On Bayesian robust regression with diverging number of predictors

Daniel Nevo

Department of Statistics
The Hebrew University of Jerusalem
Mount Scopus, Jerusalem 91905, Israel
and
Current address: Departments of Biostatistics and Epidemiology,
Harvard T.H. Chan School of Public Health
Boston, MA 02115, USA
e-mail: danielnevo@gmail.com

and

Ya'acov Ritov

Department of Statistics
The Hebrew University of Jerusalem
Mount Scopus, Jerusalem 91905, Israel
and
Department of Statistics
University of Michigan
1085 South University
Ann Arbor, MI 48109-1107
e-mail: yaacov.ritov@gmail.com

Abstract: This paper concerns the robust regression model when the number of predictors and the number of observations grow in a similar rate. Theory for M-estimators in this regime has been recently developed by several authors (El Karoui et al., 2013; Bean et al., 2013; Donoho and Montanari, 2013). Motivated by the inability of M-estimators to successfully estimate the Euclidean norm of the coefficient vector, we consider a Bayesian framework for this model. We suggest a two-component mixture of normals prior for the coefficients and develop a Gibbs sampler procedure for sampling from relevant posterior distributions, while utilizing a scale mixture of normal representation for the error distribution. Unlike M-estimators, the proposed Bayes estimator is consistent in the Euclidean norm sense. Simulation results demonstrate the superiority of the Bayes estimator over traditional estimation methods.

1. Introduction

When fitting a linear regression model to data, estimators robust to outliers are often desired. One popular approach to achieve robustness to outliers is to

use M-estimators under a penalty function other than the quadratic one. This methodology is often termed as "robust regression". Classical results for robust regression are that the M-estimator of the coefficients vector is consistent and normally distributed (see Huber, 2011, Chap. 7). These results were obtained for the case p, the number of predictors, is fixed or grows slowly with the number of observations, n. The case where p grows faster than n have been drawing a lot of a attention. In that scenario, a popular approach is to consider penalization based estimation methods, e.g., the Lasso (Tibshirani, 1996).

We consider a different scenario. Assume that $p < n$, yet p grows at the same rate as n. That is, $p/n \to \kappa$ for some positive constant $\kappa < 1$. This scenario was first recognized as an interesting one by Huber (1973). It is, however, only with the emergence of "big data" that researchers have begun to investigate the robust regression model under this regime. Asymptotic distribution and variance calculations were recently developed for M-estimators in this regime (El Karoui et al., 2013; El Karoui, 2013; Donoho and Montanari, 2013). Arguably the main result is that while the obtained estimator is normally distributed, its variance differs from Huber's classical results. Bean et al. (2013) have further shown that, unlike the classical $p \ll n$ scenario, the optimal M-estimator, in terms of efficiency, is not obtained by maximizing the log density of the errors. One more striking result is that for Double-Exponential errors, for κ larger than approximately 0.3, least squares regression is superior to least absolute deviations regression.

We argue in the current paper that M-estimation might be the wrong approach to the robust regression model in the $p/n \to \kappa$, $0 < \kappa < 1$ regime. Our main motivation is as follows. When using M-estimators, the estimation error of a single coefficient is in the usual order of $n^{-1/2}$ (El Karoui et al., 2013). However, the error accumulated over the coefficient vector does not vanish when $n \to \infty$. We will further argue that when the signal and the noise are of the same asymptotic order, only a few of the true coefficient values can be larger (in their absolute value) than $n^{-1/2}$. Putting it together, the estimation error of small coefficients, which are the majority, is larger than their actual value (in absolute value) when using M-estimators. Thus, apparent large effects can be actually microscopic.

Recognizing this characteristic of the problem, shrinkage methods may be a better fit in this robust regression model. Shrinkage can be achieved by using either the aforementioned regularization methods or by using a Bayesian approach with appropriate scaling of the prior hyperparameters. In this paper we consider the latter option.

It is well known that shrinkage can be achieved using a Bayesian methodology. Most of the discussion is centered in the normal error model. The James-Stein estimator, (James and Stein, 1961), is an empirical Bayes estimator (Efron and Morris, 1973); Ridge regression estimator is identical to what we get if assuming the regression coefficients are iid with normal prior, and a maximum *a posteriori* (MAP) estimator is used; and if we replace the normal prior with Laplace distribution prior we get the Lasso. The Laplace prior for the coefficients is actively researched. An efficient Gibbs sampler was suggested by Park and Casella

(2008). However, the Lasso attracts criticism from a Bayesian point of view, since the full posterior distribution of the coefficients vector does not attain the same risk rate as the posterior mode (Castillo, Schmidt-Hieber and Van der Vaart, 2015). Another, more recent, prior proposal is the horseshoe prior (Carvalho, Polson and Scott, 2009, 2010), which has some appealing properties, at least when the design matrix is the identity (van der Pas, Kleijn and van der Vaart, 2014).

One way to incorporate the intuition stemming form our earlier descriptions in a modeling framework is to assume that the coefficients can be separated to two groups: small-value and large-value coefficients (in their absolute size). This perspective aligns with existing Bayesian variable selection literature, where priors are assigned hierarchically. First, coefficients are separated into two groups, and then a prior distribution is determined according to the group assignment. Often, though not necessarily, one of the priors is the degenerate distribution at zero. A leading example for this framework is SVSS, Stochastic Search Variable Selection (George and McCulloch, 1993, 1997). An alternative approach is to have a prior distribution on the number of non-zero coefficients, to choose these coefficients uniformly, and then to have a prior on the non-zero coefficients (Castillo et al., 2012),

In this paper we suggest a full Bayesian approach for the robust regression model when $p/n \to \kappa$, $0 < \kappa < 1$. We choose the prior distributions and hyperparameters such that our prior knowledge on the design is taken into account. We then utilize a scale mixture of normal representation of the error distribution to construct a reasonably fast Gibbs Sampler.

The rest of the paper is organized as follows. In Section 2, we present notation and model assumptions, and claim that M-estimation should not be used in this p-close-to-n regime. In Section 3, we introduce an hierarchical Bayesian model and then, in Section 4, we present a Gibbs sampler for parameter estimation. Detailed example is given in Section 5 where we also present simulation results. Section 6 offers conclusion remarks. Proofs are given in Section 7.

