Electronic Journal of Statistics
Vol. 17 (2023) 769–797
ISSN: 1935-7524
https://doi.org/10.1214/23-EJS2113

Statistical inference via conditional Bayesian posteriors in high-dimensional linear regression

Teng Wu

Department of Statistics, University of Illinois, Urbana Champaign
e-mail: tengwu2@illinois.edu

Naveen N. Narisetty

Department of Statistics, University of Illinois, Urbana Champaign
e-mail: naveen@illinois.edu

Yun Yang

Department of Statistics, University of Illinois, Urbana Champaign
e-mail: yy84@illinois.edu

Abstract: We propose a new method under the Bayesian framework to perform valid inference for low dimensional parameters in high dimensional linear models under sparsity constraints. Our approach is to use surrogate Bayesian posteriors based on partial regression models to remove the effect of high dimensional nuisance variables. We name the final distribution we used to conduct inference "conditional Bayesian posterior" as it is a sur-rogate posterior constructed conditional on quasi posterior distributions of other parameters and does not admit a fully Bayesian interpretation. Un-like existing Bayesian regularization methods, our method can be used to quantify the estimation uncertainty for arbitrarily small signals and therefore does not require variable selection consistency to guarantee its validity. Theoretically, we show that the resulting Bayesian credible intervals achieve desired coverage probabilities in the frequentist sense. Methodologically, our proposed Bayesian framework can easily incorporate popular Bayesian regularization procedures such as those based on spike and slab priors and horseshoe priors to facilitate high accuracy estimation and inference. Numerically, our proposed method rectifies the uncertainty underestimation of Bayesian shrinkage approaches and has a comparable empirical performance with state-of-the-art frequentist methods based on extensive simulation studies and a real data analysis.

MSC2020 subject classifications: Primary 62J05.
Keywords and phrases: Bayesian inference, Bayesian regularization, high dimensional linear model, sparsity, uncertainty quantification.

Received June 2021.

Naveen N. Narisetty gratefully acknowledges partial funding support from NSF grants DMS-1811768 and CAREER-1943500. Y. Yang's research was supported in part by NSF DMS-2210717.

Contents

1. Introduction

1.1. Problem setting

High dimensional covariates are prevalent in many modern scientific research areas where the number of covariates p may exceed the number of observations n. For performing statistical inference in high dimensional models, it is common to make sparsity assumptions that the response variable depends only on a small number of the covariates to make inference feasible. In the last few decades, many statistical methods, under both frequentist and Bayesian paradigms, have been proposed to perform parameter estimation and variable selection under sparsity assumptions. These methods make use of convex and non-convex penalization or shrinkage and sparsity inducing priors to perform sparse estimation.

In many practical problems, beyond point estimation of parameters, it is quite important to conduct statistical inference based on this estimator via constructing confidence intervals or performing hypothesis testings. For instance, to estimate the effect of a treatment in medical studies, researchers need to incorporate a large number of control variables such as patients' demographic and clinical features, which makes it a high dimensional problem. In this paper, we consider the scenario where the researchers are interested in performing inference on a specific variable or a small set of variables of interest, and want to

conduct inference by controlling for other high dimensional covariates as nuisance variables. Consider the following high dimensional linear model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\vartheta} + \boldsymbol{Z}\boldsymbol{\eta} + \boldsymbol{\varepsilon}, \tag{1.1}$$

where $\boldsymbol{Y}$ is an n-dimensional response variable, $(\boldsymbol{X}, \boldsymbol{Z})$ is the $n \times (q + d)$ design matrix collecting $(q + d)$ covariate vectors, $\boldsymbol{\vartheta}$ is the q-dimensional parameter of interest associated with the q covariates in $\boldsymbol{X}$, $\boldsymbol{\eta}$ is the d-dimensional nui-sance parameter associated with $\boldsymbol{Z}$, and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I_n)$ is the random noise vector with variance σ^2. Our goal is to conduct valid statistical inference un-der the Bayesian framework for the low dimensional parameter vector $\boldsymbol{\vartheta}$ in the high dimensional linear model, where the dimension q is small but allowed to grow slowly with n and d such that $q = o(\sqrt{n/\log d})$ while d is allowed to be substantially larger than n.

1.2. Existing approaches on high dimensional linear regression

Classical high dimensional estimators such as Lasso (Tibshirani, 1996), MCP (Zhang et al., 2010), SCAD (Fan and Li, 2001) focus on estimation and do not provide natural ways to obtain confidence intervals for $\boldsymbol{\vartheta}$. Indeed, Knight and Fu (2000) showed that Lasso has intractable asymptotic distributions and cannot be directly used to perform inference. Vanilla bootstrap and subsampling techniques also fail to work due to the non-continuity of the limiting distribution (Knight and Fu, 2000). To obtain valid estimation intervals for $\boldsymbol{\vartheta}$ in model (1.1), a class of debiased estimators have been proposed in the frequentist framework. These methods typically assume an extra layer of low-dimensional structure between random designs $\boldsymbol{X}$ and $\boldsymbol{Z}$. Some well-known methods utilizing this approach include: Debiased Lasso estimators (Zhang and Zhang, 2014, Javanmard and Montanari, 2014, Van de Geer et al., 2014), post-double-selection method (Belloni, Chernozhukov and Hansen, 2014), double machine learning (Chernozhukov et al., 2016). These methods lead to root-n consistent estimators that are asymptotically normally distributed and admit valid confidence intervals.

Bayesian methods have the advantage of providing a natural way of uncertainty quantification through posterior distributions. It is often desired that Bayesian credible sets have the same nominal coverage probability in the frequentist sense. For Gaussian sequence models, Castillo and Nickl (2013) used wavelet based priors and constructed credible sets with frequentist coverage through Bernstein-von Mises theorems. van der Pas, Szabó and van der Vaart (2017) studied the uncertainty quantification using Horseshoe prior for Gaussian sequence model. However, they require a "self-similarity" assumption, which is similar to the beta-min condition and excludes the bad regime where over-shrinkage occurs. In the context of high-dimensional linear regression, Bayesian regularization techniques of utilizing sparsity inducing prior distributions are often used. Commonly used priors include Laplace prior (Gelman et al., 2013), horseshoe prior (Carvalho, Polson and Scott, 2009, van der Pas, Szabó and

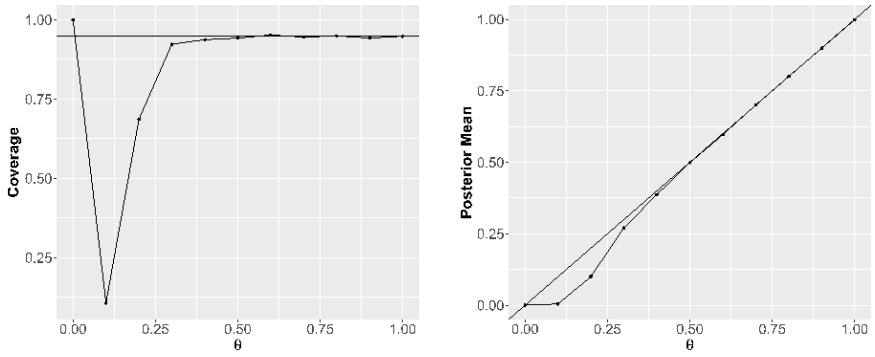
van der Vaart, 2017) and spike and slab priors (George and McCulloch, 1993, Ishwaran et al., 2005, Castillo and Szabó, 2020). Belitser and Ghosal (2020), Belitser and Nurushev (2020) and Castillo and Szabó (2020) consider empirical Bayesian approaches for Bayesian high dimensional estimation and uncertainty quantification. However, they do not address our goal of obtaining precise and valid confidence intervals for a selected parameter of interest—their Bayesian credible sets for the entire parameter vector have asymptotic coverage probability tending to one as the sample size grows. For spike and slab regression, Castillo et al. (2015) show that the joint posterior distribution of regression coefficients associated with important variables can be well approximated by a normal distribution centered at the least squares estimator of the reduced model, which makes it possible to conduct valid inference based on the posterior distribution. Song and Liang (2017) provide a similar Bernstein von-Mises type result for Bayesian high dimensional linear regression using a general class of shrinkage priors. However, such oracle properties (Castillo et al., 2015, Song and Liang, 2017) only hold under a very strong beta-min condition requiring the coefficients for all active covariates to be significantly large.

1.3. Invalid uncertainty quantification of Bayesian shrinkage approaches

In the common and realistic situations where some of the true regression coefficients are non-zero but small, shrinkage priors tend to shrink those small coefficients to zero, and the resulting posterior distribution tends to underestimate the uncertainty and the frequentist coverage probabilities of the induced credible intervals are substantially lower than their nominal levels.

We use a toy simulation study to demonstrate the impact of such over-shrinkage issue. Data are generated from model (1.1). We fix $\boldsymbol{\eta}^{\square} = \{0, 0, 2, 0, \ldots, 0\}$ and let $\vartheta^{\square}$ take values from $\{0, 0.1, \ldots, 1\}$. Other details of the data generating processes can be found in Section 4.1. We directly perform spike and slab regression to construct the 95% credible intervals for different values of ϑ based on the posterior distribution. Figure 1 shows the average posterior mean and empirical coverage of the credible intervals based on 1000 replications. Notice that when the signal is zero, spike and slab regression shrink the posterior to zero provides super-efficient credible intervals. However, for small but non-zero signals, the substantial biases in the posterior mean estimators may cause the credible intervals to have coverage probabilities lower than their nominal levels.

This over-shrinkage issue can be explained by the following heuristic argument. To analyze the coverage probability, we adopt a frequentist setting, where we use $\boldsymbol{\vartheta}^{\square}$ to denote the true value of the parameter $\boldsymbol{\vartheta}$ in the data generating model. For other parameters, we also add a $\square$ in the superscript to indicate their corresponding true values. $\lfloor \bullet \rfloor_q$ is used to denote the vector $\ddot{u}_q$ norm for $q \geq 1$. For a usual Bayesian shrinkage approach, the posterior distribution based on



(a) Empirical coverage vs signal strength (b) Posterior mean vs signal strength

Fig 1. *Toy simulation example using spike and slab regression. Left panel: Empirical coverage probabilities for Bayesian Posterior credible intervals shows undercoverage for weak signals. Right panel: Posterior mean estimator shows substantial bias for weak signals*

model (1.1) takes the following form:

$$\pi_B(\boldsymbol{\vartheta}, \boldsymbol{\eta} | \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) \propto \exp\left(-\frac{\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\vartheta} - \boldsymbol{Z}\boldsymbol{\eta}\|_2^2}{2\sigma^2}\right) \pi_{\boldsymbol{\vartheta}}(\boldsymbol{\vartheta}) \pi_{\boldsymbol{\eta}}(\boldsymbol{\eta}),$$

where $\pi_{\boldsymbol{\vartheta}}(\boldsymbol{\vartheta})$, $\pi_{\boldsymbol{\eta}}(\boldsymbol{\eta})$ are some shrinkage priors. To analyze the marginal posterior of $\boldsymbol{\vartheta}$, we start with the conditional distribution of $\boldsymbol{\vartheta}$ given all other parameters,

$$\pi(\boldsymbol{\vartheta} \mid \boldsymbol{\eta}, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) \propto N\left(\boldsymbol{\mu}_s, \sigma^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1}\right) \pi_{\boldsymbol{\vartheta}}(\boldsymbol{\vartheta}), \quad \text{with}$$

$$\boldsymbol{\mu}_s(\boldsymbol{\eta}) = \boldsymbol{\vartheta}^* + (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{\varepsilon} + (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Z}(\boldsymbol{\eta}^* - \boldsymbol{\eta}), \tag{1.2}$$

where $N(\mu, \Sigma)$ represents the normal distribution with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$. Since commonly used shrinkage priors will make $\boldsymbol{\eta}$ concentrate around $\boldsymbol{\eta}^*$ (i.e. Bayesian estimation consistency), the center of the marginal distribution of $\boldsymbol{\vartheta}$ will be roughly $\boldsymbol{\mu}_s(\boldsymbol{\eta})$ with $\boldsymbol{\eta}$ replaced by its posterior mean $\hat{\boldsymbol{\eta}}$. Notice that the leading term $\boldsymbol{\vartheta}^* + (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{\varepsilon}$ in $\boldsymbol{\mu}_s(\boldsymbol{\eta})$ is the maximum likelihood estimator of $\boldsymbol{\vartheta}^*$ with $\boldsymbol{\eta} = \boldsymbol{\eta}^*$[1]. The additional bias term $(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Z}(\boldsymbol{\eta}^* - \boldsymbol{\eta})$ has a typical order of $O_p(\sqrt{s \log d/n})$ in the high dimensional setting[2], which dominates $(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{\varepsilon} = O_p(\sqrt{q/n})$. This large bias cannot be captured by the $O(1/\sqrt{n})$ spread of the marginal posterior, which explains the low coverage probabilities under weak signals in Figure 1a. To explain why the credible intervals from Bayesian shrinkage posteriors are valid when signals are strong[3], notice that, the prediction error $\|\boldsymbol{Z}(\boldsymbol{\eta}^* - \hat{\boldsymbol{\eta}})\|_2$ of nuisance parameter $\boldsymbol{\eta}$ is reduced to $O_p(1/\sqrt{n})$ (refer to Fig 1b), so that the extra

[1] Not knowing $\boldsymbol{\eta}$ will inflate the variance

[2] The typical fitting error is $\|\boldsymbol{Z}(\boldsymbol{\eta}^* - \hat{\boldsymbol{\eta}})\|_2 = O_p(\sqrt{s \log d})$ for Bayesian regularized regressions, where s is the sparsity for $\boldsymbol{\eta}^*$. This makes the bias term $(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Z}(\boldsymbol{\eta}^* - \hat{\boldsymbol{\eta}}) = O_p(\sqrt{s \log d/n})$.

