Statistica Sinica **20** (2010), 167-181

VARIABLE SELECTION AND COEFFICIENT ESTIMATION VIA REGULARIZED RANK REGRESSION

Chenlei Leng

National University of Singapore

Abstract: The penalized least squares method with some appropriately defined penalty is widely used for simultaneous variable selection and coefficient estimation in linear regression. However, the efficiency of least squares (LS) based methods is adversely affected by outlying observations and heavy tailed distributions. On the other hand, the least absolute deviation (LAD) estimator is more robust, but may be inefficient for many distributions of interest. To overcome these issues, we propose a novel method termed the regularized rank regression (R^3) estimator. It is shown that the proposed estimator is highly efficient across a wide spectrum of error distributions. We show further that when the adaptive LASSO penalty is used, the estimator can be made consistent in variable selection. We propose using a score statistic-based information criterion for choosing the tuning parameters, which bypasses density estimation. Simulations and data analysis both show that the proposed method performs well in finite sample cases.

Key words and phrases: Composite quantile regression, lars, lasso, rank regression, viariable selection.

1. Introduction

Consider the linear regression problem described by the model

$$y_i = a^0 + \mathbf{x}_i^T \boldsymbol{\beta}^0 + \varepsilon_i, \quad i = 1, \ldots, n,$$

for independent and identically distributed observations $\{y_i, \mathbf{x}_i\}$, where ε_i follow some distribution with pdf f and cdf F, $y_i \in \mathbb{R}$ is a univariate response variable, and $\mathbf{x}_i = (\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,p})^T$ is a vector of p covariates. We assume that $\mathbf{x}_i$ and ε_i are independent. We are interested in estimating the unknown vector $\boldsymbol{\beta}^0 = (\beta_1^0, \ldots, \beta_p^0)^T$ and identifying any nonzero components. For notational purposes, the set of nonzero entries in $\boldsymbol{\beta}^0$ is labeled as $\mathcal{A}$ with $\mathcal{A} = \{k : \beta_k^0 \neq 0\}$.

Recently, a number of approaches, formulated to simultaneously estimate $\mathcal{A}$ and $\boldsymbol{\beta}$ via the penalized likelihood method, have gained increasing popularity. See, for example, the nonnegative garrote (Breiman (1995)), LASSO (Tibshirani (1996)), SCAD (Fan and Li (2001)) and LARS (Efron, Hastie, Johnstone and Tibshirani (2004)). A comprehensive account of these recent developments can be found in Fan and Li (2006).

Since such approaches aim to build sparse models without sacrificing accuracy, the oracle property, emphasized by Fan and Li (2001), is particularly relevant. Procedures possessing this property basically allow one to estimate the unknown coefficient vector $\boldsymbol{\beta}$ as if the set $\mathcal{A}$ were known in advance. More precisely, we say a variable selection and coefficient estimation procedure π is an oracle estimator if the estimator $\hat{\boldsymbol{\beta}}(\pi)$ has the properties (1) Selection consistency: $\mathrm{P}(\{k : \hat{\beta}_k(\pi) \neq 0\} = \mathcal{A}) \to 1$; (2) Estimation efficiency: $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\pi) - \hat{\boldsymbol{\beta}}_{\mathcal{A}}) \to N(\mathbf{0}, \Sigma_{\mathcal{A}\mathcal{A}})$, where $\Sigma_{\mathcal{A}\mathcal{A}}$ is the asymptotic variance of the MLE fitted to the submodel using the covariates in $\mathcal{A}$.

Considerable attention has been given to the least squares (LS) method. When combined with a suitably defined penalty function such as the adaptive LASSO penalty or the SCAD penalty on the regression coefficients, these estimators are shown to be consistent variable selection estimators (Fan and Li (2001), and Zou (2006)). However, the efficiency of the estimators via LS cannot be guaranteed. It is now well understood that the least squares method is sensitive to outliers, and is much less efficient if the error distribution has heavier tails than the normal distribution. In particular the asymptotic efficiency, introduced later, of the LS estimator is zero if the error distribution is Cauchy (Lehmann (1983)). On the other hand, formulation of a fully efficient estimator requires estimation of the unknown pdf f, which creates extra technical difficulties.

At first glance, a natural alternative seems to be the least absolute deviation (LAD) estimator (Wang, Li and Jiang (2007)), which can be more robust when f deviates from the normal. Nevertheless, the efficiency of the LAD compared to the MLE is proportional to the density at the median. For the Gaussian error case, the distribution of the greatest interest, this quantity is only 0.637. And, worse still, the efficiency can be arbitrarily small if $f(0)$ is close to zero (Hettmansperger and McKean (1998)).