2. Achilles heel of M-estimators when $p/n \to \kappa \in (0, 1)$

We start with notations. We use $\| \cdot \|, \| \cdot \|_1$ and $\| \cdot \|_\infty$ for the Euclidean norm, the ℓ_1 norm and the maximum norm of a vector, respectively. Throughout the paper we consider the model

$$Y^{(n)} = X^{(n)} \beta^{(n)} + \epsilon^{(n)}, \tag{1}$$

where $\epsilon^{(n)}$ is a vector of i.i.d random variables with a known density function $f_\epsilon(\cdot; \theta)$ characterized by θ, an unknown parameter, $X^{(n)}$ is a matrix of random predictors, $\beta^{(n)}$ is an unknown parameter vector we wish to estimate. To improve clarity, we henceforth omit the superscript indicating that all model components given in (1) depends on n (θ excluded). We denote X_i^T for the ith row of X. X and ϵ are assumed to be independent. We denote β^0 for the true value of β. For

a given penalty function ρ, the M-estimator of β, $\hat{\beta}^\rho$, is defined as

$$\hat{\beta}^\rho = \arg\min_\beta \sum_{i=1}^n \rho(Y_i - X_i^T \beta). \tag{2}$$

If ρ is convex, one could alternatively solve the equation

$$\sum_{i=1}^n X_i^T \psi(Y_i - X_i^T \beta) = 0, \quad \psi(\cdot) := \rho'(\cdot).$$

Huber's classical result (1973) is that if $p^2/n \to 0$, then $\sqrt{n}(\hat{\beta}^\rho - \beta^0)$ is asymptotically normal with a covariance matrix

$$\frac{E(\psi^2(\epsilon_1))}{E^2(\psi'(\epsilon_1))} \lim_{n\to\infty} (X^T X)^{-1}.$$

If it is further assumed that $E(\epsilon_1) = 0$, then by using general M-estimation theory it can be shown that this result holds for $p/n \to 0$. Portnoy (1984, 1985) derived consistency and asymptotic normality of M-estimators in the robust regression model under weaker assumptions. See also Maronna and Yohai (1981).

We now claim that in the model described above, a robust regression model where the number of predictors and the number of observations are similar, M-estimators have undesirable properties. But first, we present our model assumptions:

(M1) $\lim_{n\to\infty} \frac{p}{n} = \kappa \in (0,1)$.

(M2) The rows of X are i.i.d $N(0,\Sigma)$ for a covariance matrix sequence $\Sigma = \Sigma_p$. Furthermore, assume the eigenvalues of Σ are bounded away from zero, for all p.

(M3) $\epsilon_i, i = 1, 2, ..., n$ are i.i.d mean zero random variables with a density function f_ϵ. $\ell_\epsilon = \log f_\epsilon$ is concave, bounded from above and has three bounded derivatives, such that $\sup_{|t|<M} \ell_\epsilon''(t) < 0$ for any $M < \infty$.

(M4) f_ϵ is symmetric and for the function $g_\epsilon(u) = f_\epsilon(\sqrt{u})$ we have, for $u > 0$ and $k = 1, 2, \ldots$,

$$\left(-\frac{d}{du}\right)^k g(u) \geq 0.$$

Assumption (M4) would be exploited when we will consider the Bayesian formulation of the problem. We note in passing that this assumption is fulfilled by rich family of distributions, such as Student's T distribution, the Laplace distribution, and the more inclusive Exponential power family (Andrews and Mallows, 1974; West, 1987).

We now take the frequentist point of view, which we will later, in the next section, replace in a Bayesian perspective. When estimating a multidimensional parameter, a loss function is needed to aggregate over the different components. A natural loss function is the ℓ_2 of the estimation errors. The following proposition motivates our discussion. It is based on the result of El Karoui et al. (2013) which is stated without a clear statement of the conditions.

Proposition 1. *Let assumptions* (M1)–(M3) *and assume Result 1 in El Karoui et al. (2013) holds. Let $\hat{\beta}^\rho$ be the M-estimator defined in* (2) *with respect to a non-linear convex function ρ. Then $\|\hat{\beta}^\rho - \beta^0\| = O_p(1)$*

A proof using the results of El Karoui et al. (2013) is given in Section 7. Proposition 1 implies that the so-called ℓ_2-consistency cannot be achieved by M-estimators under this regime.

However, we now claim M-estimators might be the wrong approach here. Consider specifically the problem arising when the signal and the noise are of the same asymptotic order, i.e, when $X_i^T \beta^0 = O_p(1)$. For a moment, let $\Sigma = I$. Then, $X_i^T \beta^0 = O_p(1)$ implies $\|\beta^0\| = O_p(1)$. If $\Sigma \neq I$, then since Σ is known, the last statement holds if taking $\tilde{X}_i = X_i \Sigma^{-1/2}$ instead of X_i. Informally, having the number of predictors in the same scale as the number of observations while considering a finite signal-to-noise ratio, that does not vanish as n grows, implies additional assumptions on β^0 structure. For example, not too many components of β^0 can be much larger than $n^{-1/2}$ (in their absolute value), otherwise the signal would be stronger than assumed. On the other hand, if β^0 is concentrated very close to zero for all n, then the signal is too week as $X_i^T \beta^0$ is too small.

M-estimation is invariant to translation of β^0, but some β^0 values are less expected then others, as we argued in the preceding paragraph. Therefore, another approach, which exploits that knowledge on the parameter vector β^0 is desirable. Since we know that many of the true coefficients are smaller than $n^{-1/2}$ (in their absolute value) we could potentially gain better estimates if we shrink some coefficients towards zero. This can be done using regularization based methods, or alternatively, using a Bayesian approach in a way that shrinkage is encouraged by the specified prior distribution. In this paper, we choose to take the latter option, and in the next section we develop such a Bayesian hierarchical model.

3. A Bayesian model

A Bayesian model for the robust regression involves at least one more level of parameters. Assume model (1) holds with a density function f_ϵ that depends on a parameter θ, and obeys assumptions (M3)-(M4). We start with a prior for β. As we argued in the previous section, we expect many of its components to be small, while some are considerably larger. To accommodate for this, we present a mixture prior for each $\beta_j, j = 1, .., p$, with two normal mixture components. Denote $T = (t_1, ..., t_p) \in \{1, 2\}^p$ for the vector that indicates for each component j if β_j has a large variance ($t_j = 1$) or a small one ($t_j = 2$). We also denote $\phi = E(\sum_{j=1}^{p} \mathbf{1}\{t_j = 1\})$ for the number of "large" components in T. Finally, let $\delta_k^2, k = 1, 2$ be the variance in each of the mixture components. Putting all together, the following assumptions depict our prior distributions.

(P1) The prior for β is of an iid mixture of two zero-mean normal variables

$$(t_j - 1)|\phi \sim Ber(\phi/p) \qquad \beta_j|t_j, \delta_1, \delta_2 \sim N(0, \delta_{t_j}^2)$$

(P2) $\phi = o_p(n/\log n)$.
(P3) δ_1^2 is $O_p(\phi^{-1})$ and δ_2^2 is $O_p(1/n^\xi)$, for some fixed $\xi > 1$.