[3] All non-zero elements in $\boldsymbol{\vartheta}^*$ and $\boldsymbol{\eta}^*$ are larger than $O(\sqrt{\log d/n})$.

$\log d$ term in the additional bias term can be avoided. This heuristic analysis shows that Bayesian shrinkage posteriors capture the estimation uncertainty for strong signals but not weak signals.

1.4. Our contributions

In this paper, we propose a novel conditional Bayesian posterior framework that rectifies the uncertainty underestimation of Bayesian shrinkage approaches with theoretical justifications from a frequentist perspective. We incorporate a similar low dimensional structure between $\boldsymbol{X}$ and $\boldsymbol{Z}$ used in Belloni, Chernozhukov and Hansen (2014) and Chernozhukov et al. (2016). By utilizing the Bayesian framework, our method can conveniently incorporate prior information on the parameter of interest. This is particularly useful in practice, since it is common to have prior information on the low dimensional parameters of interest and not on the high dimensional nuisance parameters. Compared with frequentist methods such as double machine learning (Chernozhukov et al., 2016) and double selection (Belloni, Chernozhukov and Hansen, 2014) that generally require estimation in multiple stages, our conditional posterior procedure automatically propagates estimation uncertainty of the high dimensional nuisance parameters in multiple layers. To overcome the over-shrinkage issue, Hahn et al. (2018a) proposed a fully Bayesian approach that can perform inference on the treatment effect when there is high dimensional confounding. Their method places independent priors in transformed parameter space and admits an efficient sampling algorithm. However, the validity of the resulting posterior credible intervals was not theoretically validated. We show that our proposed procedure provides credible intervals that achieve desired coverage probabilities in the frequentist sense. The length of resulting Bayesian credible interval is optimal in the minimax sense as shown by Cai et al. (2017) in the sparse regime. The conditional Bayesian posterior we proposed can be viewed similarly as a quasi-Bayesian approach since we do not assume a particular data generating process for obtaining the posterior. Compared with the usual quasi-Bayesian methods (Syring and Martin, 2019), the proposed method can provide valid inference without the need for any calibration. Our proposed approach has the additional benefit of being naturally filled into semi-supervised learning setting, where a large number of unlabeled training data are available that can be incorporated to improve the estimation accuracy and boost the hypothesis testing power. Simulation studies show that our method has better performance compared with Bayesian regularization methods, and frequentist inference approaches such as debiased Lasso and double selection.

The rest of the paper is organized as follows. In Section 2, we describe our main methodology to perform inference of low dimensional parameters in high dimensional linear model and show in theory that the resulting interval estimation has valid frequentist coverage probability. Section 3 provides simulation under various settings and a real data application. Finally, Section 4 concludes the paper and discusses possible extensions.

2. Methodology and results

In this section, we propose a conditional Bayesian posterior approach to overcome the over-shrinkage issue described in Section 1.3, followed by the theoretical investigation into its validity.

2.1. Methodology

We now present our method to conduct inference for low dimensional parameters of interest in the high dimensional linear model. To motivate our method, we first consider the regression model (1.1) for a single covariate of interest with $q = 1$ and d is much smaller than n. From the standard theory of linear models, the least squares estimator $\hat{\vartheta}_{LS}$ of ϑ in jointly fitting $(\vartheta, \boldsymbol{\eta})$ can be equivalently obtained via the residual on residual regression (Velleman and Welsch, 1981). More precisely, perform a linear regression of $\boldsymbol{X}$ on $\boldsymbol{Z}$ and obtain the residuals $\tilde{\boldsymbol{X}}$. Similarly, let $\tilde{\boldsymbol{Y}}$ be the residuals when regressing $\boldsymbol{Y}$ on $\boldsymbol{Z}$. When regressing $\tilde{\boldsymbol{Y}}$ on $\tilde{\boldsymbol{X}}$, the resulting slope estimator is the same with $\hat{\vartheta}_{LS}$. The first two linear regression models can be interpreted as removing the effect of $\boldsymbol{Z}$ from $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. Let $P_z = \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T$ be the projection matrix induced by $\boldsymbol{Z}$. This equivalent procedure of reaching $\hat{\vartheta}_{LS}$ can be justified by applying the projection operator $(I - P_z)$ to both sides of model (1.1) as in the following,

$$(I - P_z)\boldsymbol{Y} = (I - P_z)\boldsymbol{X}\vartheta + (I - P_z)\boldsymbol{\varepsilon}.$$

This classical idea can be generalized to high dimensional settings with some additional assumptions. Under high dimensional settings, we can not obtain the residuals via projection. However, an approximated version of residuals can still be obtained by performing regularized regression. We consider performing Bayesian regularized regression under the following working models:

$$\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\omega}, \quad \boldsymbol{\omega} \sim N(\mathbf{0}, \sigma_1^2 I_n) \tag{2.1}$$

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\varphi} + \boldsymbol{\nu}, \quad \boldsymbol{\nu} \sim N(\mathbf{0}, \sigma_2^2 I_n), \tag{2.2}$$

where $\boldsymbol{\gamma}$, $\boldsymbol{\varphi}$ are sparse d-dimensional parameters, $\boldsymbol{\omega}$, $\boldsymbol{\nu}$ are independent random noise vectors, and $\sigma_1^2, \sigma_2^2 > 0$ are corresponding residual variances. To assure $E[\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{Z}]$ stays the same with model (1.1), we require the newly introduced parameters to satisfy $\boldsymbol{\varphi} = \boldsymbol{\gamma}\vartheta + \boldsymbol{\eta}$. Models (2.1) and (2.2) serve the purpose of removing the effect of $\boldsymbol{Z}$ from $\boldsymbol{X}$ and $\boldsymbol{Y}$.

A similar sparsity structure between $\boldsymbol{X}$ and $\boldsymbol{Z}$ as our model (2.1) is a common assumption made in the literature. In fact, the sparsity assumption on the regression coefficient γ when we perform $X \sim Z$ regression is equivalent to the sparsity of the X column in the joint precision matrix of (X, Z) (Peng et al., 2009), a fact that is heavily used for devising computationally efficient regression based methods for estimating sparsity of high dimensional precision matrices (Peng et al., 2009, Khare, Oh and Rajaratnam, 2015). The working

models we assume in Equations (2.1)-(2.2) are based on additional modeling assumptions which are not necessarily assumed for a high dimensional model targeting estimation alone. However, this is a common phenomenon for high dimensional inference with similar assumptions being made for existing methods including the Double Selection method (Belloni, Chernozhukov and Hansen, 2014) which explicitly considers two stage models similar to (2.1) and (2.2) along with sparsity assumptions on the coefficients γ and φ (see their Section 2.1 for more detailed discussion on these assumptions). The debiased Lasso estimator proposed by Zhang and Zhang (2014) applied Lasso regression separately for each column $\boldsymbol{X}_j$ on the rest of the predictors $\boldsymbol{X}_{-j}$ to estimate "an inverse" to the sample covariance matrix of the design. This procedure assumed model (2.1) and sparsity of the regression parameters γ (see their Equation (9)). Van de Geer et al. (2014) and Chernozhukov et al. (2016) also made similar assumptions for obtaining valid inference methods.

Model (2.2) serves the purpose of extracting the residuals from regressing $\boldsymbol{Y}$ on $\boldsymbol{Z}$, so that the residual on residual regression idea can be generalized to the high dimensional context. More specifically, we use $\tilde{\boldsymbol{X}} = \boldsymbol{X} - \boldsymbol{Z}\gamma$ to denote the residual vector of regressing $\boldsymbol{X}$ on $\boldsymbol{Z}$ and $\tilde{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{Z}\varphi$ to denote the residual vector of regressing $\boldsymbol{Y}$ on $\boldsymbol{Z}$. Conditional on γ and φ, the residual on residual regression motivates the following linear model as our third working model

$$\boldsymbol{Y} - \boldsymbol{Z}\varphi = (\boldsymbol{X} - \boldsymbol{Z}\gamma)\vartheta + \boldsymbol{\varepsilon}, \qquad \text{where } \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 I_n). \tag{2.3}$$

Note that the models given by Equation (1.1) and (2.3) are equivalent with each of them corresponding to a different reparametrization of the parameters due to $\eta = \varphi - \gamma\vartheta$. Therefore, the interpretation of the linear model and the parameter ϑ remains unchanged.

With the three working models (2.1)–(2.3), we propose the following surrogate posterior, which will be referred to as the conditional Bayesian posterior distribution,

$$\begin{aligned}\pi(\vartheta, \gamma, \varphi, \sigma_1^2, \sigma_2^2 \mid \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}) \propto \\ \pi_{\tilde{\boldsymbol{Y}}-\tilde{\boldsymbol{X}}}(\vartheta \mid \gamma, \varphi, \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z})\pi_{\boldsymbol{X}-\boldsymbol{Z}}(\gamma, \sigma_1^2 \mid \boldsymbol{X}, \boldsymbol{Z})\pi_{\boldsymbol{Y}-\boldsymbol{Z}}(\varphi, \sigma_2^2 \mid \boldsymbol{Y}, \boldsymbol{Z}),\end{aligned} \tag{2.4}$$

where the three components correspond to the working models (2.1)–(2.3), with forms:

$$\pi_{\boldsymbol{X}-\boldsymbol{Z}}(\gamma, \sigma_1^2 \mid \boldsymbol{X}, \boldsymbol{Z}) \propto \frac{1}{\sigma_1^n}\exp\left\{-\frac{\|\boldsymbol{X} - \boldsymbol{Z}\gamma\|_2^2}{2\sigma_1^2}\right\}\pi_\gamma(\gamma)\pi_{\sigma^2}(\sigma_1^2), \tag{2.5}$$

$$\pi_{\boldsymbol{Y}-\boldsymbol{Z}}(\varphi, \sigma_2^2 \mid \boldsymbol{Y}, \boldsymbol{Z}) \propto \frac{1}{\sigma_2^n}\exp\left\{-\frac{\|\boldsymbol{Y} - \boldsymbol{Z}\varphi\|_2^2}{2\sigma_2^2}\right\}\pi_\varphi(\varphi)\pi_{\sigma^2}(\sigma_2^2), \tag{2.6}$$

$$\pi_{\tilde{\boldsymbol{Y}}-\tilde{\boldsymbol{X}}}(\vartheta \mid \gamma, \varphi, \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}) \propto \exp\left\{-\frac{\|\boldsymbol{Y} - \boldsymbol{Z}\varphi - (\boldsymbol{X} - \boldsymbol{Z}\gamma)\vartheta\|_2^2}{2\sigma^2}\right\}\pi_\vartheta(\vartheta), \tag{2.7}$$

where $\pi_{\boldsymbol{\gamma}}(\boldsymbol{\gamma})$, $\pi_{\boldsymbol{\varphi}}(\boldsymbol{\varphi})$ are some sparsity inducing priors, $\pi_{\sigma_1^2}(\sigma_1^2)$, $\pi_{\sigma_2^2}(\sigma_2^2)$ are inverse Gamma priors, and $\pi_{\boldsymbol{\vartheta}}(\boldsymbol{\vartheta})$ is some generic prior for $\boldsymbol{\vartheta}$. To obtain samples of $\boldsymbol{\vartheta}$ from (2.4), we simply need to sample $\boldsymbol{\gamma}$ from quasi posterior (2.5), and sample $\boldsymbol{\varphi}$ from quasi posterior (2.6). Both samples can be easily obtained by implementing Bayesian regularized regression. Finally, $\boldsymbol{\vartheta}$ can be generated conditional on $\boldsymbol{\gamma}$ and $\boldsymbol{\varphi}$ based on (2.7).

