Biometrics 000, 000–000 DOI: 000
000 0000

Weighted Wilcoxon-type Smoothly Clipped Absolute Deviation Method

Lan Wang

School of Statistics, University of Minnesota, 224 Church Street SE, Minneapolis, MN 55455, U.S.A.

email: lan@stat.umn.edu

and

Runze Li

Department of Statistics and The Methodology Center, Pennsylvania State University,

University Park, PA 16802-2111, U.S.A.

email: rli@stat.psu.edu

Summary: Shrinkage-type variable selection procedures have recently seen increasing applications in biomedical research. However, their performance can be adversely influenced by outliers in either the response or the covariate space. This paper proposes a weighted Wilcoxon-type smoothly clipped absolute deviation (WW-SCAD) method, which deals with robust variable selection and robust estimation simultaneously. The new procedure can be conveniently implemented with the statistical software R. We establish that the WW-SCAD correctly identifies the set of zero coefficients with probability approaching one and estimates the nonzero coefficients with the rate $n^{-1/2}$. Moreover, with appropriately chosen weights the WW-SCAD is robust with respect to outliers in both the $\mathbf{x}$ and y directions. The important special case with constant weights yields an oracle-type estimator with high efficiency at the presence of heavier-tailed random errors. The robustness of the WW-SCAD is partly justified by its asymptotic performance under local shrinking contamination. We propose a BIC-type tuning parameter selector for the WW-SCAD. The performance of the WW-SCAD is demonstrated via simulations and by an application to a study that investigates the effects of personal characteristics and dietary factors on plasma beta-carotene level.

Key words: Leverage points; Rank-based analysis; Oracle property; Outlier; Smoothly clipped absolute deviation; Shrinking contamination; Robust estimation; Robust model selection; Wilcoxon method.

1. Introduction

In biomedical research, statisticians often need to analyze data sets with a non-normally distributed response variable and/or many covariates that potentially contain multiple high leverage points. This often imposes serious problems for variable selection and the subsequent inference. Existing work on robust variable selection are mostly robust best-subset procedures, such as robust AIC or BIC, see Ronchetti (1985), Hurvich and Tsai (1990), Burman and Nolan (1995), Ronchetti and Staudte (1994), Ronchetti, Field and Blanchard (1997), Wisnowski et al. (2003) and Müller and Welsh (2005), among others. The best-subset type procedures are computationally intensive even for moderately large number of covariates; and are known to have inherent instability (Brieman, 1996) due to their discrete nature. Moreover, these approaches in general are only robust against outliers in the response space but are still sensitive to high-leverage points. This paper introduces a novel unified framework called the weighted Wilcoxon-type smoothly clipped absolute deviation method (WW-SCAD, for short) for automatic robust variable selection and robust estimation that can effectively handle the above concerns.

In Section 4, we analyzed a data set from a study investigating the effects of personal characteristics and dietary factors on plasma beta-carotene level. It has been observed that low plasma concentrations of beta-carotene might be associated with increased risk of developing certain types of cancer. Due to the nature of the study, many patients have rather low plasma beta-carotene levels. This results in a long-tailed and highly skewed distribution for the response variable (plasma beta-carotene level, ng/ml), see the histogram depicted in Figure 1(a). Also, two of the ten covariates: x_8 (number of alcoholic drinks consumed per week) and x_9 (cholesterol consumed per day) clearly contain multiple outliers, some are even quite extreme, as revealed by their boxplots in Figure 1(b). This leads us to propose a procedure that is robust on both the covariate and response spaces to analyze this data

set. Some covariates may not have effects on the plasma beta-carotene level. Thus, it is of great interest to further develop variable selection procedure in robust statistical modeling. From the analysis in section 4.2, the newly proposed robust variable selection procedure reduces the median prediction error on the validation data to about 68% of that given by its nonrobust alternative.

[Figure 1 about here.]

The WW-SCAD procedure is motivated by recent developments in shrinkage-type variable selection procedures such as LASSO (Tibshirani, 1996) and SCAD (Fan and Li, 2001). Distinguished from the robust subset-type procedures, the WW-SCAD simultaneously selects covariates and estimates parameters by minimizing an objective function which is the sum of the weighted Wilcoxon-type dispersion function and the smoothly clipped absolute deviation (SCAD) penalty function, see Section 2.2. The penalty term shrinks the estimated small coefficients to zero, thus serves the purpose of variable selection.

The WW-SCAD is robust against outliers in both the $\mathbf{x}$ and y directions with appropriately chosen weights. This is different from the LAD-LASSO procedure based on the least absolute deviation regression (Wang, Li and Jiang, 2007) and the penalized composite quantile regression (Zou and Yuan, 2007), which provide a certain degree of protection against outliers in the response space but are vulnerable to high leverage points. We provide theoretical justification for the robustness of the WW-SCAD by studying its performance under shrinking local contamination. Under the local contamination, we reveal that the WW-SCAD still identifies zero coefficients with probability approaching one and estimates nonzero coefficients with a bias bounded in $(\mathbf{x}, y)$ when the weights are appropriately chosen.

The WW-SCAD with constant weights leads to an important special case that is closely related to the classical Wilcoxon inference based on Jaeckel's (1972) dispersion function with Wilcoxon scores. In this case, with a proper tuning parameter the resulted estimator possesses

the oracle property (Fan and Li, 2001) and often significantly improves the efficiency of the LS-SCAD (least-squared procedure with SCAD penalty) in the presence of heavy-tailed errors. The tuning parameter in the WW-SCAD controls the model complexity and plays an important role in the variable selection procedure. In practice, it is desirable to select the tuning parameter using a data-driven method. We propose a BIC-type tuning parameter selector and show that with probability tending to one, the WW-SCAD with the BIC-selector can identify the most parsimonious correct model.

