Statistica Sinica **16**(2006), 411-424

REGULARIZED OPTIMIZATION IN STATISTICAL LEARNING: A BAYESIAN PERSPECTIVE

Bin Li and Prem K. Goel

The Ohio State University

Abstract: Regularization plays a major role in modern data analysis, whenever non-regularized fitting is likely to lead to over-fitted model. It is known that most regularized optimization problems have Bayesian interpretation in which the prior plays the role of the regularizer. In this paper, we consider the issue of sensitivity of the regularized solution to the prior specification within the Bayesian perspective. We suggest a class of flat-tailed priors for a general likelihood function for robust Bayesian solutions, in the same spirit as the *t-distribution being suggested as a flat-tail prior for normal likelihood.* Results are applied to a family of regularized learning methods and group LASSO. In addition, the consistency issue for LASSO is discussed within this framework.

Key words and phrases: Bayesian robustness, bridge regression, flat-tailed prior, group LASSO, LASSO, regularized optimization.

1. Introduction

Regularized optimization, which plays an important role in both statistical and machine learning problems, can often be described as:

$$\hat{\beta}(\lambda) = arg \min_{\beta} L(Y, X\beta) + \lambda J(\beta), \tag{1.1}$$

where $L(Y, X\beta)$ is a non-negative loss (cost) function, convex in β, $J(\beta)$ is a non-negative convex penalty, and λ is the non-negative tuning parameter. Many popular methods such as ridge regression (Hoerl and Kennard (1970)), LASSO (Tibshirani (1996)), SVM (Vapnik (1998)), SCAD (Fan and Li (2001)), Group LASSO (Yuan and Lin, (2004)) and Elastic Net (Zou and Hastie (2005)) fall into this category. It is known that most of these problems have Bayesian interpretation, such that the loss function is interpreted as the negative log-likelihood, the penalty term as the negative log-prior density and the regularized solution corresponds to the global maxima of the posterior distribution. In regularized optimization, the optimal choice of the tuning parameter is usually based on some cross-validation. This setup corresponds to Empirical Bayes technique with a hierarchical prior specification, i.e., given the hyperparameter, the prior belongs to some robust family of distributions, and the data provide guidance in choosing the hyperparameter.

Generally, robust Bayesian analysis focuses on the sensitivity of Bayesian answers to the inputs in the analysis: the model, the prior distribution, the loss function, or any combinations thereof. Since a Bayesian decision maker incorporates the prior information in the study, which is often viewed as subjective, the sensitivity analysis of priors becomes the most important issue in robust Bayesian analysis.

There are three main approaches in sensitivity analysis of priors. First, and most common in practice, is *informal* sensitivity analysis, in which the analysis is repeated simply on different prior distributions and the resulting inferences compared. Second, *global* sensitivity analysis, which concerns the range of Bayesian answers as the prior varies over a certain class, see e.g., Moreno (2000). The third approach is *local* sensitivity analysis, which concerns the rate of change in inference as the input varies, see e.g., Gustafson, Srinivasan and Wasserman (1996), Gustafson (2000) and Sivaganesan (2000). The key requirement in both global and local approaches is to choose an appropriate class of prior distributions. Berger (1994) mentions four criteria in choosing a class of priors: (1) easy to elicit and interpret; (2) easy to handle computationally; (3) large enough to reflect prior uncertainty; (4) extendable to higher dimensions.

In high-dimensional problems (dimensionality corresponds to the number of parameters, p, in the model), it may not be feasible to elicit a subjective prior distribution, unless a low-dimensional structure is assumed. One way to bypass this problem is to use some inherently robust prior in the analysis, to make it easier to build in the robustness at the beginning (before observing the data) than to check the sensitivity of the inference after observing the data. Many studies (Box and Tiao (1973), Berger (1985), O'Hagan (1988), Fan and Berger (1992)) have shown that use of flat-tailed priors tends to be robust and they suggest using t-distribution for the normal likelihood. However, there seems to be no general result about the flat-tailed property in Bayesian literature. In this paper, we investigate a class of flat-tailed priors for a general likelihood function in the same spirit as the ' t-distribution suggested as a flat-tail prior for normal likelihood,' for robust Bayesian solution using squared error loss.

In this paper, Bayesian inference is based on the posterior mode, similar to that in the regularized optimization setup. Furthermore, we focus on the case where p is finite and does not depend on the sample size n. The remaining part of this paper is organized as follows. In Section 2, we suggest a criterion characterizing the tail behavior of a distributions and formalize the robustness property in terms of the relative tail behaviors of the likelihood and the priors.

