The Korean Journal of Applied Statistics (2012) DOI: http://dx.doi.org/10.5351/KJAS.2012.25.5.829
25*(5), 829–835*

On the Use of Adaptive Weights for the F_∞-Norm Support Vector Machine

Sungwan Bang[1] · Myoungshic Jhun[2]

[1]Department of Mathematics, Korea Military Academy; [2]Department of Statistics, Korea University

(Received July 19, 2012; Revised August 27, 2012; Accepted September 13, 2012)

Abstract

When the input features are generated by factors in a classification problem, it is more meaningful to identify important factors, rather than individual features. The F_∞-norm support vector machine(SVM) has been developed to perform automatic factor selection in classification. However, the F_∞-norm SVM may suffer from estimation inefficiency and model selection inconsistency because it applies the same amount of shrinkage to each factor without assessing its relative importance. To overcome such a limitation, we propose the adaptive F_∞-norm (AF_∞-norm) SVM, which penalizes the empirical hinge loss by the sum of the adaptively weighted factor-wise L_∞-norm penalty. The AF_∞-norm SVM computes the weights by the 2-norm SVM estimator and can be formulated as a linear programming(LP) problem which is similar to the one of the F_∞-norm SVM. The simulation studies show that the proposed AF_∞-norm SVM improves upon the F_∞-norm SVM in terms of classification accuracy and factor selection performance.

Keywords: Adaptive weight, F_∞-norm penalty, factor selection, feature selection, support vector machine.

1. Introduction

Consider a linear binary classification problem with a training data set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, where $\boldsymbol{x}_i \in R^p$ is p-dimensional input features for the i^{th} observation and $y_i \in \{-1, +1\}$ denotes its class label. When given new input features $\boldsymbol{x} \in R^p$, its classification is performed through the construction of a hyperplane $f(\boldsymbol{x}) = \beta_0 + \boldsymbol{x}^T\boldsymbol{\beta}$, where β_0 is an intercept and $\boldsymbol{\beta} \in R^p$ is a coefficient vector for the input feature. We can assign one of two possible classes to it through the classification rule $\text{sign}(f(\boldsymbol{x}))$.

The standard 2-norm SVM estimates the coefficients of the hyperplane $(\hat{\beta}_0, \hat{\boldsymbol{\beta}})$ by maximizing the margin between the training points for the positive and negative classes (Vapnik, 1995; Hastie *et al.*, 2001)

$$\arg\max_{\beta_0, \boldsymbol{\beta}} \frac{1}{||\boldsymbol{\beta}||_2}, \tag{1.1}$$

$$\text{subject to } \; y_i\left(\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}\right) \geq 1 - \xi_i, \quad \xi_i \geq 0, \; \forall\, i, \quad \text{and} \quad \sum_{i=1}^n \xi_i \leq s,$$

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2010-0009204).
[2]Corresponding author: Professor, Department of Statistics, Korea University, Seoul 136-701, Korea.
E-mail: jhun@korea.ac.kr

where $\xi_i\,(i=1,\ldots,n)$ are slack variables and s is a prespecified positive number that controls the overlap between the two classes. It can be shown that the optimization problem (1.1) of the 2-norm SVM can be equivalently expressed as the following 'loss + penalty' formulation (Vapnik, 1995; Hastie *et al.*, 2001)

$$\left(\hat{\beta}_0,\hat{\boldsymbol{\beta}}\right)^{L_2}=\arg\min_{\beta_0,\boldsymbol{\beta}}\sum_{i=1}^{n}\left[1-y_i\left(\beta_0+\boldsymbol{x}_i^T\boldsymbol{\beta}\right)\right]_+ +\lambda||\boldsymbol{\beta}||_2^2, \tag{1.2}$$

where the loss function $(1-yf)_+$ is called hinge loss and the subscript "+" indicates the positive part ($z_+=\max(z,0)$). The tuning parameter λ controls the trade-off between loss and penalty, and there is a one-to-one correspondence between s in (1.1) and λ in (1.2).