Rank-based statistical procedures have wide applications in biomedical research due to their robustness and high efficiency; see Jin et al. (2003), Jung and Ying (2003), Mahfoud and Randles (2005), Rosner, Glynn and Lee (2006a, 2006b), Heller (2007), Datta and Satten (2008), Wang and Zhao (2008) and the references therein. However, the aforementioned work mainly focuses on estimation and hypothesis testing. Our proposal therefore extends rank-based nonparametric analysis to the important area of variable selection.

The rest of the paper is structured as follows. In the next section, we introduce the WW-SCAD procedure and discuss its implementation via the software package R. In Section 3, we establish the asymptotic normality and consistency of selection, and provide justification for robustness by considering the asymptotic distribution under local contamination. Furthermore, we introduce a BIC-type procedure for selecting the tuning parameter. In Section 4, we demonstrate the performance of the WW-SCAD by Monte Carlo studies and apply it to analyze the plasma beta-carotene level data set. Section 5 summarizes the paper.

2. Weighted Wilcoxon-type Smoothly Clipped Absolute Deviation Method

Consider a linear regression model

$$\mathbf{Y} = \alpha \mathbf{1}_n + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ is an $n \times 1$ vector of responses, α is the intercept, $\mathbf{1}_n$ is an $n \times 1$ vector of ones, $\mathbf{X}$ is an $n \times d$ matrix of covariates which without loss of generality is assumed to be centered, $\boldsymbol{\beta}$ is a $d \times 1$ vector of unknown parameters, and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of independent, identically distributed random errors with probability density function $f(\cdot)$. We assume that some components of $\boldsymbol{\beta}$ are zero in the true model. The goal of our work is to identify the zero coefficients consistently and robustly, and to estimate the nonzero coefficients efficiently and robustly.

2.1 *The WW-SCAD*

The penalized weighted Wilcoxon method estimates $\boldsymbol{\beta}$ by minimizing

$$n^{-1} \sum_{i<j} b_{ij} |e_i - e_j| + n \sum_{j=1}^{d} p_\lambda(|\beta_j|),$$

where the b_{ij}'s are positive and symmetric weights, $e_i = Y_i - \mathbf{x}_i'\boldsymbol{\beta}$ with $\mathbf{x}_i$ being the ith row of $\mathbf{X}$, $p_\lambda(\cdot)$ is a penalty function and λ is a tuning parameter controlling the complexity of the model. In Section 3.4, we propose a data-driven method to select λ. In our asymptotic analysis, we write λ as λ_n to emphasize its dependence on the sample size n.

Directly minimizing $n^{-2} \sum_{i<j} b_{ij} |e_i - e_j|$, a weighted version of Gini's mean difference measure of variability, yields the generalized rank estimator (GR estimator), see Sievers (1983), Naranjo and Hettmansperger (1994), Chang, McKean, Naranjo, and Sheather (1999), Terpstra, McKean and Naranjo (2001), among others. When b_{ij} are constant, minimizing $n^{-2} \sum_{i<j} b_{ij} |e_i - e_j|$ is equivalent to minimizing Jaeckel's (1972) Wilcoxon-type dispersion function $\sqrt{12} \sum_{i=1}^{n} \left[\frac{R(Y_i - \mathbf{x}_i'\boldsymbol{\beta})}{n+1} - \frac{1}{2} \right] (Y_i - \mathbf{x}_i'\boldsymbol{\beta})$, where $R(Y_i - \mathbf{x}_i'\boldsymbol{\beta})$ denotes the rank of $Y_i - \mathbf{x}_i'\boldsymbol{\beta}$ among $Y_1 - \mathbf{x}_1'\boldsymbol{\beta}, \ldots, Y_n - \mathbf{x}_n'\boldsymbol{\beta}$.

Fan and Li (2001) provided deep insights into the principles of choosing an appropriate penalty function. They proposed the smoothly clipped absolute deviation (SCAD) penalty

function, which satisfies $p_\lambda(0) = 0$ and has the first-order derivative

$$p'_\lambda(\theta) = \lambda \left\{ I(\theta \leq \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda) \right\} \tag{1}$$

for some $a > 2$ and $\theta > 0$. Following Fan and Li (2001), we will use $a = 3.7$ throughout this paper. Recently, Zou and Li (2007) proposed a local linear approximation to the SCAD penalty, which retains the same asymptotic properties and at the same time significantly improves the computational efficiency of Fan and Li's LS-SCAD. Adopting this idea, we propose a WW-SCAD procedure for robust simultaneous variable selection and estimation. Formally, the WW-SCAD method estimates $\boldsymbol{\beta}$ by minimizing

$$n^{-1} \sum_{i<j} b_{ij} |e_i - e_j| + n \sum_{j=1}^{d} p'_\lambda(|\beta_j^0|)|\beta_j|, \tag{2}$$

where $p'_\lambda(\cdot)$ is defined in (1), $\sum_{j=1}^{d} p'_\lambda(|\beta_j^0|)|\beta_j|$ is the linearized SCAD penalty (Zou and Li, 2007) and $\boldsymbol{\beta}^0$ is an initial estimator, which we set to be the unpenalized weighted Wilcoxon estimator. Unlike the objective function for the LS-SCAD, the objective function defined in (2) is convex in $\boldsymbol{\beta}$.

We complete this subsection with a brief discussion of estimating the intercept parameter α. Since (2) is invariant to a location change, α cannot be estimated simultaneously with $\boldsymbol{\beta}$. Instead, α is estimated based on $e_i(\widehat{\boldsymbol{\beta}}) = Y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}$, $i = 1, \ldots, n$. A common practice is to estimate α by the median of the $e_i(\widehat{\boldsymbol{\beta}})$'s, see for example Section 3.5.2 of Hettmansperger and McKean (1998).

