Bayesian Variable Selection for Skewed Heteroscedastic Response

Libo Wang[1], Yuanyuan Tang[1], Debajyoti Sinha[1], Debdeep Pati[1], and Stuart Lipsitz[2]

[1] Department of Statistics, Florida State University, Tallahassee, FL 32306

[2] Brigham and Women's Hospital

Abstract

In this article, we propose new Bayesian methods for selecting and estimating a sparse coefficient vector for skewed heteroscedastic response. Our novel Bayesian procedures effectively estimate the median and other quantile functions, accommodate non-local prior for regression effects without compromising ease of implementation via sampling based tools, and asymptotically select the true set of predictors even when the number of covariates increases in the same order of the sample size. We also extend our method to deal with some observations with very large errors. Via simulation studies and a re-analysis of a medical cost study with large number of potential predictors, we illustrate the ease of implementation and other practical advantages of our approach compared to existing methods for such studies.

Keywords: Bayesian consistency; median regression; sparsity

1. Introduction

Large number of possible predictors and highly skewed heteroscedastic response are often major challenges for many biomedical applications. Selection of an optimal set of covariates and subsequent estimation of the regression function are important steps for scientific conclusions and policy decisions based on such studies. For example, previous analyses of Medical Expenditure Panel Study (Natarajan et al., 2008; Cohen, 2003) testify to the highly skewed and heteroscedastic nature of the main response of interest, total health care expenditure in a year. Also, it is common in such studies to have a small proportion of patients with either very high or very low medical costs. Popular classical sparse-regression methods such as Lasso (Least absolute shrinkage operator) by Tibshirani (1996) and Efron et al. (2004), and later related methods of Fan and Li (2001), Zou and Hastie (2005), Zou (2006) and MCP (Zhang, 2010) assume Gaussian (or, at least symmetric) response density with common variance. Limited recent literature on consistent variable selection for non-Gaussian response includes Zhao and Yu (2006) under common variance assumption, Bach (2008) under weak conditions on covariate structures, and Chen et al. (2014) under skew-t errors. However, none of these methods deal with estimation of quantile function for heteroscedastic response frequently encountered in complex biomedical studies. Many authors including Koenker (2005) argue effectively against focusing on mean regression for skewed heteroscedastic response. Our simulation studies demonstrate that directly modeling skewness and heteroscedasticity, particularly in presence of analogous empirical evidence, leads to better estimators of quantile functions for finite samples compared to existing methods which ignore skewness and heteroscedasticity.

Bayesian methods for variable selection have some important practical advantages including incorporation of prior information about sparsity, evaluation of uncertainty about the final model, interval estimate for any coefficient of interest, and evaluation of the relative importance of different coefficients. Asymptotic properties of Bayesian variable selection methods when the number of potential predictors, p, increases as a function of sample size n have received lot of attention recently in the literature. Traditionally, to select the important variables out of $(X_1, \ldots, X_p)$, a two component mixture prior, also referred to as "spike and slab" prior, (Mitchell and Beauchamp, 1988; George and McCulloch, 1993, 1997) is placed on the coefficients $\beta = (\beta_1, \ldots, \beta_p)$. These mixture priors include a discrete mass, called a "spike", at zero to characterize the prior probability of a coefficient being exactly zero (that is, not including the corresponding predictor in the model) and a continuous density called a "slab", usually centered at null-value zero, representing the prior opinion when the coefficient is non-zero. Following Johnson and Rossell (2010), when the continuous density of the slab part of a spike and slab prior has value 0 at null-value 0, we will call it a non-local mixture prior. Continuous analogues of local mixture priors are being proposed recently by Park and Casella (2008); Carvalho et al. (2010); Bhattacharya et al. (2014) among others. Bondell and Reich (2012) presented the selection consistency of joint Bayesian credible sets. However, current Bayesian variable selection methods usually focus on mean regression function for models with symmetric error density and common variance.

Johnson and Rossell (2010) recently showed a startling selection inconsistency phenomenon for using several commonly used mixture priors, including local mixture (spike and slab prior with non-zero value at null-value 0 of the slab density) priors, when p is larger than the order of $\sqrt{n}$. To address this for mean regression with sparse β, they advocated the use of non-local mixture density presenting continuous "slab" density with value 0 at null-value 0 because these priors, called non-local mixture priors here, obtain selection consistency when the dimension p is $O(n)$. Castillo et al. (2014) provided several conditions to ensure selection consistency even when $p \gg n$. However, none of these Bayesian methods specifically deal with skewed and heteroscedastic response, contamination of few observations with large errors and variable selection for median and other quantile functionals.

In this article, we accommodate skewed and heteroscedastic response distribution using transform-both-sides model (Lin et al., 2012) with sparsity inducing prior for the vector of regression coefficients. Our key observation is that, under such models with generalized Box-Cox transformation (Bickel and Doksum, 1981), even a local mixture prior on after-transform regression coefficients induces non-local priors on the original regression function for certain choices of the transformation parameter. Similar to moment and inverse moment non-local priors in Johnson and Rossell (2010), this method enables clear demarcation between the signal and the noise coefficients in the posterior leading to consistent posterior selection even when $p = O(n)$. Our use of standard local priors on the transformed regression coefficients facilitates straightforward posterior computation which can be implemented in publicly available softwares. We later extend this model to accommodate cases when the observations are contaminated with a small number of observations with very large (or small) errors. Our approaches are shown to out-perform well-known competitors in simulation studies as well as for analyzing and interpreting a real-life medical cost study.

2. Bayesian variable selection model

2.1 Transform-both-Sides Model

For the skewed and heteroscedastic response Y_i for $i = 1, \ldots, n$, we assume the transform-both-sides model (Lin et al., 2012)

$$g_\eta(Y_i) = g_\eta(x_i^T \beta) + e_i \; , \tag{1}$$

where $\beta = (\beta_1, \ldots, \beta_p)'$, x_i is the observed p-dimensional covariate vector, $g_\eta(y)$ is the monotone power transformation (Bickel and Doksum, 1981), an extension of the Box-Cox power family,

$$g_\eta(y) = \frac{y|y|^{\eta-1} - 1}{\eta}, \tag{2}$$

with unknown parameter $\eta \in (0, 2)$, and e_i's are independent mean 0 errors with common symmetric density function f_e and variance σ^2. The transformation $g_\eta(y)$ in (2) is monotone with derivative $g'_\eta(y) = |y|^{\eta-1} \geq 0$. Model (1) can be expressed as a linear model

$$Y_i \; = \; x_i^T \beta \; + \; \epsilon_i \; , \tag{3}$$

where ϵ_i has a skewed heteroscedastic density with median 0 because $P[\epsilon_i > 0] = 1/2$, and approximate variance is $\sigma^2 |x_i^T \beta|^{2-2\eta}$. Hence, the median of the skewed and heteroscedastic response Y_i in (1) is $x_i^T \beta$. In this article, we consider a Gaussian $\mathrm{N}(0, \sigma^2)$ density for f_e in (1). Later in §4, we consider other densities to accommodate a heavy tail for f_e.

As discussed in §1, independent spike and slab priors for every β_j constitute a popular choice to induce sparsity in β. However, for the model of (1), a prior for β should depend on the transformation parameter η since η has a significant effect on the range and scale of Y_i

(approximate variance $\sigma^2|x_i^T\beta|^{2-2\eta}$). Based on this argument, we specify an conditional mixture prior for $g_\eta(\beta_j)$ given η using a "local" $\phi(\cdot;0,\sigma_\beta^2)$ density for the "slab" when $g_\eta(\beta_j)$ is non-zero with discrete prior probability $(1-\pi_0)$, where $\phi(\cdot;\mu,v^2)$ is the Gaussian density with mean μ and variance v^2. This conditional mixture prior for $g_\eta(\beta_j)$ given η (a local mixture prior according to definition of Johnson and Rossell (2010)) results in a conditional mixture prior

$$f_\beta(\beta_j \mid \eta) = \pi_0\delta_0(\beta_j) + (1-\pi_0)\phi(g_\eta(\beta_j);0,\sigma_\beta^2)|g'_\eta(\beta_j)| \quad (4)$$

for β_j given η independently for $j = 1, \cdots, p$, where $\pi_0 \in [0,1]$ is the probability of β_j being zero, $\delta_0(\cdot)$ is the discrete measure at 0. When $\eta > 1$, $g'_\eta(0) = 0$ and hence the resulting unconditional marginal prior for the nonzero β_j in (4) turns out to be a non-local mixture prior of Johnson and Rossell (2010). However, the prior of transformed $g_\eta(\beta_j)$, a mixture of discrete measure at 0 and $\phi(\cdot;0,\sigma_\beta^2)$ density, is a local mixture prior. This is demonstrated via the plots of two resulting unconditional priors of β_j when $\eta = 0.5$ and $\eta = 1.5$ in Figure 1. Our simulation study in §5 shows that the model selection and estimation procedures for our Bayesian method perform substantially better than competing methods when $\eta > 1$ (case with non-local unconditional prior for β_j) compared to, say, when $\eta = 0.5 \in (0,1)$ (case with a local prior for β given η). Thus, heteroscedasticity and the possibly non-local property of $\pi(\beta \mid \eta)$ come as a bi-product of the transform-both-sides model of (1). This implicit non-local mixture prior modeling of unconditional β_j may be the reason for some desirable asymptotic properties of our method even when $p = O(n)$ (discussed in §3). However, our methods' ability to use a local prior for $g_\eta(\beta_j)$ significantly reduces computational complexity of the associated Markov chain Monte Carlo (MCMC) algorithms, while facilitating the desirable asymptotic property.