The standard 2-norm SVM has been successfully applied to various classification areas due to its great flexibility and high level of classification accuracy; however, the 2-norm SVM classifier cannot automatically select input features. In classification problem, the feature selection plays important role in model building: excluding important features may produce severely biased estimation results; however, including irrelevant features may make it difficult to interpret the resultant model and reduce its classification accuracy. To accomplish the goal of automatic feature selection in the SVM classifier, Zhu *et al.* (2003) considered the 1-norm SVM by replacing the ridge penalty in (1.2) (Hoerl and Kennard, 1970) with the lasso penalty (Tibshirani, 1996)

$$\left(\hat{\beta}_0,\hat{\boldsymbol{\beta}}\right)^{L_1}=\arg\min_{\beta_0,\boldsymbol{\beta}}\sum_{i=1}^{n}\left[1-y_i\left(\beta_0+\boldsymbol{x}_i^T\boldsymbol{\beta}\right)\right]_+ +\lambda||\boldsymbol{\beta}||_1. \tag{1.3}$$

To overcome the lack of oracle property of the lasso penalty, Zou (2007) suggested the hybrid SVM by adopting the adaptive lasso penalty (Zou, 2006)

$$\left(\hat{\beta}_0,\hat{\boldsymbol{\beta}}\right)^{H}=\arg\min_{\beta_0,\boldsymbol{\beta}}\sum_{i=1}^{n}\left[1-y_i\left(\beta_0+\boldsymbol{x}_i^T\boldsymbol{\beta}\right)\right]_+ +\lambda\sum_{j=1}^{p}\left|\hat{\beta}_j^{L_2}\right|^{-\gamma}\left|\beta_j\right|, \tag{1.4}$$

where the weights $|\hat{\beta}_j^{L_2}|^{-\gamma}$ are computed by the 2-norm SVM and $\gamma>0$ is some pre-specified number.

In this paper, we consider a classification problem in which the input features of the classifier are generated by factors; subsequently, the model is best interpreted in terms of significant factors. In such a situation, it is more meaningful to identify important factors, rather than individual features. In a regression problem, some regularization methods have been developed for automatic factor selection that include the group lasso (Yuan and Lin, 2006), the adaptive group lasso (Wang and Leng, 2008), the penalized method using composite absolute penalties (Zhao *et al.*, 2009), and the adaptive sup-norm regularized quantile regression (Bang and Jhun, 2012a). In order to automatically and simultaneously select significant factors to estimate the SVM classifier, Zou and Yuan (2008) proposed the F_∞-norm SVM that penalizes the empirical hinge loss by the sum of the factor-wise L_∞-norm penalty; however, the F_∞-norm SVM may suffer from estimation inefficiency and model selection inconsistency in the same way as the lasso and the 1-norm SVM. As a remedy, we suggest adopting the idea of the adaptive lasso (Zou, 2006) and propose the adaptive F_∞-norm SVM, which allows different amounts of shrinkage to be imposed on different factors according to their relative importance.

The rest of the paper is organized as follows. In Section 2, we introduce the adaptive F_∞-norm SVM and show that the adaptive F_∞-norm SVM can be formulated as a linear programming(LP)

problem. In Section 3, we evaluate the proposed method through simulation studies. Section 4 contains the concluding remarks.