2.2 *Computation*

An appealing feature of the WW-SCAD is that its computation can be conveniently carried out using the statistical software R. Our algorithm is similar to that of the LAD-LASSO (Wang, Li and Jiang, 2007). The key observation is that minimizing (2) can be achieved by fitting an L_1 regression model based on the pseudo observations $(\mathbf{x}_k^*, Y_k^*)$, $k = 1, \ldots, \frac{n(n-1)}{2} + d$. The first $n(n-1)/2$ pseudo observations correspond to $(b_{ij}(\mathbf{x}_j - \mathbf{x}_i), b_{ij}(Y_j - Y_i))$, for

$1 \le i < j \le n$, and the last d pseudo observations correspond to $(p'_\lambda(|\beta_j^0|)\psi_j, 0)$, where ψ_j is the d-dimensional vector with the jth component being one and all other components being zeros.

The unpenalized weighted Wilcoxon estimator β_j^0 can obtained using the function *wwfit* in the R software developed by Terpstra and McKean (2005) for rank regression (downloadable from http://www.stat.wmich.edu/mckean/HMC/Rcode). And the L_1 regression model can be fitted using the R package *quantreg* by Roger Koenker for quantile regression.

Remark. It is worth emphasizing that in order to achieve practical robustness it is not sufficient to merely have a robust objective function. The algorithm itself is also of critical importance. For shrinkage-type procedures, special algorithms are often used for the implementation and most of these algorithms are sensitive to outlier contamination. As an example, if we approximate the WW-SCAD objective function quadratically and then apply the LARS algorithm (Efron et al., 2004), the resulting procedure is likely to be still sensitive to outliers.

3. Asymptotic Properties

3.1 *Notations and assumptions*

We consider Mallows-type weights b_{ij} that possibly depend on the covariates in the form $b_{ij} = b(\mathbf{x}_i, \mathbf{x}_j)$. In the simulations and data analysis, we adopt the GR weights (Chang, et al., 1999, Terpstra and McKean, 2005): $b_{ij} = h(\mathbf{x}_i)h(\mathbf{x}_j)$, where

$$h(\mathbf{x}_i) = \min\left\{1, \frac{b}{(\mathbf{x}_i - \widehat{\mu})' S^{-1} (\mathbf{x}_i - \widehat{\mu})}\right\},$$

with $(\widehat{\mu}, S)$ being the robust minimum volume ellipsoid (MVE) estimator of the location and scatter (Rousseeuw and van Zomeren, 1990), and b being the 95th percentile of $\chi^2(d)$.

Following the notations in Naranjo and Hettmansperger (1994), let $\mathbf{W}$ be an $n \times n$ matrix

of elements w_{ij}, where

$$w_{ij} = \begin{cases} -n^{-1}b_{ij}, & i \neq j \\ n^{-1}\sum_{k\neq i} b_{ik}, & i = j. \end{cases}$$

Assume $n^{-1}\mathbf{X}'\mathbf{W}\mathbf{X} \xrightarrow{p} \mathbf{C}$, $n^{-1}\mathbf{X}'\mathbf{W}^2\mathbf{X} \xrightarrow{p} \mathbf{V}$ and $n^{-1}\mathbf{X}'\mathbf{X} \xrightarrow{p} \mathbf{\Sigma}$, where $\mathbf{C}$, $\mathbf{V}$ and $\mathbf{\Sigma}$ are positive definite matrices:

$$\begin{aligned} \mathbf{C} &= \frac{1}{2}\iint (\mathbf{x}_2 - \mathbf{x}_1)(\mathbf{x}_2 - \mathbf{x}_1)' b(\mathbf{x}_1, \mathbf{x}_2) dM(\mathbf{x}_2) dM(\mathbf{x}_1), \\ \mathbf{V} &= \int \{(\mathbf{x}_2 - \mathbf{x}_1) b(\mathbf{x}_2, \mathbf{x}_1) dM(\mathbf{x}_2)\} \{(\mathbf{x}_2 - \mathbf{x}_1) b(\mathbf{x}_2, \mathbf{x}_1) dM(\mathbf{x}_2)\}' dM(\mathbf{x}_1), \\ \mathbf{\Sigma} &= \frac{1}{2}\iint (\mathbf{x}_2 - \mathbf{x}_1)(\mathbf{x}_2 - \mathbf{x}_1)' dM(\mathbf{x}_2) dM(\mathbf{x}_1), \end{aligned}$$

and $M(\mathbf{x})$ denotes the cumulative distribution function of $\mathbf{x}$.

We denote the true value of $\boldsymbol{\beta}$ by $\boldsymbol{\beta}_0 = (\beta_{10}, \ldots, \beta_{d0})' = (\boldsymbol{\beta}_{10}', \boldsymbol{\beta}_{20}')'$. Without loss of generality, we assume that $\boldsymbol{\beta}_{20} = \mathbf{0}$ and that the elements of $\boldsymbol{\beta}_{10}$ are all nonzero. We also assume the dimension of $\boldsymbol{\beta}_{10}$ is s $(1 \leq s \leq d)$. Let $\mathbf{X}_1$ be the first s columns of $\mathbf{X}$ that correspond to $\boldsymbol{\beta}_{10}$, and write

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix}, \quad \mathbf{V} = \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{bmatrix} \quad \text{and} \quad \mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix}.$$

In addition to the above, we assume that the error density function $f(\cdot)$ is absolutely continuous with finite Fisher information, i.e., $\int \{f(x)\}^{-1} f'(x)^2 dx < \infty$. And $\mathbf{X}$ and $\mathbf{WX}$ both satisfy Huber's condition, a sufficient and necessary condition for the least-squares estimator to have an asymptotic normal distribution; see condition (D.2) of Hettmansperger and McKean (1998). Under these conditions, the unpenalized WW estimator is $\sqrt{n}$-consistent for $\boldsymbol{\beta}_0$ and asymptotically normal.

