Variable Selection Incorporating Prior Constraint Information into Lasso

Shurong Zheng, Guodong Song and Ning-Zhong Shi
School of Mathematics & Statistics, Northeast Normal University, P.R.China

Abstract

We propose the variable selection procedure incorporating prior constraint information into lasso. The proposed procedure combines the sample and prior information, and selects significant variables for responses in a narrower region where the true parameters lie. It increases the efficiency to choose the true model correctly. The proposed procedure can be executed by many constrained quadratic programming methods and the initial estimator can be found by least square or Monte Carlo method. The proposed procedure also enjoys good theoretical properties. Moreover, the proposed procedure is not only used for linear models but also can be used for generalized linear models(*GLM*), Cox models, quantile regression models and many others with the help of Wang and Leng (2007)'s LSA, which changes these models as the approximation of linear models. The idea of combining sample and prior constraint information can be also used for other modified lasso procedures. Some examples are used for illustration of the idea of incorporating prior constraint information in variable selection procedures.

Keywords: lasso; linear models; prior constraint information; sample information; variable selection;

1. Introduction

In practice, a number of variables are included into an initial regression analysis, but many of them may not be significant to the response variables and should be excluded from the final model in order to increase the accuracy of prediction and interpretation. Variable selection is fundamental in statistical modeling. The least absolute shrinkage and selection operator (*LASSO*) (Tibshirani

Address for correspondence: Ning-Zhong Shi, School of Mathematics & Statistics, Northeast Normal University, Changchun City 130024, P.R.China
Email: shinz@nenu.edu.cn

1996) is a useful and well-studied approach to the problem of variable selection (Knight and Fu 2000; Fan and Li 2001; Leng et al. 2006; Wang et al. 2007a; Yuan and Lin 2007). It shrinks some coefficients and sets others to 0, and hence tries to retain the good features of both subset selection and ridge regression. Moreover, lasso's major advantage is its simultaneous execution of both parameter estimation and variable selection. In particular, allowing an adaptive amount of shrinkage for each regression coefficient results in an estimator which is as efficient as oracle (Zou 2006; Wang et al. 2007b; Wang and Leng 2007). About the computational techniques, please see Osborne et al. (2000), Efron et al. (2004), Rosset (2004), Zhao and Yu (2004) and Park and Hastie (2006).

In spite of that, in variable selection or the estimation of regression coefficients, except for sample information, some prior constraint information can be known. Constraints can be expressed as $\mathbf{g}(\beta) \leq 0$ including equalities and inequalities where $\mathbf{g}(\cdot)$ are k-dimensional linear or nonlinear functions (see Rao and Toutenburg 1995; Silvapulle and Sen 2005). In fact, the common simple order $\beta_1 \leq \cdots \leq \beta_p$; tree order $\beta_i \leq \beta_p$ for $i = 1, \cdots, p-1$; umbrella order $\beta_1 \leq \cdots \leq \beta_l \geq \cdots \geq \beta_p$ or more generally $A\beta \leq \mathbf{a}$ are only the special cases of $\mathbf{g}(\beta) \leq 0$. All these constraints have very important applications in biomedical studies, life science, econometrics and social research etc. For example, in many biomedical studies, treatment groups in a clinical trial many be formulated according to increasing levels of dosage of a drug and the severity of disease in patients. In econometrics, the homogeneity of degree zero of a demand equation implies that the price and income elasticities add up to zero, whereas the negativity of the substitution matrix in consumer demand theory requires that all latent roots of the substitution matrix should be nonpositive. Stahlecker (1987) shows a variety of examples from the field of economics (such as input-output models), where the constraints for the parameters are so-called workability conditions of the form $\beta_i \geq 0$ or $\beta_i \in (a_i, b_i)$ or $E(y_t|X) \leq a_t$. Literature deals with this problem under the generic term constrained least squares (see Judge and Takayama 1966; Dufour 1989; Geweke 1986; Moors and van Houwelingen 1987; Rao and Toutenburg P75 1995). Dorfman and McIntosh (2001) show that imposing the curvature conditions on a system of demand equations improves the MSEs on estimated elasticities from 2 to 50% depending on the signal-to-noise ratio and the sample size. For researchers, it will increase the efficiency of variable selection and parameter estimation to effectively combine the sample and prior information because prior information tells us a narrower region to select these variables and estimate these parameters.