In Section 3, we study the robustness properties for bridge regression, a special family of penalized regression of a penalty function, $\sum|\beta_j|^\gamma$ with $\gamma \geq 1$. The robustness property for the group LASSO solution is discussed in Section 4, followed by a discussion of the consistency issue for the LASSO solution in Section 5. Discussion and future work are presented in Section 6. The proof of Theorem 1 is postponed to Section 7.

2. Sensitivity Analysis of Flat-tailed Priors

First, a criterion that characterizes the tail behavior of a distribution is suggested. Second, two notions of Bayesian robustness are proposed. Then, the main result of Bayesian robustness is shown.

2.1. Characterizing distribution's tail behavior

A natural way to describe the tail behavior of a distribution $\pi(\theta)$ is

$$\frac{\pi(\theta+\Delta)-\pi(\theta)}{\Delta}, \tag{2.1}$$

where $\theta, \theta+\Delta \in \Theta$. The main drawback of (2.1) is that, when θ is in the tail region, (2.1) tends to be close to zero. A more meaningful criterion is to use the relative change

$$\frac{1}{\pi(\theta)} \times \frac{\pi(\theta+\Delta)-\pi(\theta)}{\Delta}. \tag{2.2}$$

If the multivariate density $\pi(\theta)$ is differentiable with respect to θ, (2.2) suggests an alternative way measure the tail behavior of a distribution by

$$\tau[\pi(\theta)] = \Big\|\frac{\partial \log \pi(\theta)}{\partial \theta}\Big\|_2. \tag{2.3}$$

Plausibly, the quantity in (2.3) should be small for a flat-tailed prior. Table 1 provides $\tau[\pi(\theta)]$ for four well-known multivariate distributions.

Table 1. $\tau[\pi(\theta)]$ for four well-known distributions.

Distribution	$\pi(\theta) \propto$	$\tau[\pi(\theta)]$
Gaussian	$\exp[-\sum_{j=1}^p \frac{\theta_j^2}{2\sigma^2}]$	$\frac{\|\|\theta\|\|_2}{\sigma^2}$
Laplacian	$\exp[-\sum_{j=1}^p \frac{\|\theta_j\|}{a}]$	$\frac{\sqrt{p}}{a}$, $\theta_j \neq 0$, $j = 1, \ldots, p$
t-distribution	$\prod_{j=1}^p [1+\frac{\theta_j^2}{\nu_j}]^{-\frac{\nu_j+1}{2}}$	$\left\{\sum_{j=1}^p \frac{(\nu_j+1)^2\theta_j^2}{(\nu_j+\theta_j^2)^2}\right\}^{\frac{1}{2}}$
Exponential power	$\exp[-\frac{1}{2}\sum_{j=1}^p \|\frac{\theta_j}{\sigma}\|^\gamma]$	$\frac{\gamma}{2\sigma^\gamma}\left\{\sum_{j=1}^p \|\theta_j\|^{2(\gamma-1)}\right\}^{\frac{1}{2}}$

From Table 1, we can see that when $||\theta||_2$ is large, the tail behaviors of these distributions in terms of $\tau[\pi(\theta)]$ are substantially different. For the normal distribution, $\tau[\pi(\theta)]$ increases as the θ goes away from the center. For the Laplacian distribution, $\tau[\pi(\theta)]$ is a constant with respect to θ. For the t-distribution, $\tau[\pi(\theta)]$ is bounded above by a constant, and will go to zero if all θ_j's go away from the origin simultaneously. The exponential power family, which includes the Gaussian and Laplacian distributions as two special cases, shows the different tail behaviors in terms of $\tau[\pi(\theta)]$ for different γ.

2.2. Two notions of Bayesian robustness

Let $Y = (y_1, \ldots, y_n)$ be the observed data and $l(\theta|Y)$ the log-likelihood function. Write $\hat{\theta}$ for the maximum likelihood estimate (MLE), θ^* for the mode of the prior distribution $\pi(\theta)$, $\tilde{\theta}$ for the posterior mode, and let $||\hat{\theta} - \theta^*||_2 = C$. Note that, whenever the prior and/or the posterior distributions are multimodal, θ^* and $\tilde{\theta}$ denote the global maxima of the prior and the posterior respectively.