3.2 *Asymptotic properties of the WW-SCAD*

Theorem 1 below presents the asymptotic property of the WW-SCAD as a simultaneous model selection and parameter estimation tool, and its proof is given in the Web Appendix.

THEOREM 1: *Assume the regularity conditions in Section 3.1. If* $\lambda_n \to 0$ *and* $\sqrt{n}\lambda_n \to \infty$ *as* $n \to \infty$*, then the WW-SCAD estimator* $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_{10}', \widehat{\boldsymbol{\beta}}_{20}')'$ *must satisfy that* $P(\widehat{\boldsymbol{\beta}}_{20} = \mathbf{0}) = 1$*, and*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{10} - \boldsymbol{\beta}_{10}) \to N_s(\mathbf{0}, \tau^2 \mathbf{C}_{11}^{-1} \mathbf{V}_{11} \mathbf{C}_{11}^{-1}),$$

where $\tau = [\sqrt{12} \int f^2(u)du]^{-1}$.

The case with constant weight $b_{ij} \equiv 1$ is particularly important due to its simplicity and its close connection with the familiar Wilcoxon inference. In this case, we have $\mathbf{C}_{11} = \mathbf{V}_{11} = \boldsymbol{\Sigma}_{11} = \mathbf{X}_1' \mathbf{X}_1$, thus we have the following corollary.

COROLLARY 1: *Assume the conditions in Theorem 3.1, then when* $b_{ij} \equiv 1$*,* $\widehat{\boldsymbol{\beta}}$ *satisfies:* $P(\widehat{\boldsymbol{\beta}}_{20} = \mathbf{0}) = 1$ *and* $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{10} - \boldsymbol{\beta}_{10}) \to N_s(\mathbf{0}, \tau^2 \boldsymbol{\Sigma}_{11}^{-1})$.

Corollary 1 suggests that the Wilcoxon-SCAD, with $b_{ij} \equiv 1$ and a properly chosen tuning parameter, possesses the oracle property (Fan and Li, 2001). That is, with probability approach one, the WW-SCAD can correctly identify the nonzero coefficients, and estimate them as efficiently as the unpenalized WW rank regression does as if the true model were known in advance. Moreover, the WW-SCAD can be more efficient than the LS-SCAD for estimating $\boldsymbol{\beta}_{10}$ in the presence of heavier-tailed errors. It is easy to show that the asymptotic relative efficiency is $ARE = 12\sigma^2 \left[\int f^2(u)du\right]^2$.

Remark 1. This asymptotic relative efficiency is the same as that of the one-sample Wilcoxon test with respect to the t-test. It is well known in the literature of rank analysis that the ARE is as high as 0.955 for normal error distribution, and can be significantly higher than 1 for many heavier-tailed distributions. For instance, $ARE = 1.5$ for the double exponential distribution, and $ARE = 1.9$ for the t distribution with 3 degrees of freedom. For symmetric error distributions with finite Fisher information, this asymptotic relative efficiency is known to have a lower bound equal to 0.864.

Remark 2. The asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}_{10}$ in Corollary 1 can be shown to be equivalent to that in Theorem 2.1 of Zou and Yuan (2007) for composite quantile regression when K, the number of quantiles, goes to infinity. The composite quantile regression, however, is more computationally involved. Its objective function involves a mixture of quantile objective functions at different quantiles (the suggested value of K for practical use is 19). As a result, besides the regression parameters one also needs to estimate K additional parameters corresponding to K different quantiles of the error distribution.

With nonconstant weights b_{ij}, the WW-SCAD still consistently selects the correct model; however the asymptotic covariance matrix of $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{10} - \boldsymbol{\beta}_{10})$ is slightly different from what one would obtain if the true model were known in advance. This can be seen by observing that in $\mathbf{C}_{11} = \mathbf{X}_1'\mathbf{W}\mathbf{X}_1$ (and similarly $\mathbf{V}_{11}$), the matrix $\mathbf{W}$ involves all d covariates; while if the true model were known then $\mathbf{W}$ would only use s covariates. Indeed, for the WW-SCAD to work as a model selection criterion, it is necessary to allow the weights to depend on all candidate covariates.

3.3 *Asymptotics under local shrinking contamination*

Now we study the robustness property of the WW-SCAD. For robust estimation and hypothesis testing, the influence function approach offers a convenient and essential way to investigate the local robustness (Hampel, 1974, Hampel et al., 1986). This approach, however, is not adequate in the current setting as we perform variable selection and parameter estimation simultaneously.

Following the spirit of influence function approach, we directly study the performance of the WW-SCAD under infinitesimal contamination. Specifically, we consider the following local shrinking contamination as in Heritier and Ronchetti (1994):

$$H_n^*(\mathbf{x}, y) = \left(1 - \frac{\delta}{\sqrt{n}}\right) H(\mathbf{x}, y) + \frac{\delta}{\sqrt{n}} \Delta_{(\mathbf{x}^*, y^*)}, \tag{3}$$

where $H(\mathbf{x}, y)$ is the joint cumulative distribution function of the underlying distribution without contamination, $\Delta_{(\mathbf{x}^*,y^*)}$ represents the point mass at $(\mathbf{x}^*, y^*)$ and δ is a constant.

THEOREM 2: *Assume the regularity conditions in Section 3.1. If* $\lambda_n \to 0$ *and* $\sqrt{n}\lambda_n \to \infty$ *as* $n \to \infty$*, then under local shrinking contamination (3), the WW-SCAD estimator* $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_{10}', \widehat{\boldsymbol{\beta}}_{20}')'$ *must satisfy that* $P(\widehat{\boldsymbol{\beta}}_{20} = \mathbf{0}) = 1$*, and*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{10} - \boldsymbol{\beta}_{10}) \to N_s(\eta, \tau^2 \mathbf{C}_{11}^{-1} \mathbf{V}_{11} \mathbf{C}_{11}^{-1}),$$

where $\eta = \delta[2F(y^* - \mathbf{x}^*\boldsymbol{\beta}_0) - 1] \int b(\mathbf{x}^*, \mathbf{x})(\mathbf{x}^* - \mathbf{x}) dM(\mathbf{x})$.