This paper proposes a procedure to combine prior and sample information into lasso and hopes to obtain more accurate variable selection and parameter estimation. The idea of combining prior constraint and sample information can be shown by the black region in Figure 1. It shows that when we know some prior information of parameters, then variable selection will be executed in a narrower black region AEFD not in a wide region ABCD. It will increase the efficiency of choosing the true model correctly. Moreover, our procedure incorporating prior constraint information is not only used for linear models but also can be used for generalized linear models, Cox models and quantile regression models with the help of Wang and Leng (2007)'s LSA, which changes these models as the approximation of linear models. In fact, prior constraint information can be also used for other modified lasso procedures, e.g. Tibrashini et al. (2005)'s fused lasso and the modified lasso procedure for an adaptive amount of shrinkage for each regression coefficient (Zou 2006; Wang et al. 2007b; Wang and Leng 2007) etc.

The paper is organized as follows: Section 2 introduces variable selection procedure combining sample and prior constraint information in lasso and other modified lasso procedures. Main theoretical properties are discussed in Section 3. Section 4 discusses degrees of freedom of the lasso procedure incorporating prior constraint information. The proposed procedure is illustrated by some examples in Section 5. Section 6 gives a short discussion.

2. Variable Selection Combining Sample and Prior Constraint Information into Lasso

2.1 Variable Selection Combining Sample and Prior Constraint Information into Lasso in Linear Models

We first consider variable selection incorporating prior constraint information into lasso in linear models:

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^T \beta\right)^2 \quad \text{subject to} \quad \sum_j |\beta_j| \leq s \quad \text{and} \quad \mathbf{g}(\beta) \leq \mathbf{0}$$

or

$$\hat{\beta}_s = Arg\left\{\min_{\beta} \left(\mathbf{Y} - \mathbf{X}^T\beta\right)^T \left(\mathbf{Y} - \mathbf{X}^T\beta\right) \quad \text{subject to} \quad \sum_j |\beta_j| \leq s \quad \text{and} \quad \mathbf{g}(\beta) \leq \mathbf{0}\right\}. \qquad (1)$$

where $\mathbf{Y} = (y_1, \cdots, y_n)^T$, $\mathbf{X} = (\mathbf{x}_1^T, \cdots, \mathbf{x}_n^T)^T$ and $\mathbf{g}(\cdot)$ are linear or nonlinear functions. That is, the modified lasso objective function is as follows

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \beta)^2 + \lambda^{(1)} \sum_{j} |\beta_j| + (\lambda^{(2)})^T \mathbf{g}(\beta)$$

where $\lambda^{(1)}$ and $\lambda^{(2)} = (\lambda_1^{(2)}, \cdots, \lambda_k^{(2)})^T$ are tuning parameters. The tuning parameters can be obtained by estimating the prediction error for the procedure incorporating prior constraint information into lasso by cross-validation (CV) as described in chapter 17 of Efron and Tibshirani (1993) or generalized cross-validation (GCV). The prediction error of prediction term $\hat{\eta}(\mathbf{X})$ of CV is given by

$$PE = E\{Y - \hat{\eta}(\mathbf{X})\}^2.$$

Then the value $\hat{s}$ yielding the lowest estimated PE is selected.