To overcome some of the above issues, Zou and Yuan (2008) proposed the composite quantile regression (CQR) estimator by averaging K quantile regressions. They showed that CQR is selection consistent and can be more robust in various circumstances. Note that CQR can be seen as a special case of the weighted sum of quantile functions (Koenker (2005, Chap. 5.5)). In this paper, we propose an equally efficient estimator compared with CQR, referred to as the R^3 (Regularized Rank Regression) estimator. R^3 is a simple yet attractive estimator that combines properties of LAD and LS. The basic idea is to take pairwise differences among observations and then to fit the differences by LAD. This procedure is followed by applying the penalized least squares method via the adaptive LASSO penalty (Zou (2006), and Zhang and Lu (2007)). The LAD step can be seen as a rank-based estimator via Wilcoxon scores, to minimize the residual dispersion function given in Jaeckel (1972). The rank estimator is

generally more robust than LS and more efficient than the LAD estimator, and it is consistent and asymptotically normal. In addition, efficiency need not be sacrificed when the rank estimator is used in lieu of LAD or LS (Hettmansperger and McKean (1998)). Along with its simplicity, this formulation permits us to take advantage of the fast LARS algorithm (Efron et al. (2004)) for computing all the solutions on the regularization path. We show that when the penalty parameter is appropriately chosen, the R^3 method is consistent in variable selection, and its asymptotic efficiency is the same as the oracle CQR. When compared to the LS estimator, its efficiency is never below 0.864, which sharpens the bound 0.703 given in Zou and Yuan (2008). We propose further the use of a novel score statistic-based information criterion to choose the penalty parameter. It is shown that this criterion gives consistent results in terms of variable selection. At the same time, the proposed approach gives estimates with the same asymptotic covariance as that of the rank estimator as if the true model were known.

Some aspects of our approach echo to a degree those of the least squares approximation (LSA) in Wang and Leng (2007), who investigated a generalization of the LARS formulation in general parametric models. Rank-based variable selection is an active field of research. In independent work, Johnson and Peng (2008) and Wang and Li (2009) studied rank-based regression with SCAD. These approaches possess the oracle property. However it seems difficult, although not impossible, to develop a path-following algorithm for SCAD penalized methods, because of the nonconvexity of the penalty function (Fan and Li (2001)). Our approach, on the other hand, makes use of the fast Lars-Lasso algorithm in Efron et al. (2004) and can be quickly implemented to obtain the solution path.

The rest of the paper is organized as follows. Section 2 introduces the R^3 procedure and gives its asymptotic properties. The formulation of R^3 permits us to develop an efficient path-following algorithm to compute the entire solution path of the estimator, which is given in Section 3. In the same section, we propose a score statistic-based information criterion to choose the tuning parameter. We present some simulations and two data analyses in Section 4. Concluding remarks can be found in Section 5. A set of R functions are freely downloadable from `http://www.stat.nus.edu.sg/~stalc`. Online supplement to this paper can be found on `http://www.stat.sinica.edu.tw/statistica`.

2. The Regularized Rank Regression Estimator

Let $y_{ij} = y_i - y_j$ and $\mathbf{x}_{ij} = \mathbf{x}_i - \mathbf{x}_j$. We propose the initial estimator

$$\tilde{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} D(\boldsymbol{\beta}), \quad \text{where} \quad D(\boldsymbol{\beta}) = (2n)^{-1} \sum_{i,j=1}^{n} \left| y_{ij} - \mathbf{x}_{ij}^T \boldsymbol{\beta} \right|, \tag{2.1}$$

followed by an updated estimator obtained by minimizing the regularized least squares objective function

$$\hat{\boldsymbol{\beta}}_\lambda = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})^T \frac{X^T X}{n}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}) + \lambda \sum_{k=1}^{p} \lambda_k |\beta_k|, \tag{2.2}$$

where λ and λ_k are positive tuning parameters and X is the design matrix. The individual weights on entries of $\boldsymbol{\beta}$ are introduced to alleviate the effect that a uniform penalization parameter for all the coefficients may fail to identify a consistent yet efficient model (Zou (2006)). We can also take the SCAD penalty in the R^3 formulation and all the theoretical properties will continue to hold. The initial estimate mimics LAD by taking $\mathbf{x}_{ij}$ and y_{ij} as the observations, while the updating estimator is typical of a penalized least squares problem. Note that the objective function $D(\boldsymbol{\beta})$ is free of the intercept α from linear regression. The objective function $D(\boldsymbol{\beta})$ can be seen as Jaeckel's (1972) rank dispersion function for Wilcoxon scores

$$\tilde{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}(n)^{-1} \sum_{i=1}^{n} \left\{ \mathrm{R}(y_i - \mathbf{x}_i^T\boldsymbol{\beta}) - \frac{(n+1)}{2} \right\} \Big(y_i - \mathbf{x}_i^T \boldsymbol{\beta} \Big), \tag{2.3}$$

where $\mathrm{R}(z_i)$ is the rank of z_i among $z_1, \ldots, z_n$. For this reason, we refer to $\tilde{\boldsymbol{\beta}}$ as the Rank Regression (R^2) estimator and $\hat{\boldsymbol{\beta}}_\lambda$ as the regularized rank regression (R^3) estimator. Note that the R^2 estimator is scale and affine equivariant. A simple estimator of α^0 can be taken as the median of the residuals $\hat{\alpha} = \mathrm{med}\{y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}\}$, once we have evaluated $\hat{\boldsymbol{\beta}}$.

To study the asymptotic properties of R^3, we make the following assumptions: (A1) $C_n = n^{-1}X^TX = n^{-1}\sum_{i=1}^n \mathbf{x}_i^T\mathbf{x}_i \to_P C$ as $n \to \infty$; (A2) $\max_{i=1,\ldots,n} \|\mathbf{x}_i\|/\sqrt{n} \to_p 0$ as $n \to \infty$; (A3) $\mathrm{E}_\varepsilon\{D(\boldsymbol{\beta})\}$ is uniquely minimized by $\boldsymbol{\beta}^0 \in \mathbb{R}^p$, where the expectation is with respect to the distribution of ε. Assumptions A1 and A2 are routinely made in the linear regression modeling literature. Assumption A3 states that $\mathrm{E}_\varepsilon\{D(\boldsymbol{\beta})\}$ is uniquely minimized by $\boldsymbol{\beta}^0$. The asymptotic properties of the R^2 estimator is given in the following theorem.