The proof of Theorem 2 is given in the Web Appendix. Theorem 2 indicates that under the local contamination (3), the WW-SCAD can still correctly identifies the set of zero coefficients with probability tending to one; but the contamination introduces a bias η in estimating the nonzero coefficients. Note that $[2F(y^* - \mathbf{x}^*\boldsymbol{\beta}_0) - 1] \int b(\mathbf{x}^*, \mathbf{x})(\mathbf{x}^* - \mathbf{x}) dM(\mathbf{x})$ is also the core part of the influence function of the unpenalized weighted Wilcoxon estimator. The bias η is bounded in y^*. With proper choice of weights b_{ij}, such as the GR weights introduced in Section 3.1, η is also bounded in $\mathbf{x}^*$ (Naranjo and Hettmansperger, 1994, Chang, et al. 1999).

3.4 *Data-driven tuning parameter selection*

The tuning parameter λ controls the model complexity and plays a critical role in the WW-SCAD procedure. It is desirable to select λ automatically by a data-driven method. Here we propose to select λ for the WW-SCAD by minimizing

$$BIC_\lambda = \log\left(n^{-2} \sum_{i<j} b_{ij} |(Y_i - \mathbf{x}_i' \widehat{\boldsymbol{\beta}}_\lambda) - (Y_j - \mathbf{x}_j' \widehat{\boldsymbol{\beta}}_\lambda)| \right) + df_\lambda \log(n)/n \tag{4}$$

over an interval $[0, \lambda_{max}]$, where $\widehat{\boldsymbol{\beta}}_\lambda$ is the WW-SCAD estimator with tuning parameter λ, and df_λ is the number of nonzero components in $\widehat{\boldsymbol{\beta}}_\lambda$. It is assumed that $\lambda_{max} \to 0$ as $n \to \infty$. We refer to this approach as the BIC-selector, and denote the selected λ by $\hat{\lambda}_{BIC}$. It is

worth noting that the BIC-selector is different from the traditional BIC best subset variable selection procedure.

To introduce the property of the BIC tuning parameter selector, we next define some notations. We use $S = \{j_1, \ldots, j_{d^*}\}$, the set of the indices of the covariates in the model, to denote a given candidate model. Let S_T denote the true model, let S_F denote the full model, and let S_λ denote the set of the indices of the covariates selected by WW-SCAD with tuning parameter λ.

For a given candidate model S, let $\boldsymbol{\beta}_S$ be the vector of parameters. The i-th coordinate of $\boldsymbol{\beta}_S$ is set to be zero if $i \notin S$. Further, define $L_n^S = n^{-2} \sum_{i<j} b_{ij} |(Y_i - \mathbf{x}_i' \widehat{\boldsymbol{\beta}}_S) - (Y_j - \mathbf{x}_j' \widehat{\boldsymbol{\beta}}_S)|$, where $\widehat{\boldsymbol{\beta}}_s$ is the unpenalized weighted Wilcoxon estimator for model S. To demonstrate that the BIC-selector can identify the true model consistently, we assume

(1) for any $S \subset S_F$, $L_n^S \xrightarrow{p} L^S$ for some $L^S > 0$, where $\xrightarrow{p}$ means converges in probability;
(2) for any $S \not\supset S_T$, we have $L^S > L^{S_T}$.

Note that L_n^S is the objective function to obtain the weighted Wilcoxon estimator when model S is used. Conditions (1) and (2) are standard for investigating parameter estimation under model misspecification, see White (1981). Let $R(\boldsymbol{\beta}) = 0.5 \iint b(\mathbf{x}_1, \mathbf{x}_2) |(y_1 - \mathbf{x}_1' \boldsymbol{\beta}) - (y_2 - \mathbf{x}_2' \boldsymbol{\beta})| dH(\mathbf{x}_1, y_1) dH(\mathbf{x}_2, y_2)$. Then for the true model S_T, $L^{S_T} = R(\boldsymbol{\beta}_0)$ where $\boldsymbol{\beta}_0$ is the true parameter and minimizes $R(\boldsymbol{\beta})$ under the full model; and for a general model S, $L^S = R(\boldsymbol{\beta}_{0s})$ where $\boldsymbol{\beta}_{0s}$ minimizes $R(\boldsymbol{\beta})$ under model S.

THEOREM 3: *Assume the conditions above and the regularity conditions in Section 3.1, then* $P(S_{\hat{\lambda}_{BIC}} = S_T) \to 1$.

The proof of Theorem 3 is given in the Web Appendix. Theorem 3 indicates that $\hat{\lambda}_{BIC}$ leads to a WW-SCAD estimator which consistently yields the true model. The verification of

this theorem is similar to that in Wang, Li and Tsai (2007), in which the authors proposed a novel BIC-selector for the SCAD penalized least squares procedures.

4. Numerical Examples

4.1 *Simulation study*

In the literature, the LS-SCAD has been compared with the nonnegative garrote (Breiman, 1995), the LASSO, and the best subset variable selection procedures such as AIC or BIC, see for example, Fan and Li (2001) and Zou and Li (2007). Our simulations are designed to demonstrate the robustness and the efficiency of the WW-SCAD, compared with the LS-SCAD which is computed with the BIC tuning parameter selector of Wang, Li and Tsai (2007). We also compare with the benchmark oracle procedure, which sets the estimate of zero coefficients to be zero and estimates the nonzero coefficients by excluding the covariates of zero coefficients.