When the error density f_e in (1) is Gaussian $\phi(\cdot;\mu,\sigma^2)$ with mean 0 and variance σ^2, we can specify a hierarchical Bayesian model using the prior $\pi(\beta \mid \eta)$ in (1) along with known priors for σ^2 and η,

$$\sigma^2 \sim \text{IGa}(a,b), \quad \eta/2 \sim \text{Beta}(c_1,d_1) \quad (5)$$

and a hyperprior for the hyperparameters (π_0, σ_β^2). For computational simplicity, in §3, we

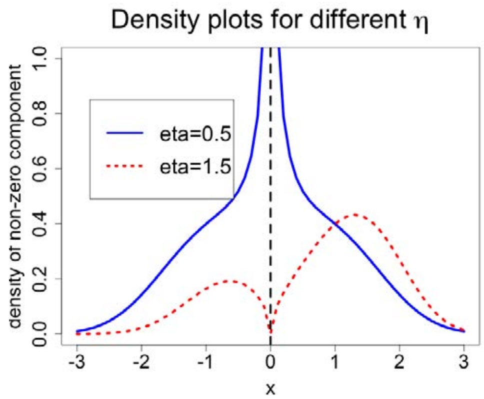

Figure 1: Density Plot for Different η

establish variable selection consistency of this hierarchical Bayesian model using (1) along with the non-local prior (4) on β and the prior for σ^2 in (5).

3. Consistent variable selection for large p

Unlike the Gaussian likelihood with mixture priors of Johnson and Rossell (2010), our Bayesian model described in (1) - (5) does not admit a closed form expression of the marginal likelihood. We use appropriate bounds of the marginal likelihood to obtain the desired Bayesian consistency results. Here we only present a brief outline of our assumptions, developments and practical implications of our theoretical results. Supporting results and details of the proofs are deferred to §A amd §B. For brevity of exposition, we consider a design matrix which has $O(n)$ contrasts for estimating the coefficients. Considering n to be even, let the first column of X be $(1, -1, 1, -1, \ldots, 1, -1)^T$ and all the remaining $p-1$ columns be $\mathbf{1}_n$, the $(n \times 1)$ vector with each component being equal to 1. Johnson and Rossell (2010) consider design matrices X with $c_1 n \leq \lambda_1(X^T X) \leq \lambda_p(X^T X) \leq c_2 n$ for identifiability, where λ_1 and λ_p are respectively the smallest and largest eigen-values of $X^T X$. For our chosen X, the eigen-values of any $(n \times 2)$ sub-matrix formed by the first column and any other column of X are $\{n, n\}$. This means only β_1 is identifiable, and it is meaningful to consider only β_1 to be non-zero.

For any vector $\beta \in \mathbb{R}^p$, $S_\beta = \{j : \beta_j \neq 0\} \subset \{1, 2, \ldots, p\}$, the set of non-zero components of β is called the support of β. For brevity, we use $S_0 \equiv S_{\beta_0}$ for support of β_0, the true value of the parameter vector β. Under an optimal η, the marginal likelihood

$m_S(\mathbb{Y})$ for $\beta_S = \{\beta_j : j \in S\}$ is

$$m_S(\mathbb{Y}) = \int \prod_{i=1}^{n} \phi\{0; g_\eta(y_i) - g_\eta(x_{iS}^T \beta_S), \sigma^2\}\pi(\beta_S)\pi(\sigma^2) d\beta d\sigma^2 \ , \tag{6}$$

where the projection $\pi(\beta_S)$ of β is derived from (4) as

$$\pi(\beta_S) = \prod_{j \in S_\beta} \phi(g_\eta(\beta_j); 0, \sigma_\beta^2) \left| g_\eta'(\beta_j) \right| . \tag{7}$$

Based on definitions of key concepts in (6) and (7), we state our main theorem on selection consistency of our Bayesian method even when p is of the order $O(n)$.

Theorem 1 *When the observations are generated from* (1) *for a fixed* η*, the true sparse* β_0 *satisfies* $S_0 = \{1\}$*, and model* (1) *has priors of* (4) *and* (5) *for the optimal* η *and with any* $\pi_0 \in (0, 1)$*, then for* $p \le n/\log 4$*, the posterior probability* $P(S_\beta = S_0 \mid D_n) \to 1$ *almost surely as* $n, p \to +\infty$.

The detailed proof of Theorem 1, given in §A, is a non-trivial extension of the proof of Theorem 1 in Johnson and Rossell (2010) which used non-local priors to obtain variable selection consistency. Unlike them, we use a local priors of $g_\eta(\beta)$ to induce a possible non-local prior for β given η in (4). To the best of our knowledge, this is the first result on Bayesian selection consistency when the response distribution is skewed and heteroscedastic.

4. Accommodating extremely large errors

Presence of few observations with extremely large errors and their influences on final analysis for various application areas have been emphasized by many authors including Hampel et al. (2011). The assumption of Gaussian error density f_e in (1) may not be valid due to the presence of a small number of observations with large errors even after optimal Box-Cox transformation. To address this, we extend the model (1) to a random location-shift model with

$$g_\eta(Y_i) = g_\eta(x_i^T \beta) + \gamma_i + e_i \ , \tag{8}$$

where γ_i is nonzero if the ith observation has large error, and zero otherwise. We assume the vector $\gamma = (\gamma_1, \ldots, \gamma_n)^T$ to be sparse to ensure only a small probability of the response having a large error after transformation. Similar idea of location-shift model, however, with un-transformed response, is popular in the recent literature on robust linear models (for example, She and Owen (2011) and McCann and Welsch (2007)). To ensure that $g_\eta(x_i^T\beta)$ is the mean and median of $g_\eta(Y_i)$, we require the mean and median of $\gamma_i + e_i$ to be zero, that is, we need a symmetric distribution for γ_i. For this purpose, we use another spike-and-slab mixture prior $f_\gamma(\gamma_i) = \pi_\gamma\delta_0 + (1-\pi_\gamma)\phi(\gamma_i; 0, \sigma_\gamma^2)$ independently for $i = 1, \ldots, n$, where $0 < \pi_\gamma < 1$.

To induce a heavy-tailed error density after transformation, we also consider another extension of the model (1) as

$$g_\eta(Y_i) = g_\eta(x_i^T\beta) + U_i^{-1/2}e_i \qquad with \quad U_i \sim H(\cdot \mid \nu)\ , \tag{9}$$

where $H(\cdot \mid \nu)$ is a positive mixing distribution indexed by a parameter ν and e_i's are again independent $\mathrm{N}(0, \sigma^2)$. This class of heavy-tailed error distributions of (9) is called normal independent (NI) family (Lange and Sinsheimer, 1993). We consider three kinds of heavy tailed distribution, Student's-t, slash and contaminated normal (CN) respectively, for the marginal error density in (9) using the following specific choices of $H(\cdot \mid \nu)$ (Lachos et al., 2011): χ_ν^2/ν distribution with possibly non-integer $\nu > 2$, $H(u \mid \nu) = u^\nu$ for $u \in [0, 1]$, and discrete $H(u \mid \nu)$ with $P[\rho < 1] = 1 - P[\rho = 1] = \nu$. For student-t error, we use the prior for the degrees of freedom parameter ν to be a truncated exponential on the interval $(2, \infty)$. For ν of the slash distribution marginal error, we use a Gamma(a, b) prior with small positive values of a and b with $b \ll a$. For contaminated normal marginal error, we assign Beta(ν_0, ν_1) and Beta(ρ_0, ρ_1) priors respectively for ν and ρ. In §5, we compare the performances of Bayesian analyses under these competing models for highly skewed and heteroscedastic responses.