Theorem 1. *Under conditions* A1–A3, *we have* $\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \to_d N(\mathbf{0}, 1/(12\omega^2) C^{-1})$ *for* $n \to \infty$, *where* $\omega = \int f^2(t)dt$.

Asymptotically, CQR has the same distribution as the rank regression estimator. Note that the asymptotic variance of R^2 and CQR is the same as the Hodges-Lehmann estimator. The scale multiplier $1/\{12(\int f^2(t)dt)^2\}$ is a measure of scale that is the rank analogue of the variance in the least squares procedure. The constant $(\int f^2(t)dt)^2$ indicates the height of the density of $Y_1 - Y_2$ at the origin. Defining the asymptotic relative efficiency of the R^2 to LS as $e(\tilde{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}^{LS})$,

Theorem 6.1 of Lehmann (1983) showed that $\inf_{F_S} e(\tilde{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}LS) = 0.864$ where F_S denotes cdf's which have finite Fisher information. This result tightens the bound 0.7026 given by Zou and Yuan (2008).

Note that the MLE oracle estimator is asymptotically $N(\mathbf{0}, I_f^{-1} C_{\mathcal{A}\mathcal{A}}^{-1})$ where $I_f = \int [f'(t)']^2 / f(t) dt$ is the Fisher information. The absolute efficiency of R^2, defined as the efficiency of R^2 compared to that of the MLE, is thus given by $e(\tilde{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}^{MLE}) = I_f^{-1} 12 (\int f^2(t) dt)^2$, The absolute asymptotic efficiency of R^2, together with that of R^2 compared to LS and LAD, are documented in Table 1 for a number of distributions, which is available in the online supplementary materials. Generally speaking, R^2 is almost as efficient as LS for normal errors but can be more robust for other errors; and R^2 is asymptotically much more efficient than LAD for many distributions of interest.

To simplify computation, we pre-select $\{\lambda_k\}$'s as $\lambda_k = |\tilde{\beta}_k|^{-1}$, $k = 1, \ldots, p$. For such λ_k, we define $a_n = \max\{\lambda_k : \ k \in \mathcal{A}\}$, which is of $O_p(1)$, and $b_n = \min\{\lambda_k : \ k \in \mathcal{A}^C\}$ which is of $O_p(n^{\tau/2}) \to \infty$, since the $\{\tilde{\beta}_k\}$'s are $\sqrt{n}$-consistent. The asymptotic properties of R^3 are in the following theorem.

Theorem 2. *Assume that $\sqrt{n}\lambda a_n \to 0$, and that $\sqrt{n}\lambda b_n \to \infty$. Then under Assumptions* A1−A3, *the R^3 solution $\hat{\boldsymbol{\beta}}_\lambda$ satisfies*

1. *Selection consistency: $P(\{k : \hat{\beta}_{\lambda k} \neq 0\} = \mathcal{A}) \to 1$;*
2. *Asymptotic normality: $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\lambda\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}) \to N(\mathbf{0}, 1/(12\omega^2) C^{-1}_{\mathcal{A}\mathcal{A}})$, where $\omega = \int f^2(t) dt$ and $C_{\mathcal{A}\mathcal{A}}$ is the submatrix of C whose entries correspond to the variables in $\mathcal{A}$.*

Remark 1. For model selection, the LSA in Wang and Leng (2007) requires a consistent estimator of $1/(12\omega^2)C^{-1}$, the asymptotic covariance matrix of $\tilde{\boldsymbol{\beta}}$. Here this condition is relaxed and we only need a consistent estimator of C without estimating $1/(12\omega^2)$. Therefore, we avoid density estimation for ω. This may seem a trivial point at first glance but it clearly has implications in computational implementation, especially because density estimation requires a careful choice of the bandwidth (Silverman (1986)).

3. Computation and Tuning

The initial estimate $\tilde{\boldsymbol{\beta}}$ can be efficiently computed, e.g., by the R function rq in library *quantreg*. Letting $\tilde{y} = X\tilde{\boldsymbol{\beta}}$, we can write the objective function in (2.2) as

$$\frac{1}{n}(\tilde{y} - X\boldsymbol{\beta})^T(\tilde{y} - X\boldsymbol{\beta}) + \lambda \sum_{k=1}^{p} \lambda_k |\beta_k|.$$

This is a standard penalized least squares problem, whose solution path can be computed by the fast LARS algorithm at a computational complexity equal to a single least squares fit.

Let the score statistic for $\hat{\boldsymbol{\beta}}_\lambda$ be

$$T_\lambda = 3n^{-1}G^T(\hat{\boldsymbol{\beta}}_\lambda)C_n^{-1}G^T(\hat{\boldsymbol{\beta}}_\lambda), \quad \text{where} \quad G(\hat{\boldsymbol{\beta}}_\lambda) = \frac{\partial}{\partial \boldsymbol{\beta}} D(\boldsymbol{\beta})\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_\lambda}. \tag{3.1}$$

In this definition, this partial derivative is set to zero at zero. For tuning parameter selection, we propose λ to minimize the following score information criterion (SIC): $SIC_\lambda = T_\lambda + df_\lambda \log\log(n)$, where the degrees of freedom df_λ is the number of nonzero entries in $\hat{\boldsymbol{\beta}}_\lambda$. An advantage of SIC is that we do not need to estimate the nuisance parameter ω, whose estimation requires density estimation. The use of the score statistic seems novel in the penalized model selection literature. The consistency of the proposed SIC method can be proved in a way similar to Wang and Leng (2007), and the proof is in the online supplement.

