Electronic Journal of Statistics
Vol. 13 (2019) 3082–3094
ISSN: 1935-7524
https://doi.org/10.1214/19-EJS1605

Posterior asymptotic normality for an individual coordinate in high-dimensional linear regression

Dana Yang

Department of Statistics and Data Science
Yale University
e-mail: xiaoqian.yang@yale.edu

Abstract: It is well known that high-dimensional procedures like the LASSO provide biased estimators of parameters in a linear model. In a 2014 paper Zhang and Zhang showed how to remove this bias by means of a two-step procedure. We show that de-biasing can also be achieved by a one-step estimator, the form of which inspires the development of a Bayesian analogue of the frequentists' de-biasing techniques.

Keywords and phrases: Bernstein-von Mises, high-dimensional Bayesian procedure, de-biasing.

Received January 2019.

1. Introduction

Consider the regression model

$$Y = Xb + \epsilon, \quad \epsilon \sim \mathcal{N}(0, I_n). \tag{1}$$

The design matrix X is of dimension $n \times p$. The vector $Y \in \mathbb{R}^n$ is the response and $b \in \mathbb{R}^p$ is the unknown parameter. We are particularly interested in the case where $p > n$, for which b itself is not identifiable. In such a setting identifiability can be attained by adding a sparsity constraint, an upper bound on $\|b\|_0$, the number of nonzero b_i's. That is, the model consists of a family of probability measures $\{\mathbb{P}_b : b \in \mathbb{R}^p, \|b\|_0 \leq s^*\}$, and the observation Y is distributed $\mathcal{N}(Xb, I_n)$ under $\mathbb{P}_b$.

We are interested in posterior inference on the vector b, when Y is actually distributed $N(X\beta, I_n)$ for some true sparse β. Throughout this paper the notation β is reserved for the truth, a p-dimensional deterministic vector. The notation b stands for the random vector with marginal distribution μ (*a.k.a.* the prior) and conditional distribution μ_Y (*a.k.a.* the posterior) given Y.

If p were fixed and X were full rank, classical theorems (the Bernstein-von Mises theorem, as in [1, page 141]) gives conditions under which the posterior distribution of b is asymptotically normal centered at the least squares estimator, with covariance matrix $(X^TX)^{-1}$ under $\mathbb{P}_\beta$.

The classical theorem fails when $p > n$. Although sparse priors have been proposed that give good posterior contraction rates [2] [3], posterior normality of b is only obtained under strong signal-to-noise ratio (SNR) conditions, such as those of Castillo el al. [2, Corollary 2], which forced the posterior to eventually have the same support as β. Effectively, their conditions reduce the problem to the classical, fixed dimensional case. However that is arguably not the most interesting scenario. Without the SNR condition, Castillo et al. [2, Theorem 6] pointed out that under the sparse prior, the posterior distribution of b behaves like a mixture of Gaussians.

There is hope to obtain posterior normality results without the SNR condition if one considers the situation where only one component of b is of interest, say b_1, without loss of generality. All the other components are viewed as nuisance parameters. As shown by Zhang and Zhang [4] in a non-Bayesian setting, it is possible to construct estimators that are efficient in the classical sense that

$$\hat{\beta}_1 = \beta_1 + \frac{X_1^T \epsilon}{\|X_1\|_2^2} + o_p\left(\frac{1}{\sqrt{n}}\right). \tag{2}$$

Here and subsequently $o_p(\cdot)$ is a shorthand for a stochastically small order term under $\mathbb{P}_\beta$ and X_i denotes the i'th column of X. Similarly X_{-i} denotes the $n \times (p-1)$ matrix formed by all columns of X except for X_i. For $J \subset [p]$ denote by b_J the vector $(b_j)_{j \in J}$ in $\mathbb{R}^{|J|}$. Write b_{-1} for $b_{[p]\setminus\{1\}}$. The $\|\cdot\|_2$ norm on a vector refers to the Euclidean norm.

Approximation (2) is useful when $\|X_1\|_2 = O(\sqrt{n})$, in which case the expansion (2) implies weak convergence [5, page 171]:

$$\|X_1\|_2(\hat{\beta}_1 - \beta_1) \rightsquigarrow \mathcal{N}(0,1) \quad \text{under } \mathbb{P}_\beta.$$

Such behavior for $\|X_1\|$ is obtained with high probability when the entries of X are generated *i.i.d.* from the standard normal distribution. More precisely, Zhang and Zhang [4] proposed a two-step estimator $\hat{\beta}_1^{(ZZ)}$ that satisfies (2) under some regularity assumptions on X and no SNR conditions. The exact form of the estimator $\hat{\beta}_1^{(ZZ)}$ will be given in section 2.1. Zhang and Zhang required the following behavior for X.