Notice that Equations (2.1), (2.2) and (2.3) are only working models and do not constitute a valid data-generating model. They are always misspecified: the residuals $\boldsymbol{\omega}$, $\boldsymbol{\nu}$ and $\boldsymbol{\varepsilon}$ do not have i.i.d. component since the three models are related by relationship $\boldsymbol{\nu} = \boldsymbol{\omega}\boldsymbol{\vartheta}^{\square} + \boldsymbol{\varepsilon}$. Although our model is a over-determined system and has the potential to be overparameterized, our theoretical results in Section 2.3 show that the conditional posterior in (2.4) can be used to construct pseudo credible sets for $\boldsymbol{\vartheta}$ that have coverage probabilities achieving their nominal level in the frequentist sense.

When we are interested in performing inference for a multivariate covariate vector so that the dimension of $\boldsymbol{X}$ is $q > 1$, then the model given by (2.1) would be

$$\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\omega}, \quad \boldsymbol{\omega} \square N(\mathbf{0}, I_n \otimes \Sigma), \tag{2.8}$$

where $\otimes$ denotes Kronecker product. The residual term $\boldsymbol{\omega}$ is now an $n \times q$ matrix. This working model treats the different columns of $\boldsymbol{\omega}$ to be independent as we use a diagonal Σ for estimation of $\boldsymbol{\gamma}$. However, for the theoretical analysis, we will allow dependency between the columns of $\boldsymbol{\omega}$.

2.2. Comparison with Bayesian shrinkage approaches

To illustrate how our method rectifies the undercoverage of Bayesian shrinkage approaches in interval estimation, we consider the following similar heuristic analysis to show the asymptotic normality of the conditional posterior of $\boldsymbol{\vartheta}$. Our conditional Bayesian posterior in (2.4) can be simplified to

$$\pi(\boldsymbol{\vartheta}, \boldsymbol{\varphi}, \boldsymbol{\gamma}, \sigma_1^2 \mid \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) \propto \exp\left\{-\frac{\|\boldsymbol{Y} - (\boldsymbol{X} - \boldsymbol{Z}\boldsymbol{\gamma})\boldsymbol{\vartheta} - \boldsymbol{Z}\boldsymbol{\varphi}\|_2^2}{2\sigma^2}\right\} \frac{1}{\sigma_1^n} \exp\left\{-\frac{\|\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\varphi}\|_2^2}{2\sigma_1^2}\right\} \pi_{\boldsymbol{\vartheta}}(\boldsymbol{\vartheta})\pi_{\boldsymbol{\varphi}}(\boldsymbol{\varphi})\pi_{\sigma_1}(\sigma^2).$$

We can also derive the conditional distribution of $\boldsymbol{\vartheta}$ given all other parameters,

$$g(\boldsymbol{\vartheta} \mid \boldsymbol{\varphi}, \boldsymbol{\gamma}, \sigma^2, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) \propto N\left(\boldsymbol{\mu}_q, \sigma^2 \right)(\boldsymbol{X} - \boldsymbol{Z}\boldsymbol{\gamma})^T (\boldsymbol{X} - \boldsymbol{Z}\boldsymbol{\gamma})^{*-1} \pi_{\boldsymbol{\vartheta}}(\boldsymbol{\vartheta}),$$

with

$$\mu_q(\boldsymbol{\varphi}, \boldsymbol{\gamma}) = \boldsymbol{\vartheta}^{\square} + (\boldsymbol{X} - \boldsymbol{Z}\boldsymbol{\gamma})^T (\boldsymbol{X} - \boldsymbol{Z}\boldsymbol{\gamma})^{*-1} (\boldsymbol{X} - \boldsymbol{Z}\boldsymbol{\gamma})^T \boldsymbol{\varepsilon} + (\boldsymbol{X} - \boldsymbol{Z}\boldsymbol{\gamma})^T (\boldsymbol{X} - \boldsymbol{Z}\boldsymbol{\gamma})^{*-1} (\boldsymbol{X} - \boldsymbol{Z}\boldsymbol{\gamma})^T \boldsymbol{Z}(\boldsymbol{\varphi}^{\square} - \boldsymbol{\varphi}).$$

With appropriate shrinkage priors, $\boldsymbol{\varphi}$ and $\boldsymbol{\gamma}$ will concentrate on their corresponding true parameter values $\boldsymbol{\varphi}^{\square}$ and $\boldsymbol{\gamma}^{\square}$. Consequently, the marginal distribution

of ϑ will center on $\mu_q(\tilde{\varphi}, \tilde{\gamma})$ where $\tilde{\varphi}$ and $\tilde{\gamma}$ are parameter values close to φ^* and γ^*. Notice that, the leading term $\vartheta^* + (X^* Z^{\gamma})^T (X^* Z^{\gamma})^{-1} (X^* Z^{\gamma})^T \varepsilon$ is the maximum likelihood estimator of ϑ from model (2.3) for $\gamma = \gamma^*$. To ana-lyze the additional bias term $(X^* Z^{\tilde{\gamma}})^T (X^* Z^{\gamma})^{-1} (X^* Z^{\tilde{\gamma}})^T Z(\varphi^* - \tilde{\varphi})$, notice that

$$(X^* Z^{\tilde{\gamma}})^T Z(\varphi^* - \tilde{\varphi}) \le |\varphi^* - \tilde{\varphi}|_1 |Z^T (X^* Z^{\gamma})|_\infty. \tag{2.9}$$

For the first term in inequality (2.9), a usual posterior concentration implies $|\varphi^* - \tilde{\varphi}|_1 = O_p(s_2 \sqrt{\log d/n})$ where s_2 is the sparsity for φ^*. The second term $Z^T (X^* Z^{\gamma})$ is close to $Z^T (X^* Z^{\gamma^*}) = Z^T \omega^*$, which is of order $O_p(\sqrt{n \log d})$. To summarize, the third term in $\mu_q(\tilde{\varphi}, \tilde{\gamma})$ is of order $O_p(s_2 \log d/n)$, which is negligible to the order $O_p(1/\sqrt{n})$ of the second term, when $s_2 \log d \ne \sqrt{n}$. This is different from the analysis in the introduction of $\mu_s(\eta)$ for the usual Bayesian regularized regression, since our conditional Bayesian posterior can utilize the orthogonal structure between $X^* Z^{\gamma}$ and Z. Consequently, the marginal distribution of ϑ still approaches to a normal distribution regardless of the signal strength of ϑ^*. This gives some heuristic understanding on why our model is capable of providing valid posterior credible intervals without requiring variable selection consistency.

2.3. Theoretical results

In this section, we show the theoretical results of our proposed method by presenting a Bernstein von-Mises type theorem. Bernstein von-Mises type results state that the posterior distribution of a parameter in a smooth finite-dimensional model can be approximated by a normal distribution if the number of observations tends to infinity. This kind of results has been widely stud-ied in Bayesian literature to justify the use of Bayesian credible intervals as valid confidence intervals in the frequentist space (Bontemps et al., 2011, Kleijn et al., 2012). In particular, the theorem we present asserts that the conditional Bayesian posterior of ϑ converges in total variation to a normal distribution which is centered at the maximum likelihood estimator in the frequentist sense with a standard deviation of order $1/\sqrt{n}$. We first discuss some assumptions required for our theorem. We say a p-dimensional vector β to be s sparse if $\sum_{i=1}^{p} \mathbf{1}\{\beta_i = 0\} = s$.

Assumption 1. *Assume there is a vector γ^* such that the residual vector $\omega^* = X^* Z^{\gamma^*}$ satisfies that:*

1. *There exists a $q \times q$ positive definite matrix Σ such that $|\omega^{*T} \omega^*/n - \Sigma|_2 \le M_1 \sqrt{q/n}$ on set E_1 with $P(\mathrm{E}_1) > 1 - n^{-c_1}$,*
2. *and for each $j = 1, \ldots, q$, $|Z^T \omega_j^*|_\infty \le M_2 \sqrt{n \log d}$ on set E_2 with $P(\mathrm{E}_2) > 1 - d^{-c_2}$,*

where ω_j^* is the j-th column of ω^* and c_1, c_2, M_1, M_2 are some positive constant.

Assumption 2. $\|\vartheta^*\|_1 \le M_0 q$ *for some* $M_0 > 0$. *Each column of* γ^* *is at most* s_1 *sparse and* η^* *is at most* s_2 *sparse.* s_1 *and* s_2 *satisfy*

$$s_1 + s_2 = o\left(\frac{\sqrt{n}}{q \log d}\right).$$

Assumption 1 allows for both fixed and random design of (X, Z). In the case of random design, Assumption 1 is implied by sub-Gaussianity of ω^*, where Σ can be taken as the common covariance matrix of each row of ω^*. Sparsity assumptions are commonly used in high dimensional literature, similar assumptions are made by Belloni, Chernozhukov and Hansen (2014) etc. Sparsity assumptions are imposed to obtain consistent estimations of γ^* and φ^*. Since X is allowed to have more than one column, the sparsity assumption is made on each column of γ^*. Notice that we do not require different columns of γ^* to have the same support. In particular, if the supports are the same, Bayesian multivariate regression methods can be applied which encourage a common sparsity pattern by imposing a global shrinkage prior (Bai and Ghosh, 2018). This extra layer of sparsity structure can be easily incorporated with a suitable prior. An upper bound in the sparsity Assumption 2 is necessary in the sense that, in order to construct adaptive confidence intervals that have the optimal length of the parametric rate $n^{-1/2}$ without knowing the exact sparsity level, we need to restrict ourselves to the ultra sparse regime, where the sparsity level is $o(n^{1/2}/\log d)$ (Cai et al., 2017).

Our next assumption concerns the posterior concentration for γ and φ, as well as the contraction rate the fitting error, which are satisfied by many Bayesian regularized methods, see Theorem 2 in Castillo et al. (2015), Theorem 8 in Ročková and George (2018) and Theorem 2.1-2.2 in Song and Liang (2017).

Assumption 3. *The marginal posterior distribution of* γ *and* φ *based on models (2.1) and (2.2) satisfy the following concentration properties:*

1. *We assume that the posterior of* γ *concentrates as*

$$\Pi_{X,Z}\left(\max\left\{\frac{\|\gamma_j - \gamma_j^*\|_1}{s_1}, \frac{\|\gamma_j - \gamma_j^*\|_2}{s_1^{1/2}}, \frac{\|Z(\gamma_j - \gamma_j^*)\|_2}{n^{1/2} s_1^{1/2}}\right\} \ge M_3 \sqrt{\frac{\log d}{n}} \,\middle|\, X, Z\right) \le M_4 d^{-c_3},$$

   for $j = 1, \ldots, q$ *on set* E_3, *with* $P(\mathrm{E}_3) \ge 1 - qd^{-c_3}$,

2. *We assume that the posterior of* φ *concentrates as*

$$\Pi_{Y,Z}\left(\max\left\{\frac{\|\varphi - \varphi^*\|_1}{s_1 + s_2}, \frac{\|\varphi - \varphi^*\|_2}{(s_1 + s_2)^{1/2}}\right\} \ge M_3 \sqrt{\frac{\log d}{n}} \,\middle|\, Y, Z\right) \le M_4 d^{-c_4},$$

   on set E_4, *with* $P(\mathrm{E}_4) \ge 1 - d^{-c_4}$,

where M_3, M_4, c_3 and c_4 are some positive constants.