2. Methodology

2.1. The adaptive F_∞-norm SVM

Suppose that the input features of the SVM classifier are generated by G factors, that is, the features $\boldsymbol{x}^T = (x_1, \ldots, x_p)$ are grouped into G factors as $\boldsymbol{x}^T = (\boldsymbol{x}_{(1)}^T, \ldots, \boldsymbol{x}_{(G)}^T)$, where $\boldsymbol{x}_{(j)}^T = (x_{j1}, \ldots, x_{jp_j})$ is a group of p_j features for $j = 1, \ldots, G$ and $\sum_{j=1}^G p_j = p$. Then the classifier can be represented by

$$f(\boldsymbol{x}) = \beta_0 + \boldsymbol{x}^T\boldsymbol{\beta} = \beta_0 + \sum_{j=1}^G \boldsymbol{x}_{(j)}^T\boldsymbol{\beta}_{(j)}, \tag{2.1}$$

where $\boldsymbol{\beta}_{(j)} = (\beta_{j1}, \ldots, \beta_{jp_j})^T \in R^{p_j}$ is the coefficient vector associated with the j^{th} factor. In such a situation, it is more reasonable to automatically and simultaneously select significant factors, rather than individual derived features. To accomplish the goal of automatic factor selection in the SVM classifier, Zou and Yuan (2008) proposed the F_∞-norm SVM, which penalizes the empirical hinge loss by the sum of the factor-wise L_∞-norm penalty

$$\left(\hat{\beta}_0, \hat{\boldsymbol{\beta}}\right)^{F_\infty} = \arg\min_{\beta_0, \boldsymbol{\beta}} \sum_{i=1}^n \left[1 - y_i\left(\beta_0 + \sum_{j=1}^G \boldsymbol{x}_{i,(j)}^T\boldsymbol{\beta}_{(j)}\right)\right]_+ + \lambda\sum_{j=1}^G \left|\left|\boldsymbol{\beta}_{(j)}\right|\right|_\infty, \tag{2.2}$$

where the infinite norm is defined by

$$\left|\left|\boldsymbol{\beta}_{(j)}\right|\right|_\infty = \max\left\{\left|\beta_{j1}\right|, \ldots, \left|\beta_{jl}\right|, \ldots, \left|\beta_{jp_j}\right|\right\}, \quad j = 1, \ldots, G. \tag{2.3}$$

Owing to the nature of the L_∞-norm, the F_∞-norm SVM is able to simultaneously select significant features in a grouped manner, hence it is a more appropriate tool for factor selection than the 1-norm SVM in (1.3).

In the F_∞-norm SVM (2.2), the same tuning parameter λ is used for each factor-wise L_∞-norm penalty; subsequently, the same amount of shrinkage is imposed on each factor without assessing its relative importance. In this paper, we suggest that different factors should be penalized differently according to their relative importance. In a typical linear regression setting, some researchers have suggested that an excessive penalty applied to important variables can degrade estimation efficiency and may influence model selection consistency (Fan and Li, 2001; Leng *et al.*, 2006; Yuan and Lin, 2007; Zou, 2006). Ideally, small penalties should be used on the significant factors that we want to retain in the model; however, large penalties should be imposed on irrelevant factors in order to eliminate them from the final SVM classifier.

In order to further improve upon the F_∞-norm SVM, we adopt the idea of the adaptive lasso from Zou (2006). Suppose that we first fit the 2-norm SVM classifier using all available input features. Then we suggest using the 2-norm SVM estimator $\boldsymbol{\beta}_{(j)}^{L_2}$ $(j = 1, \ldots, G)$ to construct the adaptively weighted F_∞-norm penalty and propose the following adaptive F_∞-norm (AF_∞-norm) SVM

$$\left(\hat{\beta}_0, \hat{\boldsymbol{\beta}}\right)^{\mathrm{AF}_\infty} = \arg\min_{\beta_0, \boldsymbol{\beta}} \sum_{i=1}^n \left[1 - y_i\left(\beta_0 + \sum_{j=1}^G \boldsymbol{x}_{i,(j)}^T\boldsymbol{\beta}_{(j)}\right)\right]_+ + \lambda\sum_{j=1}^G \left|\left|\hat{\boldsymbol{\beta}}_{(j)}^{L_2}\right|\right|_\infty^{-\gamma} \left|\left|\boldsymbol{\beta}_{(j)}\right|\right|_\infty. \tag{2.4}$$