Remark 2. Mimicking the Bayesian information criterion (BIC), we can define an alternative information criterion (Wang and Leng (2007)) as $12\omega^2(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^0)^T C_n(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^0) + df_\lambda \log(n)/n$. However, in this formulation, density estimation is required to obtain ω. An alternative criterion is $(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^0)^T C_n(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^0) + df_\lambda \log(n)/n$, but this information criterion does not perform well in small sample cases as the results can be sensitive to the value of ω (results not shown). Although the proposed SIC method avoids density estimation for ω, density estimation is still mandatory for statistical inference of the estimated coefficients. In practice, we use a consistent estimator of ω for statistical inference on the estimated coefficients (Terpstra and McKean (2005)). An alternative induced smoothing approach for inference in this context can be found in Brown and Wang (2005, Sec. 5). This approach can be applied after the model is identified by R^3. This line of research will be pursued elsewhere.

4. Simulation and Data Analysis

We study the finite sample performance of R^3 in this section via a simulation study and two data analyses: one on the Boston housing data set and the other on the Diabetes data set. The analyses imply that the R^3 method provides more accurate estimates when the normality assumption on the error distribution fails and, at the same time, provides comparable results to those of the penalized least squares method when the error distribution comes from the normal.

4.1. Simulation

For comparison purposes, we include in this section the adaptive LASSO estimates by using LS and LAD as the loss function, respectively. More precisely, these estimators are obtained by minimizing $L(\boldsymbol{\beta}) + \lambda \sum_{k=1}^p \lambda_k |\beta_k|$, where for LS, $L(\boldsymbol{\beta})$ is the least squares criterion, and for LAD, $L(\boldsymbol{\beta})$ is the least absolute

deviation loss. We also compare R^3 with CQR (Zou and Yuan (2008)), the SCAD estimator (Johnson and Peng (2008), and Wang and Li (2009)), and another penalized rank (PR) estimator suggested by two anonymous referees which solves $\min_\beta (2n)^{-1} \sum_{i,j=1}^n |y_{ij} - X_{ij}^T\beta| + \lambda \sum_{j=k}^p \lambda_k |\beta_k|$. The parameters $\{\lambda_k\}$ are set in a fashion simialar to those for R^3. Following Zou and Yuan (2008), we fit the model using the generated data set and choose the tuning parameter via an independent validation set as an alternative to SIC. This is referred to as cross validation and, for the methods compared, we denote them as R^3-CV, LAD-CV, LS-CV and CQR-CV. These methods are compared to R^3-SIC and PR-SIC, where SIC is used for tuning. Note that the simulation scheme is in favor of cross validation as an additional data set is used for choosing the tuning parameters. For CQR, we follow the suggestion of Zou and Yuan (2008) by taking the number of quantiles to be 19 so quantiles at $5\%, 10\%, \ldots, 95\%$ are used. The simulation setup follows Example 1 in Wang and Li (2009). Specifically, we simulated 500 data sets consisting 100 training observations from the following linear model $y = \mathbf{x}^T\boldsymbol{\beta} + \sigma\varepsilon$, where $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)$ and $\mathbf{x}$ follows a multivariate normal distribution $N(\mathbf{0}, \Sigma_\mathbf{x})$ with $(\Sigma_\mathbf{x})_{ij} = 0.5^{|i-j|}$. To choose λ via cross validation, 100 additional observations were generated from the model for each experiment. We considered three error distributions: the standard normal, the t distribution with 3 degrees of freedom (t_3) and a contaminated normal distribution with 10% outliers from the standard Cauchy distribution.

Since we knew the true model, we could compute the model error given by $\mathrm{ME} = \mathrm{E}\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)^T \Sigma_\mathbf{x} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\}$. In Table 3.1, we report the median relative model errors (MRME) of various methods compared to the unpenalized rank-based estimator, the average number of correct zeros (C), the average number of incorrect zeros (IC), and the percentage of correct models identified. The results for SCAD are taken from Wang and Li (2009).

A few observations can be made from Table 3.1. First, the proposed R^3 method was highly efficient for all the distributions under consideration. Its efficiency compared to LS and LAD was high. The LS-CV method performed worst for t_3 and contaminated normal distributions. Second, the proposed SIC method for choosing λ performed satisfactorily and conformed to the asymptotic results. The cross validation based variable selection procedures (LS-CV, LAD-CV and CQR-CV) did not seem to give consistent model selection results. This confirms the findings in Leng, Lin and Wahba (2006). A related discussion on why cross validation fails in model selection can be found in Wang, Li and Tsai (2007). Additionally, the SIC method seemed to outperform the BIC approach in Wang and Li (2009) in terms of variable selection and median model error. This may be due to the fact that T_λ essentially follows a χ^2 distribution for overfitted models. Third, because of improved performance in terms of variable selection,

Table 3.1. Estimation performance.