In this paper the Bayesian inference is based on $\tilde{\theta}$, and its robustness property is examined with respect to the effect of the nominal prior. For any prior distribution, when C is small, $\tilde{\theta}$ is a compromise between $\hat{\theta}$ and θ^*, and data adaptiveness is not a critical issue. However, when C is large, the nominal prior is not compatible with the observed data, and we wish the posterior mode to be data-adaptive for moderate or large sample size, mainly because the data generating mechanism, or the likelihood, is assumed to be correct. This data-adaptive property can be measured by the distance between $\tilde{\theta}$ and $\hat{\theta}$, e.g., $||\hat{\theta} - \tilde{\theta}||_2$, as well as by the relative distance between $\tilde{\theta}$ and $\hat{\theta}$, compared to the distance between $\tilde{\theta}$ and θ^*, i.e., the ratio

$$r(\tilde{\theta}, \theta^*, \hat{\theta}) = \frac{||\hat{\theta} - \tilde{\theta}||_2}{||\theta^* - \tilde{\theta}||_2}. \tag{2.4}$$

Based on these two measures for data-adaptiveness, we define the notions of weak and strong Bayesian robustness as follows.

Definitions: Weak and Strong Bayesian Robustness

Weak Bayesian Robustness: the relative distance $r(\tilde{\theta}, \theta^*, \hat{\theta}) \to 0$ as $C \to \infty$.

Strong Bayesian Robustness: the distance between $\tilde{\theta}$ and $\hat{\theta}$, $||\hat{\theta} - \tilde{\theta}||_2$, remains bounded, no matter how large C is.

By the triangle inequality for the L_2 norm, it is easy to see that strong Bayesian robustness implies weak Bayesian robustness, while the reverse is not true. For weak Bayesian robustness, $||\hat{\theta} - \tilde{\theta}||_2$ can still go to infinity as $C \to \infty$, even though the relative distance $r(\tilde{\theta}, \theta^*, \hat{\theta})$ goes to zero. Therefore, strong

Bayesian robustness is preferable vis-*à*-vis its stronger data-adaptive property, but it requires more 'flat-tailedness' in the prior.

2.3. Bayesian robustness

Our main results of Bayesian robustness are given in Theorem 1, under the following assumptions.

Assumption 1. The support of $l(\theta|Y)$ and $\pi(\theta)$ is R^p and both of them are everywhere differentiable with respect to θ. The MLE $\hat{\theta}$ is unique and is an interior point of R^p.

Assumption 2A. There exists a $P > 0$, such that for some positive constants A_2,

$$\left\|\frac{\partial l(\theta|Y)}{\partial\theta}\right\|_2 \leq A_2||\theta-\hat{\theta}||_2^P \quad \forall\,\theta\in\mathcal{R}^p. \tag{2.5}$$

Assumption 2B. There exists a $P > 0$, such that for some positive constants A_1,

$$A_1||\theta-\hat{\theta}||_2^P \leq \left\|\frac{\partial l(\theta|Y)}{\partial\theta}\right\|_2 \quad \forall\,\theta\in\mathcal{R}^p. \tag{2.6}$$

Given that for some $P > 0$, the log-likelihood function $l(\theta|Y)$ has tail behavior according to Assumption 2, we define three classes of priors characterized by different tail behaviors relative to the log-likelihood tails. Generally, the priors in the first class have the 'least flat' tails, while priors in the third class have the 'most flat' tails.

Definitions: Prior tail-behavior relative to the likelihood tail

The prior distribution $\pi(\theta)$, with $\tau[\pi(\theta)]$ defined in (2.3), is said to belong to Class 1, 2 and 3, respectively, according to the following criteria.

Class 1. There exists a $Q \geq P$, such that for some constant B_1,

$$B_1\,||\theta-\theta^*||_2^Q \leq \tau[\pi(\theta)], \quad \forall\,\theta\in\mathcal{R}^p. \tag{2.7}$$

Class 2. There exists a Q, $0 < Q < P$, such that for some constant B_2,

$$\tau[\pi(\theta)] \leq B_2\,||\theta-\theta^*||_2^Q\,, \quad \forall\,\theta\in\mathcal{R}^p, \tag{2.8}$$

$$\tau[\pi(\theta)] \to \infty \quad \text{as} \quad ||\theta-\theta^*||_2 \to \infty. \tag{2.9}$$

Class 3. $\tau[\pi(\theta)]$ is bounded above by a constant:

$$\tau[\pi(\theta)] \leq B_3 \quad \forall\,\theta\in\mathcal{R}^p. \tag{2.10}$$

Theorem 1. *Under Assumption* 1 *and the weak and strong Bayesian robustness defined earlier, the inference based on the prior $\pi(\theta)$ has the following properties.*

1. *If $l(\theta|Y)$ satisfies Assumption 2A and $\pi(\theta)$ is in Class 1, the weak Bayesian robustness property does not hold.*
2. *If $l(\theta|Y)$ satisfies Assumption 2B and $\pi(\theta)$ is in Class 2, the weak Bayesian robustness property holds, but the strong property does not hold.*
3. *If $l(\theta|Y)$ satisfies Assumption 2B and $\pi(\theta)$ is in Class 3, the strong Bayesian robustness property holds.*

Robustness properties in Theorem 1 hold for all sample sizes n and dimensions p. In fact, using a flat-tailed prior is motivated by choosing some inherently robust prior class in advance, so a robustness property that does not depend on n or p is preferred. However, in Section 5, the effect of sample size n on the upper bound for $||\tilde{\theta} - \hat{\theta}||_2$ is utilized to discuss the consistency properties of the LASSO solution.

