A scalable Bayesian method for estimating a fixed number of coordinates in the high-dimensional sparse linear regression model

Ismaël Castillo [1] & Alice L'Huillier[2] & Kolyan Ray [3] & Luke Travis [4]

[1] *LPSM, Sorbonne Université, France, ismael.castillo@upmc.fr*
[2] *LPSM, Sorbonne Université, France, alice.lhuillier@sorbonne-universite.fr*
[3] *Department of Mathematics, Imperial College London, United Kingdom, kolyan.ray@imperial.ac.uk*
[4] *Department of Mathematics, Imperial College London, United Kingdom, luke.travis15@imperial.ac.uk*

Résumé. Dans ce travail, on se place dans le contexte du modèle de régression linéaire en grande dimension sous contrainte de sparsité. On souhaite obtenir un intervalle de confiance pour une coordonnée (ou un nombre fixe de coordonnées) du vecteur de régression par une méthode bayésienne. Pour cela, on définit une loi a priori sur les vecteurs sparses qui cible la coordonnée d'intérêt. L'échantillonage du posterior correspondant étant coûteux, on propose une approximation variationnelle du posterior qui va, elle aussi, cibler la coordonnée d'intérêt. D'un côté, on étudie cette méthode du point de vue théorique en montrant un résultat de type Berstein-von Mises pour l'approximation variationnelle sous des hypothèses de compatibilité sur la matrice de design. Ensuite, la méthode est implémentée sur des données simulées et montre de bonnes performances dans différents cadres, tout en ayant un temps de calcul comparable aux autres méthodes fréquentistes.

Mots-clés. intervalle de confiance, modèle de régression linéaire en grande dimension, sparsité, approximation variationnelle, Berstein-von Mises.

Abstract. We study the problem of providing confidence intervals for one or a fixed number of coordinates in the high-dimensional linear regression model under sparsity constraints via a Bayesian approach. We define an appropriate sparse prior on the whole regression vector that targets the coordinates of interest. Simulating from the actual posterior distribution in this setting may be computationally intensive. To overcome this difficulty, we propose a variational approximation to the posterior designed to target the coordinates of interest. This tractable approximation can then be used to construct confidence intervals. We give theoretical guarantees for the proposed method by deriving a Berstein-von Mises theorem for the variational approximation under compatibility conditions on the design matrix. The method is implemented on simulated data illustrating that our Bayesian procedure can be computed in comparable time to other frequentist methods while exhibiting favourable performance in a number of different settings.

Keywords. confidence intervals, high-dimensional linear regression model, sparsity, variational approximation, Berstein-von Mises.

1 Introduction

In this work we consider the high-dimensional linear regression model

$$Y = X\beta^0 + \varepsilon, \tag{1}$$

where $Y \in \mathbb{R}^n$, X is a given deterministic $n \times p$ design matrix with $p > n$, $\beta^0 \in \mathbb{R}^p$ is the unknown parameter, and $\varepsilon \sim \mathcal{N}_n(0, I_n)$ is a Gaussian noise. Here we are interested in the case where β^0 is sparse : only a small fraction of coefficients β_i^0 are non-zero. Our goal is to infer one coordinate or more generally a fixed number of coordinates of the regression vector β^0 by providing (asymptotic) confidence intervals.

The works [7], [5] and [3] addressed this problem by constructing confidence intervals from estimators built from the LASSO.

In a Bayesian setting, the problem of estimating one coordinate, say β_1^0, has been studied by [6]. In [6], the author proposes an appropriate model-selection type prior and derives a Berstein-von Mises (BvM) type result for the posterior induced on the first coordinate. This implies that one can construct confidence intervals for β_1^0 from the posterior by considering the credible intervals.

When using model selection type prior, as in [6], the posterior may be challenging to sample from for large dimensions n and p and hence the difficulty to approximate credible intervals. In this work, our goal is to develop a Bayesian method scalable to high-dimensional settings.

A popular approach to develop scalable Bayesian methods is variational Bayes where one computes the best approximation of the posterior within a class of simpler distributions. Then this (tractable) approximation is used in place of the posterior to conduct the inference. We will use this approach to develop our method as we briefly describe now.

First, we borrow from the work of [6] to define a prior that gives a posterior from which one is able to construct confidence intervals. As in [6], the posterior obtained is hard to simulate from directly. Our strategy consists in proposing a variational approximation of the posterior relying on the mean-field variational approximation studied in [4] and using it to conduct inference on β_1^0.