We focus on examining the performance of the WW-SCAD in terms of model complexity and model errors (ME) defined by

$$\mathrm{ME}(\widehat{\boldsymbol{\beta}}) = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)' E(\mathbf{x}_1 \mathbf{x}_1')(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0). \tag{5}$$

Example 1. As in Tibshirani (1996) and Fan and Li (2001), data are generated from

$$Y_i = \mathbf{x}_i' \boldsymbol{\beta} + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, 100, \tag{6}$$

where $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)$ and $\mathbf{x}_i = (x_1, \ldots, x_8)' \sim N_8(0, \Omega)$, in which the (i, j)th element of Ω equals $0.5^{|i-j|}$ for $1 \leq i, j \leq 8$. We consider three different error distributions: the standard normal distribution, the t distribution with 3 degrees of freedom, and a contaminated standard normal distribution with 10% outliers from the standard Cauchy distribution. For each case, we conduct 500 simulations.

Simulation results are summarized in Table 1, in which we report the average number of correct 0's (the average number of the five true zero coefficients that are correctly estimated

to be zero) and the average number of incorrect 0's (the average number of the three true zero coefficients that are incorrectly estimated to be zero). We also report the proportion of correctly fitted models. To evaluate the lack-of-fit of the selected model, we report the relative model error RME=ME/$\text{ME}_{\text{WilFull}}$, where $\text{ME}_{\text{WilFull}}$ is the ME for fitting the full model with unpenalized Wilcoxon rank regression.

[Table 1 about here.]

¿From Table 1, we can see that the median of the RME of the WW-SCAD is close to that of the WW_{oracle}, the weighted Wilcoxon estimator from the oracle procedure. In terms of model error, the performance of the WW-SCAD is similar to the LS-SCAD for normal error, but much better than the LS-SCAD for both t_3 error and contaminated normal error in terms of model error. And the WW-SCAD gives significantly higher percentage of correctly fitted 0's compared to the LS-SCAD.

Example 2. We now investigate the effect of outliers in the $\mathbf{x}$ direction on model selection. For this purpose, we consider the same regression model (6) with the standard normal random errors. We consider a contamination of the covariate $\mathbf{x}$ by replacing a random 5% of $\mathbf{x}$ with $\mathbf{x} + \mathbf{e}$, where $\mathbf{e} = (e_1, \ldots, e_8)'$, with e_3 having a $N(0, 5)$ distribution and all the other e_i's having independent $N(0, 1)$ distributions. For the WW-SCAD procedure, we consider both the Wilcoxon weights and the GR weights.

In this example, the relative model error is defined as RME=ME/$\text{ME}_{\text{GRFull}}$, where $\text{ME}_{\text{GRFull}}$ is the model error obtained by fitting the full model with the unpenalized weighted Wilcoxon procedure and the GR weights. We use the weighted Wilcoxon procedure with the GR weights under the true mode as the benchmark here. The simulation results are summarized in Table 2, from which we observe that the GR weights lead to model selection procedures robust to outliers in the $\mathbf{x}$ direction; in contrast the performance of the LS-SCAD is adversely affected.

We also note that the WW-SCAD with the Wilcoxon weights is not as seriously affected as the LS-SCAD but does not perform as well as the WW-SCAD with the GR weights.

[Table 2 about here.]

4.2 *Analysis of plasma beta-carotene level data*

Observational studies have suggested that low plasma concentration of beta-carotene might be associated with increased risk of developing certain types of cancer. We consider a data set from a cross-sectional study that consists of 273 female patients who had an elective surgical procedure during a three-year period to biopsy or remove a lesion of the lung, colon, breast, skin, ovary or uterus that was found to be non-cancerous (Nierenberg et al., 1989). The response variable y is the plasma beta-carotene level (ng/ml) and there are ten covariates: x_1 is age, x_2 is smoking status (1=never, 2=former smoker, 3=current smoker), x_3 is quetelet (weight/height^2), x_4 denotes vitamin use (1=yes, fairly often, 2=yes, not often, 3=no), x_5 is the number of calories consumed per day, x_6 is grams of fat consumed per day, x_7 is grams of fiber consumed per day, x_8 is number of alcoholic drinks consumed per week, x_9 is cholesterol consumed (mg per day) and x_{10} is dietary beta-carotene consumed (mcg per day).

As revealed by Figure 1, the distribution of y is highly skewed, while x_8 and x_9 contain some obvious outliers. One may suggest log transform the response variable. However, our preliminary analysis indicates that the log transformed y is still nonnormal. And it becomes even harder to find an appropriate transformation for x_8, which is on ordinal scale. Since the transformation may not remove the outliers and often brings additional issues for interpretability, we choose to analyze the variables on their original scale.

We use the first 200 observations as a training data set to select and fit the model, and use the rest as a validation data set to evaluate the prediction ability (measured by the median absolute prediction error) of the selected model. The λ values selected by the BIC criterion are 1.249, 2.834 and 3.028 for the LS-SCAD, the WW-SCAD with the Wilcoxon

score, and the WW-SCAD with the GR score, respectively. The resulting estimated models are displayed in Table 3. The LS-SCAD does not exclude any of the ten candidate covariates from the selected model. The WW-SCAD with either the Wilcoxon score or the GR score fits a much more succinct model that includes x_5 and x_{10}, and suggests that increased plasma beta-carotene level is associated with increased dietary intake of beta-carotene and reduced number of calories consumed per day. The WW-SCAD with the Wilcoxon score also includes x_9.

In terms of the prediction performance on the validation data, the WW-SCAD with either the Wilcoxon score or the GR score yields a much smaller median absolute prediction error. The median absolute prediction error of the WW-SCAD with the GR score is only 68% of that given by the LS-SCAD.

[Table 3 about here.]