Theorem 1 (Normal Approximation to Posterior)**.** *If Assumption 1-3 hold, $\Pi(\boldsymbol{\vartheta} \mid \Delta)$ converges to U in total variation with probability at least $1 - n^{-c_1} - (q+2)d^{-c_3}$. More specifically, the following inequality holds with probability at least $1 - n^{-c_1} - (q+2)d^{-c_3}$:*

$$\|\Pi(\boldsymbol{\vartheta} \mid \Delta) - U\|_{TV} \leq C\,\frac{q(s_1+s_2)\log d}{n^{1/2}} + 2M_3 d^{-c_3},$$

where $\|\cdot\|_{TV}$ denote the total variation difference between two measures, $U \sim N\big(\hat{\boldsymbol{\vartheta}}^{\mathcal{D}} + (\boldsymbol{\omega}^{\mathcal{D}T}\boldsymbol{\omega}^{\mathcal{D}})^{-1}\boldsymbol{\omega}^{\mathcal{D}T}\boldsymbol{\varepsilon}),\ \sigma^2(\boldsymbol{\omega}^{\mathcal{D}T}\boldsymbol{\omega}^{\mathcal{D}})^{-1}\big)$ is the distribution of the maximum likelihood estimator of $\boldsymbol{\vartheta}$ when $\boldsymbol{\gamma}^{\mathcal{D}}$ is known, and C is some positive constant.

Remark 1. *For the case where we have an additional n_1 unlabeled pairs $(\boldsymbol{X}, \boldsymbol{Z})$, the convergence result in Assumption 3 can be modified to*

$$\Pi_{\boldsymbol{X},\boldsymbol{Z}}\left(\max\left\{\frac{\|\boldsymbol{\gamma}_j - \boldsymbol{\gamma}^{\mathcal{D}}\|_1}{s},\ \frac{\|\boldsymbol{\gamma}_j - \boldsymbol{\gamma}^{\mathcal{D}}\|_2}{s_1^{1/2}}\right\} \geq M_3\sqrt{\frac{\log d}{n+n_1}} \,\middle|\, \boldsymbol{X}, \boldsymbol{Z}\right) \leq M_4 d^{-c_3}, \quad (1)$$

which will improve the estimation for $\boldsymbol{\gamma}$. However, the convergence rate in Theorem 1 is not affected.

Remark 2. *Assumption 2 can be relaxed to ℓ_q sparse settings, which is a weaker version of sparsity that does not require exact zeros (see Ye and Zhang (2010) for definition of ℓ_q sparse). Many statistical methods for high dimensional linear model are shown to enjoy similar concentration properties under ℓ_q sparsity and usual sparsity assumption. For simplicity, we will stick with Assumption 2 and 3 for our main results and proofs, but all the steps would go through if we have consistent posterior samples that concentrate in a similar rate under ℓ_q sparse assumptions.*

Remark 3. *In this paper, we are considering a high dimensional setting where the dimension d is considered to be comparable or much larger than the sample size n. More specifically, we are considering the regime where $d \geq Cn$. For a general dimension d that might be fixed or slowing growing with n, the same assumptions and theoretical statements will remain valid when the dimension d is replaced with the maximum of d and n, denoted by $(d \vee n)$.*

Theorem 1 asserts that the posterior of $\boldsymbol{\vartheta}$ can be well approximated by a Gaussian distribution whose covariance matches the covariance of the maximum likelihood estimator of $\boldsymbol{\vartheta}$ when $\boldsymbol{\gamma}^{\mathcal{D}}$ is known, and the credible interval obtained from the Bayesian procedure provides valid frequentist coverage. Let $q = 1$, $\hat{q}^B_{\alpha/2}$ and $\hat{q}^B_{1-\alpha/2}$ be the $(\alpha/2)^{th}$ and $(1-\alpha/2)^{th}$ percentile of $\Pi(\vartheta \mid \Delta)$, the following corollary characterize how the convergence result in Theorem 1 affects the error rate in the coverage probability of the Bayesian credible interval.

Corollary 1. *The $(1-\alpha)$-credible interval has the correct frequentist coverage, and we have*

$$P^{\mathcal{D}}\left(\vartheta^{\mathcal{D}} \in (\hat{q}^B_{\alpha/2}, \hat{q}^B_{1-\alpha/2})\right) - (1-\alpha) = O(q(s_1+s_2)\log d/n^{1/2} + qd^{-c_3} + n^{-c_1}).$$

In the case where q is larger than 1, for any $\boldsymbol{a} \in \mathrm{R}^q$, the credible intervals for $\boldsymbol{a}^T\boldsymbol{\vartheta}$ can be constructed similarly.

2.4. Comparison with existing bayesian regularized regression

In this section, we compare our theoretical results with existing Bayesian methods based on sparsity inducing priors. Let $\boldsymbol{W} = (\boldsymbol{X}, \boldsymbol{Z})$, and $\boldsymbol{\beta} = (\boldsymbol{\vartheta}, \boldsymbol{\eta})$. Spike and slab regression results in a joint posterior that can be expressed as a mixture,

$$\pi_{SnS}(\boldsymbol{\beta} \mid \Delta) = \sum_{\xi \subset \{1,\ldots,q+d\}} \pi(\xi \mid \Delta)\pi(\boldsymbol{\beta}_\xi \mid \boldsymbol{W}_\xi, \boldsymbol{Y})\mathbf{1}\{\boldsymbol{\beta}_{\xi^c} = \mathbf{0}\},$$

where ξ is the subset model, $\boldsymbol{\beta}_\xi$ and $\boldsymbol{\beta}_{\xi^c}$ denote the coefficient vectors included in and excluded from subset model ξ, and $\boldsymbol{W}_\xi$ denote the sub-design matrix that corresponds to subset model ξ. If the true model ξ^* can be correctly selected with high probability, $\pi(\xi^* \mid \Delta) \to 1$, under additional regularity conditions (Song and Liang, 2017), the posterior can be approximated by the following normal distribution.

$$\Pi_{SnS}(\boldsymbol{\beta} \mid \Delta) \approx N(\boldsymbol{\beta}_{\xi^*}; \hat{\boldsymbol{\beta}}_{\xi^*}, (\boldsymbol{W}_{\xi^*}^T\boldsymbol{W}_{\xi^*})^{-1}) \otimes \delta_0(\boldsymbol{\beta}_{(\xi^*)^c}),$$

where δ_0 is the Dirac measure. Similar shape approximation results have been shown in Song and Liang (2017) for Bayesian regression based on global shrinkage priors.

The above results suggest that when $\boldsymbol{\vartheta}$ is indeed an active parameter with a sufficiently large magnitude, the posterior distribution for $\boldsymbol{\vartheta}$ can be well approximated by the corresponding normal distribution as if we knew the true model. However, when $\boldsymbol{\vartheta}^*$ is zero, the posterior distribution for $\boldsymbol{\vartheta}$ degenerates to zero and ends up with super-efficient estimation intervals. For a small signal that does not satisfy beta-min condition, the posterior underestimates the uncertainty and the frequentist coverage probabilities are substantially lower than the nominal levels. Intuitively, a minimum signal strength (beta-min) condition is required for variable selection consistency, and a valid credible interval can be constructed with a consistent variable selection approach for such strong signals under this condition. Without a beta-min condition, it is well known (Fu and Knight, 2000) that the asymptotic distribution of a typical sparsity inducing method, such as LASSO, will have point mass at zero; for Bayesian regularized regression, the posterior has the overly conservative tendency of shrinking all small but nonzero signals to zero. Song and Liang (2017) provided theoretical insights into this issue in their Section 2.3. Hahn et al. (2018b) also discussed this issue of over-shrinking the signal induced by high-dimensional regularization prior distributions. The over-shrinkage issue is also consistent with the numerical results in our toy example in Section 1.3. More theoretical discussion on the over-shrinkage issue can be found in Section 2.3 of Song and Liang (2017). In comparison, our method does not depend on whether the parameter of interest $\boldsymbol{\vartheta}$ is zero or not, and the theoretical result in Theorem 1 holds for all $\boldsymbol{\vartheta}^*$.

3. Numerical studies

In this section, we demonstrate the advantage of using the proposed method through various simulation studies followed by a real data application.

3.1. Simulation

We first investigate the performance of our proposed method under different simulation settings. The simulation results demonstrate that our proposed method can provide correct interval inference even in situations where commonly used Bayesian regularization methods do not work. Compared with commonly used frequentist methods, our method is also observed to have less bias, more accurate coverage probabilities and better robustness under various simulation settings. The data are generated from the following model:

$$\boldsymbol{Y} = \boldsymbol{X}\vartheta^{\square} + \boldsymbol{Z}\boldsymbol{\eta}^{\square} + \boldsymbol{\varepsilon},$$

$$\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\gamma}^{\square} + \boldsymbol{\omega},$$

where $\boldsymbol{X}$ is an n-dimensional vector and $\boldsymbol{Z} \square N(\mathbf{0}, \Sigma)$ is $n \times d$ dimensional matrix. We use Toeplitz covariance matrix $\Sigma_{ij} = 0.8^{|i-j|}$ as default unless otherwise stated. The error terms $\boldsymbol{\varepsilon} \square N(\mathbf{0}, I_n)$ and $\boldsymbol{\omega} \square N(\mathbf{0}, I_n)$ are independent of $\boldsymbol{Z}$. We consider $(n, 1 + d) = (100, 100), (200, 500)$ and let ϑ take values from $\{0, 0.1, \ldots, 1\}$. We calculate the empirical coverage probabilities of 95% confidence/credible intervals for different values $\vartheta^{\square}$ based on 1000 Monte Carlo simulations. We consider the following methods for performance comparison in our simulation studies:

1. **HS:** We perform Bayesian regression using horseshoe prior for shrinkage and construct credible interval of ϑ based on posterior distribution (Carvalho, Polson and Scott, 2009). We use the default implementation in R package bayeslm (Hahn, He and Lopes, 2018, 2019).
2. **Laplace:** We perform Bayesian regression using Laplace prior for shrinkage (also known as Bayesian Lasso (Gelman et al., 2013)) and construct credible interval of ϑ based on the posterior distribution. We use the default implementation in R package bayeslm (Hahn, He and Lopes, 2018, 2019).
3. **SnS:** We perform Bayesian regression using spike and slab prior (George and McCulloch, 1993) and construct credible interval of ϑ based on the posterior distribution. We use the default implementation in R package BoomSpikeSlab (Scott, 2021).
4. **DeLasso:** We obtain the debiased Lasso estimator and its Wald interval (Zhang and Zhang, 2014, Van de Geer et al., 2014). We use the default implementation in R package hdi (Dezeure et al., 2015).
5. **DS:** We obtain the double selection estimator and its Wald interval (Belloni, Chernozhukov and Hansen, 2014). We use the default implementation in R package hdm (Chernozhukov, Hansen and Spindler, 2016).

6. **DML:** We obtain the double machine learning estimator and its Wald interval without performing sample splitting (Chernozhukov et al., 2016). We use 10-fold cross-validation to fit Lasso estimator for two-stage regression model.
7. **CBP + HS:** We implement the proposed conditional Bayesian posterior method with horseshoe prior to generate samples for $\boldsymbol{\varphi}$ and $\boldsymbol{\gamma}$. For horseshoe regressions, we adopt the default implementation from R package bayeslm. We use Jeffery's prior $\pi(\vartheta) \propto 1$ in expression (2.7).
8. **CBP + SnS:** We implement the proposed conditional Bayesian posterior method with spike and slab prior to generate samples for $\boldsymbol{\varphi}$ and $\boldsymbol{\gamma}$. For spike and slab regressions, we adopt the default implementation from R package BoomSpikeSlab. We use Jeffery's prior $\pi(\vartheta) \propto 1$ in expression (2.7).

We first compare the performance of different methods under sparse $\boldsymbol{\gamma}$. We let $\boldsymbol{\gamma}^{\square} = (2, 1, 0, \ldots, 0)$ and set $\boldsymbol{\eta}^{\square} = (0, 0, 2, 0, \ldots, 0)$. Fig. 2 provides the comparisons of different methods in terms of coverage and interval length. The results demonstrate that our proposed methods achieve more precise coverage than other methods. In particular, Bayesian methods based on horseshoe prior and spike and slab priors show a similar performance. When $\vartheta = 0$ or relatively small, both methods tend to shrink ϑ to zero and provide super narrow intervals. This ends up with super-efficient intervals when $\vartheta = 0$. However, for relatively small signals, the coverage probabilities are very low. The resulting posterior from Bayesian Lasso fails to quantify the uncertainty in ϑ and the overall performance is the worst among all the methods. Due to this reason, we will exclude this method in all the following discussions. The resulting intervals from debiased Lasso undercover for all signal strengths due to a dominating bias term. Double selection method provides intervals that undercover larger signals, which is likely due to the strong correlation between the significant and insignificant predictors in $\boldsymbol{Z}$, which cast more difficulty during model selection for model (2.1) and (2.2). This shows the advantage of using the proposed method, since the validity of our resulting credible intervals does not rely on model selection consistency. The intervals based on double machine learning achieve similar coverage probabilities and average interval lengths compared to our method in this setting. Our proposed methods together with double machine learning tend to have wider intervals compared with intervals resulting from Bayesian methods based on horseshoe priors or spike and slab priors, despite that all estimation intervals achieve correct coverage probabilities for large signals. This is the price we pay for achieving uniformly correct coverage.