In the following, we introduce how to choose the tuning parameters from CV in detail. Similarly, GCV can be used to choose the tuning parameters. l-fold CV is one of the methods to choose the tuning parameters s. l-fold CV is to split the n patterns into a training set of size $n - l$ and a test of size l. l-fold CV averages the squared error on the left-out pattern over all the possible ways of obtaining such a partition. The advantage is that all the data can be used for training - none has to be held back in a separate test set. Take $l = 1$ for an example. Let

$$\widehat{\beta}_s^{(-j)} = Arg\left\{\min_{\beta} \sum_{i=1, i\neq j}^{n} (y_i - \mathbf{x}_i^T \beta)^2 \quad \text{subject to} \quad \sum_{h} |\beta_h| \le s \quad \text{and} \quad \mathbf{g}(\beta) \le \mathbf{0}\right\} \tag{2}$$

where $\widehat{\beta}_s^{(-j)}$ is the estimation on the training data $\mathbf{x}_1, \cdots, \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, \cdots, \mathbf{x}_n$ for $j = 1, \cdots, n$ from the procedure incorporating prior constraint information into lasso. Let $PE_s = \sum_{j=1}^{n} \left(y_j - \mathbf{x}_j^T \widehat{\beta}_s^{(-j)}\right)^2$ be the estimated prediction error of 1-fold CV given the tuning parameter s. Then the chosen tuning parameters s is as follows

$$\hat{s} = Arg\left\{\min_{s} PE_s\right\} = Arg\left\{\min_{s} \sum_{j=1}^{n} \left(y_j - \mathbf{x}_j^T \widehat{\beta}_s^{(-j)}\right)^2\right\}$$

where $\hat{s}$ minimizes the estimated prediction error. Then the simultaneous parameter estimation and variable selection incorporating prior constraint information is as follows

$$\widehat{\beta}_{\hat{s}} = Arg\left\{\min_{\beta} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \beta)^2 \quad \text{subject to} \quad \sum_{h} |\beta_h| \le \hat{s} \text{ and } \mathbf{g}(\beta) \le \mathbf{0}\right\}. \tag{3}$$

Remark 1. (*Algorithm*)We know that the most important thing for obtaining $\widehat{\beta}_{\hat{s}}$ is to compute $\widehat{\beta}_s^{(-j)}$. If there are no constraints on the parameters, many well developed procedures can be used to find the solution for

$$\min_{\beta} \sum_{i=1, i\neq j}^{n} (y_i - \mathbf{x}_i^T \beta)^2 \quad \text{subject to} \quad \sum_h |\beta_h| \leq s.$$

For example, quadratic programming (Tibshirani 1996), the shooting algorithm (Fu 1998), local quadratic approximation (Fan and Li 2001) and lease angle regression (*LARS*) (Efron et al. 2004). When there are prior constraint information, the above procedures can not be directly used for (2). But if some modifications are made for these algorithms, (2) may be solved by them. It will be an interesting topic for us in the future. In fact, many quadratic programming methods can be used to find the solution for (2) (see Dantig and Eaves 1974). The solution of the quadratic programming does not yield a sparse solution. If a tolerance is set, the small parameter estimate can be regarded as 0.

Remark 2. (*Initial Estimator*) In fact, the OLS estimator may be regarded as the initial estimator. But in order to obtain more accurate estimator, Monte Carlo method can be used for the initial estimator of (1) or (2). The optimal problem (1) can be written as

$$\begin{aligned}(\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) &= \beta^T\mathbf{X}^T\mathbf{X}\beta - 2\beta^T\mathbf{X}^T\mathbf{Y} + \mathbf{Y}^T\mathbf{Y} \\ &= (\beta-\mu)^T\Sigma^{-1}(\beta-\mu) + \mathbf{Y}^T\left(I - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)\mathbf{Y}\end{aligned}$$

with known $\mu_{p\times 1} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$ and $\Sigma_{p\times p} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}$. That is,

$$\widehat{\beta}_s = Arg\left\{\min_{\beta}(\beta-\mu)^T\Sigma^{-1}(\beta-\mu) \quad \text{subject to} \quad \sum_{h=1}^{p}|\beta_h| \leq s \text{ and } \mathbf{g}(\beta) \leq \mathbf{0}\right\}$$