5. Summary

We propose a novel robust framework called WW-SCAD for simultaneous variable selection and parameter estimation. This new procedure can be conveniently implemented using the statistical software R. It is much less computationally intensive compared with the best subset type procedures. With appropriately chosen weights, the WW-SCAD procedure can effectively handle outliers in both the $\mathbf{x}$ and y directions. Moreover, it loses very little efficiency with normal data and can be much more efficient than the LS-SCAD at the presence of heavier-tailed random errors. Although we have focused on studying the SCAD penalty, without any difficulty our method can be extended with other penalty functions.

Acknowledgements

The authors would like to thank the Co-Editor for constructive comments that substantially improved an earlier draft, and thank Dr. John Dziak for his help on presentation. Wang's research is supported by National Science Foundation grant DMS-0706842; and Li's research is supported by National Institute on Drug Abuse, NIH, 1 R21 DA024260, and National Science Foundation grant DMS-0348869.

Supplementary Materials

Wed Appendix referenced in Section 3 and the R codes for implementing the WW-SCAD procedure are available under the Paper Information link at the the Biometrics website `http://www.tibs.org/biometrics`.

References

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* **37**, 373-384.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics* **24**, 2350-2383.

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association* **80**, 580-598.

Burman, P. and Nolan, D. (1995). A general Akaike-type criterion for model selection in robust regression. *Biometrika* **82**, 877-886.

Chang, W. H., McKean J. W., Naranjo, J. D., and Sheather, S. J. (1999). High-breakdown rank regression. *Journal of the American Statistical Association* **94**, 205219.

Datta, S. and Satten, G. A. (2008). A signed-rank test for clustered data. *Biometrics*, to appear.

Efron, B., Hastie, T., Johnstone. I., and Tibshirani. R. (2004). Least angle regression. *The Annals of Statistics* **32**, 407-499.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348-1360.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association* **69**, 383-393.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York: John Wiley.

Heller, G. (2007). Smoothed rank regression with censored data. *Journal of the American Statistical Association* **102**, 552-559.

Heritier, S., and Ronchetti, E. (1994). Robust bounded-influence tests in general parametric models. *Journal of the American Statistical Association* **89**, 897-904.

Hettmansperger, T. P. and McKean, J. W. (1998). *Robust Nonparametric Statistical Methods*. London: Arnold.

Hurvich, C. M. and Tsai, C-L. (1990). Model selection for least absolute deviations regression in small samples. *Statistics & Probability Letters* **9**, 259-265.

Jaeckel, L. A. (1972). Estimating regression coefficients by minimizing the dispersion of the residuals. *The Annals of Mathematical Statistics* **43**, 1449-1458.

Jin, Z., Lin, D. Y., Wei, L. J., and Ying, Z. (2003). Rank-based inference for the accelerated failure time model. *Biometrika* **90**, 341-353.

Jung, S. H. and Ying, Z. L. (2003). Rank-based regression with repeated measurements data. *Biometrika* **90**, 732-740.

Mahfouda, Z. R. and Randles, R. H. (2005). Practical tests for randomized complete block designs. *Journal of Multivariate Analysis* **96**, 73-92.

Müller, S. and Welsh, A. H. (2005). Outlier robust model selection in linear regression.

Journal of the American Statistical Association **100**, 1297-1310.

Naranjo, J. D. and Hettmansperger, T. P. (1994). Bounded influence rank regression. *Journal of the Royal Statistical Society, Series B* **56**, 209-220.

Nierenberg, D. W., Stukel, T. A., Baron, J. A., Dain, B. J., and Greenberg, E. R. (1989). Determinants of plasma levels of beta-carotene and retinol. *American Journal of Epidemiology* **130**, 511-521.

Ronchetti, E. (1985). Robust model selection in regression. *Statistics & Probability Letters* **3**, 21-23.

Ronchetti, E. and Staudte, R. G. (1994). A robust version of Mallows' C_p. *Journal of the American Statistical Association* **89**, 550-559

Ronchetti, E., Field, C., and Blanchard, W. (1997). Robust linear model selection by cross-validation. *Journal of the American Statistical Association* **92**, 1017-1023.

Rosner, B., Glynn, R. J., and Lee, M-L. T. (2006a). The Wilcoxon signed rank test for paired comparisons of clustered data. *Biometrics* **62**, 185192

Rosner, B., Glynn, R. J., and Lee, M-L. T. (2006b). Extension of the rank sum test for clustered data: two-group comparisons with group membership defined at the subunit level. *Biometrics* **62**, 12511259.

Rousseeuw, P. J. and van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association* **85**, 633-639.

Sievers, G. L. (1983). A weighted dispersion function for estimation in linear models. *Communications in Statistics, Theory and Methods* **12**, 11611179.

Terpstra, J. and McKean, J. (2005). Rank-Based Analyses of Linear Models using R. *Journal of Statistical Software*, Volume 14, Issue 7.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B* **58**, 267-288.

Wang, H., Li, G., and Jiang, G. (2007). Robust regression shrinkage and consistent variable selection via the LAD-LASSO. *Journal of Business & Economics Statistics* **25**, 347-355.

Wang, H., Li, R., and Tsai, C. L. (2007). Tuning parameter selector for SCAD. *Biometrika* **94**, 553-568.

Wang, L. and Li, R. (2007). Weighted Wilcoxon-type smoothly clipped absolute deviation method. Technical report # 665, School of Statistics, University of Minnesota.