By using the solution to (2.4), the fitted AF_∞-norm SVM classifier is $\hat{f}(\boldsymbol{x}_i) = \hat{\beta}_0^{\mathrm{AF}_\infty} + \sum_{j=1}^{G} \boldsymbol{x}_{i,(j)}^T \hat{\boldsymbol{\beta}}_{(j)}^{\mathrm{AF}_\infty}$, and the classification rule is $\mathrm{sign}(\hat{f}(\boldsymbol{x}))$. The factor-wise adaptively weighted L_∞-norm penalty in (2.4) has been considered as the common variable selection in multiple quantile regression (Bang and Jhun, 2012b) as well as in the multicategory SVM (Zhang *et al.*, 2008). It is interesting to note that in the context of quantile regression, Bang and Jhun (2012a) also suggested using a penalty term similar to the one of (2.4) to automatically select significant factors. When each individual feature is considered as a factor, that is, $p_1 = \cdots = p_G = 1$, the AF_∞-norm SVM reduces to the hybrid SVM described in (1.4). Therefore, the AF_∞-norm SVM is a generalization of the hybrid SVM; the AF_∞-norm SVM has the capability to automate factor selection in the model fitting, whereas the hybrid SVM contains no information on the factors.

2.2. Computing algorithm

In this section, we show that the optimization problem (2.4) can be formulated as a linear programming(LP) problem. To derive the LP formulation of the AF_∞-norm SVM, we introduce n slack variables such that

$$\xi_i = \left[1 - y_i \left(\beta_0 + \sum_{j=1}^{G} \boldsymbol{x}_{i,(j)}^T \boldsymbol{\beta}_{(j)}\right)\right]_+, \quad i = 1, 2, \ldots, n. \tag{2.5}$$

With such notation, the optimization problem (2.4) can be expressed as

$$\begin{aligned} &\arg\min_{\beta_0, \boldsymbol{\beta}} \sum_{i=1}^{n} \xi_i + \lambda \sum_{j=1}^{G} \left|\left|\hat{\boldsymbol{\beta}}_{(j)}^{L_2}\right|\right|_\infty^{-\gamma} \left|\left|\boldsymbol{\beta}_{(j)}\right|\right|_\infty. \\ &\text{subject to } \; y_i \left(\beta_0 + \sum_{j=1}^{G} \boldsymbol{x}_{i,(j)}^T \boldsymbol{\beta}_{(j)}\right) \geq 1 - \xi_i \quad \text{and} \quad \xi_i \geq 0, \; \forall\, i. \end{aligned} \tag{2.6}$$

To further simplify the optimization problem (2.6), let the new variable $M_j = ||\boldsymbol{\beta}_{(j)}||_\infty$ for $j = 1, 2, \ldots, G$ and write β_0 as $\beta_0^+ - \beta_0^-$ and $\boldsymbol{\beta}_{(j)}$ as $\boldsymbol{\beta}_{(j)}^+ - \boldsymbol{\beta}_{(j)}^-$, where $\beta_0^+ \geq 0$, $\beta_0^- \geq 0$, $\boldsymbol{\beta}_{(j)}^+ = (\beta_{j1}^+, \ldots, \beta_{jp_j}^+)^T \geq \mathbf{0}$, and $\boldsymbol{\beta}_{(j)}^- = (\beta_{j1}^-, \ldots, \beta_{jp_j}^-)^T \geq \mathbf{0}$. Then the adaptively weighted F_∞-norm penalty term $\lambda \sum_{j=1}^{G} ||\hat{\boldsymbol{\beta}}_{(j)}^{L_2}||_\infty^{-\gamma} ||\boldsymbol{\beta}_{(j)}||_\infty$. can be linearly reformulated with some linear inequality constraints. By using these variables and notations, the optimization problem (2.4) and (2.6) can be equivalently expressed as