or

$$\widehat{\beta}_s = Arg\max_{\beta} -\frac{1}{2}\left[(\beta-\mu)^T\Sigma^{-1}(\beta-\mu) + \log(|\Sigma|)\right] \quad \text{subject to} \quad \begin{cases}\sum_{j=1}^{p}|\beta_j| \leq s \\ \mathbf{g}(\beta) \leq \mathbf{0}\end{cases} \tag{4}$$

where

$$l(\beta) = -\frac{1}{2}(\beta-\mu)^T\Sigma^{-1}(\beta-\mu) - \frac{1}{2}\log(|\Sigma|) \tag{5}$$

is just the log-density of $N(\mu,\Sigma)$ regarding β as a random variable. Randomly draw $m = 100000$ samples $\mathbf{Z}_1, \cdots, \mathbf{Z}_m$ from $N(\mu,\Sigma)$ where $\mathbf{Z}_j = (Z_{1j}, \cdots, Z_{pj})^T$ for $j = 1, \cdots, m$. Set Z_{old} as the initial estimator which satisfies

$$Z_{old} = Arg\left\{ \max_{j=1,\cdots,m} l(Z_j) \quad \text{subject to} \quad \sum_{h=1}^{p} |Z_{hj}| \le s \text{ and } \mathbf{g}(Z_j) \le \mathbf{0} \right\}.$$

2.2 Variable Selection Combining Sample and Prior Constraint Information into Other Modified Lassos

The limitation of lasso is that all the regression coefficients share the same amount of shrinkage $\min_\beta \sum_{i=1}^n \left(y_i - \mathbf{x}_i^T\beta\right)^2 + \lambda \sum_{j=1}^p |\beta_j|$. Then Wang et al. (2007b) extend the lasso to the modified lasso* criterion which allows for different tuning parameters for different coefficients

$$\min_\beta \sum_{i=1}^n \left(y_i - \mathbf{x}_i^T\beta\right)^2 + \sum_{j=1}^p \lambda_j|\beta_j|.$$

In order to combining the sample and prior constraint information, variable selection procedure can be executed as follows

$$\min_\beta \sum_{i=1}^n \left(y_i - \mathbf{x}_i^T\beta\right)^2 + \sum_{j=1}^p \lambda_j|\beta_j| + \phi^T\mathbf{g}(\beta)$$

which not only uses the prior information but also overcomes the limitation of the traditional lasso procedure.

Similarly, the prior constraint information can be incorporated into Tibshirani et al. (2005)'s fused lasso which encourages sparsity in their differences, i.e. flatness of the coefficient profiles β_j as a function of j.

2.3 Variable Selection Combining Sample and Prior Constraint Information into Lasso in Nonlinear Models

The proposed variable selection procedure can not be directly used for nonlinear models, e.g. generalized linear models; Cox models and quantile regression models etc. But with the help of Wang and Leng (2007)'s *LSA*, the proposed variable selection procedure can be used for these nonlinear models. LSA regards

$$(\beta - \tilde{\beta})^T\hat{\Sigma}^{-1}(\beta - \tilde{\beta}) \tag{6}$$

as the least square approximation of the original loss $n^{-1}L_n(\beta)$ where $\tilde{\beta}$ is the unpenalized estimator obtained by minimizing $L_n(\beta)$, $\hat{\Sigma}^{-1} = n^{-1}\ddot{L}_n(\tilde{\beta})$ and $\ddot{L}_n(\cdot)$ is the second derivatives of the loss function $L_n(\cdot)$. The expression (6) is similar to the log-density $l(\beta)$ in (5). So it is clear that the lasso procedure incorporating prior constraint information can also be used for variable selection in nonlinear models with the help of the least squares approximation.

3. Some Theoretical Properties

In this section, we derive some theoretical results for the lasso combining the sample and prior constraint information that are analogous to those for the lasso and fused lasso (Knight and Fu (2000); Tibshirani et al (2005)). The penalized least squares criterion is

$$\sum_{i=1}^{n}(y_i - X_i^T\beta)^2 + \lambda_n^{(1)}\left(\sum_{j=1}^{p}|\beta_j| - s\right) + \mathbf{g}(\beta)^T\lambda_n^{(2)}$$

with $\beta = (\beta_1, \cdots, \beta_p)^T$ and $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})^T$, and the Lagrange multipliers $\lambda_n^{(1)}$ and $\lambda_n^{(2)}$ are functions of the sample size n. Let the optimal solution be $\widehat{\beta}_n$.