Error Distribution	Method	No. of Zeros C	No. of Zeros IC	Correct Fit (%)	MRME (%)
Normal	SCAD	4.42	0	68.5	43.8
	CQR-CV	4.30	0	62.4	44.8
	PR-SIC	4.94	0	96.4	33.2
	LS-CV	3.93	0	52.0	39.7
	LAD-CV	4.05	0	54.2	55.3
	R^3-CV	4.02	0	54.4	37.0
	R^3-SIC	4.93	0	94.4	35.1
t_3	SCAD	4.46	0	73.0	40.5
	CQR-CV	4.30	0	71.2	41.7
	PR-SIC	4.94	0	94.6	36.8
	LS-CV	4.08	0	57.2	58.1
	LAD-CV	4.03	0	54.4	45.6
	R^3-CV	4.13	0	60.8	37.1
	R^3-SIC	4.93	0	93.6	36.4
Contaminated Normal	SCAD	4.48	0	67.5	40.6
	CQR-CV	4.53	0	68.0	44.8
	PR-SIC	4.93	0	93.0	38.9
	LS-CV	3.92	0	53.2	71.2
	LAD-CV	4.04	0	55.0	48.6
	R^3-CV	4.00	0	58.4	39.5
	R^3-SIC	4.94	0	94.2	37.9

R^3-SIC gave smaller model errors compared to R^3-CV. Finally, the performance of R^3-SIC and PR-SIC were similar, as has been discussed for general linear models in Wang and Leng (2007).

To illustrate the accuracy of the asymptotic variance formula, we estimated ω using the method in Terpstra and McKean (2005). As in Fan and Li (2001), the median absolute deviation divided by .6745, denoted by SD in Table 3.2, of 500 estimated coefficients in 500 simulations can be regarded as the true standard error. The median of the 500 estimated SD's, denoted by SD_m, measures the performance of the standard errors. The estimated mean bias is also included in this table. We can see clearly that the R^3 estimates were nearly unbiased and that the inference procedure was satisfactory.

We make a few remarks on the computational issue for R^3 and PR. Both R^3 and PR require one to compute the R^2 estimator $\tilde{\beta}$ in order to get λ_k. As correctly pointed out by an anonymous referee, a path-following algorithm for the PR estimator can be derived in a way similar to the quantile regression in Li and Zhu (2008). However, PR deals with a penalized LAD problem with $n(n-1)/2$ observations while R^3 only deals with a penalized least square problem with n observations. More severely, the data pairs $(\mathbf{x}_{ij}, y_{ij})$ are no longer independent

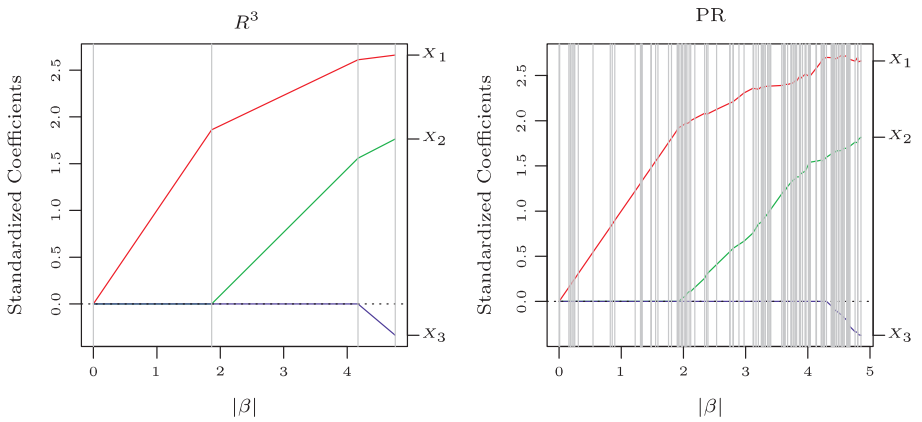

Figure 3.1. The solution paths for a simple three-dimensional example. Any line segment between two adjacent gray lines is linear.

Table 3.2. Accuracy of inference. N: standard normal, CN: contaminated normal.

ε	β_1			β_2			β_5		
	Bias	SD	SD_m	Bias	SD	SD_m	Bias	SD	SD_m
N	-0.008	0.122	0.121	0.028	0.124	0.123	0.017	0.107	0.106
t_3	0.010	0.150	0.153	0.024	0.168	0.155	0.017	0.136	0.135
CN	0.002	0.153	0.152	0.023	0.145	0.152	0.022	0.127	0.134

and this complicates the path-following algorithm for PR. Indeed, we observe empirically that for the simulation study in this section, it took PR many more steps than R^3. To demonstrate this, we generated 20 observations from a simple three dimensional example $y = 3x_1 + 1.5x_2 + 0x_3 + \varepsilon$, where $(x_1, x_2, x_3)^T$ was generated as before and ε was $N(0, 1)$. In Figure 3.1, we plot the whole solution paths for the two algorithms, where R^3 requires three steps and PR requires 145 steps. The computational demand for PR increases for problems with larger sample sizes. Alternatively, we can fit PR on a fixed grid. For each grid value, we need to minimize a penalized LAD problem with $n(n-1)/2$ observations, which again can be slower than R^3. We conclude that R^3 is computationally more desirable than PR.

4.2. Boston housing data

The Boston Housing data set was analyzed by Harrison and Rubinfeld (1978) who wanted to find out whether "clean air" had an influence on house prices. This data set is available in the R package mlbench with 14 variables and 506 cases. The response of interest here is the logarithm of the median value of owner occupied homes (LMV). The other 13 independent variables include the following: per capita crime rate by town (CRIM); proportion of residential land

Table 3.3. Estimates for the Boston Housing data.