Assumption 1. *Let* $\gamma_i = X_1^T X_i / \|X_1\|_2^2$, *and* $\lambda_n = \sqrt{\frac{\log p}{n}}$. *There exists a constant* $c_1 > 0$ *for which*

$$\max_{2 \le i \le p} |\gamma_i| \le c_1 \lambda_n.$$

In addition, $\max_{i \le p} \|X_i\|_2 = O(\sqrt{n})$.

Assumption 2. *(REC($3s^*, c_2$)) There exist constants* $c' > 0$ *and* $c_2 > 2$ *for which*

$$\kappa(3s^*, c_2) = \min_{\substack{J \subset [p], \\ |J| \le 3s^*}} \inf_{\substack{b \neq 0, \\ \|b_{J^C}\|_1 \le c_2 \|b_J\|_1}} \frac{\|Xb\|_2}{\sqrt{n}\|b_J\|_2} > c' > 0. \tag{3}$$

Assumption 3. *The model dimension satisfies*

$$s^* \log p = o(\sqrt{n}).$$

Remark 1. *Assumption 2 is known as the restricted eigenvalue condition [6, page 1710] required for penalized regression estimators such as the LASSO estimator [7, page 1] and the Dantzig selector [8, page 1] to enjoy optimal l^1 and l^2 convergence rates. Note that assumption 2 forces $\|X_i\|_2 > c'\sqrt{n}$ for all $i \leq p$. Therefore assumptions 1 and 2 imply that the lengths of all columns of X are of order $\Theta(\sqrt{n})$.*

Remark 2. *Assumptions 1 and 2 are satisfied with high probability when the $n \times p$ entries of X are generated i.i.d. from a sub-Gaussian random variable with a fixed sub-Gaussian parameter. Assumption 1 can be easily proved via the Markov inequality. For the proof of assumption 2 see Mendelson et al. [9] and Zhou [10].*

Zhang and Zhang [4] provided the following theorem.

Theorem 1 ([4, Section 2.1, 3.1])**.** *Under assumptions 1, 2 and 3, the estimator $\hat{\beta}_1^{(ZZ)}$ has expansion* (2)*.*

The goal of this paper is to give a Bayesian analogue for Theorem 1, in the form of a prior distribution on b such that as $n, p \to \infty$, the posterior distribution of b_1 starts to resemble a normal distribution centered around an estimator in the form of (2). We provide the following bias corrected version of the sparse prior proposed by Gao, van der Vaart, and Zhou [3].

The bias corrected prior distribution on $b \in \mathbb{R}^p$:

1. *Let the sparsity level s of b_{-1} obey the probability mass function $\pi(s) \propto \frac{\Gamma(s)}{\Gamma(s/2)} \exp(-2Ds \log \frac{e(p-1)}{s})$ for a positive constant D.*
2. *Denote the projection matrix onto span(X_1) by H. Write $W = (I-H)X$. Let $S|s \sim Unif\,(Z_s := \{S \subset \{2, ..., p\} : |S| = s, W_S$ is full rank$\})$.*
3. *Given S, let b_S have density $f_S(b_S) \propto \exp(-\eta\|W_S b_S\|_2)$ for a positive constant η. Set $b_{S^c} = 0$.*
4. *Let $b_1|b_{-1} \sim \mathcal{N}(-\sum_{i\geq 2}\gamma_i b_i, \sigma_n^2)$ where $\sigma_n^2 \gg \|\beta\|_1\lambda_n/\|X_1\|_2$ and γ_i, λ_n are as defined in assumption 1.*

The following is the main result of this paper.

Theorem 2. *Under assumptions 1, 2 and 3, for each constant η, there exists a large enough constant $D > 0$ for which the prior distribution on b described above gives a posterior distribution of $\|X_1\|_2(b_1 - \hat{\beta}_1)$ that satisfies*

$$\left\|\mathcal{L}\left(\|X_1\|_2(b_1 - \hat{\beta}_1)|Y\right) - \mathcal{N}(0,1)\right\|_{BL} \to 0 \text{ in } \mathbb{P}_\beta, \tag{4}$$

where $\hat{\beta}_1$ is an estimator of β_1 with expansion (2)*.*

Here $\|\cdot\|_{BL}$ denotes the bounded Lipschitz norm, which metrizes the topology of weak convergence [11, page 323]. The bounded Lipschitz norm between two probability measures P and Q on $\mathcal{X}$ is defined as $\|P - Q\|_{BL} = \sup_f |Pf - Qf|$

where the supremum is over all functions $f : \mathcal{X} \to [-1, 1]$ with Lipschitz constant at most 1.

An estimator $\hat{\beta}_1$ with expansion (2) is the appropriate centering for the posterior distribution of b_1 given Y. To see that, take the two-sided α-credible interval as an example.