We derive theoretical guarantees for the proposed method. We also investigate its performance on simulated data and compare it with that of the methods proposed by [7] and [3]. Finally, we show that the method can be extended to infer a subset of k coordinates. However, to be concise, we will not present this last part of our work in this abstract.

Notations. For $\beta \in \mathbb{R}^p$ we denote $\beta_{-1} = (\beta_i)_{i=2}^p \in \mathbb{R}^{p-1}$ the vector of the last $p-1$ coordinates of β. We denote $X_i \in \mathbb{R}^n$ the i^{th} column of X and X_{-1} the matrix $(X_2, \ldots, X_p) \in \mathbb{R}^{n\times(p-1)}$ consisting of the last $p-1$ columns of X. The Lebesgue measure on $\mathbb{R}$ is denoted by Λ. The Laplace distribution on $\mathbb{R}$, denoted $\mathrm{Lap}(\mu, \sigma)$, has density proportional to $x \to e^{-|x-\mu|/\sigma}$. The Kullback-Leibler divergence between two probability distributions P and Q is denoted by $KL(P, Q)$.

2 Methodology

Prior's construction

To build a sparse prior on $\beta \in \mathbb{R}^p$ that targets β_1^0 , we borrow two ideas used in [6] for the prior's construction. First, we put a sparse prior only on β_{-1} instead of a sparse prior on the whole vector β as is usually considered for estimation of the whole regression vector β_0 (see e.g. [1] [2]). Then, on the first coordinate β_1, we consider a continuous distribution. Here, we use the second idea considered in [6]. Define $\gamma_i := X_1^T X_i / \|X_1\|_2^2$ for $i = 2, \dots, p$ the rescaled correlation between the i^{th} and 1^{st} columns of the design matrix X and

$$\beta_1^* := \beta_1 + \sum_{i=2}^{p} \gamma_i \beta_i. \tag{2}$$

In the prior, we sample independently β_1^* and β_{-1} with β_1^* sampled from a continuous distribution on $\mathbb{R}$ and β_{-1} sampled from a sparse prior. As we will see in the following, this allows for a simple characterisation of the posterior. The sparse prior we consider on β_{-1} is the model selection prior defined as follows.

Definition 1 (Model selection prior)**.** For $d \geq 1$, ν a probability distribution on $\{0, \dots, d\}$ and $\lambda > 0$, the *model selection* prior on $u \in \mathbb{R}^d$, denoted $MS_d(\nu, \lambda)$, is defined in the following hierarchical manner:

1. The sparsity s of u is distributed according to ν.
2. The active set S of u, given s, is uniform on the $\binom{d}{s}$ subsets of $\{1, \dots, d\}$ of size s.
3. $u_i | S \overset{ind}{\sim} \begin{cases} \text{Lap}(\lambda), & i \in S, \\ \delta_0, & i \notin S, \end{cases}$
   for δ_0 the Dirac mass at 0.

The popular spike-and-slab prior where $\beta_i \sim q \cdot \text{Lap}(\lambda) + (1 - q) \cdot \delta_0$ independently with δ_0 the Dirac mass at 0 and $q \in [0, 1]$ fits within Definition 1 with $\nu \sim \text{Binomial}(d, q)$. We now define the prior we consider on the full parameter β.

Definition 2. For ν a probability distribution on $\{0, \dots, p-1\}$, $\lambda > 0$ and g a positive density with respect to Lebesgue measure Λ, consider the following prior distribution Π on $\beta \in \mathbb{R}^p$:

$$\begin{aligned} \beta_{-1} &\sim MS_{p-1}(\nu, \lambda) \\ \beta_1 | \beta_{-1} &\sim g(\cdot + \sum_{i=2}^{p} \gamma_i \beta_i) d\Lambda. \end{aligned} \tag{3}$$

Note that under this prior, we indeed have β_{-1} independent of β_1^* (cf (2)) and $\beta_1^* \sim g$ where g is a continuous distribution on $\mathbb{R}$.

We consider two particular examples of g for the prior (3).

Example 1 (Laplace prior). For $\sigma_n > 0$, take $g = g_n$ the $\mathrm{Lap}(0, \sigma_n)$ distribution. Then the prior reduces to $\beta_{-1} \sim MS_{p-1}(\nu, \lambda)$, $\beta_1|\beta_{-1} \sim \mathrm{Lap}(-\sum_{i=2}^p \gamma_i \beta_i, \sigma_n)$.