Method	Least Squares			PLS	R^3
	Coef$\times 10^2$	SE$\times 10^2$	Z Value	Coef$\times 10^2$	Coef$\times 10^2$
CRIM	-1.027	(0.132)	-7.81	-1.010	-0.819
ZN	0.117	(0.055)	2.13	0.086	0
INDUS	0.247	(0.246)	1.00	0	0
CHAS	10.089	(3.449)	2.93	9.746	4.625
NOX	-77.840	(15.289)	-5.09	-69.896	-42.280
RM	9.083	(1.673)	5.43	9.114	15.781
AGE	0.021	(0.053)	0.40	0	0
DIS	-4.909	(0.798)	-6.15	-4.894	-2.923
RAD	1.427	(0.266)	5.37	1.259	0.698
TAX	-0.063	(0.015)	-4.16	-0.052	-0.037
PTRATIO	-3.827	(0.524)	-7.31	-3.755	-3.411
B	0.041	(0.011)	3.85	0.039	0.052
LSTAT	-2.904	(0.203)	-14.30	-2.882	-2.511

zoned for lots over 25,000 sq.ft (ZN); proportion of non-retail business acres per town (INDUS); Charles river dummy variable (= 1 if tract bounds river; 0 otherwise. CHAS); nitric oxides concentration (parts per 10 million, NOX); average number of rooms per dwelling (RM); proportion of owner-occupied units built prior to 1940 (AGE); weighted distances to five Boston employment centers (DIS); index of accessibility to radial highways (RAD); full-value property-tax rate per $10,000$ (TAX); pupil-teacher ratio by town (OTRATIO); $1000(\text{bk}-0.63)^2$ where bk is the proportion of blacks by town (B); proportion of population that has a lower status (LSTAT).

In fitting a linear regression model to the data set, we start with the usual least squares (LS) fit and list the fitted coefficients in Table 3.3. The penalized LS (PLS) estimates with the adaptive LASSO penalty are also presented in the same table. In view of the simulation results, the tuning parameter λ is chosen by minimizing the BIC criterion $BIC_\lambda = \log(RSS_\lambda/n) + df_\lambda \log(n)/n$, where RSS_λ is the residual sum of squares for the model fitted with the parameter λ. But the normality assumption on the error distribution may be poor, as suggested by the density plot and the quantile-quantile (Q-Q) plot for the LS residuals in Figure 3.2 (a) and (b). In fact, a one-sample Kolmogorov-Smirnov test on the residuals is highly significant (p-value = 0.01).

These two sets of estimates are compared to the proposed R^3 estimates via SIC. From Table 3.3, we see that the PLS produces estimates very close to the LS ones and shrinks two insignificant coefficients (ZN and AGE) to zero. The proposed R^3 gives three zero coefficients (ZN, AGE and INDUS). The R^3 estimates are visually different from those of LS and PLS. As a matter of fact,

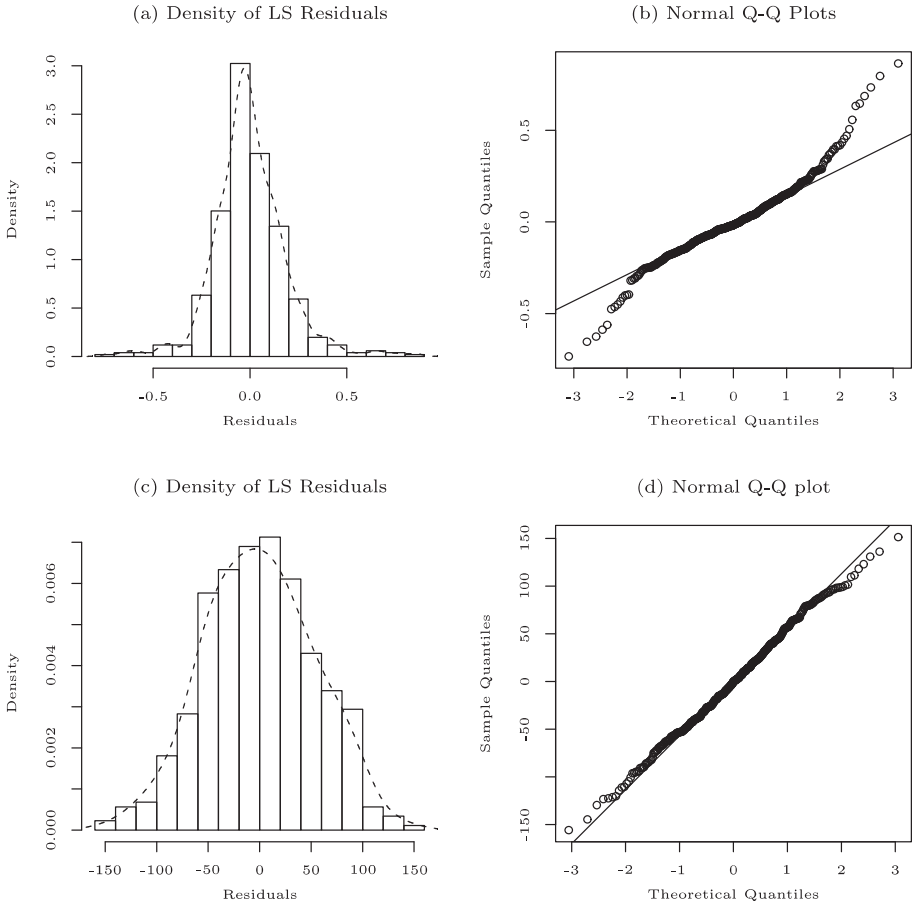

Figure 3.2. Checking normality assumptions. (a): Histogram of the LS residuals for the Boston Housing dataset; (b): Q-Q plot of the LS residuals for the Boston Housing dataset; (c): Histogram of the LS residuals for the Diabetes dataset; (d): Q-Q plot of the LS residuals for the Diabetes dataset.

four estimated coefficients (NOX, RM, DIS and RAD) for R^3 are not within twice the standard errors of the corresponding LS estimates.