In this paper, the definitions of weak/strong Bayesian robustness and Assumption 2 are based on the L_2 norm. However, the only property of the L_2 norm used in the proof of Theorem 1 is the triangle inequality. Therefore, the results of Theorem 1 can be generalized to other proper norms.

The motivation for the regularized optimization in high-dimensional problems is to avoid overfitting to noisy data by restricting the solution to a subset of the original space. Different penalty function $J(\beta)$ in (1.1) corresponds to different subsets, and choosing the tuning parameter λ corresponds to selecting the regularized solution in that subset. From the Bayesian perspective, choosing the subset as our solution space is simply selecting the class of prior distributions. The more prior information (constraints) we have, the smaller the subset we focus on in which to find our solution. Intuitively, a prior with strong/weak Bayesian robustness property can be described as follows: *no matter what the unregularized solution, the regularized solution in that subspace will not be too far away from it in a specified sense.*

Consider the linear regression model with p explanatory variables:

$$Y = X\beta + \epsilon, \quad \epsilon \sim N(0, I), \tag{2.11}$$

where X is a $n \times p$ matrix with full column rank. Let ν_1 and ν_2 be the smallest and largest eigenvalues of $X^T X$, $\hat{\beta}^o$ be the ordinary least square (OLS) solution. Then we have

$$\nu_1 \|\beta - \hat{\beta}^o\|_2 \leq \left\| \frac{\partial l(\beta|Y)}{\partial \beta} \right\|_2 \leq \nu_2 \|\beta - \hat{\beta}^o\|_2, \tag{2.12}$$

which indicates that for the normal likelihood case, both Assumption 2A and 2B hold with P equal to 1. Suppose the prior distribution for β is multivariate independent t. From Table 1, we see that the Student t-prior belongs to Class 3 and Assumption 1 is trivial to check. Thus, by Theorem 1, we have the following result.

Corollary 1. *For the linear regression model in* (2.11), *suppose X has full column rank. The inference based on the multivariate independent t-prior distribution has the strong Bayesian robustness property.*

Sometimes, in order to achieve a sparse solution, the penalty term, $J(\beta)$, is not differentiable at the origin and Assumption 1 is not satisfied. However, in some cases, we can still investigate the robustness properties of the estimates, via arguments similar to those used in proving Theorem 1.

3. Bridge Regression Family

Frank and Friedman (1993) introduced bridge regression and it was further discussed by Fu (1998). Bridge regression is a family of penalized regressions, which minimize squared error loss subject to the constraint $\sum |\beta_j|^\gamma \leq t$ with $\gamma \geq 0$, i.e.,

$$\hat{\beta}^{bridge} = arg \min_\beta \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^\gamma. \tag{3.1}$$

Here, we consider the bridge regression with $\gamma \geq 1$ (when $\gamma < 1$, the penalty function is not convex) and X of full column rank. Note that, when $\gamma = 1$, bridge regression is same as the LASSO, and when $\gamma = 2$, bridge regression becomes ridge regression.

From the Bayesian perspective, the bridge regression solution can be viewed as the mode of the posterior density

$$\pi(\beta|Y) \propto exp\Big\{ -\frac{1}{2}\Big[(Y - X\beta)^T (Y - X\beta) + \lambda \sum_{j=1}^{p} |\beta_j|^\gamma\Big]\Big\}, \tag{3.2}$$

with the prior in the exponential power family. From Table 1, we know that

$$\tau[\pi(\beta)] \propto \Big[\sum_j \beta_j^{2(\gamma-1)}\Big]^{\frac{1}{2}}. \tag{3.3}$$