5. Simulation Studies

Simulation model with no outliers: We use different simulation models to compare our Bayesian methods under model (1) with LASSO (Tibshirani, 1996) and the penalized quantile methods (Koenker, 2005). From each simulation model, we simulated 50 replicated datasets of sample size 50. For both simulation studies, the observations are sampled from the model (1) with $e_i \sim \mathrm{N}(0, \sigma_0^2)$ with $\sigma_0^2 = 1$. The hyperparameters for priors in (5) are set as $a = 2, b = 2, c_1 = 1$ and $d_1 = 1$. The tuning parameters for LASSO and penalized quantile regression are selected via a grid search based on the 5-fold cross-validation. We compare the estimators from different methods based on following criteria: the mean masking proportion M (fraction of undetected true $\beta_j \neq 0$), the mean swamping proportion S (fraction of wrongly selected $\beta_j = 0$), and the joint detection rate JD (fraction of simulations with 0 masking). We also compare the goodness-of-fit of estimation methods using an influence measure $L/L^* - 1$, where

$$L = \sum_{i=1}^{n} (g_{\eta_0}(y_i) - g_{\eta_0}(x_i^T \hat{\beta}))^2/(2\sigma_0^2) - n/2 \log(2\pi\sigma_0^2) + (\eta_0 - 1) \sum_{i=1}^{n} \log(|y_i|). \quad (10)$$

is the log-likelihood under $(\hat{\beta}, \eta_0, \sigma_0)$ and L^* is the same log-likelihood under $(\beta_0, \eta_0, \sigma_0)$, and $(\beta_0, \eta_0, \sigma_0)$ are the known true parameter values (of the simulation model). The results of our study using simulated data from TBS model (1) with different values of η are displayed in Table 1.

Table 1: Results of simulation studies for using different methods of analysis:

Simulation model of (1) with $\eta_0 = 0{\cdot}5, p = 8, \beta_0 = (3, 1{\cdot}5, 0, 0, 2, 0, 0, 0)$.

Method used	$L/L^* - 1$	# of non-zeros	M(%)	S(%)	JD(%)
TBS-SG	-0·02	3·16	0	3·2	100
Penalized Quantile	0·04	5·84	0	56·8	100
LASSO	0·04	4·98	0	3·96	100
TBSt-SG	-0·01	3·16	0	3·2	100
TBSS-SG	-0·02	3·16	0	3·2	100
TBSCN-SG	-0·02	3·14	0	2·8	100

Simulation model of (1) with $\eta_0 = 1{\cdot}8, p = 8, \beta_0 = (3, 1{\cdot}5, 0, 0, 2, 0, 0, 0)$.

Method used	$L/L^* - 1$	# of non-zeros	M(%)	S(%)	JD(%)
TBS-SG	-0·06	3·02	0	0·4	100
Penalized Quantile	0·04	5·82	0	56·4	100
LASSO	0·66	4·56	0	3·12	100
TBSt-SG	-0·05	3	0	0	100
TBSS-SG	-0·06	3	0	0	100
TBSCN-SG	-0·06	3	0	0	100

M: masking proportion (fraction of undetected true $\beta_j \neq 0$); S: swamping proportion S (fraction of wrongly selected β_j with true value 0); JD: joint detection rate.

In Table 1, we compare our Bayesian TBS model (1) with prior (4) for β (called TBS-SG in short) to frequentist methods of penalized quantile and LASSO. From the results in Table 1, it is evident that our TBS-SG method provides better results than competing frequentist methods in terms of average number of non-zeros and swamping error rate. We also compare TBS-SG method with other Bayesian TBS models with heavy tailed normal independence (NI) error in (1). These competing NI models in (9) include TBSt-SG model (in short) with t distribution for $H(\cdot \mid \nu)$, TBSS-SG model (in short) with slash distribution for $H(\cdot \mid \nu)$ and TBSCN-SG model (in short) with contaminated normal distribution for $H(\cdot \mid \nu)$. All our Bayesian methods have "SG" in their end of acronym to indicate the spike Gaussian prior of (4) for β. TBS models accommodating heavy tailed response perform the best in competing models with ideal masking, swamping and joint outlier detection rates. Both our methods and frequentist methods provide comparable performances based on average $L/L^* - 1$ values, although the $L/L^* - 1$ values from penalized quantile estimates using different datasets are highly variable. All methods perform desirable with respect to masking and joint detection. Also, we found that our Bayesian methods provide better results when true η value is $\eta_0 = 1.8$, compared to $\eta_0 = 0.5$.

To compare the performances for different η, we set $p = 20$ and the number of non-zero coefficient to be 12. Denote by $(x)_k$ the vector formed by appending k copies of x. Consider case i) $\beta_0 = \{(2)_{12}, (0)_8\}$, case ii) $\beta_0 = \{(-10)_6, (4)_6, (0)_8\}$, case iii) $\beta_0 = \{(-10)_{10}, (4)_2, (0)_8\}$, case iv) $\beta_0 = \{(-10)_2, (-4)_2, (-2)_2, (2)_2, (4)_2, (10)_2, (0)_8\}$, case v) $\beta_0 = \{(-10)_6, (2)_6, (0)_8\}$, case vi) $\beta_0 = \{(-10)_2, (-8)_2, (-6)_2, (-4)_2, (-2)_2, (2)_2, (0)_8\}$. We use only TBSCN-SG model for analysis because these three TBS models accommodating heavy tailed response have similar performance.

From the results in Table 2, we can clearly see that for all the cases, all the four methods perform better when $\eta_0 = 1.8$ compared to when $\eta_0 = 0{\cdot}5$, with respect to average number of non-zeros, masking, swamping and joint detection rate. This can be explained by the fact that when $\eta_0 = 1{\cdot}8$, we expect the posterior draws of η to be close to $1{\cdot}8$ which corresponds to a non-local prior for β (see Figure 1). When the range of signals is large and when there are many groups of small coefficients (see case (iv) and case (vi)), all of the methods do not perform well. Considering only variable selection results (average number of non-zeros), our TBS model clearly out performs penalized quantile method and LASSO.

Studies using simulation model with outliers and heavy-tailed distribution: Our simulation models are similar to previous simulation model of (1) except that a few of the observations are now have large errors even after transformation. Although the Bayesian TBS methods with NI error in (9) do not provide the identification and estimation of these

Table 2: Results of simulation studies for using different methods of analysis when the true model is (1) with $p = 20$:

		TBS-SG		Penalized Quantile		LASSO		TBSCN-SG	
	Measurement	η_0=0·5	η_0=1·8	η_0=0·5	η_0=1·8	η_0=0·5	η_0=1·8	η_0=0·5	η_0=1·8
Case i)	$L/L^* - 1$	-0·05	2·72	0·02	0·7	0·07	-82·81	-0·04	2·74
	# of non-zeros	13·6	12·02	16·36	15·58	14·02	12·24	13·74	12·02
	M (%)	0	0	0·5	0	0	0	0	0
	S(%)	20	0·25	55·25	44·75	25·75	3	21·75	0·25
	JD(%)	100	100	94	100	96	100	100	100
Case ii)	$L/L^* - 1$	-0·03	0·03	0·05	-0·24	0·1	-1282	-0·02	0·09
	# of non-zeros	13·26	12	17·12	15·02	14·96	12·14	13·24	12
	M (%)	0·67	0	0·67	0	0	0	8·33	0
	S(%)	16·75	0	65	37·75	37·5	1·75	16·75	0
	JD(%)	94	100	94	100	96	100	92	100
Case iii)	$L/L^* - 1$	-0·03	-0·06	0·03	-0·15	0·11	-2736	-0·02	0·04
	# of non-zeros	12·94	12	16·94	14·52	14·7	12·24	12·92	12
	M (%)	0·33	0	0·33	0	0·83	0	0·5	0
	S(%)	12·25	0	62·25	31·5	35	3	12·25	0
	JD(%)	96	100	96	100	90	100	94	100
Case iv)	$L/L^* - 1$	-0·01	0·09	0·05	-0·22	0·09	-689·6	-0·01	0·13
	# of non-zeros	12·58	12	16·42	15·32	14·2	12·08	12·18	12
	M (%)	6·17	0	3·67	0	5·67	0·1	7·67	0
	S(%)	16·5	0	60·75	41·5	36	2	13·75	0
	JD(%)	46	100	64	100	54	92	34	100
Case v)	$L/L^* - 1$	0·01	0·04	0·05	-0·21	0·1	-1072	0·01	0·09
	# of non-zeros	11·32	12	16·12	14·96	13·56	12·04	11·16	12
	M (%)	11·17	0	5	0	7·17	0·83	12	0
	S(%)	8·25	0	59	37	30·25	1·75	7·5	0
	JD(%)	24	100	62	100	34	90	14	100
Case vi)	$L/L^* - 1$	-0·01	0·10	0·06	-0·27	0·12	-763·3	0·01	0·13
	# of non-zeros	12·1	12	15·62	14·96	13·52	12·14	11·76	12
	M (%)	6·67	0	4·83	0	8·33	1·33	8·33	0
	S(%)	11·25	0	52·5	37	31·5	3·75	9·5	0
	JD(%)	44	100	62	100	36	84	36	100