Both PLS and R^3 solutions are piecewise linear functions of λ and can be efficiently computed via the LARS algorithm. The two solution paths are presented in Figure 3.3. As this figure shows, the solution paths of R^3 and PLS are quite different, especially for the variables RM and RAD. The magnitude of some estimated regression coefficients of the two methods, eg LSTAT and NOX, is also visually different in the figure.

In order to compare the predictive performance of PLS and R^3, we randomly divided the data set into 10 subsets which are about equally sized. Each time,

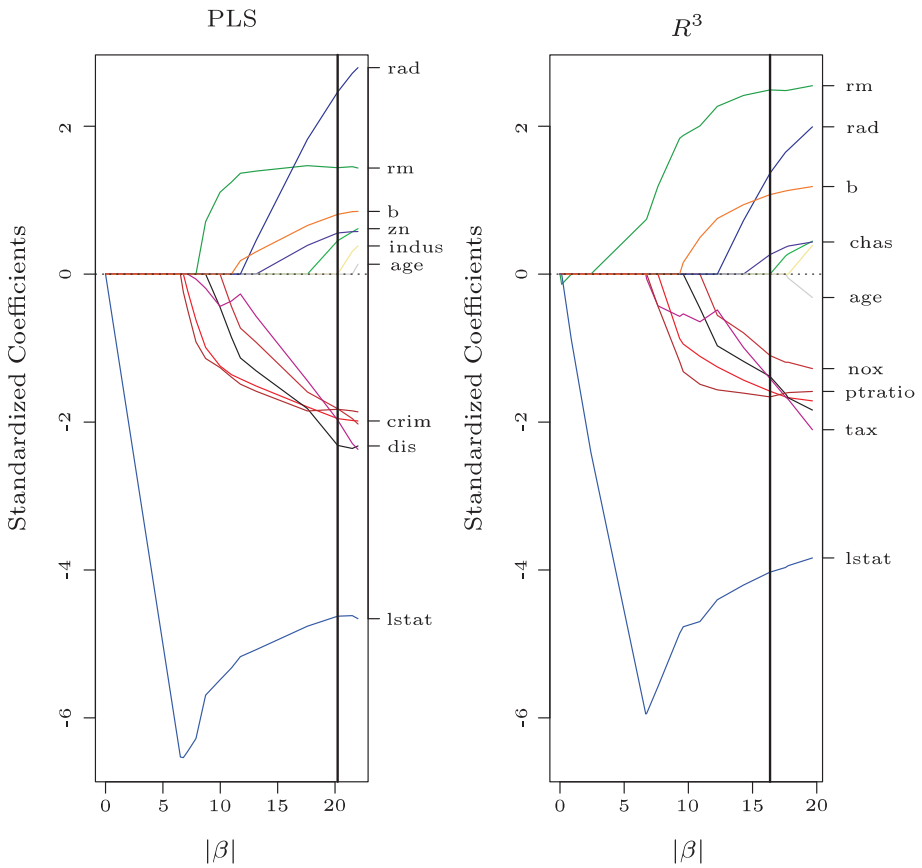

Figure 3.3. The solution paths for the Boston housing data set. The coefficients are standardized by the corresponding standard errors of the covariates. The vertical black lines indicate the chosen models.

we used 9 subsets of the observations for model fitting, and the remainder as the test set. We calculated the mean absolute errors for the test set in terms of $|y - \hat{\alpha} - \mathbf{x}^T\hat{\boldsymbol{\beta}}|$ for PLS and R^3 estimates, respectively. The average error and model size over the 10 runs are summarized in Table 3.4. Pairwise t-tests suggest that R^3 produced significantly smaller errors (p-value $= 0.02$) and models with fewer number of covariates (p-value < 0.01). The comparison results are not surprising due to the previous observation that the normal error assumption is poor. Therefore, we conclude that the R^3 method is the preferred method for analyzing this data set.

4.3. Diabetes data

We examine the diabetes data set in Efron et al. (2004). This data set con-

Table 3.4. Cross validation results. Standard errors are in parentheses.

	Boston Housing		Diabetes	
Variable	Absolute Error	Model Size	Absolute Error	Model Size
PLS	0.1372 (0.0096)	12.4 (0.31)	44.42 (1.07)	6.4 (0.16)
R^3	0.1334 (0.0097)	10.3 (0.15)	44.84 (1.18)	5.7 (0.15)

Table 3.5. Estimates for the Diabetes data.