In addition, it is easy to check that when $\gamma > 1$, the log-prior density is differentiable with respect to β. By the Hölder inequality, we have

$$\Big[\sum_j \beta_j^{2(\gamma-1)}\Big]^{\frac{1}{2}} \geq D_1\, (||\beta||_2)^{\gamma-1} \text{ given } \gamma \geq 2, \tag{3.4}$$

$$\Big[\sum_j \beta_j^{2(\gamma-1)}\Big]^{\frac{1}{2}} \leq D_2\, (||\beta||_2)^{\gamma-1} \text{ given } 1 < \gamma < 2, \tag{3.5}$$

where D_1 and D_2 are two positive constants. Now, (3.4) and (3.5) imply that the bridge regression prior with $\gamma \geq 2$ and $1 < \gamma < 2$ belong to Class 1 and 2, respectively. Therefore, by Theorem 1, we know that the bridge solution with $\gamma \geq 2$ does not have the weak Bayesian robustness property, whereas for $1 < \gamma < 2$ it has the weak, but not the strong, Bayesian robustness property.

When $\gamma = 1$, the bridge regression prior is the Laplacian distribution, which is not differentiable at $\beta_j = 0$, $j = 1, \ldots, p$. Thus, Theorem 1 cannot be applied in the LASSO case. However, $l(\beta|Y)$ is differentiable everywhere and the LASSO solution satisfies

$$\left| \frac{\partial l(\theta|Y)}{\partial \beta_j} |_{\hat{\beta}^{lasso}} \right| \leq \frac{\lambda}{2}, \quad \forall\, j = 1, \ldots, p. \tag{3.6}$$

Then, by (2.12), we have

$$\|\hat{\beta}^{lasso} - \hat{\beta}^{o}\|_2 \leq \frac{\lambda\sqrt{p}}{2\nu_1}. \tag{3.7}$$

Thus, the LASSO solution has strong Bayesian robustness property. The above discussion leads to the following result.

Theorem 2. *Suppose the matrix X has full column rank. The bridge regression satisfies the following.*

1. *If $\gamma \geq 2$, the weak Bayesian robustness property does not hold.*
2. *If $1 < \gamma < 2$, the weak Bayesian robustness property holds, but the strong property does not hold.*
3. *If $\gamma = 1$, the strong Bayesian robustness property holds.*

Pericchi and Smith (1992) obtained the exact solution of posterior mean under a Gaussian likelihood and Laplacian prior in one dimension. Interestingly, they found that the absolute distance between posterior mean and data-based weighted average is bounded by a constant. Therefore, the posterior mean has a property similar to strong Bayesian robustness in this particular situation.

4. Group LASSO

Consider the regression model with J factors:

$$Y = \sum_{j=1}^{J} X_j \beta_j + \epsilon, \tag{4.1}$$

where Y is a length n vector, $\epsilon \sim N(0, \sigma^2 I)$, X_j is a $n \times p_j$ matrix corresponding to the jth factor, and β_j is a coefficient vector of size p_j, $j = 1, \ldots, J$. Given positive definite matrices $K_1, \ldots, K_J$, the group LASSO estimate is defined as the solution to

$$(Y - \sum_{j=1}^{J} X_j \beta_j)^T (Y - \sum_{j=1}^{J} X_j \beta_j) + \lambda \sum_{j=1}^{J} (\beta_j^T K_j \beta_j)^{\frac{1}{2}}, \tag{4.2}$$

where $\lambda \geq 0$ is a tuning parameter.

Define $Z_j = X_j K_j^{-1/2}$ and $\eta_j = K_j^{1/2} \beta_j$ for $j = 1, \ldots, J$. Define $Z = (Z_1, \ldots, Z_J)$ and $\eta^T = (\eta_1^T, \ldots, \eta_J^T)$. Then, (4.2) can be represented as

$$(Y - Z\eta)^T (Y - Z\eta) + \lambda \sum_{j=1}^{J} (\eta_j^T \eta_j)^{\frac{1}{2}}. \tag{4.3}$$

Like LASSO, the posterior distribution corresponding to the group LASSO is not differentiable everywhere. However, it can be shown that the group LASSO solution $\hat{\beta}^{glasso}$ satisfies

$$\left| \frac{\partial l(\eta|Y)}{\partial \eta_{ij}} \,|_{\hat{\beta}^{glasso}} \right| \leq \frac{\lambda}{2}, \quad i = 1, \ldots, p_j, \quad j = 1, \ldots, J, \tag{4.4}$$

where η_{ij} is the ith element of η_j and $l(\eta|Y)$ is the Gaussian log-likelihood. Then, (4.4) implies that the group LASSO estimate has the strong Bayesian robustness property.

5. Revisit LASSO

Consider the linear regression model

$$y_i = x_i \beta + \epsilon_i, \quad \epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2), \quad i = 1, \ldots, n. \tag{5.1}$$

The following theorem gives us a probabilistic bound for the L_2 distance between the LASSO estimate and the true value β.