M: masking proportion; S: swamping proportion S; JD: joint detection rate.

observations, we wonder whether they ensure robust variable selection and estimation of β, particularly in comparison to the Bayesian method using random location-shift model of (8).

For the sake of brevity of the presentation, we omit the tables for results of simulation studies using data simulated from models (8) and (9), and only summarize the results here. When we use the simulation model (8) with $\eta_0 = 0{\cdot}5, p = 8, \gamma_{(1:2)} = 8, \gamma_3 = -8, \gamma_{(4:50)} = 0$, and $\beta_0 = (3, 1{\cdot}5, 0, 0, 2, 0, 0, 0)$, our Bayesian method with model (8) obtains $3{\cdot}32$ non-zero γ_i's on average. The masking (M), swamping (S) and joint detection (JD) rates are $1{\cdot}33\%$, $0{\cdot}77\%$ and 98%. Also, our method provides $3{\cdot}22$ non-zero estimated β_j on average with the masking, swamping and joint detection rates of $0{\cdot}67\%$, $4{\cdot}8\%$ and 98% respectively.

For Simulation 4, we choose $\eta_0 = 1{\cdot}8, p = 8, \gamma_{(1:2)} = 8, \gamma_3 = -8, \gamma_{(4:50)} = 0$ and $\beta_0 = (3, 1{\cdot}5, 0, 0, 2, 0, 0, 0)$. For our Bayesian method with model (8), we have $3{\cdot}28$ non-zero γ_i's on average with the masking, swamping and joint detection rates of 0%, $0{\cdot}6\%$ and 100%. Also, our method provides $3{\cdot}22$ non-zero estimated β_j on average with the masking, swamping and joint detection rates of 0%, 0% and 100% respectively.

When we simulate data from model (9), the Bayesian method leads to results similar to the results obtained for using simulation model (8). However, only the Bayesian method using (8) provides the identification of the observations with errors of large magnitude. In practice, identification of such observations will facilitate further investigations regarding their measurement accuracy, influence on inference and other exploratory diagnostics. All of our Bayesian models provide better results than the penalized quantile regression and LASSO with respect to average number of non-zeros, masking, swamping and joint detection.

6. Analysis of medical expenditure study

Our motivating study is the Medical Expenditure Panel Survey (Cohen, 2003; Natarajan et al., 2008), called the MEPS study in short, where the response variable is each patient's 'total health care expenditures in the year 2002'. Previous analyses of of this study (Natarajan et al., 2008) suggest that the variance of the response is a function of the mean (heteroscedasticity). Often in practice, medical cost data are typically highly skewed to the right, because a small percentage of patients may accumulate extremely high costs compared to other patients, and the variance of total cost tends to increase as the mean increases.

In this article, we focus only on one large cluster because every cluster of MEPS study has different sampling design. After removing only a few patients with missing observations, we have 173 patients and 24 potential predictors including age, gender, race, disease history, etc. The minimum cost is 0 and the maximum is \$79660, with a mean \$4584 and median \$1342. For the convenience of computation, we standardize the response (cost) and five potential predictors of the patient: age in 2002, highest education degree attained, perceived health status, body mass index (BMI), and ability to overcome illness (OVERCOME). Rest of the potential predictors are binary variables with values 0 and 1. We analyze this study using our proposed Bayesian models and compare the results with the penalized quantile regression method of Koenker (2005). For Bayesian methods, we use our transform-both-sides model (1), the model of (8) with sparse large errors (TBSO-SG in short) and the model of (9) with contaminated normal marginal error. For each method, we compute an observed residual $y_{i0} - x_i^T\hat{\beta}$, where y_{i0} is the observed un-transformed response and $x_i^T\hat{\beta}$ is the estimated median. The Q-Q plots for the residuals are in Figure 2.

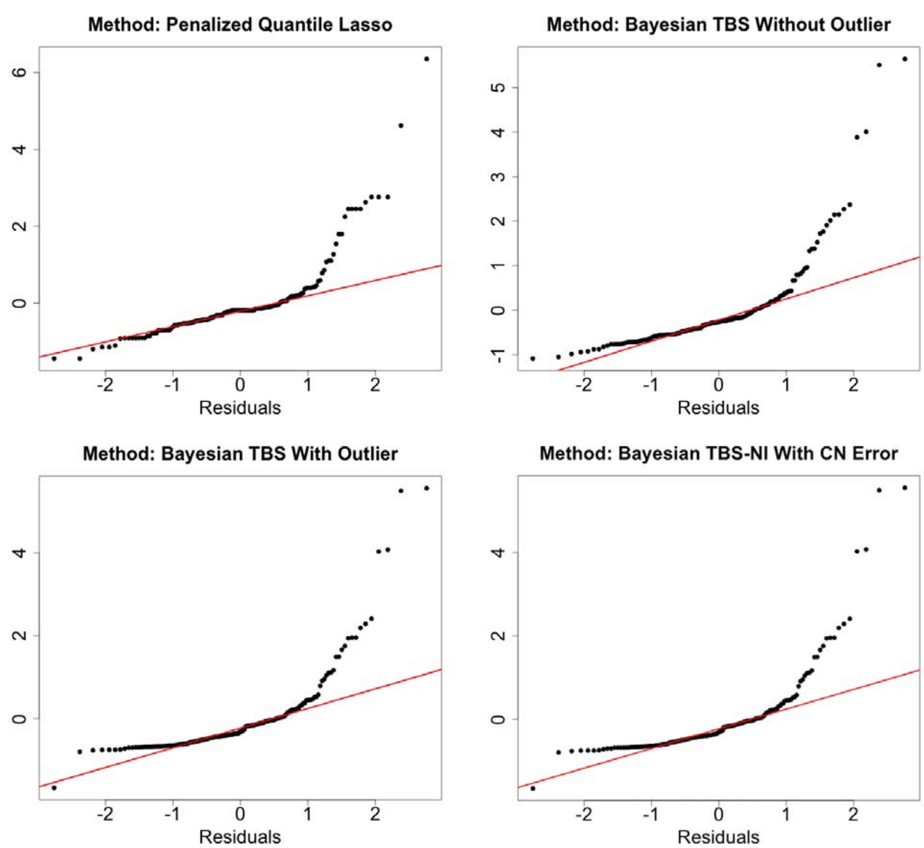

Figure 2: Q-Q plots of observed residuals obtained from 4 methods

From Q-Q plots in Figure 2, it is obvious that the normality assumption about untransformed response is untenable. We now compare the goodness-of-fit of the three Bayesian methods to evaluate their abilities from handling skewness and heteroscedasticity. For this purpose, we use the residual $g_{\hat{\eta}}(y_{i0}) - g_{\hat{\eta}}(x_i^T \hat{\beta})$ for the Bayesian TBS-SG model of (1) and the TBSCN-SG model of (9), and the residual $g_{\hat{\eta}}(y_{i0}) - g_{\hat{\eta}}(x_i^T \hat{\beta}) - \hat{\gamma}_i$ for Bayesian TBSO-SG model of (8), and then display their Q-Q plots in Figure 3. It is evident from the Q-Q plots that TBSO-SG model of (8) has the best justification to use it for Bayesian analysis, and TBSCN-SG model of (9) also performs well except may be for some observations in both tails.