We now demonstrate the performance of our proposed method under semiparametric learning setting. The way we formulate our conditional Bayesian posterior (2.4) suggests we can sample $\boldsymbol{\gamma}$ and $\boldsymbol{\varphi}$ independently. We use the same setting with the above sparse $\boldsymbol{\gamma}$ case and fix the signal size at $\vartheta = 1$. We report the simulation results of our conditional Bayesian posterior with horseshoe regression when there are additional unlabeled pairs $(\boldsymbol{X}, \boldsymbol{Z})$. We consider the case $(n, 1+d) = (100, 100), (300, 100)$, and then there are 500 or 1000 more unlabeled

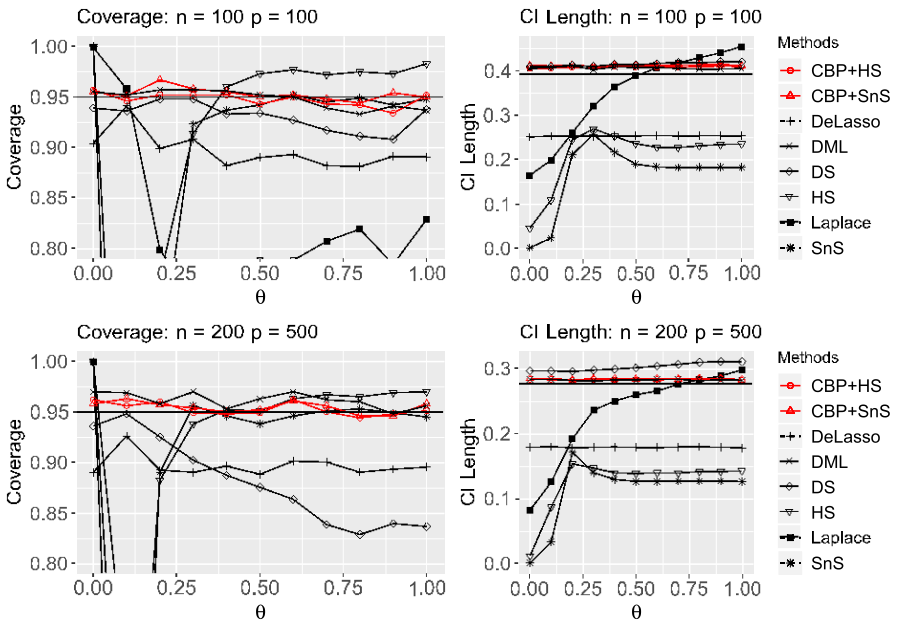


Fig 2. *The proposed methods (red solid) achieve nominal coverage probability under sparse* $\boldsymbol{\gamma}$

data. The results are summarized in Table 1. The plugin method stands for the case that we calculate the sample mean of $\boldsymbol{\gamma}$ and $\boldsymbol{\varphi}$ when generating samples ϑ using model (2.3). The results for the debiased Lasso, Bayesian Lasso and horseshoe regression are also provided for comparison. It can be seen that when there are more unlabeled observations, the length of the interval will become slightly smaller.

We also investigate in a case when $\boldsymbol{\omega} \sim N(\mathbf{0}, \sigma_1^2 \boldsymbol{I}_n)$ and $\sigma_1^2 = 0.1$. This means model (2.1) has a smaller variability, and $\boldsymbol{X}$ is highly correlated with $\boldsymbol{Z}$. The rest of the data generating process is the same as the sparse $\boldsymbol{\gamma}$ case. The results in Fig. 3 show that the proposed method with spike and slab regression achieves the most satisfactory coverage across all signal strengths. Double machine learn-ing provides slightly narrower intervals and the proposed method with horseshoe regression provides slightly wider intervals, thus resulting in slight undercover-age and overcoverage, respectively. The performance of double selection method is better compared with the result in Fig. 2 and the resulting confidence inter-vals have similar coverage probabilities and average interval length across all signal strengths. This is due to the fact that small σ_1^2 provides a large signal to noise ratio, making the variable selection step easier for model (2.1). Bayesian regression based on spike and slab priors shows similar performance to results in Fig. 2. The resulting credible intervals show over-shrinkage for smaller sig-nals and provide empirical coverage probabilities close to the nominal level for large signals. All other methods fail to provide estimation intervals that give

Table 1
Inference on ϑ for semi-parametric setting

$n = 100$	DeLasso	HS	Laplace	CBP+HS	Plugin	500 more	1000 more
Estimation	1.039	1.108	0.959	0.987	0.998	0.987	0.987
Bias	0.039	0.108	−0.041	−0.013	−0.002	−0.013	−0.013
Var of Est.	0.003	0.034	0.013	0.010	0.010	0.010	0.010
MSE	0.005	0.045	0.015	0.010	0.010	0.010	0.010
Coverage(95/%)	0.790	0.812	0.952	0.958	0.938	0.956	0.960
Length of CI	0.280	0.490	0.515	0.414	0.400	0.411	0.410
$n = 300$	DeLasso	HS	Laplace	CBP+HS	Plugin	500 more	1000 more
Estimation	1.017	1.004	0.985	0.993	0.996	0.993	0.993
Bias	0.017	0.004	−0.015	−0.007	−0.004	−0.007	−0.007
Var of Est.	0.003	0.004	0.004	0.003	0.003	0.003	0.003
MSE	0.004	0.004	0.004	0.003	0.003	0.003	0.003
Coverage(95/%)	0.862	0.954	0.932	0.956	0.956	0.954	0.956
Length of CI	0.176	0.238	0.245	0.230	0.228	0.230	0.230

satisfactory coverage probabilities.

We further examine the robustness of our method under model misspecification. We investigate the case when the homoscedastic error assumption does not hold. We still use the same data generating process for $\boldsymbol{Z}$ and set $\boldsymbol{\eta}^{\square} = (0, 0, 2, 0, \ldots, 0)$. However, for $\boldsymbol{X}$, we let

$$X_i = 2Z_{i1} + Z_{i2} + w_i(1 + Z_{i1}), \text{ for } i = 1, \ldots, n,$$

where Z_{ij} are the i, jth entry of the design matrix $\boldsymbol{Z}$. In this case, the homoscedastic assumption in the error term $\boldsymbol{\omega}$ is violated for the partial regression model (2.1). The simulation results under such model misspecification are summarized in Fig. 4. Horseshoe regression and spike and slab regression still fail to provide reasonable credible intervals for small signals. Similar to the result in Fig. 2, double selection estimators have larger bias due to the strong dependence structure and the resulting confidence intervals have low coverage probabilities. Different from previous results, double machine learning in this case fails to provide intervals with valid coverage probabilities. A closer investigation shows that the estimated variance is smaller than the variance of the actual double machine learning estimator, which results in the undercoverage. The two conditional Bayesian posterior methods provide good coverage results and demonstrate robustness.

Next, we compare the performance of different methods when the sparsity assumption for $\boldsymbol{\gamma}$ is violated. We set $\boldsymbol{\gamma}^{\square} = 2 \bullet (1/2^2, 1, 1/3^2, 1/4^2, \ldots, 1/d^2)$ and $\boldsymbol{\eta}^{\square} = (1, 0, \ldots, 0)$. The rest of the data generating process is the same as the default setting. Notice that, in this case, $\boldsymbol{\gamma}^{\square}$ is no longer a sparse vector while still $\ddot{u}_1$ sparse. The simulation results are summarized in Fig. 5. Similar to the sparse $\boldsymbol{\gamma}^{\square}$ case (Fig. 2), the proposed methods still result in credible intervals that achieve nominal coverage probabilities across different signal strengths. The performance of the credible intervals from horseshoe regression and spike and slab regression only catch when the signal strength is large enough. For small signals, the over-shrinkage issues still exist. The confidence intervals based on debiased Lasso still show undercoverage for all signal strengths due to the large

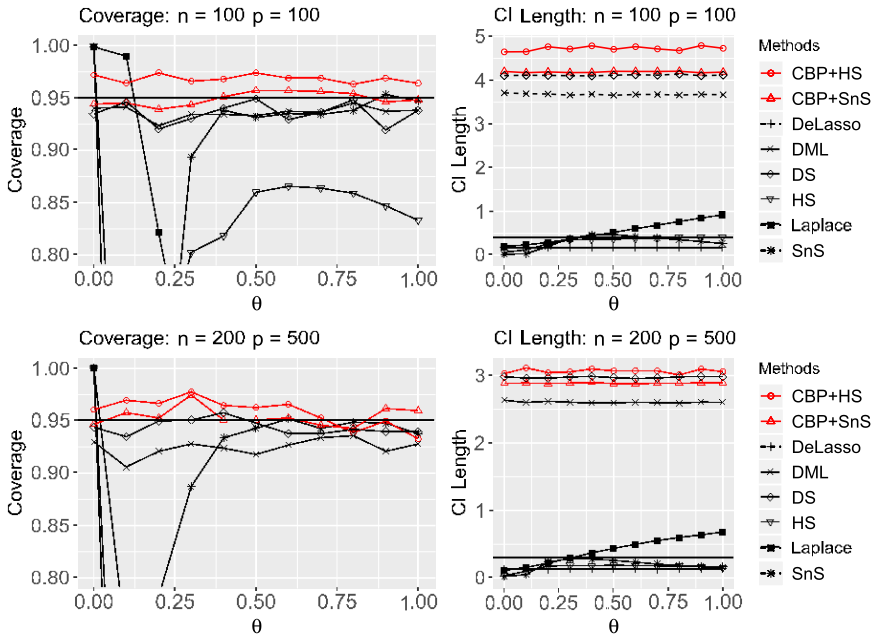


Fig 3. *The proposed methods (red solid) achieve nominal coverage probability under small* σ_1^2

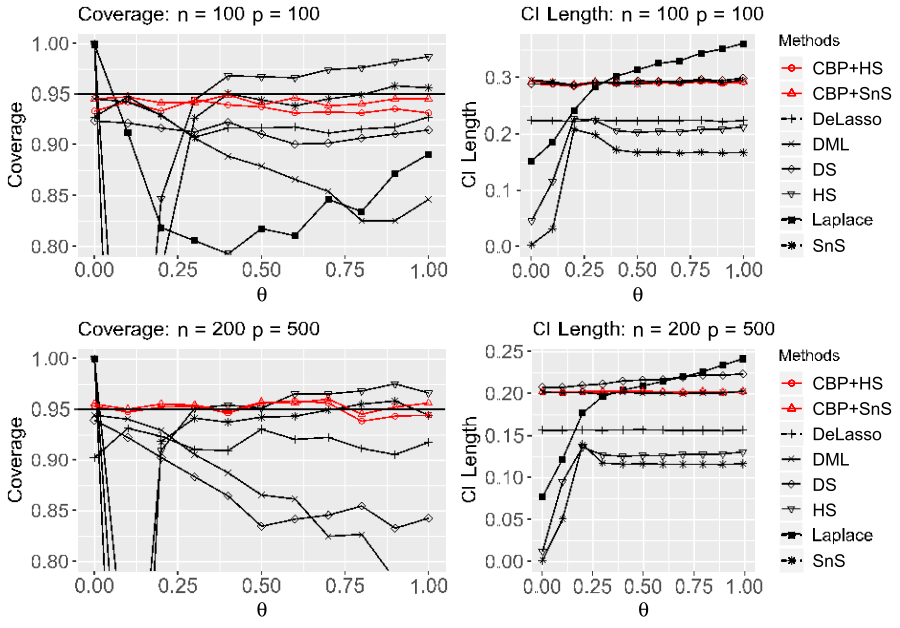


Fig 4. *The proposed methods (red solid) demonstrate reasonable robustness when homoscedastic error assumption does not hold*

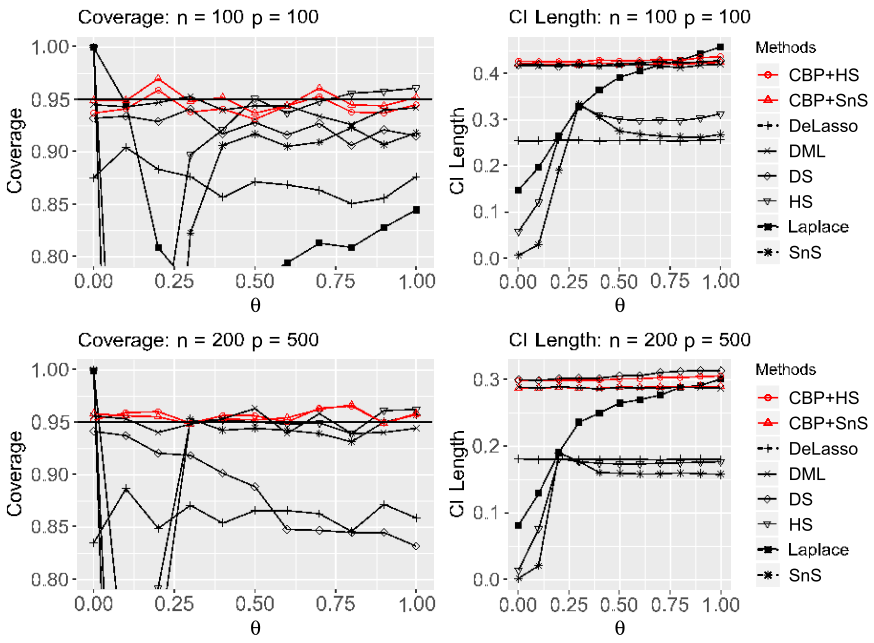


Fig 5. *The proposed methods (red solid) achieve nominal coverage probability under dense* γ

bias. Double selection still fails to provide intervals with satisfactory coverage probabilities for large signals due to the correlation structure. Double machine learning provides satisfactory interval estimations that perform similarly to the proposed methods.