Recall that μ_Y stands for the posterior distribution of b given Y. It is an easy consequence of (4) that (by taking a sequence of bounded Lipschitz functions approaching an indicator function):

$$\left|\mu_Y \left\{ \|X_1\|_2 \left| b_1 - \hat{\beta}_1 \right| \le \Phi^{-1}\left(\frac{1+\alpha}{2}\right) \right\} - \alpha \right| \to 0 \text{ in } \mathbb{P}_\beta, \quad \text{or}$$
$$\mu_Y \left\{ b_1 \in \left[\hat{\beta}_1 - \frac{\Phi^{-1}((1+\alpha)/2)}{\|X_1\|_2}, \hat{\beta}_1 + \frac{\Phi^{-1}((1+\alpha)/2)}{\|X_1\|_2} \right] \right\} = \alpha + o_p(1).$$

On the other hand, for any estimator $\hat{\beta}_1$ with expansion (2), under the assumption that $\|X_1\|_2 = O(\sqrt{n})$,

$$\mathbb{P}_\beta \left\{ \beta_1 \in \left[\hat{\beta}_1 - \frac{\Phi^{-1}((1+\alpha)/2)}{\|X_1\|_2}, \hat{\beta}_1 + \frac{\Phi^{-1}((1+\alpha)/2)}{\|X_1\|_2} \right] \right\} = \alpha + o(1).$$

That is, the Bayesian's credible interval and the frequentist's confidence interval are both $\left[\hat{\beta}_1 - \Phi^{-1}((1+\alpha)/2)/\|X_1\|_2, \hat{\beta}_1 + \Phi^{-1}((1+\alpha)/2)/\|X_1\|_2\right]$, which covers the truth β_1 roughly α proportion of the time. In other words, Theorem 2 implies that the Bayesian inference on b_1 and frequentist inference on β_1 are aligned in the asymptotics.

We would like to point out that although our Bayesian analogue of bias correction matches the frequentist's treatment in terms of statistical performance, the from of posterior distribution involves up to 2^p integrations and is therefore very expensive to compute.

The paper is organized as follows. We begin by discussing the frequentists' de-biasing techniques in section 2.1, including the two-step procedure developed by Zhang and Zhang [4] and a one-step estimator. We show that the one-step estimator also achieves de-biasing. In section 2.2 we use the form of the one-step estimator to illustrate the intuition behind the construction of the bias corrected prior distribution. The proof of our main result Theorem 2 is given in section 3.

2. Main results

2.1. How does de-biasing work?

This section describes the main idea behind the construction of the two-step debiasing estimator proposed by [4]. An estimator is proposed to provide another way of interpreting the two-step procedure. The success of these estimators inspired us to design a prior distribution that achieves de-biasing under the same set of assumptions.

In sparse linear regression, penalized likelihood estimators such as the LASSO are often used and tend to give good global properties, such as control of the l_1 loss:

$$\mathbb{P}_\beta\left\{\|\tilde{\beta}-\beta\|_1 > Cs^*\lambda_n\right\} \to 0 \text{ as } n,p\to\infty \text{ for some } C>0, \tag{5}$$

where λ_n is as defined in assumption 1. For example, Bickel et al. [6, Theorem 7.1] showed that under the REC condition (assumption 2) the LASSO estimator satisfies (5).

In general, penalized likelihood estimators introduce bias for the estimation of individual coordinates. To eliminate this bias, Zhang and Zhang [4] proposed a two-step procedure using the following idea. First find a $\tilde{\beta}$ that satisfies (5), perhaps via a LASSO procedure. Then define

$$\hat{\beta}_1^{(ZZ)} = \arg\min_{b_1\in\mathbb{R}} \left\|Y - X_{-1}\tilde{\beta}_{-1} - b_1X_1\right\|_2^2. \tag{6}$$

Remark 3. *The estimator given by* (6) *is not exactly the same as the one that appears in [4]. Note that* $\hat{\beta}_1^{(ZZ)}$ *can be equivalently written as*

$$\hat{\beta}_1^{(ZZ)} = \tilde{\beta}_1 + \frac{X_1^T(Y-X\tilde{\beta})}{X_1^TX_1}.$$

Compare with the estimator proposed by Zhang and Zhang [4] which takes the form

$$\hat{\beta}_1 = \tilde{\beta}_1 + \frac{Z_1^T(Y-X\tilde{\beta})}{Z_1^TX_1}, \tag{7}$$

where Z_1 *is some pre-calculated vector, typically obtained by running penalized regression of* X_1 *on* X_{-1} *and taking the regression residual. Getting a Bayesian analogue for* (7) *may be possible. But we choose to present our findings on the simpler version* (6) *to better illustrate the idea behind the prior design.*

The estimator in (6) essentially penalizes the size of all coordinates except the one of interest. Under assumptions 1, 2 and 3, the two-step estimator $\hat{\beta}_1^{(ZZ)}$ is asymptotically unbiased with expansion (2).