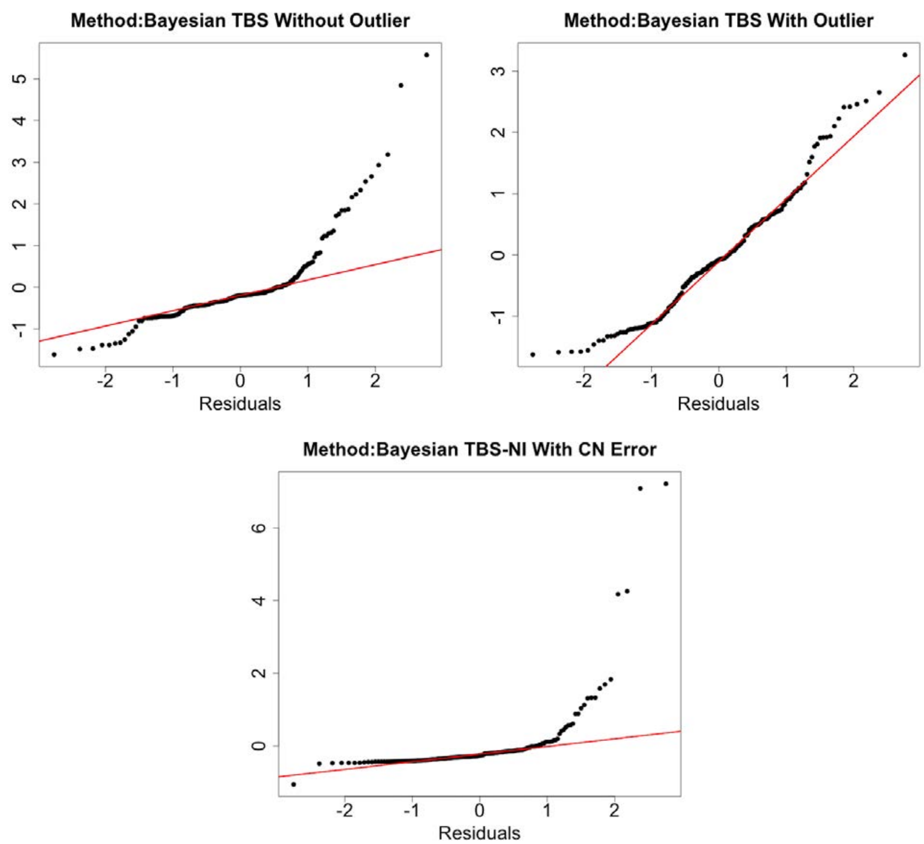

Figure 3: Q-Q plots for residuals of transformed responses obtained from 3 Bayesian models

Using our Bayesian model of (8), we find large posterior evidence of effects of OVERCOME variable with posterior mean=$-0{\cdot}17$ and 95% credible interval $(-0{\cdot}21, -0{\cdot}12)$, stroke with posterior mean=$0{\cdot}92$ and 95% credible interval $(0{\cdot}66, 1{\cdot}23)$, and medication with posterior mean=$-0{\cdot}35$ and 95% credible interval $(-0{\cdot}41, -0{\cdot}24)$. Model(9) identifies these same predictors of model (8) with slightly different interval estimates. Model (1) also identifies three predictors with large posterior evidence of effects: perceived health status, stroke and the indicator of major ethnic group. Stroke is the only variable with large posterior evidence of effects in all three models. Even though penalized quantile regression based analysis selects a larger number of predictors compared to the number of predictors selected by our models, the only statistically significant variable from quantile regression analysis is age (estimate of $0{\cdot}08$ with standard error $0{\cdot}03$). This may be explained by the larger estimated standard errors of the estimates from quantile regression compared to the posterior standard deviations of the corresponding parameters obtained via Bayesian analysis.

In order to better understand the prediction performance on observed data, we present a scatter plot with overlaid quantile lines for each Bayesian method in Figure 4. For each method, we display scaled $x_i^T\hat{\beta}$ and scaled observed response y_{i0}, along with estimated 25th percentile and 75th percentile curves using $g_{\hat{\eta}}^{-1}\{g_{\hat{\eta}}(x_i^T\beta) + Z_\alpha^*\}$, where Z_α^* is estimated α-percentile of $f_e(\cdot)$. Figure 4 shows that the method using (8) explains the observed data better than methods using (1) and (9). The observations with large errors identified by analysis using (8) are marked by asterisk signs in the second plot. We find that all the observations identified by (8) are outside the estimated interquartile ranges. It shows that our transform-both-sides model of (8) is successful in handling data with skewness, heteroscadesticity as well as very large errors in few subjects. Model of (9) is also able to handle skewness and heteroscadesticity but is not able to identify observations with extremely large errors.

We also use posterior predictive loss approach (Gelfand and Ghosh, 1998) to evaluate the prediction accuracy under each Bayesian method. We compute the prediction errors of our Bayesian methods by $\sum_{i=1}^n \mathrm{E}[\{g_{\hat{\eta}}(y_{i0}) - g_{\hat{\eta}}(x_i^T\hat{\beta})\}^2|D]$ from model (1), and by $\sum_{i=1}^n E[\{g_{\hat{\eta}}(y_{i0}) - g_{\hat{\eta}}(x_i^T\hat{\beta}) - \gamma_i\}^2|D]$ for model (8) using MCMC approximation, where D is the observed dataset. The average prediction error from model (8) is $15 \cdot 03$, which is considerably better than model (1) with average prediction error $174 \cdot 45$.

7. Discussion

In this article, we propose Bayesian variable selection methods for skewed and heteroscedastic response. The methods are highly suitable for modeling, computation, analysis and in-

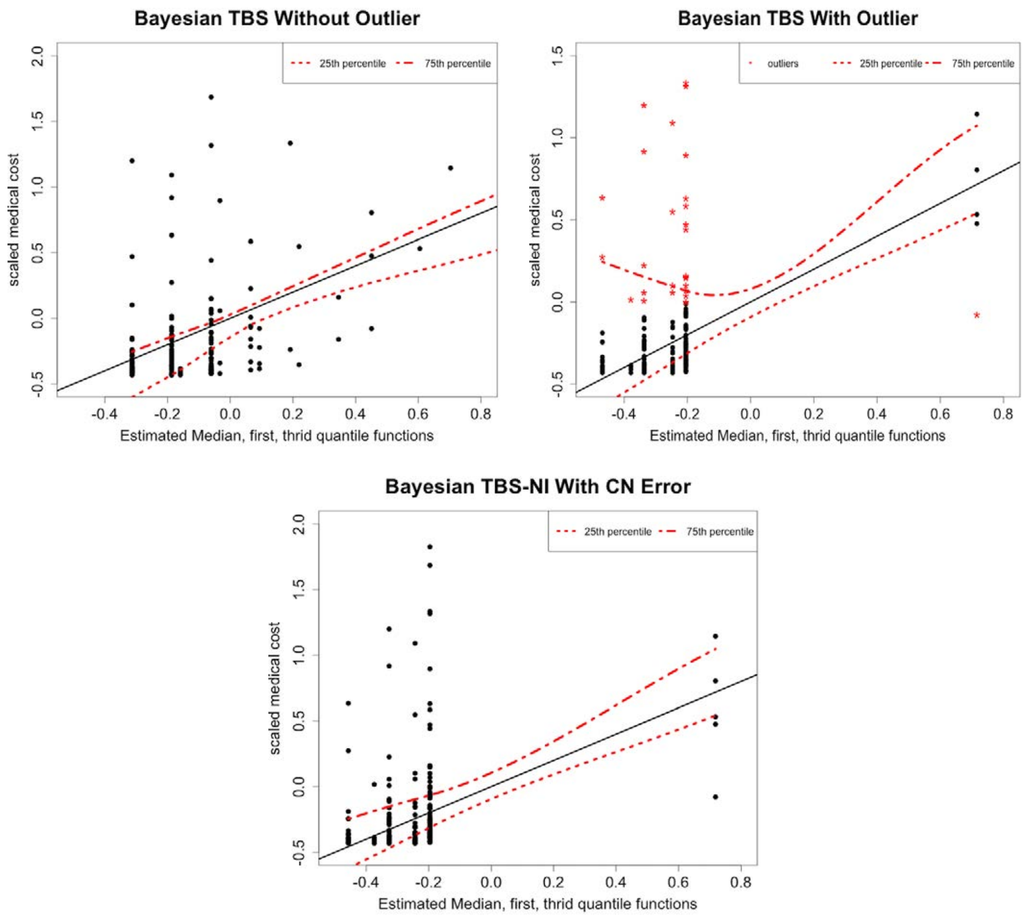

Figure 4: Scatter plots of scaled observed responses and quantile regression functions obtained from 3 Bayesian models

terpretation of real-life health care cost studies, where we aim to determine and estimate effects of a sparse set of explanatory variables for health care expenditures out of a large set of potential explanatory variables. Simulation results indicate a better performance of our Bayesian methods compared to existing frequentist quantile regression tools. Also, our Bayesian approaches provide flexible and robust estimations to incorporate a wide variety of practical situations. The advantages of our Bayesian methods include their practical and easy implementation using standard statistical software. In the appendix, we prove the consistency of variable selection even when the number of potential predictors p is comparable to, however, smaller than n. The proofs are only provided for a special case of the covariate matrix and when the transform parameter η is known. Proof for a more general case can be obtained following a similar, but more tedious mathematical arguments.