$$\begin{aligned} &\arg\min_{\beta_0, \boldsymbol{\beta}} \sum_{i=1}^{n} \xi_i + \lambda \sum_{j=1}^{G} \left|\left|\hat{\boldsymbol{\beta}}_{(j)}^{L_2}\right|\right|_\infty^{-\gamma} M_j \\ &\text{subject to } \; y_i \left(\beta_0^+ - \beta_0^- + \sum_{j=1}^{G} \boldsymbol{x}_{i,(j)}^T \left(\boldsymbol{\beta}_{(j)}^+ - \boldsymbol{\beta}_{(j)}^-\right)\right) \geq 1 - \xi_i \quad \text{and} \quad \xi_i \geq 0, \; \forall\, i, \\ &\qquad M_j \geq \beta_{jl}^+ + \beta_{jl}^-, \; \forall\, j, l, \\ &\qquad \beta_0^+ \geq 0, \; \beta_0^- \geq 0 \quad \text{and} \quad \beta_{jl}^+ \geq 0, \; \beta_{jl}^- \geq 0, \; \forall\, j, l. \end{aligned} \tag{2.7}$$

This LP formulation of the AF_∞-norm SVM is similar to the one of the F_∞-norm SVM, but the key difference between these two LP formations is that the AF_∞-norm SVM uses the adaptive weights for penalizing different factors according to their relative importance in the objective function (2.7);

however, the F_∞-norm SVM uses the equal weight for each factor. In the present study, we used the *lpSolve* package provided in R to implement the above LP problem and our R code is available for interested readers upon request.

3. Simulation Studies

In this section, simulation studies were conducted to evaluate the finite sample performance of the proposed AF_∞-norm SVM method. For comparison purpose, we also included the 2-norm SVM, the 1-norm SVM, the hybrid SVM, and the F_∞-norm SVM in the simulation. Two simulated models were considered, which are similar to those in Zou and Yuan (2008). For each of the simulated models, we generated 100 training observations, along with 100 independent observations for the validation set. The models were fitted on training data only, and the validation set was used to select the tuning parameter λ. To evaluate the classification accuracy for each method, we also independently generated 10,000 observations as a test set. The model selection performance was measured by the number of correctly selected factors and features(NC), the number of incorrectly selected factors and features(NIC), and the number of times that the true model is correctly identified. To assess the sampling variability, this procedure was repeated 100 times independently. In the tables, the classification error, NC, and NIC were reported on averages over 100 runs. The numbers given in parentheses are the standard deviations of the classification errors.

3.1. Example 1

In this example, we first generated 8 factors $z_1, \ldots, z_8$ from the standard normal distribution with $\text{Cov}(z_j, z_{j'}) = 0.5^{|j-j'|}$. In addition, 40 random variables $w_1, \ldots, w_{40}$ were independently generated from the standard normal distribution. Then the 40 covariates were obtained by

$$
\begin{aligned}
&x_{1l} = 2^{-\frac{1}{2}}(z_1 + w_l), \quad \text{for } l = 1, \ldots, 6, && x_{2l} = 2^{-\frac{1}{2}}(z_2 + w_{l+6}), \quad \text{for } l = 1, \ldots, 4,\\
&x_{3l} = 2^{-\frac{1}{2}}(z_3 + w_{l+10}), \quad \text{for } l = 1, \ldots, 6, && x_{4l} = 2^{-\frac{1}{2}}(z_4 + w_{l+16}), \quad \text{for } l = 1, \ldots, 5,\\
&x_{5l} = 2^{-\frac{1}{2}}(z_5 + w_{l+21}), \quad \text{for } l = 1, \ldots, 4, && x_{6l} = 2^{-\frac{1}{2}}(z_6 + w_{l+25}), \quad \text{for } l = 1, \ldots, 5,\\
&x_{7l} = 2^{-\frac{1}{2}}(z_7 + w_{l+30}), \quad \text{for } l = 1, \ldots, 4, && x_{8l} = 2^{-\frac{1}{2}}(z_8 + w_{l+34}), \quad \text{for } l = 1, \ldots, 6.
\end{aligned}
$$