Assumption (P2) implies that ϕ grows with n, but yet it is much smaller than p. Note that under assumptions (P1)–(P3) we have

$$E(\|\beta\|^2|\delta_1, \delta_2, \phi) = \phi\delta_1^2 + (p - \phi)\delta_2^2 = O_p(1),$$

so together with the Assumption (M2) we get that $X_i^T\beta = O_p(1)$.

Working prior distributions can be taken for ϕ, δ_1 and δ_2, adding another level of hierarchy to the model. In practice, the parameters of these prior distributions are chosen such that assumptions (P2) and (P3) are fulfilled. In Section 5, we present such an example. Alternatively, ϕ, δ_1 and δ_2 may be taken as known values that obeys Assumptions (P1)–(P3). The prior distribution for β as specified in Assumptions (P1)–(P3) reflects our knowledge on β when we assume the signal and the noise in our model are of the same order. This prior implies that only a small part of the coefficients can be larger than the $n^{-1/2}$ threshold. This can be stated formally using Chebyshev's inequality; See Lemma 1 in Section 7. Of course, one can think of other prior distributions for β having this property.

As for θ, we assume a prior distribution $q(\theta)$, where with some abuse of notation, q always denotes a density, and the particular relevant density would be clear from its argument. Unlike other model parameters, this prior does not change with n.

Before moving to the estimation procedure, we present a theoretical result showing that unlike M-estimators, a Bayes estimator for β can achieve ℓ_2-consistency. As stated before, M-estimators in our regime are consistent when considering each coordinate of the vector separately, but not when considering the parameter vector as a whole. The following theorem shows that Bayesian estimator in the discussed model is consistent (in the Euclidean norm sense) for the parameter vector.

Theorem 1. *Consider the model* (1) *and assume* (M1)–(M3) *and* (P1)–(P3). *Let* β^0 *be the true value of* β. *Let* $\hat\beta^*$ *be a Bayes estimator with respect to the posterior distribution* $q(\beta|Y, X)$ *and a loss function* L *of the form* $L(\beta, \beta^0) = L(\|\beta - \beta^0\|)$, *for a bounded* L. *Then* $\|\hat\beta^* - \beta^0\| \xrightarrow{p} 0$ *as* $n \to \infty$,

where $\xrightarrow{p}$ denotes convergence in probability. The proof is given in Section 7.

4. Sampling from the posterior distribution

Modern Bayesian statistics relies on the ability to sample from the posterior distribution. This may pose a challenge especially if the parameter space is

high dimensional. Assumption (M4) aids us in construction of a Gibbs sampler (Geman and Geman, 1984), with block sampling available for the full conditional of β. Under the condition in Assumption (M4), we can write f_ϵ as a scale mixture of normal distribution, Andrews and Mallows (1974),

$$f_\epsilon(t;\theta) = \int_0^\infty \frac{1}{\sigma}\varphi\left(\frac{t}{\sigma}\right)q(\sigma^2|\theta)d\sigma^2 \tag{3}$$

where φ is the density function of a standard normal random variable. This implies that while $q(Y|X,\beta,\theta)$ is $\prod_{i=1}^n f_\epsilon(Y_i - X_i^T\beta;\theta)$, $q(Y|X,\beta,\sigma^2)$ is the density of independent normally distributed random variables with mean zero and variance σ_i^2. The mixing distribution $q(\sigma^2|\theta)$ can be identified in some cases (Andrews and Mallows, 1974; West, 1987). If f_ϵ is the density of a Laplace distribution, as in the example we present in Section 5, then $q(\sigma^2|\theta) \propto \exp(-\theta^2\sigma^2/2)$. This representation adds n parameters as an individual σ_i^2 is artificially introduced to the model for each observation i. However, it allows for direct sampling from all the full conditionals of all the parameters, and a Gibbs sampler can be used. This Gibbs sampler resembles the one suggested for the Bayesian Lasso (Park and Casella, 2008), especially for the case presented in our example in Section 5 when the errors have a Laplace distribution. Note however that in their case the scale mixture of normals representation is taken for the prior β and the errors are normally distributed, where here we apply this representation to the errors, and the prior for β is a mixture of two normal distributions.

We now present the Gibbs sampler when Φ, δ_1 and δ_2 are known hyperparameters. In Section 5, we demonstrate how standard conjugate priors can be used for these parameters in a concrete example. The Gibbs sampler iterates between the full conditionals of $\theta, \sigma_1^2, ..., \sigma_n^2, t_1, ..., t_p$ and β. Starting with θ, one can sample from $q(\theta|\sigma_1^2...,\sigma_n^2) \propto q(\theta)\prod_{i=1}^n q(\sigma_i^2|\theta)$. The mixing distribution $q(\sigma_i^2|\theta)$ is determined by the error distribution $f_\epsilon(t;\theta)$, and depending on the prior $q(\theta)$, we may have a conjugate family (as in our Section 5 example). Alternatively, we may use Metropolis-Hastings step for θ only. Since θ is one dimensional, we expect such a step to marginally affect the computation time of the Bayes estimator.

Moving to each σ_i^2, we have

$$q(\sigma_i^2|Y_i,X_i,\beta,\theta) \propto \frac{1}{\sigma}\varphi\left(\frac{Y_i - X_i^T\beta}{\sigma_i}\right)q(\sigma_i^2|\theta) \tag{4}$$

and depending on the mixture distribution this may be a known distribution to sample from (see again Section 5).

Next, let $\Gamma = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2)$ and let $V = \text{diag}(\delta_{t_1}^2, \delta_{t_2}^2, \ldots, \delta_{t_p}^2)$. The full conditional of β is a multivariate normal with mean μ and variance A^{-1} where $A = X^T\Gamma^{-1}X + V^{-1}$ and $\mu = A^{-1}X^T\Gamma^{-1}Y$. This is the main advantage of the scale mixture of normals representation for f_ϵ; block sampling of β can be used. Finally, denote $p_k = P(t_j = k|\beta_j, \phi, \delta_1, \delta_2)$ for the full conditional of

each t_j. We have

$$[p_1 \quad p_2] \propto \left[\frac{1}{\delta_1} \varphi \left(\frac{\beta_j}{\delta_1} \right) \frac{\phi}{p} \quad \frac{1}{\delta_2} \varphi \left(\frac{\beta_j}{\delta_2} \right) \frac{p - \phi}{p} \right].$$