References

Bach, F. R. (2008), Bolasso: model consistent lasso estimation through the bootstrap, *in* 'Proceedings of the 25th International Conference on Machine learning', ACM, pp. 33–40.

Bhattacharya, A., Pati, D., Pillai, N. and Dunson, D. (2014), 'Dirichlet-laplace priors for optimal shrinkage', *JASA* .

Bickel, P. J. and Doksum, K. A. (1981), 'An analysis of transformations revisited', *JASA* **76**(374), 296–311.

Bondell, H. D. and Reich, B. J. (2012), 'Consistent high-dimensional bayesian variable selection via penalized credible regions', *JASA* **107**(500), 1610–1624.

Carvalho, C., Polson, N. and Scott, J. (2010), 'The horseshoe estimator for sparse signals', *Biometrika* **97**(2), 465–480.

Castillo, I., Schmidt-Hieber, J. and van der Vaart, A. W. (2014), 'Bayesian linear regression with sparse priors', *ArXiv e-prints* .

Chen, L., Pourahmadi, M. and Maadooliat, M. (2014), 'Regularized multivariate regression models with skew-t error distributions', *Journal of Statistical Planning and Inference* **149**, 125 – 139.

Cohen, S. B. (2003), 'Design strategies and innovations in the medical expenditure panel survey', *Medical Care* **41**(7), III.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004), 'Least angle regression', *The Annals of statistics* **32**(2), 407–499.

Fan, J. and Li, R. (2001), 'Variable selection via nonconcave penalized likelihood and its oracle properties', *JASA* **96**(456), 1348–1360.

Gelfand, A. E. and Ghosh, S. K. (1998), 'Model choice: A minimum posterior predictive loss approach', *Biometrika* **85**(1), 1–11.

George, E. I. and McCulloch, R. E. (1997), 'Approaches for Bayesian variable selection', *Statistica Sinica* **7**(2), 339–373.

George, E. and McCulloch, R. (1993), 'Variable selection via Gibbs sampling', *JASA* **88**(423), 881–889.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (2011), *Robust Statistics: the Approach Based on Influence Functions*, Vol. 114, Wiley.

Johnson, V. and Rossell, D. (2010), 'On the use of non-local prior densities in Bayesian hypothesis tests', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**(2), 143–170.

Koenker, R. (2005), *Quantile Regression*, Vol. 38, Cambridge University Press.

Lachos, V. H., Bandyopadhyay, D. and Dey, D. K. (2011), 'Linear and nonlinear mixed-effects models for censored hiv viral loads using normal/independent distributions', *Biometrics* **67**(4), 1594–1604.

Lange, K. and Sinsheimer, J. S. (1993), 'Normal/independent distributions and their applications in robust regression', *Journal of Computational and Graphical Statistics* **2**(2), 175–198.

Lin, J., Sinha, D., Lipsitz, S. and Polpo, A. (2012), 'Semiparametric bayesian survival analysis using models with log-linear median', *Biometrics* **68**(4), 1136–1145.

McCann, L. and Welsch, R. E. (2007), ‘Robust variable selection using least angle regression and elemental set sampling’, *Computational Statistics & Data Analysis* **52**(1), 249–257.

Mitchell, T. J. and Beauchamp, J. J. (1988), ‘Bayesian variable selection in linear regression’, *JASA* **83**(404), 1023–1032.

Natarajan, S., Lipsitz, S. R., Fitzmaurice, G., Moore, C. G. and Gonin, R. (2008), ‘Variance estimation in complex survey sampling for generalized linear models’, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **57**(1), 75–87.

Park, T. and Casella, G. (2008), ‘The Bayesian lasso’, *JASA* **103**(482), 681–686.

She, Y. and Owen, A. B. (2011), ‘Outlier detection using nonconvex penalized regression’, *JASA* **106**(494), 626–639.

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

Zhang, C.-H. (2010), ‘Nearly unbiased variable selection under minimax concave penalty’, *The Annals of Statistics* pp. 894–942.

Zhao, P. and Yu, B. (2006), ‘On model selection consistency of lasso’, *J. Mach. Learn. Res.* **7**, 2541–2563.

Zou, H. (2006), ‘The adaptive lasso and its oracle properties’, *JASA* **101**(476), 1418–1429.

Zou, H. and Hastie, T. (2005), ‘Regularization and variable selection via the elastic net’, *Journal of the Royal Statistical Society, Series B* **67**, 301–320.

Appendices

A. Proof of Theorem 1

Observe that, $P(S_\beta = S_0 \mid D_n) = m_{S_0}(\mathbb{Y})\pi(S_0)/\{\sum_{S\subset\{1,2,\ldots,p\}} m_S(\mathbb{Y})\pi(S)\}$. Here $\pi(S_0) = \pi_0(1-\pi_0)$ and $\pi(S) = \pi_0^{|S|}(1-\pi_0)^{p-|S|}$. Depending on the nature of S, we have different estimates for $m_S(\mathbb{Y})$. Denote by $\dot{z}_i = \left|g_\eta^{'}(y_i)\right|$. We will provide the bounds for $m_S(\mathbb{Y})$ for $p=2$ which then can be easily generalized to arbitrary p in view of the bounds in Lemma 2. For $S=\phi$,

$$\begin{aligned} m_S(\mathbb{Y}) &= \frac{b^a}{\Gamma(a)}\int\prod_{i=1}^n \phi(0; g_\eta(y_i)-g_\eta(0),\sigma^2)\left|g_\eta^{'}(y_i)\right|(\sigma^2)^{-(a+1)}\exp(-b/\sigma^2)d\sigma^2 \\ &= C_n(\mathbb{Y})\frac{\Gamma(n/2+a)}{\left\{\sum(z_i+1/\eta)^2/2+b\right\}^{\frac{n}{2}+a}}, \end{aligned}$$

where $C_n(\mathbb{Y}) = \left(1/\sqrt{2\pi}\right)^n\prod_{i=1}^n|\dot{z}_i|\, b^a/\Gamma(a)$. Similarly, for $S=\{2\}$,

$$m_S(\mathbb{Y}) = \frac{C_n(\mathbb{Y})}{\sqrt{n}\sigma_\beta}\frac{\Gamma\{(n-1)/2+a\}}{(\sum z_i^2/2 - n\bar{z}^2/2+b)^{(n-1)/2+a}}.$$

For $S=\{1\}$, define $Z_e = \{2,4,\ldots,n\}$ and $Z_o=\{1,3,\ldots,n-1\}$

$$m_S(\mathbb{Y}) = \frac{C_n(\mathbb{Y})}{\sqrt{n}\sigma_\beta}\frac{\Gamma\{(n-1)/2+a\}}{\left\{\sum_{j\in Z_o} z_j^2/2 + \sum_{j\in Ze}(z_j+2/\eta)^2/2 - \left(\sum_{j\in Z_o} z_j - \sum_{j\in Ze} z_j - n/\eta\right)^2/2n + b\right\}^{(n-1)/2+a}}$$

where $t_1 = g_\eta(\beta_1), -t_1-2/\eta = g_\eta(-\beta_1)$.