The binary response y was generated by a logistic model with $P(y = 1) = \exp(f(\boldsymbol{x}))/(1+\exp(f(\boldsymbol{x})))$ and $P(y = -1) = 1 - P(y = 1)$, where the true hyperplane is given by

$$
\begin{aligned}
f(\boldsymbol{x}) = {} & 1.2x_{11} - 0.8x_{12} + 1.6x_{13} + 1.5x_{14} - 3x_{15} + x_{16} + x_{21} - 0.9x_{22} - 1.1x_{23} - 1.3x_{24}\\
& + 1.5x_{61} + 2x_{62} - x_{63} - 2.5x_{64} + 3x_{65}.
\end{aligned}
$$

This model has 3 important factors and 15 important covariates. Table 3.1 summarizes the performance of each method for the above simulated model. As expected, the adaptively regularized methods (*i.e.*, the hybrid-SVM and the AF_∞-norm SVM) were superior to their non-adaptive counterparts in terms of classification error and model selection performance. We can see that the number of incorrectly selected factors for the proposed AF_∞-norm SVM is smaller than that of the Hybrid SVM (while number of incorrectly selected variables for the Hybrid SVM is smaller than that of the proposed AF_∞-norm SVM); however, the proposed AF_∞-norm SVM outperformed the others in terms of classification error.

Table 3.1. Simulation results for Example 1

Method	Test error	No. of factors selected		No. of variables selected		No. of times
		NC	NIC	NC	NIC	for true model
2-norm SVM	0.2034 (0.0187)	3.00	5.00	15.00	25.00	0
1-norm SVM	0.1857 (0.0203)	3.00	4.78	13.38	13.40	0
Hybrid SVM	0.1815 (0.0203)	3.00	3.93	12.18	7.46	0
F_∞-norm SVM	0.1696 (0.0161)	3.00	4.17	15.00	21.13	0
AF_∞-norm SVM	0.1561 (0.0168)	2.99	2.86	14.96	14.90	6

The numbers in parentheses are standard deviations.

Table 3.2. Simulation results for Example 2

Method	Test error	No. of factors selected		No. of variables selected		No. of times
		NC	NIC	NC	NIC	for true model
2-norm SVM	0.1399 (0.0114)	2.00	13.00	20.00	40.00	0
1-norm SVM	0.1471 (0.0110)	2.00	3.19	9.09	4.09	0
Hybrid SVM	0.1494 (0.0113)	2.00	2.96	8.03	3.48	0
F_∞-norm SVM	0.1262 (0.0105)	2.00	1.30	20.00	13.00	24
AF_∞-norm SVM	0.1239 (0.0090)	2.99	0.42	20.00	4.20	66

The numbers in parentheses are standard deviations.

3.2. Example 2

In this example, 5 random variables $z_1, \ldots, z_5$ were independently generated from a standard normal distribution. In addition, 60 standard normal variables $\{\varepsilon_j\}_{j=1}^{60}$ were generated. Then the 60 covariates were obtained by

$$x_{1l} = z_1 + \varepsilon_l, \quad \text{for } l = 1, \ldots, 10, \qquad x_{2l} = z_2 + \varepsilon_{l+10}, \quad \text{for } l = 1, \ldots, 10,$$
$$x_{3l} = z_3 + \varepsilon_{l+20}, \quad \text{for } l = 1, \ldots, 10, \qquad x_{4l} = z_4 + \varepsilon_{l+30}, \quad \text{for } l = 1, \ldots, 10,$$
$$x_{5l} = z_5 + \varepsilon_{l+40}, \quad \text{for } l = 1, \ldots, 10, \qquad x_{j1} = \varepsilon_j, \quad \text{for } j = 6, 7, \ldots, 15.$$