We show in the next theorem that the same asymptotic behavior can be obtained in a single step. The idea of penalizing all coordinates but one is seen more clearly here. By leaving one term out of the LASSO penalty, de-biasing is achieved. This observation inspired us to construct our bias corrected prior (see section 2.2) such that the parameter of interest is not penalized.

Theorem 3. *Define*

$$\hat{\beta} = \arg\min_{b\in\mathbb{R}^p}\left(\|Y-Xb\|_2^2 + \eta_n\sum_{i\geq 2}|b_i|\right).$$

Under assumptions 1, 2 and 3, if η_n *is a large enough multiple of* $n\lambda_n$*, the one-step de-biasing estimator* $\hat{\beta}$ *achieves* l_1 *control* (5) *and de-biasing of the first* *coordinate simultaneously. The estimator for β_1 satisfies*

$$\hat{\beta}_1 = \beta_1 + \frac{X_1^T \epsilon}{\|X_1\|_2^2} + o_p\left(\frac{1}{\sqrt{n}}\right). \tag{8}$$

Proof. We will first show that $\hat{\beta}$ satisfies (5). It is well known that when the penalty involves all coordinates of b, then the bound on the l_1 norm is true [6, Theorem 7.1]. It turned out that leaving one term out the of penalty does not ruin that property.

As in the proof of [6, Theorem 7.1], we compare the evaluation of the penalized log-likelihood function at $\hat{\beta}$ and the truth β using the definition of $\hat{\beta}$.

$$\left\|Y - X\hat{\beta}\right\|_2^2 + \eta_n \left\|\hat{\beta}_{-1}\right\|_1 \leq \|Y - X\beta\|_2^2 + \eta_n \|\beta_{-1}\|_1.$$

Plug in $Y = X\beta + \epsilon$, the above is reduced to

$$\left\|X(\hat{\beta} - \beta)\right\|_2^2 \leq 2\sum_{i \leq p} \xi_i(\hat{\beta}_i - \beta_i) + \eta_n \left(\|\beta_{-1}\|_1 - \left\|\hat{\beta}_{-1}\right\|_1\right),$$

where $\xi_i = X_i^T \epsilon$. With high probability $|\max_{i \leq n} \xi_i| \leq R = C_2 n\lambda_n$, in which case we have

$$\left\|X(\hat{\beta} - \beta)\right\|_2^2 \leq 2R\left\|\hat{\beta} - \beta\right\|_1 + \eta_n \left(\|\beta_{-1}\|_1 - \left\|\hat{\beta}_{-1}\right\|_1\right). \tag{9}$$

From here we can bound $\|\beta_{-1}\|_1 - \|\hat{\beta}_{-1}\|_1$ by $\|(\hat{\beta} - \beta)_{-1}\|_1$ using the triangle inequality. But since $\beta_{S_C} = 0$, we can obtain a much tighter bound:

$$\begin{aligned}\|\beta_{-1}\|_1 - \left\|\hat{\beta}_{-1}\right\|_1 &\leq \left\|(\hat{\beta} - \beta)_{S\backslash\{1\}}\right\|_1 - \left\|\hat{\beta}_{S^C\backslash\{1\}}\right\|_1 \\ &= \left\|(\hat{\beta} - \beta)_{S\backslash\{1\}}\right\|_1 - \left\|(\hat{\beta} - \beta)_{S^C\backslash\{1\}}\right\|_1.\end{aligned}$$

Combine with (9) to deduce that

$$\left\|X(\hat{\beta} - \beta)\right\|_2^2 \leq (\eta_n + 2R)\left\|(\hat{\beta} - \beta)_{S\cup\{1\}}\right\|_1 - (\eta_n - 2R)\left\|(\hat{\beta} - \beta)_{S^C\backslash\{1\}}\right\|_1.$$

By choosing η_n to be a large enough multiple of $n\lambda_n$, we have

$$\left\|X(\hat{\beta} - \beta)\right\|_2^2 \leq c_3 n\lambda_n \left\|(\hat{\beta} - \beta)_{S\cup\{1\}}\right\|_1 - c_4 n\lambda_n \left\|(\hat{\beta} - \beta)_{S^C\backslash\{1\}}\right\|_1 \tag{10}$$

for some positive constants c_3, c_4 with $c_3/c_4 \leq 2 < c_2$. Since $\|X(\hat{\beta} - \beta)\|_2$ is always nonnegative, the inequality above implies

$$\left\|(\hat{\beta} - \beta)_{S^C\backslash\{1\}}\right\|_1 \leq \frac{c_3}{c_4}\left\|(\hat{\beta} - \beta)_{S\cup\{1\}}\right\|_1. \tag{11}$$