Variable	Least Squares			PLS	R^3
	Coef	SE	Z Value	Coef	Coef
AGE	-10.0	59.7	-0.17	0	0
SEX	-239.8	61.2	-3.92	-201.7	-250.0
BMI	519.8	66.5	7.81	540.5	539.4
BP	324.4	65.4	4.96	314.5	330.9
x_5	-792.2	416.7	-1.90	-514.2	-579.3
x_6	476.7	339.0	1.41	268.7	297.1
x_7	101.0	212.5	0.48	0	0
x_8	177.1	161.5	1.10	119.6	132.1
x_9	751.3	171.9	4.37	682.5	716.3
x_{10}	67.6	66.0	1.03	0	0

sists of 442 diabetic patients with ten baseline factors age (AGE), sex (SEX), body mass index (BMI), average blood pressure (BP), and six blood serum measurements (coded as x_5 to x_{10}). The response is a quantitative measurement of disease progression one year after baseline. The aim for this data set is to build a predictive model to relate the response to the ten covariates. An LS fit of the model produces the density plot and Q-Q plot given in Figure 3.2 (c) and (d), which justifies a normality assumption for the error distribution. Indeed, a one-sample Kolmogorov-Smirnov test on the residuals is insignificant (p-value = 0.95). The fitted models for LS, PLS, and R^3 are listed in Table 3.5. We note that both PLS and R^3 choose the same model with 7 covariates. Further, the difference between the R^3 estimates and the corresponding LS estimates is small. Actually, the largest difference between these two, standardized by the standard errors of the LS estimates, is x_{10}, which is about one. These results show a close agreement between the LS, the PLS, and the R^3 estimates. In order to evaluate the predictive performance of PLS and R^3, we again applied a 10-fold cross validation technique to compute the average absolute errors. The results are summarized in Table 3.4. Pairwise t-test suggests that there is no significant difference between the errors for the two estimates (p-value = 0.19) and that the R^3 produces significantly smaller models (p-value = 0.02). This example shows that even in the case that the normal assumption based least squares method is the preferred method, the R^3 method is still competitive in terms of predictive values.

5. Conclusion

In this paper, we propose R^3 as an alternative to the penalized least squares method for simultaneous variable selection and coefficient estimation. We show that R^3 enjoys the oracle property in variable selection and is highly efficient. The simplicity of R^3 facilitates the development of a path-following algorithm. A novel score statistic-based information criterion is proposed to choose the penalty parameter. This criterion guarantees consistency in variable selection and as a result, gives better small sample prediction performance in the simulation study. Our simulation study and data analysis both show that R^3 performs well and that it may be preferred over LAD and LS. How to extend R^3 to problems with diverging dimensionality is an interesting topic left for future research (Fan and Lv (2008), and Bickel, Ritov and Tsybakov (2009)).

Acknowledgement

Leng's research is supported in part by NUS research grants. The author is grateful to Professor Bruce M. Brown for helpful discussions, and we would like to thank the Editors, an associate editor and three referees for their constructive comments.

References

Bickel, P., Ritov, Y. and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, **37**, 1705-1732.

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

Brown, B. and Wang, Y. (2005). Standard errors and covariance matrices for smooth rank estimators. *Biometrika* **92**, 149-158.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression (with discussion). *Ann. Statist.* **32**, 407-451.

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

Fan, J. and Li, R. (2006). Statistical challenges with high dimensionality: feature selection in knowledge discovery. In *Proceedings of the International Congress of Mathematicians* (Edited by M. Sanz-Sole, J. Soria, J. L. Varona and J. Verdera), Vol. III, 595-622.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. Roy. Statist. Soc. Ser. B* **70**, 849-911.

Harrison, D. and Rubinfeld, D. L. (1978). Hedonic prices and the demand for clean air. *Journal of Environmental Economics and Management* **5**, 81-102.

Hettmansperger, T. P. and McKean, J. W. (1998). *Robust Nonparametric Statistical Methods.* Arnold/Wiley, London and New York.

Jaeckel, L. (1972). Estimating regression coefficients by minimizing the dispersion of the residuals. *Ann. Math. Statist.* **43**, 1449-1458.

Johnson, B. and Peng, L. (2008). Rank-based variable selection. *J. Nonparametr. Stat.* **20**, 241-252.

Koenker, R. (2005). *Quantile Regression*. Econometric Society Monograph Series, Cambridge University Press.

Lehmann, E. L. (1983). *Theorery of Point Estimation*. Wiley, New York.

Leng, C., Lin, Y. and Wahba, G. (2006). A note on the LASSO and related procedures in model selection. *Statist. Sinica* **16**, 1273-1284.

Li, Y. and Zhu, J. (2008). L_1-norm quantile regression. *J. Comput. Graph. Statist.* **17**, 163-185.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, New York.

Terpstra, J. and McKean. J. (2005). Rank-based analysis of linear models using R. *J. Statist. Soft.* **14**.

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

Wang, H. and Leng, C. (2007). Unified lasso estimation via least squares approximation. *J. Amer. Statist. Assoc.* **102**, 1039-1048.

Wang, L. and Li, R. (2009). Wighted Wilcoxon-type smoothly clipped absolute deviation method. *Biometrics* **65**, 564-571.

Wang, H., Li, G. and Jiang, G. (2007). Robust regression shrinkage and consistent variable selection through the lad-lasso. *J. Bus. Econom. Statist.* **25**, 347-355.

Wang, H., Li, R. and Tsai, C. L (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94**, 553-568.

Zhang, H. H. and Lu, W. (2007). Adaptive-LASSO for Cox's proportional hazard model. *Biometrika* **94**, 691-703.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418-1429.

Zou, H. and Yuan, M. (2008). Composite quantile regression and the oracle model selection theory. *Ann. Statist.* **36**, 1108-1126.

Department of Statistics and Applied Probability, National University of Singapore, 117546, Republic of Singapore.

E-mail: stalc@nus.edu.sg

(Received June 2008; accepted December 2008)