We now comment on alternative posterior sampling strategies, all involve the Gibbs sampler described above. When the mixing distribution $q(\sigma^2|\theta)$ results in a full conditional $q(\sigma_i^2|Y_i, X_i, \beta, \theta)$ with no available direct sampling, the proposed Gibbs sampler must use other MCMC procedures for sampling from the full conditionals of σ_i^2. Moreover, if Assumption (M4) does not hold, then the scale mixture of normal representation (3) cannot be used. Then, sampling from

$$q(\theta|Y, X, \beta) \propto q(\theta) \prod_{i=1}^{n} f_\epsilon(Y_i - X_i\beta; \theta)$$

cannot necessarily be done in a direct sampling manner, but additional MCMC procedure, such as Metropoils-Hastings should be used. This is not, however, where most of the problem lies. Block sampling for β would have been based upon high-dimensional full conditional of β

$$q(\beta|Y, X, T, \delta_1, \delta_2) \propto \prod_{j=1}^{p} \delta_{t_j}^{-1} \varphi(\beta_j/\delta_{t_j}) \prod_{i=1}^{n} f_\epsilon(Y_i - X_i\beta; \theta).$$

but since p is large, this could be a too ambitious approach, when using Metropolis-Hastings or other MCMC procedures. So as a last resort, one could use a Gibbs sampler that uses $q(\beta_j|Y, X, T, \delta_1, \delta_2, \beta_{-j})$, where β_{-j} is the β vector without its j–th component. Both the approaches we just described are expected to be more computationally heavy than the procedure we suggested before. We remark that sampling from each $q(\sigma_i^2|Y_i, X_i, \beta, \theta)$ can be parallelized, to shorten computation time since the σ_i^2's are independent given θ and β.

5. Example

We present in this section a specific example. First we describe the model with more detail, then we move to the estimation procedure and then we present simulation results with concrete numerical values. We assume a Laplace distribution for the errors, that is,

$$f_\epsilon(t; \theta) = \frac{\theta}{2} \exp(-\theta|t|).$$

This implies that the mixing probability is Exponential, with $q(\sigma^2|\theta) \propto \exp(-\theta^2\sigma^2/2)$ (Andrews and Mallows, 1974). We take an Exponential prior distribution for θ^2 (a Gamma prior distribution would also work). The resulting posterior for θ^2 is a Gamma distribution with a shape parameter of $n/2+1$ and a scale parameter $1 + (\sum_{i=1}^{n} \sigma_i^2)/2$. Moving to δ_1 and δ_2. Let $N_k = \sum_{j=1}^{p} \mathbf{1}\{t_j = k\}$. Then given a prior distribution $q(\delta_k)$, the full conditional for δ_k is simplified to

$$q(\delta_k|\beta, T) \propto (\delta_k^2)^{-\frac{N_k}{2}} \exp\left(\frac{1}{2\delta_k^2} \sum_{\mathbf{1}\{t_j=k\}} \beta_j^2 \right) q(\delta_k).$$

We take here a conjugate Inverse-Gamma prior distribution for δ_k with a shape parameter α_k and a scale parameter γ_k. The resulting posterior is Inverse-Gamma with parameters $\alpha_k' = \alpha_k + N_k/2$ and $\gamma_k' = \gamma_k + \sum_{\mathbf{1}\{t_j=k\}} \beta_j^2/2$. An alternative prior distribution here would be the Jeffreys improper prior $q(\delta_k^2) \propto 1/\delta_k^2$. Next, a standard $Beta(\alpha_\phi, \gamma_\phi)$ prior is taken for ϕ/p so the resulting posterior is $Beta(\alpha_\phi + N_1, \gamma_\phi + N_2)$. We note here that the values taken for $\alpha_1, \alpha_2, \alpha_\phi, \gamma_1, \gamma_2$ and γ_ϕ should be on the background of Assumptions (P2)–(P3). These are working priors that reflect our assumptions on β. The dimension of β grows with n, while its ℓ_2 norm is assumed to be constant. Therefore, ϕ, δ_1^2 and δ_2^2 must change with n. This results in working priors that also change with n.

Sampling from the full conditionals of β and T remains the same as described in Section 3, and is generally the same regardless of f_ϵ, as long as the scale mixture of normal distributions representation (3) is used. The same property holds for the sampling from the full conditionals of δ_1, δ_2 and ϕ as described above.

Moving to σ_i^2, substituting $q(\sigma^2|\theta) \propto \exp(-\theta^2\sigma^2/2)$ in (4), we have

$$q(\sigma_i^2|Y_i, X_i, \beta, \theta) \propto \sigma^{-1} \exp\left(-\frac{(Y_i - X_i^T\beta)^2 + \theta^2\sigma_i^4}{2\sigma_i^2} \right).$$

Similarly to Park and Casella (2008), this implies an Inverse-Gaussian distribution for σ_i^{-2} with parameters $a = \theta^2$ and $b = \theta/|Y_i - X_i^T\beta|$, where the Inverse-Gaussian distribution defined as having the density

$$f(t) = \sqrt{\frac{a}{2\pi}} t^{-3/2} \exp\left[-\frac{a(t-b)^2}{2b^2t} \right] \mathbf{1}\{t > 0\}$$

Chhikara (1988). The full conditionals we just described are used in the proposed Gibbs sampler. We now turn to the presentation of simulation study results.

5.1. Simulation study

We present simulation results for various κ values and for $n = 500, 2500$. We used 500 and 1000 iterations for the Gibbs sampler described above. Data were simulated in the following way. First, X was simulated under $\Sigma = I$, i.e., each of the iid rows of X was simulated as $N_p(0, I)$. Then, we simulated ϕ, δ_1 and δ_2 according to the prior distributions described above. That is, a Beta distribution for ϕ/p, and an inverse-Gamma distribution for δ_1^2 and δ_2^2. We took the following hyper-hyperparameter values: $\{\alpha_\phi = 30, \gamma_\phi = 30(3\kappa\log(n) - 1)\}, \{\alpha_1 = 2, \gamma_1 = \log(n)/n\}$ and $\{\alpha_2 = 2, \gamma_2 = n^{-1.5}\}$. These values were chosen so $E(\phi) = n/(3\log(n)), E(\delta_1^2) = \log(n)/n, E(\delta_2^2) = n^{-1.5}$ and also to have prior distributions that are not too concentrated around their means. Given ϕ, T was

simulated as iid $Ber(\phi/p)$. Then, we simulated β as independent normally distributed variables where the prior variance of each β_j is $\delta_{t_j}^2$. Next, we simulate θ from an Exponential prior $q(\theta) = \exp(-\theta)\mathbf{1}\{\theta > 0\}$, and simulate ϵ as iid random variables that follow Laplace distribution with parameter θ. Finally, the observed data is $\{X, Y\}$, with $Y = X\beta + \epsilon$.