3.2. Real data application

In this subsection, we apply our proposed method to examine the effect of mother's smoking on infant birth weight. This problem has been studied in Lumley et al. (2009), and they confirmed the causal relationship through randomized trials. Here, we use 2016 Natality data from the National Vital Statistics System of Centers for Disease Control and Prevention. We perform a similar regression analysis as has been done in Wang, He and Xu (2018). We treat the infant birth weight as the response variable and study the effect of the binary treatment variable – smoking or non-smoking mother. The high dimensional control variables include father's age and race, infant's sex, plurality, infant's birth defects, infant's Apgar score, the obstetric estimate of gestation, induction of labor, admission to NICU, mother's pre-pregnancy weight, mother's weight gain during pregnancy, mother's height, and several variables that indicate complications during pregnancy, and some interaction terms between these selected features. We use the ordinary least squares estimator from the entire dataset as the ground truth, and evaluate the performance of high dimensional methods

Table 2
Comparison of different methods for real data application

$n = 100$	SnS	DS	DML	CBP+SnS	$n = 200$	SnS	DS	DML	CBP+SnS
Est	−0.007	−49.429	−44.462	−48.389	Est	−0.011	−60.141	−57.81	−62.295
SE	0.589	111.381	91.469	107.77	SE	0.441	73.737	62.738	71.48
Coverage 95%	0.000	0.953	0.915	0.968	Coverage 95%	0.000	0.954	0.928	0.959
Length of CI	2.308	436.605	363.035	422.449	Length of CI	1.729	289.042	247.44	280.198
Power	NA	0.209	0.164	0.190	Power	NA	0.322	0.302	0.330

$n = 300$	SnS	DS	DML	CBP+SnS	$n = 400$	SnS	DS	DML	CBP+SnS
Est	−0.016	−59.692	−60.41	−63.132	Est	−0.023	−58.806	−61.353	−65.711
SE	0.350	58.819	51.614	56.088	SE	0.378	50.004	44.379	47.604
Coverage 95%	0.001	0.961	0.925	0.960	Coverage 95%	0.001	0.959	0.923	0.963
Length of CI	1.372	230.568	203.149	219.862	Length of CI	1.482	196.011	174.495	186.605
Power	NA	0.388	0.414	0.445	Power	NA	0.436	0.458	0.529

based on subsamples.

Following the analysis done in Wang, He and Xu (2018), we only consider live, singleton births to Asian mothers between the ages of 18 and 45, with no more than 2 years of college education in the United States. We consider the same set of predictors used in Wang, He and Xu (2018). As an implementa-tion detail, we deleted the observations with missing response values. We also deleted two categorical variables that indicate whether the infant has Down Syndrome or suspected chromosomal disorder, which are highly correlated with intercept and cause numerical issues. The variable mother's weight gain during pregnancy is left censored at zero and right censored at 98. Excluding these observations will provide residuals with much better normality. Also, since the high dimensional methods are not designed to deal with censored observations, including these will influence the performance of the high dimensional meth-ods. We further delete some outliers based on studentized residuals to avoid their potential impact on the high dimensional approaches. After processing the data, we end up with 57341 observations and 283 covariates. The fitted linear regression model explains 46.78% of the variance of the infant birth. Based on the regression output, women who were self-reported smokers delivered infants weighing 70.24g less than the others on average with a standard error of 11.62, which is similar to others' findings.

To compare the performance of the high dimensional methods, we randomly draw subsamples with size n from the full sample. Since the dataset is highly un-balanced, with only 2% smoking observations, we sample $n/2$ from the smoking observations and $n/2$ from the nonsmoking observations to obtain a balanced subsample.

We present the estimated coefficients and report the coverage probability based on 1000 replications. The results are summarized in Table 2 for different subsample sizes. The results suggest that directly using spike and slab regres-

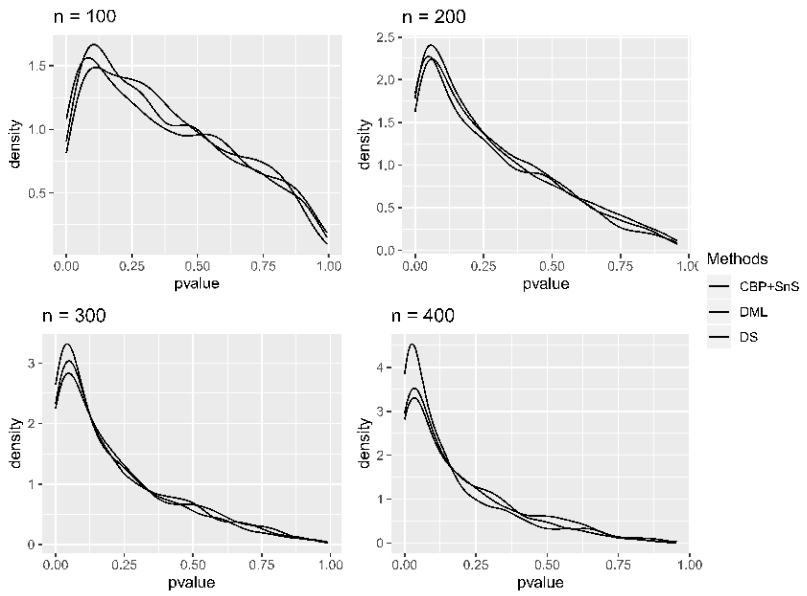


Fig 6. *Density estimation for the p-values*

sion fails to quantify the uncertainty associated with the estimator and resulting in super narrow credible intervals. For the remaining methods, double machine learning, similar to the simulation studies, results in shortest intervals among others and have coverage probabilities lower than the nominal level. The proposed conditional Bayesian posterior methods with spike and slab regression achieves smaller bias as the subsample size increases and has narrower intervals compared with double selection in general. We also report the p-values for testing whether there is a treatment effect at significance level $\alpha = 0.1$. Fig. 6 shows the kernel density estimation based on the empirical distribution of the resulting p-values. Overall, the proposed method tends to output smaller p-values in this case, which makes it more likely to assert the existence of treatment effect for the limited subsample size.

4. Technical proofs

4.1. Proof for Theorem 1

Proof. Let $\mathrm{E}^{\square} = \mathrm{E}_1 \cap \mathrm{E}_2 \cap \mathrm{E}_3 \cap \mathrm{E}_4$, we have $P(\mathrm{E}^{\square}) > 1 \square n^{\square c_1} \square d^{\square c_2} \square q d^{\square c_3} \square d^{\square c_4}$. For simplicity, assume $c_3 > \max\{c_2, c_4\}$, we have $P(\mathrm{E}^{\square}) > 1 \square n^{\square c_1} \square (q+2) d^{\square c_3}$. We will perform our analysis on this high probability set $\mathrm{E}^{\square}$.

Let $\tilde{\boldsymbol{\eta}} = \boldsymbol{\varphi} \square \boldsymbol{\gamma}\boldsymbol{\vartheta}^{\square}$, we have

$$\lfloor \tilde{\boldsymbol{\eta}} \square \boldsymbol{\eta}^{\square} \rfloor_1 = \lfloor \boldsymbol{\varphi} \square \boldsymbol{\gamma}\boldsymbol{\vartheta}^{\square} \square \boldsymbol{\eta}^{\square} \rfloor_1 = \lfloor \boldsymbol{\varphi} \square \boldsymbol{\gamma}\boldsymbol{\vartheta}^{\square} \square \boldsymbol{\eta}^{\square} + \boldsymbol{\gamma}^{\square}\boldsymbol{\vartheta}^{\square} \square \boldsymbol{\gamma}^{\square}\boldsymbol{\vartheta}^{\square} \rfloor_1$$

$$\leq \|\boldsymbol{\varphi} - \boldsymbol{\gamma}^{*}\vartheta^{*} - \boldsymbol{\eta}^{*}\|_1 + \|\boldsymbol{\gamma}^{*} - \boldsymbol{\gamma}\|_1\|\vartheta^{*}\|_1.$$

Recall that $\|\vartheta^{*}\|_1 \leq M_0 q$ for some $M_0 > 0$. Therefore, on set E^{*} we have

$$\Pi\left(\|\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}^{*}\|_1 \geq M_3((s_1 + s_2) + M_0 q s_1)\sqrt{\frac{\log d}{n}} \mid \Delta\right) \leq M_4(d^{-c_3} + d^{-c_4}).$$

Define set B by

$$B = \left\{\|\boldsymbol{\gamma} - \boldsymbol{\gamma}^{*}\|_1 \leq M_3 s_1\sqrt{\frac{\log d}{n}}, \|\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}^{*}\|_1 \leq M_3((s_1 + s_2) + M_0 q s_1)\sqrt{\frac{\log d}{n}}\right\}.$$

Notice that according to previous result, we have $\Pi(B^c|\Delta) \leq M_4(2d^{-c_3} + d^{-c_4})$, and for and set A ranging from all measurable sets in R^q,

$$\begin{aligned}
&|\Pi(\vartheta \in A \mid \Delta) - \Pi(\vartheta \in A \mid \Delta, B)| \\
=&|\Pi(\vartheta \in A \mid \Delta, B)\Pi(B \mid \Delta) + \Pi(\vartheta \in A \mid \Delta, B^c)\Pi(B^c \mid \Delta) - \Pi(\vartheta \in A \mid \Delta, B)| \\
\leq&2\Pi(B^c \mid \Delta) \leq 2M_4(2d^{-c_1} + d^{-c_2}).
\end{aligned} \tag{4.1}$$

Next, we show that

$$\sup_A |\Pi(\vartheta \in A \mid \Delta, B) - P(U \in A)| \leq C\frac{q(s_1 + s_2)\log d}{n^{1/2}},$$

for some positive constant C.

The conditional density of ϑ is given by

$$g(\vartheta|\boldsymbol{\gamma}, \boldsymbol{\varphi}, \Delta) \propto \exp\left\{-\frac{\|\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\varphi} - (\boldsymbol{X} - \boldsymbol{Z}\boldsymbol{\gamma})\vartheta\|_2^2}{2\sigma^2}\right\}\pi(\vartheta),$$

for some constant $C > 0$.