For $S = \{1, 2\}$,

$$
\begin{aligned}
m_S(\mathbb{Y}) &= \frac{b^a}{\Gamma(a)} \iiint \prod_{i=1}^{n} \phi(g_\eta(y_i) - g_\eta(x_i^T\beta), \sigma^2) \left| g_\eta^{'}(y_i) \right| \phi(g_\eta(\beta_1); 0, \sigma_\beta^2) \left| g_\eta'(\beta_1) \right| \\
&\quad \phi(g_\eta(\beta_1); 0, \sigma_\beta^2) \left| g_\eta'(\beta_1) \right| (\sigma^2)^{-a-1} \exp(-b/\sigma^2) d\sigma^2 d\beta_1 d\beta_2 \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^{n+2} \prod_{i=1}^{n} |\dot{z}_i| \frac{b^a}{\Gamma(a)} \iiint \frac{1}{\sigma^n} \exp\left[-\frac{\sum_{j\in Z_o}\{z_j - g_\eta(\beta_1+\beta_2)\}^2 + \sum_{j\in Z_e}\{z_j - g_\eta(-\beta_1+\beta_2)\}^2}{2\sigma^2} \right] \\
&\quad \times \frac{1}{\sigma_\beta^2} \exp(-t_1^2/2\sigma_\beta^2) \exp(-t_2^2/2\sigma_\beta^2) (\sigma^2)^{-a-1} \exp(-b/\sigma^2) dt_1 dt_2 d\sigma^2 \\
&= \frac{C_n(\mathbb{Y})}{\sigma_\beta^2} \int_0^\infty \int_{|\beta_1| \geq |\beta_2|} \left(\frac{1}{\sqrt{2\pi}}\right)^2 \exp\left[-\frac{\sum_{j\in Z_o}\{z_j - g_\eta(\beta_1+\beta_2)\}^2 + \sum_{j\in Z_e}\{z_j - g_\eta(-\beta_1+\beta_2)\}^2}{2\sigma^2} \right] \\
&\quad \times \exp(-t't/2\sigma_\beta^2)(\sigma^2)^{-n/2-a-1} \exp(-b/\sigma^2) dt d\sigma^2 \\
&+ \frac{C_n(\mathbb{Y})}{\sigma_\beta^2} \int_0^\infty \int_{|\beta_1| < |\beta_2|} \left(\frac{1}{\sqrt{2\pi}}\right)^2 \exp\left[-\frac{\sum_{j\in Z_o}\{z_j - g_\eta(\beta_1+\beta_2)\}^2 + \sum_{j\in Z_e}\{z_j - g_\eta(-\beta_1+\beta_2)\}^2}{2\sigma^2} \right] \\
&\quad \times \exp(-t't/2\sigma_\beta^2)(\sigma^2)^{-n/2-a-1} \exp(-b/\sigma^2) dt d\sigma^2 \\
&= I_1 + I_2. \qquad (11)
\end{aligned}
$$

where $t = (t_1, t_2)'$. From Lemma 2, when $k = 2$, $\sum_{i=1}^n \{z_i - g_\eta(\pm\beta_1 + \beta_2)\}^2 \geq \sum_{i=1}^n T_i + nt'At + 2^{\eta+1}\sum |z_i| b't$, where T_i is defined in Lemma 2. Recall that $|\beta_1| \geq |\beta_2|$ is equivalent to $|t_1 + 1/\eta| \geq |t_2 + 1/\eta|$. Accordingly, the domain of t_1 is the combination of $(-\infty, -1/\eta)$, $[-1/\eta, 0)$ and $[0, \infty)$, corresponding with the region of t_2 satisfying $|t_1 + 1/\eta| \geq |t_2 + 1/\eta|$. I_1 in (11) is can then be decomposed into three parts each of which can be upper bounded by the integral over the full domain. More precisely, I_1 can be upper-bounded as

$$
\begin{aligned}
I_1 &\leq \frac{3C_n(\mathbb{Y})}{\sigma_\beta^2} \int_0^\infty \int_{\mathbb{R}} \left(\frac{1}{\sqrt{2\pi}}\right)^2 \exp\left\{ -\frac{\sum T_i + t'(nA + \sigma^2 I/\sigma_\beta^2)t + 2^\eta \sum |z_i| b't}{2\sigma^2} \right\} \\
&= \frac{3C_n(\mathbb{Y})}{\sqrt{n}\sigma_\beta} \frac{\Gamma\{(n-1)/2 + a\}}{(\sum T_i/2 - 2^{2\eta-1} n\bar{z}_+^2 + b)^{(n-1)/2+a}} \qquad (12)
\end{aligned}
$$

where $|\Sigma| = \sigma^2/n$, $\mu' = 2^\eta \sum |z_i| b'\Sigma/\sigma^2$, and $n\bar{z}_+^2 = (\sum |z_i|)^2/n$. Similarly, I_2 in (11) is upper bounded by the same expression in the right hand side of (12). For $S = \{1, 2\}$, the marginal likelihood satisfies

$$
m_S(\mathbb{Y}) = I_1 + I_2 \leq \frac{6C_n(\mathbb{Y})}{\sqrt{n}\sigma_\beta} \frac{\Gamma\{(n-1)/2 + a\}}{(\sum T_i/2 - 2^{2\eta-1} n\bar{z}_+^2 + b)^{(n-1)/2+a}}
$$

For the numerator in the upper bound of $m_S(\mathbb{Y})$, we use a version of the Stirling's formula for $\alpha > 0$: $\Gamma(a) = \sqrt{2\pi} e^{-\alpha} \alpha^{\alpha-1/2} e^{\theta(\alpha)}$, where $0 < \theta(\alpha) < 1/12\alpha$. To tackle $m_S(\mathbb{Y})$ for $S = \{1\}$, observe that for $i \in Z_o$, $z_i \sim \mathrm{N}(g_\eta(\beta_0), \sigma_0^2)$, and for $i \in Z_e$, $z_i \sim \mathrm{N}(g_\eta(-\beta_0), \sigma_0^2)$. Denoting

$t_0 = g_\eta(\beta_0)$, we can find that $g_\eta(-\beta_0) = -t_0 - 2/\eta$. Finally, we define $\Delta = 2\max(|t_0|)$ and $\delta = 0.5\min(|t_0|)$. With these results in place, we express the expectations as $\mathrm{E}\big\{\sum(z_i+1/\eta)^2/2+b\big\} = nt_0^2/2+nt_0/\eta+n/2\eta^2+n\sigma_0^2/2+b$, $\mathrm{E}\big(\sum z_i^2/2-n\bar{z}^2/2+b\big) = nt_0^2/2+nt_0/\eta+n/2\eta^2+(n-1)\sigma_0^2/2+b$, $\mathrm{E}\big\{\sum_{j\in Z_o} z_j^2/2+\sum_{j\in Z_e}(z_j+\frac{2}{\eta})^2/2-(\sum_{j\in Z_o} z_j-\sum_{j\in Z_e} z_j-n/\eta)^2/2n+b\big\} = (n-1)\sigma_0^2/2+b$, and $\mathrm{E}(\sum T_i/2-2^{2\eta-1}n\bar{z}_+^2+b) = nt_0^2/4+n(t_0+2/\eta)^2/4-n(2^\eta+1)(\mu_o+\mu_e)/\eta-n2^{2\eta-1}(\mu_o+\mu_e)^2/4+O_p(1)+n\sigma_0^2/2+b$, where μ_o is the expectation of $|Z_o|$, μ_e is the expectation of $|Z_e|$. Recognizing $|Z_o|$ and $|Z_e|$ to be folded normal distributions, μ_o and μ_e can be expressed as $\mu_o = \sigma_0\sqrt{\frac{2}{\pi}}e^{-t_0^2/2\sigma_0^2}+t_0\{1-2\Phi(-t_0/\sigma_0)\}$, and $\mu_e = \sigma_0\sqrt{\frac{2}{\pi}}e^{-(t_0+2/\eta)^2/2\sigma_0^2} - (t_0+2/\eta)[1-2\Phi\{(t_0+2/\eta)/\sigma_0\}]$, which subsequently implies,

$$
\begin{aligned}
\mathrm{E}\big\{\sum(z_i+1/\eta)^2/2+b\big\} &= n(t_0+1/\eta)^2/2+n\sigma_0^2/2+b,\\
\mathrm{E}\big(\sum z_i^2/2-n\bar{z}^2/2+b\big) &= n(t_0+1/\eta)^2/2+(n-1)\sigma_0^2/2+b,\\
\mathrm{E}\big\{\sum_{j\in Z_o} z_j^2/2+\sum_{j\in Z_e}(z_j+2/\eta)^2/2-(\sum_{j\in Z_o} z_j &- \sum_{j\in Z_e} z_j-n/\eta)^2/2n+b\big\} = (n-1)\sigma_0^2/2+b,\\
-2^3n(t_0+1/\eta)^2/2-2^3n\sigma_0\sqrt{2/\pi}t_0-n\sigma_0^2/2+b &\leq E(\sum T_i/2-2^{2\eta-1}n\bar{z}_+^2+b)\\
\leq nt_0^2/4+n(t_0+2/\eta)^2/4 &+ n(2^\eta+1)(2/\eta+2|t_0|)/\eta+n\sigma_0^2/2+b.
\end{aligned}
$$