We used $\|\hat{\beta} - \beta\|^2$ as a performance measure of the different estimators. We therefore considered the *a posteriori* mean as the Bayesian estimator. We compared this estimator to the classical M-estimators, namely least squares and least absolute deviations estimators. We also considered standard regularization-based estimators. Let the objective function to be minimized written as

$$f(\beta) = \sum_{i=1}^{n} \rho(Y_i - X_i^T \beta) + P_\lambda(\beta)$$

where P_λ is a regularization function and λ is a tuning constant chosen here by cross-validation. We considered the following four estimators, defined by the choice of ρ and P_λ: Least squares Lasso ($\rho(x) = x^2, P_\lambda(\beta) = \|\beta\|_1$), least squares Ridge regression ($\rho(x) = x^2, P_\lambda(\beta) = \|\beta\|^2$), least absolute deviation Lasso ($\rho(x) = |x|, P_\lambda(\beta) = \|\beta\|_1$) and least absolute deviation Ridge regression ($\rho(x) = |x|, P_\lambda(\beta) = \|\beta\|^2$). The first two estimators were obtained using the algorithm described in Friedman, Hastie and Tibshirani (2010) and the latter two estimators were calculated using the algorithm described in Yi and Huang (2015) (`hqreg` package in **R**).

Table 1 presents medians of $\|\hat{\beta} - \beta\|^2$ across 1000 simulations for the estimators we just described. We report the medians and not the means due to a small number of outliers, observed mainly for the regularization based methods. For each method, the ℓ_2 error grew with κ (or with p) for a fixed n, as one may have expected. The superiority of the proposed Bayes estimator is clearly shown. For $n = 500$ taking 1000 iterations of the Gibbs sampler resulted only in minor gain comparing to using the Gibbs sampler with 500 iterations. For $n = 2500$, taking 1000 iterations was essential to improve this estimator's performance. In Table 1 of the supplementary materials (Nevo and Ritov, 2016), we compared these estimators using $\|\hat{\beta} - \beta\|_1$ as our criteria. The *a posteriori* mean and the Lasso were comparable, and preferable over the other estimators. We comment that if the goal is to minimize the $\|\cdot\|_1$ loss function, the *a posteriori* median should be used.

Considering Proposition 1, the relatively poor performance of the non regularized M-estimators was expected. As expected, for $\kappa > 0.3$ the least absolute deviations (LAD) estimator was preferable over the least squares (LS) estimator. Among the regularization based methods, however, taking the least absolute deviation as a penalty function remained superior for all κ values, both for Lasso and Ridge regression. Comparing the Lasso and Ridge regression, the former showed better performance for a fixed penalty function.

To explore the robustness of the proposed Bayesian estimator, we performed an additional study where the hyperparamters $\phi, \delta_1^2, \delta_2^2$ and θ were fixed in the data creation mechanism and did not follow the assumed priors. The results,

Medians values of $\|\hat{\beta} - \beta\|^2$ across 1000 simulations for the proposed Bayes estimator using 500 and 1000 iterations of the Gibbs sampler (Bayes$_{500}$ and Bayes$_{1000}$), the Lasso with square (LAS$_{LS}$) and absolute (LAS$_{LAD}$) deviations penalty functions, Ridge regression with square (RID$_{LS}$) and absolute (RID$_{LAD}$) deviations penalty functions and the standard least squares (LS) and least absolute deviations (LAD) estimators

κ	Bayes$_{500}$	Bayes$_{1000}$	LAS$_{LS}$	LAS$_{LAD}$	RID$_{LS}$	RID$_{LAD}$	LS	LAD
				$n = 500$				
0.1	0.080	0.083	0.119	0.103	0.115	0.099	0.328	0.252
0.2	0.104	0.095	0.127	0.115	0.136	0.128	0.643	0.584
0.3	0.118	0.111	0.145	0.130	0.163	0.154	1.192	1.216
0.4	0.129	0.129	0.165	0.150	0.175	0.168	1.955	2.193
0.5	0.141	0.137	0.173	0.161	0.195	0.187	2.877	3.367
0.6	0.139	0.148	0.178	0.166	0.197	0.188	4.572	5.449
0.7	0.151	0.147	0.181	0.166	0.199	0.194	7.319	9.139
0.8	0.154	0.155	0.191	0.178	0.213	0.207	11.830	16.177
0.9	0.162	0.155	0.192	0.177	0.211	0.206	24.849	31.944
0.95	0.169	0.168	0.203	0.190	0.228	0.222	58.334	73.344
				$n = 2500$				
0.1	0.077	0.073	0.098	0.081	0.100	0.087	0.301	0.231
0.2	0.099	0.097	0.124	0.108	0.141	0.129	0.722	0.647
0.3	0.120	0.104	0.128	0.112	0.157	0.146	1.154	1.165
0.4	0.127	0.120	0.145	0.128	0.171	0.163	1.985	2.119
0.5	0.143	0.119	0.139	0.126	0.167	0.161	3.056	3.569
0.6	0.152	0.132	0.152	0.141	0.183	0.178	4.245	5.142
0.7	0.155	0.138	0.155	0.142	0.184	0.178	6.737	8.396
0.8	0.172	0.146	0.162	0.149	0.193	0.191	12.157	15.799
0.9	0.166	0.150	0.165	0.153	0.195	0.188	26.930	35.734
0.95	0.180	0.157	0.171	0.156	0.208	0.204	58.033	72.705

presented in Table 2 of the supplementary materials, show that the Bayesian estimator remains favorable in this scenario, even though the Gibbs sampler assumes the aforementioned prior distributions.

Bayesian methods are often time consuming, especially when the target parameter is high dimensional and sampling from full conditionals is performed. The Gibbs sampler presented in this section involve direct sampling for all the parameters, with block sampling for β, without using additional MCMC steps. Table 2 compares median computation time in minutes between the different methods. Regularized least squares penalty function methods were considerably faster that the alternatives (excluding simple LS and LAD). However, as Table 1 suggests, they were also inferior to the other methods. The suggested Bayesian Gibbs sampler was comparable, and often preferable, in terms of computation time to the regularized least absolute deviations estimators.

6. Discussion

This paper provided a Bayesian alternative to frequentist robust regression when the number of predictors and the sample size are of the same order. Standard M-estimators are inconsistent when considering the error accumulated over the vector. If it is further assumed that the signal and the noise are of the same

TABLE 2

Medians computation times (minutes) across 1000 simulations for the proposed Bayes estimator using 1000 iterations of the Gibbs sampler ($Bayes_{1000}$), the Lasso with square (LAS_{LS}) and absolute (LAS_{LAD}) deviations penalty functions, Ridge regression with square (RID_{LS}) and absolute deviations (RID_{LAD}) penalty functions and the standard least squares (LS) and least absolute deviation estimators (LAD)