Therefore under assumption 2, we have

$$\left\|(\hat{\beta}-\beta)_{S\cup\{1\}}\right\|_2 \leq \frac{1}{c'\sqrt{n}}\left\|X(\hat{\beta}-\beta)\right\|_2.$$

Combine with (10) to deduce that

$$\begin{aligned}\left\|X(\hat{\beta}-\beta)\right\|_2^2 \leq & c_3 n\lambda_n \left\|(\hat{\beta}-\beta)_{S\cup\{1\}}\right\|_1 \\ \leq & c_3 n\lambda_n\sqrt{s^*+1}\left\|(\hat{\beta}-\beta)_{S\cup\{1\}}\right\|_2 \\ \leq & \frac{c_3}{c'}\sqrt{(s^*+1)\log p}\left\|X(\hat{\beta}-\beta)\right\|_2.\end{aligned}$$

Hence

$$\left\|X(\hat{\beta}-\beta)\right\|_2 \leq \frac{c_3}{c'}\sqrt{(s^*+1)\log p}.$$

Again by assumption 2, we can go back to bound the l_1 loss.

$$\begin{aligned}\left\|(\hat{\beta}-\beta)_{S\cup\{1\}}\right\|_1 &\leq \sqrt{s^*+1}\left\|(\hat{\beta}-\beta)_{S\cup\{1\}}\right\|_2 \leq \frac{1}{c'}\sqrt{\frac{s^*+1}{n}}\left\|X(\hat{\beta}-\beta)\right\|_2 \\ &\leq \frac{2c_3}{(c')^2}s^*\lambda_n.\end{aligned}$$

From (11) we have

$$\left\|\hat{\beta}-\beta\right\|_1 \leq 2\left(1+\frac{c_3}{c_4}\right)\frac{c_3}{(c')^2}s^*\lambda_n.$$

That concludes the proof of (5). To show (8), observe that the penalty term does not involve b_1.

$$\begin{aligned}\hat{\beta}_1 =& \arg\min_{b_1\in\mathbb{R}}\left\|Y-X_{-1}\hat{\beta}_{-1}-b_1X_1\right\|_2^2 \\ =& \beta_1+\sum_{i\geq 2}\gamma_i(\beta_i-\hat{\beta}_i)+\frac{X_1^T\epsilon}{\|X_1\|^2}.\end{aligned} \tag{12}$$

We only need to show the second term in (12) is of order $o_p(1/\sqrt{n})$. Bound the absolute value of that term with

$$\max_{i\geq 2}|\gamma_i|\cdot\left\|\hat{\beta}_S-\beta_S\right\|_1 \leq (c_1\lambda_n)(C_1 s^*\lambda_n),$$

by assumption 1 and the l_1 control (5). That is then bounded by $O_p(s^*\lambda_n^2) = o_p(1/\sqrt{n})$ by assumption 3. □

Remark 4. *With some careful manipulation the REC($3s^*, c_2$) condition as in assumption 2 can be reduced to REC(s^*, c_2). The proof would require an extra step establishing that $|\hat{\beta}_1-\beta_1|$ is of order $o_p(\|\hat{\beta}_S-\beta_S\|_1)+O_p(1/\sqrt{n})$.*

The ideas in the proofs for the two de-biasing estimators $\hat{\beta}_1^{(ZZ)}$ and $\hat{\beta}_1$ are similar. Ideally we want to run the regression

$$\arg\min_{b_1\in\mathbb{R}} \left\|Y - X_{-1}\beta_{-1} - b_1 X_1\right\|^2 . \tag{13}$$

That gives a perfectly efficient and unbiased estimator. However β_{-1} is not observed. It is natural to replace it with an estimator which is made globally close to the truth β_{-1} using a penalized likelihood approach. As seen in the proof of Theorem 3, most of the work goes into establishing global l_1 control (5). The de-biasing estimator is then obtained by running an ordinary least squares regression like (13), replacing β_{-1} by some estimator satisfying (5), so that the solution to the least squares optimization is close to the solution of (13) with high probability.

2.2. Bayesian analogue of de-biasing estimators

In the Bayesian regime, recall that b is the p-dimensional random vector obeying distribution μ under the prior and μ_Y under the posterior. For the Bayesian analogue to the de-biasing estimators, it is again essential to establish l_1 control on $b_{-1} - \beta_{-1}$, the deviation of b_{-1} from the truth. Such posterior contraction results were established by Castillo et al. [2] and Gao et al. [3], which already provide the preliminary steps for our Bayesian procedure. The following lemma in [3] serves as a Bayesian analogue of (5). It gives conditions under which the sparse prior proposed by Gao et al. [3] enjoys the l_1 minimax rate of posterior contraction.