For simplicity, we assume that p is fixed as $n \to \infty$ and $\mathbf{g}(\cdot)$ are differential convex functions. The following theorem adequately illustrates the basic dynamics of the lasso combining sample and prior constraint information.

Theorem 1. If $\lambda_n^{(l)}/\sqrt{n} \to \lambda_0^{(l)} (l = 1, 2)$ and

$$C = \lim_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i^T\right)$$

is non-singular, then

$$\sqrt{n}(\widehat{\beta}_n - \beta) \xrightarrow{D} \arg\min_u V(u)$$

where

$$V(u) = -2u^T W + u^T C u + \lambda_0^{(1)}\sum_{j=1}^{p}\{u_j sgn(\beta_j) I(\beta_j \neq 0)\} + |u_j| I(\beta_j = 0) + \left(\frac{\partial \mathbf{g}(\beta)}{\partial\beta}u\right)^T \lambda_0^{(2)}$$

and $\mathbf{W}$ has an $n(0, \sigma^2 C)$ distribution.

Proof.

$$\sum_{i=1}^{n}(y_i - X_i^T\beta)^2 + \lambda_n^{(1)}\left(\sum_{j=1}^{p}|\beta_j| - s\right) + \mathbf{g}(\beta)^T\lambda_n^{(2)}$$

where $\lambda_n^{(1)}$ and $\lambda_n^{(2)}$ are functions of the sample size n. Define $V_n(u)$ by

$$V_n(u) = \sum_{i=1}^{n}\{(\varepsilon_i - u^T x_i/\sqrt{n})^2 - \varepsilon_i^2\} + \lambda_n^{(1)}\sum_{j=1}^{p}\left(|\beta_j + u_j/\sqrt{n}| - |\beta_j|\right) + \left(\mathbf{g}(\beta + u/\sqrt{n}) - \mathbf{g}(\beta)\right)^T\lambda_n^{(2)}$$

with $u = (u_1, \cdots, u_p)^T$ and note that $V_n(u)$ is minimized at $\sqrt{n}(\widehat{\beta}_n - \beta)$. First note that

$$\sum_{i=1}^{n}\{(\varepsilon_i - u^T x_i/\sqrt{n})^2 - \varepsilon_i^2\} \xrightarrow{D} -2u^T W + u^T Cu$$

with finite dimensional convergence holding trivially where

$$C = \lim_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n}X_i X_i^T\right) \quad \text{and} \quad W \sim n(0, \sigma^2 C).$$

We also have

$$\lambda_n^{(1)}\sum_{j=1}^{p}\left(|\beta_j + u_j/\sqrt{n}| - |\beta_j|\right) \xrightarrow{D} \lambda_0^{(1)}\sum_{j=1}^{p}\{u_j sgn(\beta_j)I(\beta_j \neq 0)\} + |u_j|I(\beta_j = 0)$$

and

$$\left(\mathbf{g}(\beta + u/\sqrt{n}) - \mathbf{g}(\beta)\right)^T\lambda_n^{(2)} = \left(\frac{\partial\mathbf{g}(\beta)}{\partial\beta}u\right)^T\lambda_0^{(2)}.$$

Thus $V_n(u) \xrightarrow{D} V(u)$, with finite dimensional convergence holding trivially where

$$V(u) = -2u^T W + u^T Cu + \lambda_0^{(1)}\sum_{j=1}^{p}\left[u_j sgn(\beta_j)I(\beta_j \neq 0)\right] + |u_j|I(\beta_j = 0) + \left(\frac{\partial\mathbf{g}(\beta)}{\partial\beta}u\right)^T\lambda_0^{(2)}.$$

Since V_n is convex and V has a unique minimum, it follows (Geyer, 1996) that

$$\arg\min_u V_n(u) = \sqrt{n}(\widehat{\beta}_n - \beta) \xrightarrow{D} \arg\min_u V(u).$$

Theorem 2. The procedure incorporating prior constraint information into lasso will increase efficiency of selecting significant variables for responses compared with the traditional lasso procedures.