κ	$Bayes_{1000}$	LAS_{LS}	LAS_{LAD}	RID_{LS}	RID_{LAD}	LS	LAD
			$n = 500$				
0.2	1.4	0.0	0.6	0.0	0.7	0.0	0.0
0.4	2.2	0.0	3.7	0.0	4.5	0.0	0.0
0.6	2.8	0.0	6.9	0.0	6.7	0.0	0.0
0.8	5.9	0.1	20.7	0.1	18.7	0.0	0.0
0.9	7.6	0.2	28.2	0.1	22.8	0.0	0.0
0.95	8.5	0.1	34.2	0.1	26.7	0.0	0.0
			$n = 2500$				
0.2	12.2	0.2	9.2	0.3	11.0	0.0	0.6
0.4	74.6	0.7	73.8	0.9	91.8	0.1	3.0
0.6	39.7	0.5	56.4	0.5	65.3	0.1	2.5
0.8	98.2	1.5	141.2	1.0	156.2	0.2	3.9
0.9	137.8	3.0	190.9	1.2	205.9	0.2	4.2
0.95	396.6	2.6	353.4	1.4	385.8	0.6	4.6

asymptotic order, then shrinkage of many coefficients is desirable. We presented an hierarchical prior for model parameters and constructed a Bayes estimator suitable for the problem. A scale mixture of normal distribution representation for the errors' distribution allowed us to build an efficient Gibbs sampler with block sampling for the coefficients. Theorem 1 shows that under appropriate conditions, the Bayes estimator in this problem is consistent in the ℓ_2 sense. This property does not hold for M-estimators in this design.

The Bayesian estimator also outperforms regularization based methods. Although these methods, at least for the quadratic penalty function, can be computed much faster, they provide point estimate only, where the Gibbs sampler provides the entire posterior distribution that can be used to inference. This issue was also pointed out in Park and Casella (2008). It was previously shown that the asymptotic distribution of M-estimators in this design is nontrivial (El Karoui et al., 2013; Donoho and Montanari, 2013; El Karoui, 2013). In El Karoui (2013), the distribution of the M-estimator was studied as the limit of a Ridge regularized estimators. As standard asymptotic theory does not apply here, the asymptotic distribution of the Bayes estimator in this regime remains as a future challenge.

7. Proofs

7.1. Proof of Proposition 1

First, by Lemma 1 in El Karoui et al. (2013) we can write

$$\hat{\beta}^\rho - \beta^0 \overset{\mathcal{D}}{=} \|\hat{\beta}^{\rho,simp}\|\Sigma^{-1/2}u$$

where $\stackrel{\mathcal{D}}{=}$ denotes equality in distribution, and where u is a p-length vector distributed uniformly on the sphere of radius one and where

$$\hat{\beta}^{\rho,simp} = \arg\min_{\beta} \sum_{i=1}^{n} \rho(\epsilon_i - \widetilde{X}_i^T \beta).$$

with $\widetilde{X}_i^T = \Sigma^{-1/2} X_i$ being a mean zero multivariate normal vector with the identity matrix as its variance. Their lemma also asserts that $\|\hat{\beta}^{\rho,simp}\|$ and u are independent. By Result 1 in El Karoui et al. (2013), $\|\hat{\beta}^{\rho,simp}\|$ has a deterministic limit denoted here (and there) by $r_\rho(\kappa)$. They further show how $r_\rho(\kappa)$ can be found. Next, by a multivariate central limit theorem, for large enough p, $\sqrt{p}\Sigma^{-1/2}u$ is approximately $N_p(0, \Sigma^{-1})$. Thus, for large enough n we have

$$\|\sqrt{n}(\hat{\beta}^{\rho} - \beta^0)\| \stackrel{\mathcal{D}}{=} r_\rho^2(\kappa)\kappa^{-1}\|x\|^2 + o_p(1) \tag{5}$$

where $x \sim N_p(0, \Sigma^{-1})$. Let $\Sigma = T\Lambda T^T$ be the spectral decomposition of Σ. For $v \sim N_p(0, I)$ and $v \stackrel{\mathcal{D}}{=} v^*$, we have

$$\|x\|^2 = v^T T\Lambda^{-1}T^T v = (v^*)^T \Lambda^{-1}v^* \geq^{st.} \lambda_{min}^{-1}\chi_p^2$$

where $\geq^{st.}$ symbolizes larger in the stochastic sense, λ_{min} is the minimal eigenvalue of Σ, and χ_p^2 is a Chi-squared random variable with p degrees of freedom. Therefore, the right hand side of (5) is $O_p(n)$ and we are done.

7.2. Lemma 1: Presentation and proof

Denote $B_n^{C,\eta} := \frac{1}{p}\sum_j \mathbf{1}\left\{|\beta_j| > Cn^{-\eta/2}\right\}$ for the proportion of coordinates of β that are of order larger than $Cn^{-\eta/2}$. The following lemma ensures us that if the prior distribution of β admits Assumptions (P1)-(P3) then $B_n^{C,\eta}$ is not far from ϕ/p, and consequently, $B_n^{C,\eta} = O_p(\phi/p) = o_p(log(n)^{-1})$. Thus, the proportion of "large" coefficient values out of the total number of coefficients goes to zero with probability that goes to one as n grows.

Lemma 1. *Let assumptions* (M1), (P1) *and* (P2) *hold. Assume* (P3) *holds with some ξ. Then, for all $1 \leq \eta < \xi$, for all $\zeta > 0$ and for any constant C we have*

$$\lim_{n\to\infty} P\left(\left|B_n^{C,\eta} - \frac{\phi}{p}\right| > \frac{\phi}{p}\zeta\right) = 0$$

Proof. First note that $B_n^{C,\eta}$ is a mean of p independent Bernoulli random variables with success probability of

$$\nu_n^{C,\eta} = \frac{\phi}{p}2\left[\Phi\left(-\frac{C}{\sqrt{n^\eta\delta_1^2}}\right) + \Phi\left(-\frac{C}{\sqrt{n^\eta\delta_2^2}}\right)\right] + 2\Phi\left(-\frac{C}{\sqrt{n^\eta\delta_2^2}}\right)$$

$$= \frac{\phi}{p}2\Phi\left(-\frac{C}{\sqrt{n^\eta\delta_1^2}}\right) + o\left(\frac{\phi}{p}\right).$$

with Φ being the CDF of a standard normal random variable. The second equality results from $n^\eta \delta_2^2 = n^{\eta-\xi}$ (Assumption (P3)) and since $\eta < \xi$. Now, this problem is symmetric, so it is suffice to show

$$\lim_{n\to\infty} P\Big(B_n^{C,\eta} > \frac{\phi}{p}(1+\zeta)\Big) = 0.$$

By Chebyshev's inequality we have

$$P\Big(B_n^{C,\eta} > \frac{\phi}{p}(1+\zeta)\Big) \leq P\Big(|B_n^{C,\eta} - \nu_n^{C,\eta}| > \frac{\phi}{p}(1+\zeta) - \nu_n^{C,\eta}\Big)$$

$$\leq \frac{\nu_n^{C,\eta}}{p(\frac{\phi}{p}(1+\zeta) - \nu_n^{C,\eta})^2}$$

$$= \frac{\frac{\phi}{p}2\Phi\Big(-\frac{C}{\sqrt{n^\eta \delta_1^2}}\Big) + o\Big(\frac{\phi}{p}\Big)}{\frac{\phi^2}{p}\Big(1 + \zeta - 2\Phi\Big(-\frac{C}{\sqrt{n^\eta \delta_1^2}}\Big)\Big)^2 + o\Big(\frac{\phi^2}{p}\Big)}$$

and the last expression goes to zero as $n \to \infty$. □

7.3. Proof of Theorem 1

For the simplicity of the proof we will assume that $\Sigma = I$.

