on Machine Learning, pp 5646–5655. PMLR (2018)
You, C., Ormerod, J.T., Mueller, S.: On variational Bayes estimation and variational information criteria for linear regression models. Aust. New Zealand J. Stat. **56**(1), 73–87 (2014)
Zhang, C., Bütepage, J., Kjellström, H., Mandt, S.: Advances in variational inference. IEEE Trans. Pattern Anal. Mach. Intell. **41**(8), 2008–2026 (2018)