Proof. Theoretically, the general lasso procedure is as follows

$$\tilde{\beta}_{\tilde{s}} = Arg\left\{\min_{\beta}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\beta)^2 \quad \text{subject to} \quad \sum_{h}|\beta_h| \le \tilde{s}\right\}$$

where GCV or CV is used to choose the tuning parameter $\tilde{s}$ which minimizes the estimated prediction errors

$$\tilde{s} = Arg\left\{\min_{s} PE_s\right\}.$$

If the estimator $\tilde{\beta}_{\tilde{s}}$ satisfies prior constraints $\mathbf{g}(\beta_{\tilde{s}}) \le \mathbf{0}$, it means that $\tilde{\beta}_{\tilde{s}}$ clearly minimizes the estimated prediction errors in a narrower region $\mathbf{g}(\beta) \le \mathbf{0}$. That is,

$$\tilde{\beta}_{\tilde{s}} = \widehat{\beta}_{\hat{s}}$$

where $\widehat{\beta}_{\hat{s}}$ is the estimator of parameter by the lasso procedure incorporating prior constraint information in (3). Now, we take Figure 1 as an example. From Figure 1, we know that $\tilde{\beta}_{\tilde{s}}$ lies in the region $ABCD$ and minimizes the estimated prediction errors. Moreover, we know that $\tilde{\beta}_{\tilde{s}}$ lies in the region $AEFD$. It is clear that $\tilde{\beta}_{\tilde{s}}$ minimizes the estimated prediction errors in the region above the line EF. Furthermore, the true model also lies in the region above the line EF. So we obtain that if $\tilde{\beta}_{\tilde{s}}$ selects the true variables correctly, that is, the nonzero components of $\tilde{\beta}_{\tilde{s}}$ are just the significant covariates, then $\widehat{\beta}_{\hat{s}}$ also selects the true variables correctly.

If $\tilde{\beta}_{\tilde{s}}$ doesn't select significant variables correctly, some prior constraint information may bring us into a narrower region to select these variables again. It will increase the efficiency of variable selection.

4. Standard error and degrees of freedom of the lasso estimate

Since our lasso procedure combining sample and prior constraint information is a nonlinear and nondifferentiable function of the response values even for a fixed value of s, it is difficult to obtain an accurate estimate of its standard error. The problem can be solved by bootstrap approach: either s can be fixed or we may optimize over s for each bootstrap sample.

Efron et al. (2004) consider a definition of degrees of freedom using the formula of Stein (1981):

$$df(h(y)) = \frac{1}{\sigma^2}\sum_{i=1}^{n} cov(y_i, h_i) = E\left\{\sum_{i=1}^{n}\frac{\partial h(y)}{\partial y_i}\right\}$$

where $y = (y_1, \cdots, y_n)^T$ is a multivariate normal vector with mean μ and covariance I, and $h(y)$ is an estimator, an almost differential function from $\mathbb{R}^n$ to $\mathbb{R}^n$. For the lasso with orthonormal design $\mathbf{X}^T\mathbf{X} = \mathbf{I}_{p\times p}$, the degrees of freedom are the number of non-zero coefficients. Tibshirani et al.(2005) show that the natural estimate of the degrees of freedom of the fused lasso is

$$\begin{aligned} df(\hat{(y)}) &= \#\{\text{non-zero coefficient block in} \hat{\beta}\} \\ &= p - \#\{\beta_j = 0\} - \#\{\beta_j - \beta_{j-1} = 0, \beta_j, \beta_{j-1} \neq 0\} \end{aligned}$$

similarly, the natural estimate of the degrees of freedom of the lasso incorporating prior constraint information is

$$df(\hat{y}) = p - \#\{\beta_j = 0\} - \#\{\mathbf{g}(\beta) = 0\}.$$

The degrees can be used for BIC-type tuning parameter selector.