Let $\hat{\beta}^{MAP} = \arg\max_\beta q(\beta|Y, X)$ be the MAP estimator. First we will show that $\|\hat{\beta}^{MAP} - \beta^0\| = o_p(1)$, and then that $\|\hat{\beta}^{MAP} - \hat{\beta}^\star\| = o_p(1)$. We

Starting with $\hat{\beta}^{MAP}$, will first show the weaker result $\|\hat{\beta}^{MAP} - \beta^0\| = O_p(1)$. We will then build on this result to strengthen the conclusion. For any vector β, we partition β to two subvectors $\beta_\mathcal{M}$ and $\beta_{\mathcal{M}^c}$, were M is the subset of relatively large coefficients. We will then use the fact that $\|\hat{\beta}_\mathcal{M}^{MAP} - \beta_\mathcal{M}^0\| = O_p(1)$ to argue that $\|\hat{\beta}_{\mathcal{M}^c}^{MAP} - \beta_{\mathcal{M}^c}^0\| = o_p(1)$, and the latter to argue that $\|\hat{\beta}_\mathcal{M}^{MAP} - \beta_\mathcal{M}^0\| = o_p(1)$. Now, for the details.

By Taylor expansion, the log-posterior of β can be written as

$$\log q(\beta|Y, X)$$

$$= \sum_{i=1}^n \ell_\epsilon(Y_i - X_i^T\beta) + \sum_{j=1}^p \log(q(\beta_j))$$

$$= \sum_{i=1}^n \ell_\epsilon\big(\epsilon_i - X_i^T(\beta - \beta^0)\big) + \sum_{j=1}^p \log(q(\beta_j)) \tag{6}$$

$$= \sum_{i=1}^n \ell_\epsilon(\epsilon_i) + \sum_{j=1}^p \log(q(\beta_j)) - (\beta - \beta^0)^T \sum_{i=1}^n X_i \ell_\epsilon'(\epsilon_i)$$

$$+ \frac{1}{2}(\beta - \beta^0)^T \sum_{i=1}^n \ell_\epsilon''\big(\epsilon_i + a_\beta X_i^T(\beta - \beta^0)\big) X_i X_i^T(\beta - \beta^0), \tag{7}$$

for some $\alpha_\beta \in [0, 1]$.

Since the empirical distribution function converges to the cumulative distribution function there are $M < \infty$ and $\gamma < 1$ such that

$$P(\sum \mathbf{1}(|\epsilon_i| > M) < \gamma n) \to 1. \tag{8}$$

Next we want to argue that we also have

$$P\big(\sum \mathbf{1}(|\epsilon_i - X_i^T(\hat{\beta}^{MAP} - \beta^0)| > M) < \gamma n\big) \to 1. \tag{9}$$

Let $U_n \approx V_n$ if $U_n = O_p(V_n)$ and $V_n = O_p(U_n)$. Now

$$
\begin{aligned}
E|\ell(\epsilon) - \ell(0)| &= 2\int_0^\infty |\ell(x) - \ell(0)| e^{\ell(x)} dx \\
&= 2\int_0^\infty x|\ell'(\alpha_x x)| e^{\ell(x)} dx \\
&\leq 2\int_0^\infty x|\ell'(x)| e^{\ell(x)} dx \\
&= 2\int_0^\infty e^{\ell(x)} dx = 2.
\end{aligned}
\tag{10}
$$

Hence $n^{-1} \sum_{i=1}^n \ell(\epsilon_i) \xrightarrow{p} E\ell(\epsilon) > -\infty$. Thus, $\sum \ell(\epsilon_i) \approx n$. Since $\hat{\beta}^{MAP}$ is the maximizer we have that $q(\hat{\beta}^{MAP}|Y, X) - q(\beta^0|Y, X) > 0$. Consider this difference and recall also line (6). Consider first the difference in the prior contributions. If $t_j = 2$ (see Assumption (P1)) then the contribution of the prior can be made larger by making β_j smaller, but then $|\beta_j - \beta_j^0| = O_p(n^{-\xi/2})$. If $t_j = 1$ we can increase the prior contribution, but there are only $O_p(n/\log n)$ such terms. Thus, $|\sum_{j=1}^p \log(q(\beta_j^{MAP})) - \sum_{j=1}^p \log(q(\beta_j^0))| = O_p(n/\log n)$. Since $\hat{\beta}^{MAP}$ is improving over β^0 and $\sum_i^n \ell(\epsilon_i) \approx n$ we must have that for some $A > 0$, $\sum \ell(\epsilon_i - X_i^T(\hat{\beta}^{MAP} - \beta^0)) > -An$ with high probability. Since $\lim_{x\to\infty} = -\infty$, we can choose a finite M such that both (8) and (9) hold. Let

$$\mathcal{A} = \{i : |\epsilon_i| < M \text{ and } |\epsilon_i - X_i^T(\hat{\beta}^{MAP} - \beta^0)| < M\}.$$

We conclude from (8) and (9) that for large enough M, $|\mathcal{A}|$ is the same order as n. Hence

$$\sum_{i=1}^n \ell''_\epsilon\big(\epsilon_i + \alpha_{\hat{\beta}^{MAP}} X_i^T(\hat{\beta}^{MAP} - \beta^0))\big) X_i X_i^T \leq \sup_{|t| < M} \ell''_\epsilon(t) \sum_{\mathcal{A}} X_i X_i^T \tag{11}$$

in the partial order of positive semi-definite matrices. We conclude from the prior, (7), and we can verify by (11) that

$$\|\hat{\beta}^{MAP} - \beta^0\| = O_p(\sqrt{p/n}) = O_p(1). \tag{12}$$

We now build on this result to strengthen the conclusion. Let $\mathcal{M} \subseteq \{1, \ldots, p\}$ be the set of indices such that $|\beta_j^0| > n^{-\xi} \log n$. Denote by $\beta_{\mathcal{M}}^0$ and $\hat{\beta}_{\mathcal{M}}^{MAP}$ the

subvectors with indices in $\mathcal{M}$ of β^0 and $\hat{\beta}^{MAP}$, respectively. Let $\mathcal{M}^c$ the complementary set and define the corresponding subvectors and submatrix similarly.