Lemma 1. *(Corollary 5.4, [3]) Under the following prior distribution,*

1. *Let s have the probability mass function $\pi(s) \propto \frac{\Gamma(s)}{\Gamma(s/2)} \exp(-2Ds\log\frac{ep}{s})$.*
2. *Let $S|s \sim Unif(Z_s := \{S \subset \{1,...,p\} : |S| = s, X_S \text{ is full rank}\})$.*
3. *Given the subset selection S, let the coefficients b_S have density $f_S(b_S) \propto \exp(-\eta\|X_S b_S\|)$,*

if the design matrix X satisfies

$$\kappa_0((2+\delta)s^*, X) = \inf_{\|b\|_0\leq(2+\delta)s^*} \frac{\sqrt{s^*}\|Xb\|_2}{\sqrt{n}\|b\|_1} \geq c \tag{14}$$

for some positive constant c, δ, then for each positive constant η there exist constants $c_3 > 0$ and large enough $D > 0$ for which

$$\mu_Y\left\{\|b-\beta\|_1 > c_3 s^*\lambda_n\right\} \to 0 \quad \textit{in } \mathbb{P}_\beta \textit{ probability,}$$

where μ_Y denotes the posterior distribution of b given Y.

Our bias corrected prior described in section 1 is obtained by slightly modify the sparse prior of Gao et al. [3] to give good, asymptotically normal posterior behavior for a single coordinate. As discussed in the last section, classical approaches to de-biasing exploit the idea of penalizing all coordinates except the one of interest. The idea behind the construction of our bias corrected prior is to essentially put the sparse prior only on b_{-1}.

Recall that H is the matrix projecting $\mathbb{R}^n$ to $span(X_1)$. Under the model where $Y \sim \mathcal{N}(Xb, I_n)$, the likelihood function has the factorization

$$\begin{aligned}\mathcal{L}_n(b) &= \frac{1}{\sqrt{n}(2\pi)^{n/2}} \exp\left(-\frac{\|Y - Xb\|_2^2}{2}\right)\\ &= \frac{1}{\sqrt{n}(2\pi)^{n/2}} \exp\left(-\frac{\|HY - HXb\|_2^2}{2}\right)\\ &\quad \times \exp\left(-\frac{\|(I-H)Y - (I-H)Xb\|_2^2}{2}\right).\end{aligned}$$

Write $W = (I - H)X_{-1}$ and reparametrize $b_1^* = b_1 + \sum_{i\geq 2} \gamma_i b_i$ with γ_i as defined in assumption 1. The likelihood $\mathcal{L}_n(b)$ can be rewritten as a constant multiple of

$$\exp\left(-\frac{\|HY - b_1^* X_1\|_2^2}{2}\right) \exp\left(-\frac{\|(I-H)Y - Wb_{-1}\|_2^2}{2}\right).$$

The likelihood factorizes into a function of b_1^* and b_{-1}. Therefore if we make b_1^* and b_{-1} independent under the prior, they will be independent under the posterior. In the prior construction we made $b_1|b_{-1} \sim \mathcal{N}(-\sum_{i\geq 2} \gamma_i b_i, \sigma_n^2)$. Hence $b_1^* \sim \mathcal{N}(0, \sigma_n^2)$ and b_1^* is independent of b_{-1}. Note that under the prior distribution b_1 and b_{-1} are not necessarily independent.

The sparse prior put on b_{-1} is analogue to that of Gao et al. [3, section 3], using W as the design matrix in the prior construction. By lemma 1, b_{-1} is close to β_{-1} in l_1 norm with high posterior probability as long as $\kappa_o((2+\delta)s^*, W)$ is bounded away from 0.

We main result (Theorem 2) states that the prior distribution we propose has the effect of correcting for the bias, in a fashion analogous to that of the two-step procedure $\hat{\beta}_1^{(ZZ)}$. Let us first give an outline of the proof. The joint posterior distribution of b_1^* and b_{-1} factorizes into two marginals. In the X_1 direction, the posterior distribution of b_1^* is asymptotically Gaussian centered around $\frac{X_1^T Y}{\|X_1\|_2^2} = \beta_1^* + \frac{X_1^T \epsilon}{\|X_1\|_2^2}$. After we reverse the reparametrization we want the posterior distribution of b_1 to be asymptotically Gaussian centered around an efficient estimator $\hat{\beta}_1 = \beta_1 + \frac{X_1^T \epsilon}{\|X_1\|_2^2} + o_p(1/\sqrt{n})$. Therefore we need to show $b_1^* - b_1$ is very close to $\beta_1^* - \beta_1$. That can be obtained from the l_1 control on $b_{-1} - \beta_{-1}$ under the posterior. In the next section we will give the proof of Theorem 2 in detail.