Example 2 (Improper prior). Consider the prior measure $\Lambda \otimes MS_{p-1}(\nu, \lambda)$ which can be seen as the prior (3) with $g = 1$.

Posterior's characterisation

Similarly as in [6], we have a simple characterisation of the posterior induced by the prior (2). Indeed, under the posterior β_{-1} and β_1^* are independent. The posterior distribution of β_{-1} is the posterior distribution in a linear regression model induced by a model selection prior (Definition 1) and we have an explicit formula for the posterior density of β_1^*. This is formalized in Lemma 1.

Define $P \in \mathbb{R}^{n \times (n-1)}$ as a matrix whose columns are an orthonormal basis of $\mathrm{span}(X_1)^\perp$, $\check{W} = P^T X_{-1} \in \mathbb{R}^{(n-1)\times(p-1)}$ and $\check{Y} = P^T Y \in \mathbb{R}^{n-1}$.

Lemma 1. *Let* Π *be the prior* (3)*. Then under the posterior distribution,* β_{-1} *and* β_1^* *are independent. Furthermore, the posterior distribution of* β_{-1} *satisfies*

$$d\pi(\beta_{-1}|Y) \propto e^{-\frac{1}{2}\|\check{Y} - \check{W}\beta_{-1}\|_2^2} dMS_{p-1}(\nu, \lambda), \tag{4}$$

where $\check{Y} \overset{P_0}{\sim} \mathcal{N}(\check{W}\beta_{-1}^0, I_{n-1})$*. Moreover, the posterior distribution of* β_1^* *satisfies*

$$d\pi(\beta_1^*|Y) \propto e^{-\frac{1}{2}\|X_1\|_2^2 \left(\beta_1^* - \frac{X_1^T Y}{\|X_1\|_2^2}\right)^2} g(\beta_1^*) d\beta_1^*. \tag{5}$$

In the particular case of Example 2, the posterior is proper and (5) *gives that* $\beta_1^*|Y \sim \mathcal{N}\left(\frac{1}{\|X_1\|_2^2} X_1^T Y, \frac{1}{\|X_1\|_2^2}\right)$.

Now that we have a characterisation of the posterior, the next step in the bayesian method is to sample from the posterior distribution of β_1 to approximate the credible intervals. From Lemma 1, one can easily derive a way of sampling from the posterior distribution of β_1: (i) sample β_{-1} according to (4), (ii) sample β_1^* according to (5), (iii) compute $\beta_1 = \beta_1^* - \sum_{i=2}^p \gamma_i \beta_i$. Step (ii) requires to sample from a 1-dimensional distribution and can be done using standard computational tools (explicit computations or MCMC) relatively quickly. However, due to the model selection prior, step (i) is very intensive for large dimensions n and p if one uses MCMC. To overcome this difficulty, we will now consider a variational approximation of the posterior.

Variational approximation of the posterior

The difficult part in the sampling procedure described above is to sample from the posterior distribution of β_{-1}, that is, to sample from the posterior distribution in the linear regression model when one considers the model selection prior (Definition 1). In [4], the authors study

a mean-field VB approximation of the posterior distribution in this context. They show that this approximation performs well theoretically, and also provide a fast algorithm to compute it in the specific case of the spike-and-slab prior. The results of [4] along with the decoupled form of the posterior given in Lemma 1 have inspired the following approach. The idea is to approximate the posterior distribution of β_{-1} while retaining the exact posterior distribution of β_1^*.

More precisely, one approximates the posterior distribution of β_{-1} by an element of the mean-field family

$$\mathcal{Q}_{-1} = \left\{ Q_{\mu,\tau,q} = \bigotimes_{i=2}^{p} q_i \mathcal{N}(\mu_i, \tau_i^2) + (1-q_i)\delta_0 : q_i \in [0,1], \mu_i \in \mathbb{R}, \tau_i \in \mathbb{R}^+ \right\},$$

obtaining

$$\hat{Q}_{-1} = \underset{Q_{-1} \in \mathcal{Q}_{-1}}{\arg\min}\, KL(Q_{-1} || \Pi_{-1}(\cdot|Y)), \tag{6}$$