Hence the denominators in the upper bounds of $m_S(\mathbb{Y})$ are stochastically bounded as $\big\{\sum(z_i+1/\eta)^2/2+b\big\}^{n/2+a} \longrightarrow \big\{nO_p(1)\big\}^{n/2}$, $\big(\sum z_i^2/2-n\bar{z}^2/2+b\big)^{(n-1)/2+a} \longrightarrow \big\{nO_p(1)\big\}^{(n-1)/2}$, $\big\{\sum_{i\in Z_o} z_i^2/2+\sum_{j\in Z_e}(z_j+\frac{2}{\eta})^2/2-(\sum_{i\in Z_o} z_i-\sum_{j\in Z_e} z_j-\frac{n}{\eta})^2/2n+b\big\}^{(n-1)/2} \longrightarrow \big\{nO_p(1)/e\big\}^{(n-1)/2}$, and $(\sum T_i/2-2^{2\eta-1}n\bar{z}_+^2+b)^{(n-1)/2+a} \longrightarrow \big\{nO_p(1)\big\}^{(n-1)/2}$. Now we are in a position to tackle the case of general p. Proceeding similarly to (11) and (12), we have for each $|S|=k$, $0\leq k\leq p$,

$$
m_S(\mathbb{Y}) \leq \binom{p}{k}\frac{6C_n(\mathbb{Y})}{\sqrt{n}\sigma_\beta}\frac{\Gamma\{(n-1)/2+a\}}{(\sum T_i/2-k^{2\eta}n\bar{z}_+^2/2+b)^{(n-1)/2+a}},
$$

where $T_i = z_i^2-(2k^\eta/\eta+2/\eta)|z_i|$. Observing that $(\sum T_i/2-k^{2\eta}n\bar{z}_+^2/2+b)^{(n-1)/2+a} = \{k^{2\eta}nO_p(1)\}^{(n-1)/2+a}$, $P(S_\beta=S_0\mid D_n)$ is stochastically equivalent to

$$
\begin{aligned}
&\frac{\pi_0(1-\pi_0)e^{(n-1)/2}/\{nO_p(1)\}^{(n-1)/2}}{\pi_0(1-\pi_0)e^{(n-1)/2}/\{nO_p(1)\}^{(n-1)/2}+\sum\binom{p}{k}\pi(S)e^{(n-1)/2}/\{k^{2\eta}nO_p(1)\}^{(n-1)/2}}\\
\gtrsim\quad&\frac{O_p(1)}{1+\sum_{k=0}^p 6\binom{p}{k}/(e^{n/2}\sum_{k=0}^p k^{2\eta})}. \qquad (13)
\end{aligned}
$$

If p is bounded by $n/\log 4$ in (13), when $n\to\infty$,

$$
\frac{\sum_{k=0}^p 6\binom{p}{k}}{e^{n/2}\sum_{k=0}^p k^{2\eta}} \leq \frac{2^{p+1}}{e^{n/2}p(p/2)^{2\eta}} \leq \left(\frac{4^{p+1}}{e^n p^2}\right)^{1/2} \longrightarrow 0.
$$

Hence, $P(S_\beta = S_0 \mid D_n) \xrightarrow{n\to\infty} 1$ a.s.

B. Auxiliary results

In this section, we state a few auxiliary results providing bounds for terms involving $g_\eta(y_i)$ appearing in the exponential term of the likelihood. For simplicity of exposition, we first state the bound for $k = 2$.

Lemma 2 *For the case $k = 2$:*

$$\sum_{i=1}^{n}\{z_i - g_\eta(\pm\beta_1 + \beta_2)\}^2 \geq \sum_{i=1}^{n} T_i + nt'At + 2^{\eta+1}\sum |z_i|b't. \tag{14}$$

where $T_i = z_i^2 - (2^{\eta+1}+2)\,|z_i|\,/\eta$, and $t = (t_1, t_2)'$, A is a 2×2 matrix with $A(2,2) = -1$ and remaining entries 0 if $|\beta_1| \geq |\beta_2|$ and $A(1,1) = 1$ and remaining entries 0 if $|\beta_1| < |\beta_2|$, $b = (-1, 0)' \in \mathbb{R}^2$ if $t_1 \geq 0$ and as $(0, -1)'$ if $t_1 < 0$.

2. *For the case $k > 2$:*

$$\sum_{i=1}^{n}\{z_i - g_\eta(\pm\beta_1 + \beta_2 + ... + \beta_k)\}^2 \geq \sum_{i=1}^{n} T_i + nt'At + 2k^\eta\sum |z_i|b't. \tag{15}$$

with $T_i = z_i^2 - (2k^\eta+2)\,|z_i|\,/\eta$, $t_i = g_\eta(\beta_i)$and $t = (t_1, t_2, \ldots, t_k)'$. If β_j has the maximum absolute value for some $j \in (1, 2, \ldots, k)$, $A \in \mathbb{R}^{k\times k}$ is a matrix with $A_{k-j+1,k-j+1} = -1$ and remaining entries 0, $b \in \mathbb{R}^k$ with $b_j = -1$ if $t_j \geq 0$ and 1 if $t_j < 0$, all other elements of b are 0.

Proof The proof is provided for $|\beta_1| \geq |\beta_2|$ for $k = 2$. The proof $|\beta_1| < |\beta_2|$ and for general $k > 2$ follows along similar lines. Observe that,

$$\sum_{i=1}^{n}\left\{z_i - g_\eta(\pm\beta_1 + \beta_2)\right\}^2 = \sum_{i=1}^{n} z_i^2 - 2g_\eta(\pm\beta_1 + \beta_2)\sum_{i=1}^{n} z_i + n\{g_\eta(\pm\beta_1 + \beta_2)\}^2$$

If $|\beta_1| \geq |\beta_2|$, $|g_\eta(\pm\beta_1 + \beta_2)| = \eta^{-1}\,|\text{sgn}(\pm\beta_1 + \beta_2)| \pm \beta_1 + \beta_2|^\eta - 1| \leq \eta^{-1}\,|2^\eta|\beta_1|^\eta - 1| \leq$

$2^\eta |\beta_1|^\eta \eta^{-1} + \eta^{-1}$. Therefore

$$\begin{aligned} -2g_\eta(\pm\beta_1+\beta_2)\sum_{i=1}^n z_i &\geq -2\,|g_\eta(\pm\beta_1+\beta_2)|\sum_{i=1}^n |z_i| = -2\biggl(2^\eta|t_1| + 2^\eta\eta^{-1} + \eta^{-1}\biggr)\sum_{i=1}^n |z_i| \\ &= -2^{\eta+1}\sum_{i=1}^n |z_i|\,|t_1| - \biggl(2^{\eta+1}+2\biggr)|z_i|\,\eta^{-1}. \end{aligned} \quad (16)$$

Then $\sum_{i=1}^n \{z_i - g_\eta(\pm\beta_1+\beta_2)\}^2 = \sum_{i=1}^n z_i^2 - 2g_\eta(\pm\beta_1+\beta_2)\sum_{i=1}^n z_i + n\{g_\eta(\pm\beta_1+\beta_2)\}^2 \geq \sum_{i=1}^n z_i^2 - \left(\frac{2^{\eta+1}}{\eta} + \frac{2}{\eta}\right)\sum_{i=1}^n |z_i| - nt_2^2 - 2^{\eta+1}|t_1|\sum_{i=1}^n |z_i| = \sum_{i=1}^n T_i^2 + nt'At + 2^{\eta+1}b't\sum_{i=1}^n |z_i|$, where the second inequality follows from (16) and the trivial fact $n(g_\eta(\pm\beta_1+\beta_2))^2 \geq -nt_2^2$.