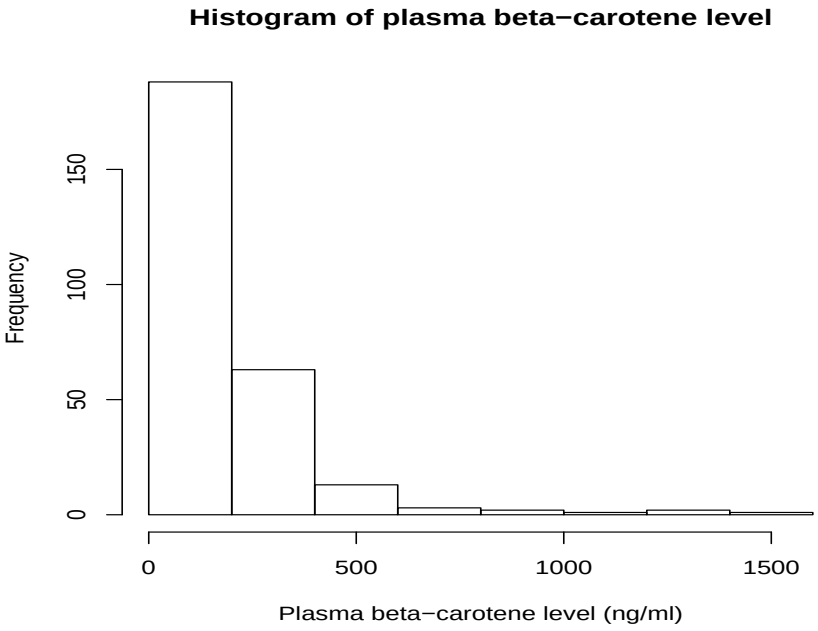
Wang, Y. G. and Zhao, Y. D. (2008). Weighted rank regression for clustered data analysis. *Biometrics* **64**, 3945.

White, H. (1981). Consequences and detection of misspecified nonlinear regression models. *Journal of the American Statistical Association* **76**, 419-433.

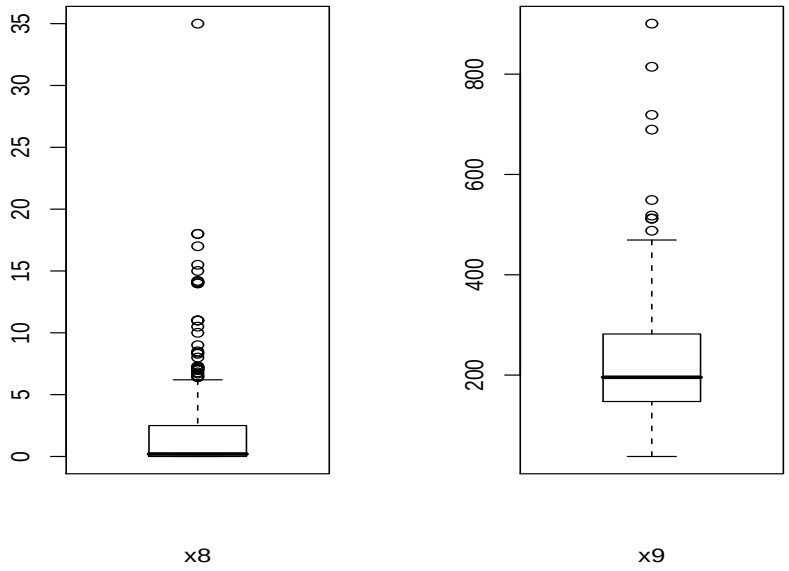
Wisnowski, J. W., Simpson, J. R., Montgomery, D. C., and Runger, G. C. (2003). Resampling methods for variable selection in robust regression. *Computational Statistics and Data Analysis* **43**, 341-355

Zou, H. and Li, R. (2007). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, to appear.

Zou, H. and Yuan, M. (2007). Composite quantile regression and the oracle model selection theory. *The Annals of Statistics*, to appear.



(a)



(b)

Figure 1: Plasma beta-carotene level data: (a) histogram of y, (b) boxplots of x_8 (number of alcoholic drinks consumed per week) and x_9 (cholesterol consumed per day)

Table 1: The simulation results are based on 500 runs. C is the average number of correct zeros; IC is the average number of incorrect zeros; Correct Fit (%) is the proportion of times the correct model is selected; and MRME is the median of relative model error.

Error Distribution	Method	No. of Zeros C	No. of Zeros IC	Correct Fit (%)	MRME (%)
normal	WW-SCAD	4.42	0	68.5	43.8
	LS-SCAD	4.32	0	60.0	40.6
	$\text{WW}_{\text{Oracle}}$	5	0	100	39.8
t_3	WW-SCAD	4.46	0	73.0	40.5
	LS-SCAD	4.33	0	63.5	64.6
	$\text{WW}_{\text{Oracle}}$	5	0	100	35.9
contaminated normal	WW-SCAD	4.48	0	67.5	40.6
	LS-SCAD	4.10	0	49.5	92.7
	$\text{WW}_{\text{Oracle}}$	5	0	100	37.0

Table 2: The simulation results are based on 500 runs. C is the average number of correct zeros; IC is the average number of incorrect zeros; Correct Fit (%) is the proportion of times the correct model is selected; and MRME is the median of relative model error.

Method	No. of Zeros		Correct Fit (%)	MRME (%)
	C	IC		
WW-SCAD (GR)	4.58	0	78.0	39.4
WW-SCAD (Wil)	4.51	0	73.0	44.4
LS-SCAD	3.61	0	39.5	100.3
$\text{WW}_{\text{Oracle}}$ (GR)	5	0	100	37.0

Table 3: Analysis of plasma beta-carotene level data. Note: The median absolute prediction error is calculated from the validation data set: median APE=median $\{|y_i - \widehat{y}_i|, i = 1, \ldots, 73\}$, where y_i is the ith response in the validation data set and $\widehat{y}_i$ is the prediction of response at $\mathbf{x}_i$ using the model chosen and fitted by the training data set.

	LS_{SCAD}	WW_{SCAD} (Wil)	WW_{SCAD} (GR)
age	2.489		
smoking status	2.561		
quetelet	-1.127		
vitamin use	-22.804		
calories	0.070	-0.009	-0.008
fat	-1.232		
fiber	4.960		
alcohol	10.353		
cholesterol	-0.110	-0.015	
dietary beta-carotene	0.026	0.021	0.022
median APE	97.902	66.742	66.609