Theorem 3. *Suppose that $(1/n)X^T X$ is non-singular with smallest eigenvalue ν. Then*

$$P\Big(||\hat{\beta}^{lasso} - \beta||_2 \leq \frac{\lambda\sqrt{p}}{2n\nu} + \epsilon\Big) \geq P\Big(\chi_p^2 \leq \frac{n\epsilon^2\nu}{\sigma^2}\Big), \tag{5.2}$$

for any $\epsilon > 0$.

Proof. Given $X^T X$ is non-singular, it is well known that the OLS estimate, $\hat{\beta}^o$ satisfies

$$(\hat{\beta}^o - \beta)^T (X^T X)(\hat{\beta}^o - \beta) \sim \sigma^2 \chi_p^2. \tag{5.3}$$

By (3.7) and the triangle inequality

$$||\hat{\beta}^{lasso} - \beta||_2 \leq ||\hat{\beta}^{lasso} - \hat{\beta}^{o}||_2 + ||\hat{\beta}^{o} - \beta||_2, \tag{5.4}$$

$$P\Big(||\hat{\beta}^{lasso} - \beta||_2 \leq \frac{\lambda\sqrt{p}}{2n\nu} + \epsilon\Big) \geq P\left((\hat{\beta}^{o} - \beta)^T(\hat{\beta}^{o} - \beta) \leq \epsilon^2\right). \tag{5.5}$$

Since

$$(\hat{\beta}^{o} - \beta)^T(X^TX)(\hat{\beta}^{o} - \beta) \geq n\nu(\hat{\beta}^{o} - \beta)^T(\hat{\beta}^{o} - \beta), \tag{5.6}$$

$$P\left((\hat{\beta}^{o} - \beta)^T(\hat{\beta}^{o} - \beta) \leq \epsilon^2\right) \geq P\left((\hat{\beta}^{o} - \beta)^T(X^TX)(\hat{\beta}^{o} - \beta) \leq n\epsilon^2\nu\right). \tag{5.7}$$

The result follows from (5.3), (5.5) and (5.7).

Corollary 2. *Suppose that $(1/n)X^TX \to \Sigma$, as $n \to \infty$, where Σ is a positive definite matrix with the smallest eigenvalue ν. Then for any $\epsilon > 0$,*

$$P\left(||\hat{\beta}^{lasso} - \beta||_2 \leq \epsilon\right) \to 1 \quad \text{as } n \to \infty. \tag{5.8}$$

Proof. The convergence of $(1/n)X^TX$ implies the convergence of the eigenvalues and the result follows from Theorem 3.

Although the LASSO solution is a consistent estimate in terms of L_2 distance, it doesn't mean that it is consistent in terms of variable selection. In fact, Leng, Lin and Wahba (2004) pointed out the inconsistency of LASSO if tuned by predictive criteria, such as cross-validation. This is akin to AIC's properties in linear regression, namely, AIC is optimal in prediction and over-consistent in variable selection. A heuristic explaination of this coincidence is the asymptotic equivalence between cross-validation and AIC established by Stone (1977).

Knight and Fu (2000) proposed a way to achieve consistency in terms of variable selection by thresholding the LASSO solution by a quantity of order n^{α} with $-1/2 < \alpha < 0$. In fact, this can be shown as a consequence of Theorem 3. Let

$$\bar{\beta}_j^{\epsilon} = \begin{cases} 0, & \text{if } |\hat{\beta}_j^{lasso}| < \epsilon, \\ \hat{\beta}_j^{lasso}, & \text{if } |\hat{\beta}_j^{lasso}| \geq \epsilon, \end{cases}$$

and $T^{\epsilon} = \{j \mid \bar{\beta}_j^{\epsilon} \neq 0\}$.

Corollary 3. *Let the conditions in Corollary 2 hold, and suppose $\epsilon \sim O(n^{\alpha})$, where $-1/2 < \alpha < 0$. Then $P(T^{\epsilon} = \{j \mid \beta_j \neq 0\}) \to 1$ as $n \to \infty$.*

Proof. It is sufficient to show that, as $n \to \infty$, $P(j \in T^{\epsilon} \mid \beta_j \neq 0) \to 1$ and $P(j \notin T^{\epsilon} \mid \beta_j = 0) \to 1$. Then must have $\lambda\sqrt{p}/(2n\nu) + \epsilon \to 0$ and $n\epsilon^2\nu/\sigma^2 \to \infty$ as $n \to \infty$, but this implies $-1/2 < \alpha < 0$.