From the frequentist prospective, we substitute $\boldsymbol{Y}$ with $\boldsymbol{X}\vartheta^{*} + \boldsymbol{Z}\boldsymbol{\eta}^{*} + \boldsymbol{\varepsilon}$, the conditional density can be simplified to

$$g(\vartheta|\boldsymbol{\gamma}, \boldsymbol{\varphi}, \Delta) \propto \exp\left\{-\frac{1}{2}(\vartheta - \boldsymbol{\mu}_1)^T\Sigma_1^{-1}(\vartheta - \boldsymbol{\mu}_1)\right\},$$

where

$$\boldsymbol{\mu}_1 = \vartheta^{*} + (\tilde{\boldsymbol{\omega}}^T\tilde{\boldsymbol{\omega}})^{-1}\tilde{\boldsymbol{\omega}}^T(\boldsymbol{\varepsilon} + \boldsymbol{Z}(\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}^{*})), \qquad \Sigma_1 = \sigma^2(\tilde{\boldsymbol{\omega}}^T\tilde{\boldsymbol{\omega}})^{-1},$$

and $\tilde{\boldsymbol{\omega}} = \boldsymbol{X} - \boldsymbol{Z}\boldsymbol{\gamma}$. Therefore,

$$g(\vartheta \mid \boldsymbol{\gamma}, \boldsymbol{\varphi}, \Delta) = N(\boldsymbol{\mu}_1, \Sigma_1).$$

Recall that $U \sim N\left(\vartheta^{*} + (\boldsymbol{\omega}^{*T}\boldsymbol{\omega}^{*})^{-1}\boldsymbol{\omega}^{*T}\boldsymbol{\varepsilon}), \sigma^2(\boldsymbol{\omega}^{*T}\boldsymbol{\omega}^{*})^{-1}\right) =: N(\boldsymbol{\mu}_2, \Sigma_2)$. To bound $|\Pi(\vartheta \in A|\Delta, B) - P(U \in A)|$, notice that

$$|\Pi(\vartheta \in A \mid \Delta, B) - P(U \in A)| = \int |\Pi(\vartheta \in A \mid \Delta, \tilde{\boldsymbol{\eta}}, \boldsymbol{\gamma}) - P(U \in A)|d\Pi(\tilde{\boldsymbol{\eta}}, \boldsymbol{\gamma} \mid \Delta)$$

$$\leq \sup_{(\eta,\gamma)} \frac{1}{2} \| \Pi(\vartheta \mid \Delta, \eta, \gamma) - P(U) \|_{TV}$$

Therefore, it suffices to find a uniform bound for the total variance distance between $\Pi(\vartheta \mid \Delta, \tilde{\eta}, \gamma)$ and U.

It is known that the total variation distance is upper bounded by K-L divergence

$$\| N(\mu_1, \Sigma_1) - N(\mu_2, \Sigma_2) \|_{TV} \leq \sqrt{\frac{1}{2} D_{KL}(N(\mu_2, \Sigma_2) \| N(\mu_1, \Sigma_1))}.$$

The K-L divergence between two multivariate normal distribution is

$$\begin{aligned} & D_{KL}(N(\mu_2, \Sigma_2) \| N(\mu_1, \Sigma_1)) \\ & = \frac{1}{2} \left((\mu_1 - \mu_2)^T \Sigma_1^{-1} (\mu_1 - \mu_2) + tr(\Sigma_1^{-1} \Sigma_2) - q + \log \det(\Sigma_1 \Sigma_1^{-1}) \right). \end{aligned} \tag{4.2}$$

We derived the bound for K-L divergence between these two distributions. First notice that on set B,

$$\begin{aligned} \| \tilde{\omega}^T \tilde{\omega} - \omega^T \omega \|_2 &= \| (\gamma - \gamma)^T Z^T Z (\gamma - \gamma) + 2 (\gamma - \gamma)^T Z^T \omega \|_2 \leq \\ & \| Z(\gamma - \gamma) \|^2 + 2 \| (\gamma - \gamma)^T Z^T \omega \|_2 \\ & \cdot \; q s_1 \log d. \end{aligned}$$

$\| Z(\gamma - \gamma) \|^2_2$ is the fitting error of the regression model between X and Z, which is assumed to be $s_1 \log d$ in Assumption 1. In the $q \times q$ matrix $(\gamma - \gamma)^T Z^T \omega$, each elements

$$(\gamma_i - \gamma_i)^T Z^T \omega_i \leq \| \gamma_i - \gamma_i \|_1 \| Z^T \omega \|_\infty \cdot \; s_1 \sqrt{\frac{\log d}{n}} \cdot \sqrt{n \log d} \cdot \; s_1 \log d. \tag{4.3}$$

Then we have $\| (\gamma - \gamma)^T Z^T \omega \|_2 \cdot \; q s_1 \log d$ on set B. According to the assumption $q s_1 \log d \cdot \sqrt{n}$. Therefore, we have

$$\| \Sigma_1^{-1} \|_2 \cdot \; \| \tilde{\omega}^T \tilde{\omega} \|_2 \leq \| \omega^T \omega \|_2 + \| \tilde{\omega}^T \tilde{\omega} - \omega^T \omega \|_2 \cdot \; n.$$

We now bound the first component in Equation (4.2), and show that on set B,

$$(\mu_1 - \mu_2)^T \Sigma_1^{-1} (\mu_1 - \mu_2) \cdot \; n \| \mu_1 - \mu_2 \|^2_2 \to 0.$$

Notice that on set B,

$$\begin{aligned} \| \mu_1 - \mu_2 \|_2 &= \| (\tilde{\omega}^T \tilde{\omega})^{-1} \tilde{\omega}^T (\varepsilon + Z(\eta - \tilde{\eta})) - (\omega^T \omega)^{-1} \omega^T \varepsilon \|_2 \\ &= \| ((\tilde{\omega}^T \tilde{\omega})^{-1} \tilde{\omega}^T - (\omega^T \omega)^{-1} \omega^T) \varepsilon + (\tilde{\omega}^T \tilde{\omega})^{-1} \omega^T Z (\eta - \tilde{\eta}) \|_2 \\ &\leq \| ((\tilde{\omega}^T \tilde{\omega})^{-1} \tilde{\omega}^T - (\omega^T \omega)^{-1} \omega^T) \varepsilon \|_2 + \| (\tilde{\omega}^T \tilde{\omega})^{-1} \omega^T Z (\eta - \tilde{\eta}) \|_2. \end{aligned}$$

We first bound $\|((\tilde{\boldsymbol{\omega}}^T\tilde{\boldsymbol{\omega}})^{-1}\tilde{\boldsymbol{\omega}}^T - (\boldsymbol{\omega}^T\boldsymbol{\omega})^{-1}\boldsymbol{\omega}^T)\boldsymbol{\varepsilon}\|_2$. Let $\Delta^T := (\boldsymbol{\omega}^T\tilde{\boldsymbol{\omega}})^{-1}\boldsymbol{\omega}^T - (\boldsymbol{\omega}^T\boldsymbol{\omega})^{-1}\boldsymbol{\omega}^T$. Since $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\boldsymbol{I})$. According to Theorem 6.3.2 and 6.3.5 in Vershynin (2018), we have for any $t \geq 0$

$$P\{\|\Delta^T\varepsilon\|_2 \geq C\|\Delta\|_F + t\} \leq \exp\left(-\frac{ct^2}{\|\Delta\|^2}\right),$$

where c and C can be viewed as fixed conditioning on X and Z. Now substituting $\delta = \exp\left(-\frac{ct^2}{\|\Delta\|^2}\right)$ with will give

$$P\left\{\|\Delta^T\varepsilon\|_2 \geq C\|\Delta\|_F + \sqrt{\frac{2\|\Delta\|_2\log(1/\delta)}{n}}\right\} \leq \delta.$$

select $\delta = n^{-c_2}$, we have $\|\Delta^T\varepsilon\|_2 \leq (C + \sqrt{2c_2})\|\Delta\|_F$ on a set C_2 with probability at least $1 - n^{-c_2}$. We continue our analysis restricted to the set $B \cap B_2$, for which the following will hold

$$\begin{aligned}
&\|((\tilde{\boldsymbol{\omega}}^T\tilde{\boldsymbol{\omega}})^{-1}\tilde{\boldsymbol{\omega}}^T - (\boldsymbol{\omega}_S^T\boldsymbol{\omega}_S)^{-1}\boldsymbol{\omega}_S^T)\boldsymbol{\varepsilon}\|_2 \\
&\cdot\|(\tilde{\boldsymbol{\omega}}^T\tilde{\boldsymbol{\omega}})^{-1}\tilde{\boldsymbol{\omega}}^T - (\boldsymbol{\omega}_S^T\boldsymbol{\omega}_S)^{-1}\boldsymbol{\omega}_S^T\|_F \\
&= \|(\tilde{\boldsymbol{\omega}}^T\tilde{\boldsymbol{\omega}})^{-1}(\tilde{\boldsymbol{\omega}} - \boldsymbol{\omega}_S)^T + [(\tilde{\boldsymbol{\omega}}^T\tilde{\boldsymbol{\omega}})^{-1} - (\boldsymbol{\omega}_S^T\boldsymbol{\omega}_S)^{-1}]\boldsymbol{\omega}_S^T\|_F \cdot \sqrt{q} \\
&\bullet \left(\sqrt{s_1}\sqrt{q\log d} + \sqrt{qs_1}\frac{\log d}{n^{2\bullet}}\sqrt{nq}\right) \\
&\cdot \frac{q^{3/2}\sqrt{s_1\log d}}{n}.
\end{aligned}$$

As for the second term

$$\|(\tilde{\boldsymbol{\omega}}^T\tilde{\boldsymbol{\omega}})^{-1}\tilde{\boldsymbol{\omega}}^T\boldsymbol{Z}(\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}_S)\|_2 \leq \|(\boldsymbol{\omega}^T\boldsymbol{\omega})^{-1}\|_2\|\boldsymbol{\omega}^T\boldsymbol{Z}(\boldsymbol{\eta} - \boldsymbol{\eta}_S)\|_2.$$

For the first part, $\|(\tilde{\boldsymbol{\omega}}^T\tilde{\boldsymbol{\omega}})^{-1}\|_2 = O(1/n)$. The second part can be bounded as follows,

$$\|\tilde{\boldsymbol{\omega}}^T\boldsymbol{Z}(\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}_S)\|_2 = \|\boldsymbol{\omega}_S^T\boldsymbol{Z}(\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}_S)\|_2 + \|(\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_S)^T\boldsymbol{Z}^T\boldsymbol{Z}(\boldsymbol{\eta} - \tilde{\boldsymbol{\eta}}_S)\|_2,$$

where

$$\|\boldsymbol{\omega}_S^T\boldsymbol{Z}(\boldsymbol{\eta} - \boldsymbol{\eta}_S)\|_2 = O(q(s_1 + s_2)\log d)),$$

following the similar argument in Equation (4.3) and

$$\|(\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_S)^T\boldsymbol{Z}^T\boldsymbol{Z}(\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}_S)\|_2 \leq \|(\boldsymbol{\gamma} - \boldsymbol{\gamma}_S)Z^T\|_2\|Z(\boldsymbol{\eta} - \boldsymbol{\eta}_S)\|_2 \cdot q(s_1 + s_2)\log d.$$

Overall, we have

$$\|(\tilde{\boldsymbol{\omega}}^T\tilde{\boldsymbol{\omega}})^{-1}\tilde{\boldsymbol{\omega}}^T\boldsymbol{Z}(\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}_S)\|_2 \cdot \frac{q(s_1 + s_2)\log d}{n}.$$

Therefore, we have

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\Sigma_1^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \cdot n\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2 \cdot_2 \frac{(q(s_1 + s_2)\log d)^2}{n} + \frac{s_1q^3\log d}{n},$$

Then we bound the term

$$tr(\Sigma_1^{-1}\Sigma_2) = tr(\tilde{\omega}^T\tilde{\omega}(\omega^T\omega)^{-1}).$$

By definition

$$\begin{aligned}&(\tilde{\omega}^T\tilde{\omega})(\omega^T\omega)^{-1}\\ &\quad = I_q + 2(\gamma^* - \gamma)^T Z^T \omega(\omega^T\omega)^{-1} + (\gamma^* - \gamma)^T Z^T Z(\gamma^* - \gamma)(\omega^{*T}\omega^*)^{-1}.\end{aligned}$$

Therefore

$$\begin{aligned}&tr(\Sigma_1^{-1}\Sigma_2) - q\\ =&tr(2(\gamma^* - \gamma)^T Z^T \omega^*(\omega^{*T}\omega^*)^{-1}) + tr((\gamma^* - \gamma)^T Z^T Z(\gamma^* - \gamma)(\omega^{*T}\omega^*)^{-1})\\ .&\sqrt{q}\lfloor(\gamma^* - \gamma)^T Z^T \omega^*(\omega^{*T}\omega^*)^{-1}\rfloor_2 + \sqrt{q}\lfloor(\gamma^* - \gamma)^T Z^T Z(\gamma^* - \gamma)(\omega^{*T}\omega^*)^{-1}\rfloor_2\\ .&\frac{q^{3/2}s_1\log d}{n}.\end{aligned}$$