for $\Pi_{-1}(\cdot|Y)$ the marginal distribution of β_{-1} in the posterior, defined by (4). We then form a full approximation of the posterior by using the exact posterior distribution of $\beta_1^*|Y$. In other words, we consider the approximation

$$\begin{aligned} \beta_{-1} \sim \hat{Q}_{-1}, \quad \beta_1^* \sim \pi(\beta_1^*|Y), \quad \beta_{-1} \perp\!\!\!\perp \beta_1^* \\ (\beta_1, \beta_{-1}) = (\beta_1^* - \sum_{i=2}^{p} \gamma_i \beta_i, \beta_{-1}), \end{aligned} \tag{7}$$

where $\pi(\beta_1^*|Y)$ is defined in (5). In the following, we denote by $\hat{\Pi}$ the distribution on β given by (7), and by $\hat{\Pi}(\beta_1)$ the marginal distribution of β_1 under $\hat{\Pi}$.

Then we are able to conduct inference on β_1^0 by simply plugging in the variational approximation (7) in the standard posterior-based approach. That is to say, denoting q_γ the γ-quantile of $\hat{\Pi}(\beta_1)$, we consider the quantile-based credible interval

$$C_\gamma := (q_{\gamma/2}, q_{1-\gamma/2}), \tag{8}$$

to give a $1-\gamma$ confidence interval for β_1^0. In practice we approximate this credible interval by simulation.

3 Theoretical guarantees

Let us now briefly describe the theoretical guarantees that we provide for the method presented in Section 2. We show that in the asymptotic regime $n, p \to \infty$, under some conditions, the quantiles-based credible interval (8) computed from the variational distribution $\hat{\Pi}(\beta_1)$ is an asymptotic confidence interval for the truth β_1^0. Thus the variational posterior-based inference is not only computable but also reliable.

This theoretical guarantee is provided by deriving a Berstein-von Mises (BvM) type result for the variational approximation $\hat{\Pi}(\beta_1)$. More precisely, defining $\hat{\beta}_1 := \beta_1^0 + \frac{X_1^T \varepsilon}{\|X_1\|_2^2}$, we show that, under some conditions, $\hat{\Pi}(\beta_1)$ resembles a Gaussian centered at $\hat{\beta}_1$ with variance $1/\|X_1\|_2^2$.

This means formally that for d_{BL} the bounded Lipschitz distance between probability distributions, we have the following result :

Theorem 1. *Under some conditions on the prior* (3)*, the design matrix X and the truth β^0, we have*

$$d_{BL}\left(\|X_1\|_2(\hat{\Pi}(\beta_1) - \hat{\beta}_1), \mathcal{N}(0,1)\right) \xrightarrow{P_0} 0. \tag{9}$$

The distributional approximation (9) implies that the quantiles-based credible interval (8) is an asymptotic confidence interval for the truth β_1^0.

Let us now describe informally how we will prove (9) and discuss briefly the conditions we will require on the prior, the truth and the design matrix.

Step 1 : By the definition (7), under the variational approximation, we have $\beta_1^* \sim \pi(\beta_1^*|Y)$. We will require g to not decrease too quickly to zero to have $\beta_1^* \approx \mathcal{N}(\frac{X_1^T Y}{\|X_1\|^2}, \frac{1}{\|X_1\|_2^2})$.

Step 2: Then one can deduce $\beta_1^* \approx \mathcal{N}(\frac{X_1^T Y}{\|X_1\|^2}, \frac{1}{\|X_1\|_2^2}) = \mathcal{N}(\hat{\beta}_1, \frac{1}{\|X_1\|_2^2}) + \sum_{i=2}^p \gamma_i \beta_i^0$. Therefore, we have

$$\beta_1 = \beta_1^* - \sum_{i=2}^p \gamma_i \beta_i \approx \mathcal{N}(\hat{\beta}_1, \frac{1}{\|X_1\|_2^2}) - \sum_{i=2}^p \gamma_i(\beta_i - \beta_i^0).$$

Consequently, if the second term is negligible with respect to the first term, namely with respect to $1/\|X_1\|_2$, then the result holds. For this, we use the bound

$$\|X_1\|_2 |\sum_{i=2}^p \gamma_i(\beta_i - \beta_i^0)| \leq \|X_1\|_2 \max_{i=2,\dots,p} |\gamma_i| \|\beta_{-1} - \beta_{-1}^0\|_1,$$

combined with a convergence rate for β_{-1} under the variational approximation and the assumption that $\max_{i=2,\dots,p} |\gamma_i|$ is relatively small with respect to this rate. To get a convergence rate under the variational approximation, we require some assumptions on the components of the model selection prior and compatibility conditions on the design matrix $\check{W}$ and on the truth β_{-1}^0 in the same spirit as in [4].