Corollary 3 suggests that thresholding LASSO estimators achieves consistency in variable selection. A detailed discussion of the asymptotic properties of estimators based on thresholding method can be found at Donoho et al. (1996).

6. Discussion and Future Work

The suggestion of using a Student t-distribution to provide a robust analysis for a Gaussian location parameter has a long history. Pericchi and Smith (1992) obtained the approximate posterior mean for the Gaussian likelihood and t-prior distribution in one dimension. Their results imply that the posterior mean corresponding to the t-prior behaves like a "trimmed" mean. In this paper, we show the robustness property of the inference from the Gaussian likelihood and t-prior. Unlike the posterior mode and MLE for the $0-1$ loss function, in general we do not have a natural data-based counterpart for the posterior mean in the case of the squared error loss.

Although Theorem 1 is useful for a large class of priors, (2.5) and (2.6) in Assumption 2A and 2B are difficult to check, except for the normal likelihood case. In our future work, we will explore the connection between the flatness of a prior and robustness for other likelihood functions. We are also investigating the robustness property of other regularizers such as the Elastic Net.

7. Proof of Theorem 1

Proof. By Assumption 1, we know that a necessary condition for $\tilde{\theta}$ to be a global maxima is that it be a stationary point where the derivative of $\log \pi(\theta|Y)$ vanishes. Hence, it is sufficient to show that Theorem 1 holds for all the stationary points. Here, we abuse the notation and write $\tilde{\theta}$ as any stationary point of $\log \pi(\theta|Y)$. By definition, for any stationary point $\tilde{\theta}$, we have

$$\frac{\partial l(\theta|Y)}{\partial \theta}|_{\tilde{\theta}} = -\frac{\partial \log \pi(\theta)}{\partial \theta}|_{\tilde{\theta}} \quad \Rightarrow \quad \Big\|\frac{\partial l(\theta|Y)}{\partial \theta}|_{\tilde{\theta}}\Big\|_2 = \tau[\pi(\tilde{\theta})]. \tag{7.1}$$

First, we consider the priors from Class 1. For any stationary point $\tilde{\theta}$, by (2.5), (2.7) and (7.1), we have

$$B_1||\tilde{\theta}-\theta^*||_2^Q \leq A_2||\tilde{\theta}-\hat{\theta}||_2^P, \quad Q \geq P > 0, \tag{7.2}$$

$$\Rightarrow \quad ||\tilde{\theta}-\theta^*||_2 \leq \left(\frac{A_2}{B_1}\right)^{\frac{1}{Q}} ||\tilde{\theta}-\hat{\theta}||_2^{\frac{P}{Q}}. \tag{7.3}$$

By the triangle inequality, we have $||\hat{\theta}-\tilde{\theta}||_2 + ||\tilde{\theta}-\theta^*||_2 \geq ||\hat{\theta}-\theta^*||_2 = C$. Now (7.3) implies that

$$||\tilde{\theta}-\hat{\theta}||_2 + \left(\frac{A_2}{B_1}\right)^{\frac{1}{Q}} ||\tilde{\theta}-\hat{\theta}||_2^{\frac{P}{Q}} \geq C. \tag{7.4}$$

Since the left hand side of (7.4) is a strictly increasing function of $||\tilde{\theta} - \hat{\theta}||_2$, it follows that, as $C \to \infty$, $||\tilde{\theta} - \hat{\theta}||_2 \to \infty$. On the other hand, (7.2) implies

$$\left(\frac{B_1}{A_2}\right)^{\frac{1}{Q}} ||\tilde{\theta} - \hat{\theta}||_2^{\frac{Q-P}{Q}} \leq r(\tilde{\theta}, \theta^*, \hat{\theta}), \tag{7.5}$$

where the left hand side term does not go to zero as $C \to \infty$.

We now consider the priors from Class 2. For any stationary point $\tilde{\theta}$, by (2.6), (2.8) and (7.1), we have

$$A_1 ||\tilde{\theta} - \hat{\theta}||_2^P \leq B_2 ||\tilde{\theta} - \theta^*||_2^Q, \quad 0 < Q < P, \tag{7.6}$$

$$\Rightarrow \quad ||\tilde{\theta} - \hat{\theta}||_2 \leq \left(\frac{B_2}{A_1}\right)^{\frac{1}{P}} ||\tilde{\theta} - \theta^*||_2^{\frac{Q}{P}}. \tag{7.7}$$

By the triangle inequality, we have

$$\left(\frac{B_2}{A_1}\right)^{\frac{1}{P}} ||\tilde{\theta} - \theta^*||_2^{\frac{Q}{P}} + ||\tilde{\theta} - \theta^*||_2 \geq ||\hat{\theta} - \theta^*||_2 = C. \tag{7.8}$$