5. Some Examples

In the following, we give three examples for illustration of the proposed procedure's practical applications in many models.

Example 1: **linear inequality constraints in linear models**

Wolak (1989) or (Silvapulle and Sen 2005 P9) consider the following double-log demand function

$$\log Q_t = \alpha + \beta_1 \log PE_t + \beta_2 \log PG_t + \beta_3 \log I_t + \gamma_1 D1_t + \gamma_2 D2_t + \gamma_3 D3_t + \epsilon_t$$

which is a linear model where

Q_t = aggregate electricity demand,

PE_t = average price of electricity to the residential sector,

PG_t = price of natural gas to the residential sector,

I_t = income per capita,

and $D1_t, D2_t, D3_t$ are seasonal dummy variables.

Prior knowledge suggests that

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} \geq \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

which are linear inequality constraints. A typical model selection question is whether or not the foregoing model provides a better fit than the simpler model

$$\log Q_t = \alpha + \gamma_1 D1_t + \gamma_2 D2_t + \gamma_3 D3_t + \epsilon_t.$$

Wolak (1989) or Wang et al. (2007b) discuss the model selection problem by a test method or by a variable selection method, respectively.

Example 2: **nonlinear inequality constraints in linear models**

Dufour (1989) considers the following econometric model

$$y_i = f(x_i, \beta) + \epsilon_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i2}^2 + \beta_5 x_{i3}^2 + 2\beta_6 x_{i2} x_{i3} + \epsilon_i.$$

This could be a production function or a unit cost function where y_i is the production or unit cost and $\{x_{i2}, x_{i3}\}$ are inputs. A problem of interest in econometrics is whether $f(x_i, \beta)$ is concave in x_i, which can be expressed by the following nonlinear inequality constraints

$$\beta_4 \leq 0, \ \beta_5 \leq 0, \ \beta_4\beta_5 - \beta_6^2 \geq 0.$$

Dufour (1989) discusses the model selection problem by a test method.

Example 3: **linear equality and inequality constraints in generalized linear models**

An assay was carried out with the bacterium *E. coli* strain 343/358(+) to evaluate the genotoxic effects of 9-aminoacridine (9-AA) and potassium chromate (KCr). Piegorsch (1990) and Silvapulle (1994) consider the following log-linear model

$$\log(1 - \pi_{ij}) = \mu + \alpha_i + \tau_j + \eta_{ij}. \tag{7}$$

to evaluate whether potassium chromate and 9-AA have a synergistic effect where $i = 1, 2$, $j = 1, \cdots, 5$ and

$$\pi_{ij} = Pr\{\text{positive response for a test unit in cell (i,j)}\}.$$

In fact, the log-linear model is just logistic regression model which is one of generalized linear models(*GLM*). To ensure that the parameters in (7) are identified, Piegorsch (1990) and Silvapulle (1994) impose the constraints $\alpha_1 = \tau_1 = \eta_{i1} = \eta_{1j} = 0$ for all (i, j) and

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \eta_{22} \\ \eta_{23} \\ \eta_{24} \\ \eta_{25} \end{pmatrix} \geq \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

which means that potassium chromate and 9-AA have a synergistic effect. The model selection problem is analyzed by a test in Piegorsch (1990), Silvapulle (1994) and Silvapulle and Sen (2005 P161).

6. Discussion

We proposed a modified lasso procedure combining prior constraint and sample information for variable selection and parameter estimation. The proposed procedure increases the efficiency of choosing the true model correctly because it executes variable selection and parameter estimation in a narrower region where the true parameters lie. The procedure may be computed by many quadratic programming methods.

Moreover, the idea of incorporating prior constraint information can be used for other lasso procedures, e.g. fused lasso and modified lasso procedure for an adaptive amount of shrinkage for each regression coefficient.