We start with estimation error of $\hat{\beta}_{\mathcal{M}^c}^{MAP}$. Let $R = \{\beta : \|\beta_{\mathcal{M}^c} - \beta_{\mathcal{M}^c}^0\| = n^{-\gamma}$ and $\beta_{\mathcal{M}} = \hat{\beta}_{\mathcal{M}}^{MAP}\}$, where $0 < \gamma < \xi - 1$ and consider

$$W = \max_{\beta \in R} \log q(\beta|Y, X) - \log q(\beta^{0,MAP}|Y, X), \tag{13}$$

where $\beta^{0,MAP}$ equals to β^0 on $\mathcal{M}^c$ and to $\hat{\beta}^{MAP}$ on $\mathcal{M}$. Substituting (7) in (13), the second term on the RHS of (7) contributes to W $-n^{\xi}n^{-2\gamma}$, the third term is at most $O_p(n^{-\gamma+1})$, while the last term is by (11) and (12) at most $O_p(n^{-2\gamma+1}) + O_p(n^{-\gamma+1})$. Thus $W < 0$, and since $\hat{\beta}_{\mathcal{M}^c}^{MAP}$ is the maximizer, $\|\hat{\beta}_{\mathcal{M}^c}^{MAP} - \beta_{\mathcal{M}^c}^0\| = o_p(1)$.

This means that the error in estimation $\beta_{\mathcal{M}^c}$ has negligible contribution. The estimating equation for $\beta_{\mathcal{M}}$ with only the $t = 1$ part of the prior is a ridge regression with $|\mathcal{M}| = o_p(n)$ variables with negligible ridge penalty, thus standard mean calculations and Assumption (M1) would show that $\|\hat{\beta}_{\mathcal{M}}^{MAP} - \beta_{\mathcal{M}}^0\| = o_p(1)$. By Assumptions (P1)–(P3) the $t = 2$ part has a relative peak at 0 of height of the ratio of the two standard deviations (Assumption (P3)) and the ratio of the mixing probabilities (Assumption (P1)). It contributes the log of the product of these two terms, $O(\log n)$, but only for components which are in an $O(n^{-\xi/2})$ of 0. But the curvature of the likelihood is $O_p(n)$ and the corresponding components of β^0 are $O_p(\phi^{-1/2})$, thus the gain in the prior cannot balance the loss in the likelihood which is of order $n\phi^{-1}$ per component which is this neighborhood of 0.

Having established that the MAP estimator is in $o(1)$ neighborhood of the truth, we consider now $\hat{\beta}^*$. Again, consider the partition to $\beta_{\mathcal{M}}$ and $\beta_{\mathcal{M}^c}$. The log-prior of any β_j is

$$\log\Big(\frac{c_1^{1/2}\phi^{3/2}}{p}e^{-c_1\phi\beta_j^2} + c_2^{1/2}n^{\xi/2}e^{-c_2 n^{\xi}\beta_j^2}\Big),$$

where $c_1, c_2 = O(1)$. Thus, for $|\hat{\beta}_j^{MAP}| < L_n n^{-\xi/2}$, for $L_n \to \infty$ slowly, the central component of the prior dominates, and the *a posteriori* mass of this ball is negligible. For $\hat{\beta}_j^{MAP} \approx \phi^{-1/2}$, we argued that the peak at 0 is much smaller than the value at $\hat{\beta}_j^{MAP}$. Since the radius of this peak is of order $n^{-\xi}$. It can be ignored. Ignoring this neighborhood of 0, the *a posteriori* is strictly concave with maximum at $\hat{\beta}_j^{MAP}$, and curvature of order n and hence $\hat{\beta}_j^* - \hat{\beta}_j^{MAP} = O_p(n^{-1/2})$. $\qquad\square$

Acknowledgements

Supplementary Material

Supplementary materials for "On Bayesian robust regression with diverging number of predictors"
(doi: [10.1214/16-EJS1205SUPP](); .pdf).

References

ANDREWS, D. F. and MALLOWS, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B (Methodological)* 99–102. MR0359122

BEAN, D., BICKEL, P. J., EL KAROUI, N. and YU, B. (2013). Optimal M-estimation in high-dimensional regression. *Proceedings of the National Academy of Sciences* **110** 14563–14568.

CARVALHO, C. M., POLSON, N. G. and SCOTT, J. G. (2009). Handling sparsity via the horseshoe. In *International Conference on Artificial Intelligence and Statistics* 73–80.

CARVALHO, C. M., POLSON, N. G. and SCOTT, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97** 465–480. MR2650751

CASTILLO, I., SCHMIDT-HIEBER, J. and VAN DER VAART, A. (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics* **43** 1986–2018. MR3375874

CASTILLO, I., VAN DER VAART, A. et al. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics* **40** 2069–2101. MR3059077

CHHIKARA, R. (1988). *The Inverse Gaussian Distribution: Theory: Methodology, and Applications* **95**. CRC Press.

DONOHO, D. and MONTANARI, A. (2013). High dimensional robust M-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields* 1–35.

EFRON, B. and MORRIS, C. (1973). Stein's estimation rule and its competitors—an empirical Bayes approach. *Journal of the American Statistical Association* **68** 117–130. MR0388597

EL KAROUI, N. (2013). Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: Rigorous results. *arXiv preprint arXiv:1311.2445*.

EL KAROUI, N., BEAN, D., BICKEL, P. J., LIM, C. and YU, B. (2013). On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences* **110** 14557–14562.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33** 1.

GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **6** 721–741.

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

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica* 339–373.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Annals of Statistics* 799–821. MR0356373

Huber, P. J. (2011). *Robust Statistics.* Springer.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium On Mathematical Statistics and Probability* **1** 361–379. MR0133191

Maronna, R. A. and Yohai, V. J. (1981). Asymptotic behavior of general M-estimates for regression and scale with random carriers. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **58** 7–20. MR0635268

Nevo, D. and Ritov, Y. (2016). Supplementary materials for "On Bayesian robust regression with diverging number of predictors". DOI: 10.1214/16-EJS1205SUPP.

Park, T. and Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association* **103** 681–686. MR2524001

Portnoy, S. (1984). Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I. Consistency. *Annals of Statistics* 1298–1309. MR0760690

Portnoy, S. (1985). Asymptotic behavior of M estimators of p regression parameters when p^2/n is large; II. Normal approximation. *Annals of Statistics* 1403–1417. MR0811499

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288. MR1379242

van der Pas, S., Kleijn, B. and van der Vaart, A. (2014). The horseshoe estimator: Posterior concentration around nearly black vectors. *Electronic Journal of Statistics* **8** 2585–2618. MR3285877

West, M. (1987). On scale mixtures of normal distributions. *Biometrika* **74** 646–648. MR0909372

Yi, C. and Huang, J. (2015). Semismooth Newton Coordinate Descent Algorithm for Elastic-Net Penalized Huber Loss and Quantile Regression. *arXiv preprint arXiv:1509.02957.*