3. Proof of Theorem 2

Since the prior and the likelihood of b_1^* are both Gaussian, the posterior distribution is also Gaussian:

$$b_1^*|Y \sim \mathcal{N}\left(\frac{\sigma_n^2}{1+\|X_1\|_2^2\sigma_n^2}X_1^TY, \frac{\sigma_n^2}{1+\|X_1\|_2^2\sigma_n^2}\right).$$

Independence of b_1^* and b_{-1} under the posterior gives that the above is also the distribution of b_1^* given Y and b_{-1}. Take $\hat{\beta}_1$ to be any estimator with expansion (2). The distribution of $\|X_1\|_2(b_1 - \hat{\beta}_1)$ given Y and b_{-1} is

$$\mathcal{N}\left(\|X_1\|_2\left(\frac{\sigma_n^2}{1+\|X_1\|_2^2\sigma_n^2}X_1^TY - \sum_{i\geq 2}\gamma_i b_i - \hat{\beta}_1\right), \frac{\sigma_n^2\|X_1\|_2^2}{1+\|X_1\|_2^2\sigma_n^2}\right). \qquad (15)$$

Note that without conditioning on b_{-1}, the posterior distribution of b_1 is not necessarily Gaussian.

The main part of the proof of Theorem 2 is to show that the bounded-Lipschitz metric between the posterior distribution of b_1 and $\mathcal{N}(\hat{\beta}_1, 1/\|X_1\|_2^2)$ goes to 0 under the truth. From Jensen's inequality and the definition of the bounded-Lipschitz norm we have

$$\begin{aligned}&\left\|\mathcal{L}(\|X_1\|_2(b_1-\hat{\beta}_1)|Y) - \mathcal{N}(0,1)\right\|_{BL}\\ \leq &\mu_Y^{b_{-1}}\left\|\mathcal{L}(\|X_1\|_2(b_1-\hat{\beta}_1)|Y, b_{-1}) - \mathcal{N}(0,1)\right\|_{BL}.\end{aligned}$$

Here $\mu_Y^{b_{-1}}$ stands for the expected value operator under the posterior distribution of b given Y. The superscript is a reminder that the operator integrates over the randomness of b_{-1}.

For simplicity denote the posterior mean and variance in (15) as ν_n and τ_n^2 respectively. The bounded-Lipschitz distance between two normals $\mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathcal{N}(\mu_2, \sigma_2^2)$ is bounded by $(|\mu_1 - \mu_2| + |\sigma_1 - \sigma_2|) \wedge 2$. Hence the above is bounded by

$$\mu_Y^{b_{-1}}(|\nu_n| \wedge 2) + \mu_Y^{b_{-1}}((|\tau_n - 1|) \wedge 2).$$

Therefore to obtain the desired convergence in (4), we only need to show

$$\mathbb{P}_\beta \mu_Y^{b_{-1}}(|\nu_n| \wedge 2) \to 0, \quad \text{and} \qquad (16)$$

$$\mathbb{P}_\beta \mu_Y^{b_{-1}}((|\tau_n - 1|) \wedge 2) \to 0. \qquad (17)$$

To show (16), notice that the integrand is bounded. Hence it is equivalent to show convergence in probability. Write

$$|\nu_n| = \frac{\sigma_n^2\|X_1\|_2^3}{1+\sigma_n^2\|X_1\|_2^2}\left(\beta_1 + \frac{X_1^T\epsilon}{\|X_1\|_2^2} + \sum_{i\geq 2}\gamma_i\beta_i\right)$$

$$
\begin{aligned}
&- \|X_1\|_2 \sum_{i\geq 2} \gamma_i b_i - \|X_1\|_2 \left(\beta_1 + \frac{X_1^T \epsilon}{\|X_1\|_2^2} + o_p\left(\frac{1}{\sqrt{n}}\right)\right) \\
\leq & \frac{\|X_1\|_2}{1+\sigma_n^2 \|X_1\|_2^2} \left| \beta_1 + \frac{X_1^T \epsilon}{\|X_1\|_2} + \sum_{i\geq 2} \gamma_i \beta_i \right| + \sum_{i \geq 2} \gamma_i (\beta_i - b_i) + o_p(1). \qquad (18)
\end{aligned}
$$

The first term is no longer random in b, and it can be made as small as we wish now that it is decreasing in σ_n. If we set $\sigma_n^2 \gg \|\beta\|_1 \lambda_n / \|X_1\|_2$, this term is of order $o_p(1)$.

For the second term, we will apply lemma 1 to deduce that this term also goes to 0 in $\mathbb{P}_\beta \mu_Y^{b_{-1}}$ probability. To apply the posterior contraction result we need to establish the compatibility assumption (14) on W.

Lemma 2. *Under assumption 1, 2, 3, the matrix $W = (I - H)X_{-1}$ satisfies*

$$
\kappa_0((2+\delta)s^*, W) = \inf_{\|b\|_0 \leq (2+\delta)s^*} \frac{\sqrt{s^*}\|Wb\|_2}{\sqrt{n}\|b\|_1} \geq c
$$

for some $c, \delta > 0$.