More work remains to be done. Efron et al. (2004)'s LARS is a good computational procedure which only needs p steps. But now it is not directly used for the lasso procedure incorporating prior constraint information. In our procedure, Monte Carlo estimator can be used for the initial estimator. How to extend LARS to the lasso procedure incorporating prior constraint information is an interesting topic for future study.

References

Dantig, G. B. and Eaves, B. C.(1974) *Studies in Optimization.* Washington: Mathematical Association of America.

Dorfman, J. and McIntosh, C.S. (2001) Imposing inequality restrictions: efficiency gains from economic theory. *Economics Letters*, **71**(2), 205-209.

Dufour, J.M. (1989) Nonlinear hypotheses, inequality restrictions and non-nested hypotheses: Exact simultaneous tests in linear regression. *Econometrics*, **57**, 335-355.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least angle regression. *The Annals of Statisitcs*, **32**, 407-489.

Efron, B. and Tibshirani, R. (1993) *An introduction to the bootstrap.* London: Chapman and Hall.

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.

Fu, W.J. (1998) Penalized regression: the bridge versus the lasso. *J. Comput. and Graph. Statist.*, **7**, 397-416.

Geyer, C. (1996) On the asymptotics of convex stochastic optimization. *Technical Report.* University of Minnesota, Minneapolis.

Geweke, J. (1986) Exact inference in the inequality constrained normal linear regression model. *J. Applied Econometrics*, **1**, 127-141.

Gill, P.E., Murray, W. and Saunders, M. A.(1997) Users guide for SqOPT 5.3: a Fortran package for large-scale linear and quadratic programming. *Technical report NA 97-4.* University of California, San Diego.

Judge, G.G. and Takayama, T. (1966) Inequality restrictions in regression analysis. *J. Amer. Statist. Assoc.*, **61**, 166-181

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

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

Moors J.J.A. and van Houwelingen, J.C. (1987) Estimation of linear models with inequality restrictions. Research Report FEW 291, Tilburg University.

Osborne, M.R., Presnell, B. and Turlach, B.A. (2000), A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis*, **20**, 389-404.

Park, M.Y. and Hastie, T. (2006) An L1 regularization-path algorithm for generalized linear models. Manuscript, Department of Statistics, Stanford University.

Piegorsch, W.W. (1990) One-sided significance tests for generalized linear models under dichotomous response. *Biometrics*, **46**, 309-316.

Rao C.R. and Toutenburg, H. (1995) *Linear models.* Springer-Verlag New York: Berlin Heidelberg

Rosset, S. (2004) *Tracking curved regularized optimization solution paths.* NIPS 2004.

Silvapulle, M.J. (1994) On tests against one-sided hypotheses in some generalized linear models. *Biometrics*, **50**, 853-858.

Silvapulle, M.J. and Sen, P.K.(2005) *Constrained statistical inference.* John Wiley & Sons, Inc., Hoboken: New Jersey.

Stahlecker, P. (1987) A prior Information und Minimax-Schätzung im linearen regressionsmodell. Athenäum, Frankfurt/M.

Stein, C. (1981) Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, **9**, 1131-1151.

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

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005) Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society, Series B.*, **67**, 91-108.

Wang, H. and Leng, C. (2007) Unified lasso estimation via least squares approximation. *Journal of the American Statistical Association*, to appear.

Wang, H., Li, G. and Jiang, G. (2007a) On the consistency of SCAD tuning parameters selector.

Biometrka, to appear.

Wang, H., Li, G.D. and Tsai, C.L. (2007b) Regression coefficient and autoregressive order shrinkage and selection via lasso. *Journal of the Royal Statistical Society, Series B.*, **69**, 63-78.

Wolak, F.A. (1989) Testing inequality constraints in linear econometric models. *Journal of Econometrics*, **41**, 205-235.

Yuan, M. and Lin, Y. (2007) On the nonnegative garrote estimator. *Journal of the Royal Statistical Society, Series B.*, To appear

Zhao, P. and Yu, B. (2004) Boosted lasso. *Technical Report* #678, Statistics, UC Berkeley.

Zou, H. (2006) The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, **101**, 1418-1429.