4 Empirical Results

In this section we assess the performance and behaviour of our proposed variational method with the improper choice of $g \equiv 1$. We refer to this method as I-SVB throughout, and compare its performance to that of two commonly used frequentist methods from [7] and [3]

Scenario (n, p, s_0, M)	Method	Cov.	MAE	Length	Time
(i)	I-SVB	**0.946**	**0.077**	0.405 ± 0.032	0.344 ± 0.032
$(100, 1000, 3, \log n)$	ZZ	0.878	0.113	0.514 ± 0.722	**0.132 ± 0.022**
	JM	0.796	0.121	0.389 ± 0.032	1.242 ± 0.022
(ii)	I-SVB	**0.948**	0.061	0.281 ± 0.022	0.304 ± 0.032
$(200, 800, 3, \log n)$	ZZ	0.926	0.064	0.293 ± 0.052	0.224 ± 0.022
	JM	0.904	0.065	0.277 ± 0.012	0.741 ± 0.022
(iii)	I-SVB	0.956	**0.057**	0.282 ± 0.022	0.305 ± 0.032
$(200, 800, 10, \log n)$	ZZ	0.918	0.064	0.292 ± 0.052	**0.223 ± 0.022**
	JM	0.908	0.064	0.277 ± 0.012	0.739 ± 0.012

Table 1: Assessing the performance of the uncertainty quantification provided by each method in 4 different scenarios. The best row of each column is in bold within each scenario.

(which we will refer to as ZZ and JM respectively). We take the improper prior because we found this to dominate the other choices in practice in terms of both the quality of uncertainty quantification and computation time.

Here we present three scenarios. We parameterise each scenario by the tuple (n, p, s_0, M), representing that the inference is being carried out from n observations of the response, with $\beta^0 \in \mathbb{R}^p$ with sparsity s_0, and where the non-zero entries of β^0 are given by M. Furthermore, the rows of X are assumed to be independently multivariate Gaussian with mean 0 and covariance Id. For each scenario, we simulate 500 sets of observations and for each set of observations compute a 95%−credible interval for each method: for the variational method, we make a large number of samples from the variational posterior and use the empirical quantiles; for the frequentist methods, we compute the confidence intervals directly. For each method we assess: (i) the coverage (the proportion of the confidence intervals which contain the true value); (ii) the mean absolute error of the centering of the confidence intervals as an estimator for the truth; (iii) the mean length of the confidence intervals; and (iv) the mean time for computation of the confidence intervals. The three scenarios we consider are given by the following choices: (i) $(100, 1000, 3, \log n)$; (ii) $(200, 800, 3, \log n)$; (iii) $(200, 800, 10, \log n)$. The results are given in Table 1.

We remark that the I-SVB method delivers coverage which is approximately 95% as intended, while its MAE and length are generally better than those of the ZZ method and the JM method. This demonstrates that our method is performing uncertainty quantification better than their frequentist counterparts in these scenarios. We also investigated scenarios where the columns of X present non trivial correlations and we found that our method performs well with respect to the two other methods in these settings. We remark finally here that the credible sets from the variational Bayesian methods can be computed in comparable time to their frequentist counterparts.

References

[1] Ismaël Castillo, Johannes Schmidt-Hieber, and Aad van der Vaart. Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5):1986 – 2018, 2015.

[2] Chao Gao, Aad W. van der Vaart, and Harrison H. Zhou. A general framework for Bayes structured linear models. *The Annals of Statistics*, 48(5):2848 – 2878, 2020.

[3] Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15(82):2869–2909, 2014.

[4] Kolyan Ray and Botond Szabo. Variational bayes for high-dimensional linear regression with sparse priors. *Journal of the American Statistical Association*, 117:1–31, 11 2020.

[5] Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166 – 1202, 2014.

[6] Dana Yang. Posterior asymptotic normality for an individual coordinate in high-dimensional linear regression. *Electronic Journal of Statistics*, 13(2):3082 – 3094, 2019.

[7] Cun-Hui Zhang and Stephanie S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 76(1):217–242, 2014.