Since the left hand side of (7.8) is a strictly increasing function of $||\tilde{\theta} - \theta^*||_2$, it follows that, as $C \to \infty$, $||\tilde{\theta} - \theta^*||_2 \to \infty$. By (2.9) and (7.1), we have

$$\left\| \frac{\partial l(\theta|Y)}{\partial \theta} |_{\tilde{\theta}} \right\|_2 = \tau[\pi(\tilde{\theta})] \to \infty \text{ as } C \to \infty, \tag{7.9}$$

which implies that $||\tilde{\theta} - \hat{\theta}||_2 \to \infty$ as $C \to \infty$. Thus the strong Bayesian robustness property does not hold. On the other hand, (7.6) implies that

$$r(\tilde{\theta}, \theta^*, \hat{\theta}) \leq \left(\frac{B_2}{A_1}\right)^{\frac{1}{P}} ||\tilde{\theta} - \hat{\theta}||_2^{\frac{Q-P}{P}}, \tag{7.10}$$

where the right hand side term of (7.10) goes to zero as $C \to \infty$. Therefore, we have the weak Bayesian robustness property.

We now consider the priors from Class 3. By (2.6), (2.10) and (7.1), we have

$$A_1 ||\tilde{\theta} - \hat{\theta}||_2^P \leq B_2 \quad \Rightarrow \quad ||\tilde{\theta} - \hat{\theta}||_2 \leq \left(\frac{B_2}{A_1}\right)^{\frac{1}{P}}. \tag{7.11}$$

Therefore, we have the strong Bayesian robustness property.

Acknowledgements

The authors would like to thank the referees for pointing out related work, and for constructive comments and suggestions that helped to improve the presentation of the paper.

References

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, New York.

Berger, J. O. (1994). An overview of robustness Bayesian analysis. *Test* **3**, 5-58.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis.* Addison-Wesley, New York.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1996). Density estimation by wavelet thresholding. *Ann. Statist.* **24**, 508-539.

Fan, J. and Li, R. Z. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-1360.

Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35**, 109-148.

Fan, T. H. and Berger, J. O. (1992). Behavior of the posterior distribution and inferences for a normal mean with t prior distributions. *Statist. Decisions* **10**, 99-120.

Fu, W. (1998). Penalized regressions: the Bridge versus the LASSO. *J. Comput. Graph. Statist.* **7**, 397-416.

Gustafson, P. (2000). Local robustness in Bayesian analysis. In *Robust Bayesian Analysis* (Edited by D. R. Insua and F. Ruggeri), 71-88. Springer-Verlag, New York.

Gustafson, P., Srinivasan, C. and Wasserman, L. (1996). Local sensitivity analysis. In *Bayesian Statistics* **5** (Edited by J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), 197-210. Oxford University Press, London.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**, 55-67.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.* **28**, 1356-1378.

Leng, C., Lin, Y. and Wahba, G. (2004). A note on the LASSO and related procedures in model selection. Technical Report 1091r.

Moreno, E. (2000). Global Bayesian robustness for some classes of prior distributions. In *Robust Bayesian Analysis* (Edited by D. R. Insua and F. Ruggeri), 45-70. Springer-Verlag, New York.

O'Hagan, A. (1988). Modelling with heavy tails. In *Bayesian Statistics* **3** (Edited by J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), 345-359. Oxford University Press, London.

Pericchi, L. R. and Smith, A. F. M. (1992). Exact and approximate posterior moments for a normal location parameter. *J. Roy. Statist. Soc. Ser. B* **54**, 793-804.

Sivaganesan, S. (2000). Global and local robustness approaches: uses and limitations. In *Robust Bayesian Analysis* (Edited by D. R. Insua and F. Ruggeri), 89-108. Springer-Verlag, New York.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. Roy. Statist. Soc. Ser. B* **39**, 44-47.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

Vapnik, V. (1998). *Statistical Learning Theory.* John Wiley, New York.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. Roy. Statist. Soc. Ser. B* **68**, 49-67.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. Ser. B* **67**, 301-320.

Department of Statistics, The Ohio State University, 1958 Neil Avenue, Cockins Hall, Room 404, Columbus, OH 43210-1247, U.S.A.

E-mail: bli@stat.ohio-state.edu

Department of Statistics, The Ohio State University, 1958 Neil Avenue, Cockins Hall, Room 404, Columbus, OH 43210-1247, U.S.A.

E-mail: goel@stat.ohio-state.edu

(Received June 2005; accepted September 2005)