For the last term, we use the approximation

$$\det(I + hM) = 1 + h \bullet tr(M) + o(h^2),$$

$$\log|\Sigma_1\Sigma_2^{-1}| \le \log^3\left(1 + M_4\frac{q^{3/2}s_1\log d}{n}\right)^4 . \frac{q^{3/2}s_1\log d}{n}.$$

Overall we have that on set B,

$$D_{KL}(N(\mu_2, \Sigma_2)||N(\mu_1, \Sigma_1)) . \frac{(q(s_1 + s_2)\log d)^2}{n},$$

which implies

$$\lfloor N(\mu_1, \Sigma_1) - N(\mu_2, \Sigma_2)\rfloor_{TV} . \frac{q(s_1 + s_2)\log d}{n^{1/2}}.$$

Therefore, we have for any measurable subset A of $\mathbb{R}^q$,

$$\begin{aligned}&|\Pi(\vartheta \in A \mid \Delta) - P(U \in A)|\\ \le&|\Pi(\vartheta \in A \mid \Delta) - \Pi(\vartheta \in A \mid \Delta, B)| + |\Pi(\vartheta \in A \mid \Delta, B) - P(U \in A)|\\ \le& C\frac{q(s_1 + s_2)\log d}{n^{1/2}} + 2M_3(2d^{-c_3} + d^{-c_4}) \text{ on set } E^-.\end{aligned}$$

□

4.2. Proof for Corollary 1

Proof. Let set $A = (-\infty, \hat{q}^B_{\alpha/2}]$. According to Theorem 1, we have

$$\left|\Phi\left(\frac{\hat{q}^B_{\alpha/2} - \vartheta}{\hat{\sigma}}\right) - \alpha/2\right| = O\left(\frac{q(s_1 + s_2)\log d}{n^{1/2}} + d^{-c_3}\right)^4 := O(\delta_n),$$

where

$$\hat{\vartheta} = \vartheta^{\square} + (\boldsymbol{\omega}^{\square T}\boldsymbol{\omega}^{\square})^{\square 1}\boldsymbol{\omega}^{\square T}\boldsymbol{\varepsilon} \text{ and } \hat{\sigma} = \frac{\sigma}{\lfloor \boldsymbol{\omega}^{\square} \rfloor_2},$$

are maximum likelihood estimators when $\boldsymbol{\gamma}^{\square}$ is known. This implies

$$\hat{q}^B_{\alpha/2} = \hat{\vartheta} + \hat{\sigma} z_{\alpha/2} + O\left(\frac{1}{\sqrt{n}}\delta_n\right).$$

Let $\hat{q}_{\alpha/2}$ be the frequentist estimator for the lower bound of the confidence interval, we have

$$\hat{q}_{\alpha/2} = \hat{\vartheta} + \hat{\sigma} z_{\alpha/2}.$$

Therefore, on set $\mathrm{E}^{\square}$, we have

$$|\hat{q}^B_{\alpha/2} \square \hat{q}_{\alpha/2}| = O(\delta_n/\sqrt{n}).$$

Similarly we have

$$|\hat{q}^B_{1\square\alpha/2} \square \hat{q}_{1\square\alpha/2}| = O(\delta_n/\sqrt{n}).$$

For the coverage of the credible interval, we have

$$\begin{aligned}
&P(\vartheta^{\square} \in (\hat{q}^B_{\alpha/2}, \hat{q}^B_{1\square\alpha/2}))\\
=&P(\vartheta^{\square} \in (\hat{q}^B_{\alpha/2}, \hat{q}^B_{1\square\alpha/2}) \cap \mathrm{E}) + O((q+2)d^{\square c_3} + n^{\square c_1})\\
\le&P(\vartheta^{\square} \in (\hat{q}_{\alpha/2} \square O(\tfrac{1}{\sqrt{n}}\delta_n), \hat{q}_{1\square\alpha/2} + O(\tfrac{1}{\sqrt{n}}\delta_n))) + O((q+2)d^{\square c_3} + n^{\square c_1})\\
=&P\left(z_{\alpha/2} \square O(\delta_n) \le \frac{\boldsymbol{\omega}^{\square T}}{\lfloor \boldsymbol{\omega}^{\square} \rfloor_2}\boldsymbol{\varepsilon} \le z_{1\square\alpha/2} + O(\delta_n)\right) + O((q+2)d^{\square c_3} + n^{\square c_1}).
\end{aligned}$$

Since $\frac{\boldsymbol{\omega}^{\square T}}{\lfloor \boldsymbol{\omega}^{\square} \rfloor_2}\boldsymbol{\varepsilon}$ follows a standard normal distribution, we have

$$P(\vartheta^{\square} \in (\hat{q}^B_{\alpha/2}, \hat{q}^B_{1\square\alpha/2})) \le \alpha + O(q(s_1+s_2)\log d/n^{1/2} + qd^{\square c_3} + n^{\square c_1}).$$

Similarly,

$$P(\vartheta^{\square} \in (\hat{q}^B_{\alpha/2}, \hat{q}^B_{1\square\alpha/2})) \ge \alpha \square O(q(s_1+s_2)\log d/n^{1/2} + qd^{\square c_3} + n^{\square c_1}).$$

Therefore, we have

$$|P(\vartheta^{\square} \in (\hat{q}^B_{\alpha/2}, \hat{q}^B_{1\square\alpha/2})) \square \alpha| = O(q(s_1+s_2)\log d/n^{1/2} + qd^{\square c_3} + n^{\square c_1}). \qquad \square$$

5. Discussion

In this paper, we propose a conditional Bayesian posterior approach to facilitate inference on low dimensional parameters in high dimensional linear models. Our approach avoids the over-shrinkage issue associated with existing Bayesian regularized regression that leads to uncertainty underestimation for small signals. Theoretical results show that our proposed method can obtain root n-consistent credible intervals that achieve the nominal coverage probabilities in

the frequentist sense regardless of the signal strength. The proposed conditional Bayesian posterior has shown better robustness compared with other frequentist and Bayesian methods under model misspecifications. The current method focuses on a small number of parameters of interest and it would be interesting to see if there is a Bayesian approach to perform root n-consistent inference for a dense transformation of a high dimensional parameter.

References

[1] Bai, R. and Ghosh, M. (2018). High-dimensional multivariate posterior consistency under global–local shrinkage priors. *Journal of Multivariate Analysis* **167** 157–170. MR3830639

[2] Belitser, E. and Ghosal, S. (2020). Empirical Bayes oracle uncertainty quantification for regression. *The Annals of Statistics* **48** 3113–3137. MR4185802

[3] Belitser, E. and Nurushev, N. (2020). Needles and straw in a haystack: robust confidence for possibly sparse sequences. *Bernoulli* **26** 191–225. MR4036032

[4] Belloni, A., Chernozhukov, V. and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies* **81** 608–650. MR3207983

[5] Bontemps, D. et al. (2011). Bernstein–von Mises theorems for Gaussian regression with increasing number of regressors. *The Annals of Statistics* **39** 2557–2584. MR2906878

[6] Cai, T. T., Guo, Z. et al. (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of Statistics* **45** 615–646. MR3650395

[7] Carvalho, C. M., Polson, N. G. and Scott, J. G. (2009). Handling sparsity via the horseshoe. In *Artificial Intelligence and Statistics* 73–80.

[8] Castillo, I. and Nickl, R. (2013). Nonparametric Bernstein–von Mises theorems in Gaussian white noise. *The Annals of Statistics* **41** 1999–2028. MR3127856

[9] Castillo, I. and Szabó, B. (2020). Spike and slab empirical Bayes sparse credible sets. *Bernoulli* **26** 127–158. MR4036030

[10] Castillo, I., Schmidt-Hieber, J., Van der Vaart, A. et al. (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics* **43** 1986–2018. MR3375874

[11] Chernozhukov, V., Hansen, C. and Spindler, M. (2016). hdm: High-Dimensional Metrics. *R Journal* **8** 185–199.

[12] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C. and Newey, W. K. (2016). Double machine learning for treatment and causal parameters Technical Report, cemmap working paper.

[13] Dezeure, R., Bühlmann, P., Meier, L. and Meinshausen, N. (2015). High-Dimensional Inference: Confidence Intervals, p-values and R-Software hdi. *Statistical Science* **30** 533–558. MR3432840

[14] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association* **96** 1348–1360. MR1946581

[15] Fu, W. and Knight, K. (2000). Asymptotics for lasso-type estimators. *The Annals of statistics* **28** 1356–1378. MR1805787

[16] Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2013). *Bayesian data analysis*. Chapman and Hall/CRC. MR3235677

[17] George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88** 881–889.

[18] Hahn, P. R., He, J. and Lopes, H. (2018). bayeslm: Efficient Sampling for Gaussian Linear Regression with Arbitrary Priors R package version 0.8.0. MR3939378

[19] Hahn, P. R., He, J. and Lopes, H. F. (2019). Efficient sampling for Gaussian linear regression with arbitrary priors. *Journal of Computational and Graphical Statistics* **28** 142–154. MR3939378

[20] Hahn, P. R., Carvalho, C. M., Puelz, D., He, J. et al. (2018a). Regularization and confounding in linear regression for treatment effect estimation. *Bayesian Analysis* **13** 163–182. MR3737947

[21] Hahn, P. R., Carvalho, C. M., Puelz, D. and He, J. (2018b). Regularization and Confounding in Linear Regression for Treatment Effect Estimation. *Bayesian Analysis* **13** 163–182. MR3737947

[22] Ishwaran, H., Rao, J. S. et al. (2005). Spike and slab variable selection: frequentist and Bayesian strategies. *The Annals of Statistics* **33** 730–773. MR2163158

[23] Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research* **15** 2869–2909. MR3277152

[24] Khare, K., Oh, S.-Y. and Rajaratnam, B. (2015). A convex pseudolikelihood framework for high dimensional partial correlation estimation with convergence guarantees. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **77** 803–825. MR3382598

[25] Kleijn, B. J., van der Vaart, A. W. et al. (2012). The Bernstein-von-Mises theorem under misspecification. *Electronic Journal of Statistics* **6** 354–381. MR2988412

[26] Knight, K. and Fu, W. (2000). Asymptotics for Lasso-type estimators. *The Annals of Statistics* **28** 1356–1378. MR1805787

[27] Lumley, J., Chamberlain, C., Dowswell, T., Oliver, S., Oakley, L. and Watson, L. (2009). Interventions for promoting smoking cessation during pregnancy. *Cochrane Database of Systematic Reviews* **3**.

[28] Peng, J., Wang, P., Zhou, N. and Zhu, J. (2009). Partial Correlation Estimation by Joint Sparse Regression Models. *Journal of the American Statistical Association* **104** 735–746. MR2541591

[29] Ročková, V. and George, E. I. (2018). The spike-and-slab Lasso. *Journal of the American Statistical Association* **113** 431–444. MR3803476

[30] Scott, S. L. (2021). BoomSpikeSlab: MCMC for Spike and Slab Regression R package version 1.2.4.

[31] Song, Q. and Liang, F. (2017). Nearly optimal Bayesian shrinkage for high dimensional regression. *arXiv preprint arXiv:1712.08964*. MR4535982

[32] Syring, N. and Martin, R. (2019). Calibrating general posterior credible regions. *Biometrika* **106** 479–486. MR3949316

[33] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58** 267–288. MR1379242

[34] Van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R. et al. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics* **42** 1166–1202. MR3224285

[35] van der Pas, S., Szabó, B. and van der Vaart, A. (2017). Uncertainty quantification for the horseshoe (with discussion). *Bayesian Analysis* **12** 1221–1274. MR3724985

[36] Velleman, P. F. and Welsch, R. E. (1981). Efficient computing of regression diagnostics. *The American Statistician* **35** 234–242.

[37] Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science* **47**. Cambridge university press. MR3837109

[38] Wang, J., He, X. and Xu, G. (2018). Debiased inference on treatment effect in a high dimensional model. *Journal of the American Statistical Association* **just-accepted** 1–000. MR4078474

[39] Ye, F. and Zhang, C.-H. (2010). Rate minimaxity of the Lasso and Dantzig selector for the l_q loss in l_r balls. *Journal of Machine Learning Research* **11** 3519–3540. MR2756192

[40] Zhang, C.-H. et al. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics* **38** 894–942. MR2604701

[41] Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76** 217–242. MR3153940