We will prove the lemma after the proof of Theorem 2.

To show (17), note that the integrand is not a random quantity. It suffices to show

$$
|\tau_n - 1| = \left| \frac{\sigma_n^2 \|X_1\|_2}{1 + \sigma_n^2 \|X_1\|_2^2} - \frac{1}{\|X_1\|_2} \right| \to 0.
$$

That is certainly true for a $\{\sigma_n\}$ sequence chosen large enough. Combine (16), (17) and the bound on the bounded Lipschitz distance, we have shown

$$
\mathbb{P}_\beta \left\| \mathcal{L}\left(\|X_1\|_2 (b_1 - \hat{\beta}_1) | Y \right) - \mathcal{N}(0,1) \right\|_{BL} \to 0.
$$

Proof of lemma 2. We will justify the compatibility assumption on W in two steps. First we will show that the compatibility assumption of the X matrix follows from the REC assumption 2. Then we will show that the compatibility constant of X and W are not very far apart.

Let us first show that under assumption 2, there exist constants $0 < \delta < 1$ and $c > 0$, for which

$$
\kappa_0((2+\delta)s^*, X) = \inf_{\|b\|_0 \leq (2+\delta)s^*} \frac{\sqrt{s^*}\|Xb\|_2}{\sqrt{n}\|b\|_1} \geq c.
$$

Denote the support of g as S. We have

$$
\begin{aligned}
\kappa_0((2+\delta)s^*, X) &\geq \inf_{\|b\|_0 \leq (2+\delta)s^*} \frac{1}{\sqrt{2+\delta}} \frac{\|Xb\|_2}{\sqrt{n}\|b_S\|_2} \\
&\geq \min_{\substack{J \subset [p], \\ |J| \leq 3s^*}} \inf_{\substack{b \neq 0, \\ \|b_{J^C}\|_1 \leq c_2 \|b_J\|_1}} \frac{\|Xb\|_2}{\sqrt{n}\|b_J\|_2}
\end{aligned}
$$

$$=\kappa(3s^*, c_2) > 0.$$

Now, under assumptions 1, 2 and 3, we will show that there exist constants $0 < \delta' < 1$ and $c' > 0$, for which

$$\kappa_0((2+\delta')s^*, W) \geq \kappa_0((2+\delta)s^*, X) + o(1).$$

For $g \in [R]^{p-1}$, we have

$$\begin{aligned}\|Wg\|_2 &= \left\| X \begin{bmatrix} 0 \\ g \end{bmatrix} - \sum_{i\geq 2} \gamma_i g_i \right\|_2 \\ &\geq \left\| X \begin{bmatrix} 0 \\ g \end{bmatrix} \right\|_2 - \lambda_n \|g_1\|_2\end{aligned}$$

by assumption 1. Deduce that

$$\begin{aligned}\kappa_0((2+\delta')s^*, W) = &\inf_{|b|_0 \leq (2+\delta)s^*} \frac{\sqrt{s^*}\|Wb\|_2}{\sqrt{n}\|b\|_1} \\ &\geq \kappa_0((2+\delta')s^* + 1, X) - \sqrt{\frac{s^*}{n}}\lambda_n \\ &= \kappa_0((2+\delta')s^* + 1, X) - \frac{\sqrt{s^* \log p}}{n}.\end{aligned}$$

The second term is of order $o(1)$ under assumption 3. □

Acknowledgement

I would like to thank David Pollard for his advice and comments. I would also like to thank the anonymous referee and associate editor for the feedback that helped improve this manuscript.

References

[1] Aad W. Van der Vaart. *Asymptotic Statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 2000. MR1652247

[2] Ismael Castillo, Johannes Schmidt-Hieber, Aad Van der Vaart, et al. Bayesian linear regression with sparse priors. *Annals of Statistics*, 43(5):1986–2018, 2015. MR3375874

[3] Chao Gao, Aad W van der Vaart, and Harrison H Zhou. A general framework for Bayes structured linear models. arXiv:1506.02174, 2015.

[4] 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. MR3153940

[5] David Pollard. *A User's Guide to Measure Theoretic Probability*, volume 8 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 2002. MR1873379

[6] Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009. MR2533469

[7] Robert Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. MR1379242

[8] Emmanuel Candes and Terence Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, 35(35):2313–2351, 2007. MR2382644

[9] Shahar Mendelson, Alain Pajor, and Nicole Tomczak-Jaegermann. Uniform uncertainty principle for Bernoulli and subgaussian ensembles. *Constructive Approximation*, 28(3):277–289, 2008. MR2453368

[10] Shuheng Zhou. Restricted eigenvalue conditions on subgaussian random matrices. *arXiv preprint* arXiv:0912.4045, 2009.

[11] R.M. Dudley. Speeds of metric probability convergence. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 22(4):323–332, 1972. MR0317364