The binary response y was generated by a logistic model with $P(y = 1) = \exp(f(\boldsymbol{x}))/(1+\exp(f(\boldsymbol{x})))$ and $P(y = -1) = 1 - P(y = 1)$, where the true hyperplane is given by

$$f(\boldsymbol{x}) = 1 + 4z_1 + z_2.$$

This model has 2 important factors and 20 important covariates. Table 3.2 reports the performance of each method. In terms of classification error and factor selection performance, the factor-wise penalized methods (*i.e.*, the F_∞-norm SVM and the AF_∞-norm SVM) worked better than their corresponding individual penalized methods. We can see that the factor-wise penalized methods always selected all important features; however the individual penalized methods (*i.e.*, the 1-norm SVM and the hybrid SVM) tended to eliminate some important features. In particular, it is shown in Table 3.2 that the AF_∞-norm SVM performed better than the F_∞-norm SVM.

4. Concluding Remarks

In this paper, we have proposed the AF_∞-norm SVM that performs simultaneous classification and factor selection. The AF_∞-norm SVM consists of two stages. At the first stage, we use the coefficients of the 2-norm SVM classifier to construct the factor-wise adaptively weighted L_∞-norm penalty. Then we solve the adaptively weighted F_∞-norm SVM by using the standard LP technique.

Simulation results show that the proposed the AF_∞-norm SVM outperforms the F_∞-norm SVM as well as the individual penalized methods (*i.e.*, the 1-norm SVM and the hybrid SVM) in terms of classification accuracy and model selection performance.

To incorporate the factor information into regularized model fitting, we can also use the group lasso penalty (Yuan and Lin, 2006; Wang and Leng, 2008) as an alternative to the L_∞-norm penalty. Considering computational efficiency, we favor the factor-wise L_∞-norm penalty because the optimization problem using the group lasso penalty is formulated as a nonlinear programming(NLP) problem. In a simulation study, we know the group information of input features, but this is not available in applications of real data. In such situations, we can employ a clustering method to cluster the features into several groups such as the hierarchical clustering and the partitioning around medoids(PAM) algorithm (Kaufman and Rousseeuw, 1990).

References

Bang, S. and Jhun, M. (2012a). Simultaneous estimation and factor selection in quantile regression via adaptive sup-norm regularization, *Computational Statistics and Data Analysis*, **56**, 813–826 .

Bang, S. and Jhun, M. (2012b). Adaptive sup-norm regularized simultaneous multiple quantiles regression, *Statistics*, accepted for publication.

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.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning*, Springer-Verlag, New York.

Hoerl, A. and Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics*, **12**, 55–67.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, New York.

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

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

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*, Springer-Verlag, New York.

Wang, H. and Leng, C. (2008). A note on adaptive group lasso, *Computational Statistics and Data Analysis*, **52**, 5277–5286.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society, Series B*, **68**, 49–67.

Yuan, M. and Lin, Y. (2007). On the nonnegative garrote estimator, *Journal of the Royal Statistical Society, Series B*, **69**, 143–161.

Zhang, H., Liu, Y., Wu, Y. and Zhu, J. (2008). Variable selection for multicategory svm via sup-norm regularization, *Electronic Journal of Statistics*, **2**, 149–167.

Zhao, P., Rocha, G. and Yu, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection, *The Annals of Statistics*, **37**, 3468–3497.

Zhu, J., Rosset, S., Hastie, T. and Tibshirani, R. (2003). 1-norm support vector machine, *Neural Information Proceeding Systems*, **16**.

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

Zou, H. (2007). An improved 1-norm SVM for simultaneous classification and variable selection, In *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics*.

Zou, H. and Yuan, M. (2008). The F_∞-norm support vector machine, *Statisitca Sinica*, **18**, 379–398.