On nonnegative garrote estimator in a linear regression model

Leila Mohammadi
EURANDOM, P.O.Box 513, 5600 MB Eindhoven, The Netherlands,
mohammadi@eurandom.tue.nl
and
Sara van de Geer
Seminar für Statistik, ETH-Zentrum, LEO D11, 8092 Zürich, Switzerland,
geer@stat.math.ethz.ch

Abstract: Subset selection regression is a frequently used statistical method. It waives some of the predictor variables and the prediction equation is based on the remaining set of variables. Subset selection is simple and it clearly reduces the variance. An other method for reducing the variance is ridge regression. Usually, subset selection is not as accurate as ridge. The problems with ridge regression are for example: 1) it is not scale invariant 2) it does not give a simple equation. We need an intermediate method which selects subsets, is stable and gains its accuracy by selective shrinking. Breiman (1995) proposed a new method, called the nonnegative (nn) garrote. In this lecture, in a linear regression model, we consider the nonnegative garrote estimator of the coefficients as introduced by Breiman (1995). This estimator shrinks the least square estimator by a parameter λ in the orthogonal case. In an especial case of λ, we prove the nn-garrote estimator is consistent and its MSE converges to zero. We also obtain the rate of convergence.

1. Introduction.

Subset selection regression is a frequently used statistical method. Suppose we are given data of the form $\{(y_n, x_{1n}, ..., x_{Mn}), n = 1, ..., N\}$. Subset selection waives some of the predictor variables $x_1, ..., x_M$ and then the prediction equation for y is based on the remaining set of variables. Subset selection is simple and it clearly reduces the variance if M is large. An other method for reducing the variance is ridge regression. In this method we assume λ to be a positive value (shrinkage parameter) and the coefficients are estimated by $(X^T X + \lambda I)^{-1} X^T Y$. Let $y = \sum_k \beta_k x_k + \epsilon$. If a few of the $\{\beta_k\}$ are nearly zero and the rest are large, then subset selection gives more accurate prediction than ridge regression. If it is not the case, then ridge regression acts better. Thus usually, subset selection is not as accurate as ridge. The problems with ridge regression are for example: 1) it is not scale invariant 2) it does not give a simple equation. As it is known, we need an intermediate method which selects subsets, is stable and gains its accuracy by selective shrinking.

Breiman (1995) proposed a new method, called the nonnegative (nn) garrote.

Consider the regression model

$$y_i = \sum_{j=1}^{m} \beta_j x_{ij} + \epsilon_i, \quad i = 1, ..., n,$$

where $\epsilon_1, ..., \epsilon_n$ are independent with mean zero. The nn-garrote estimator for β_j is $\hat{\beta}_{j,G} = c_j \hat{\beta}_j$ where $\hat{\beta}_j$ is the OLS (ordinary least square) estimator for β_j and where c_j is a shrinkage parameter selected by minimizing

$$\sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} (y_i - \sum_{j=1}^{m} c_j \hat{\beta}_j x_{ij})^2$$

under constraints $c_j \geq 0$ and $\sum_{j=1}^{m} c_j \leq s$. Clearly when s decreases, more of the $\{c_j\}$ become zero and the remaining nonzero $\{c_j\}$ are shrunken. The nn-garrote eliminates some variables, shrinks others, and is stable and scale invariant.

Breiman experimented subset selection and nn-garrote on two well-known data sets. These experiments showed nn-garrote has the mean prediction error and mean model error less than subset selection. In experiments on real and simulated data, the nn-garrote produces lower prediction error than ordinary subset selection. It is also comparable to ridge regression. The tests showed that subset selection is unstable, ridge is very stable and the nn-garrote is intermediate.

Now consider the regression model

$$y_i = \sum_{j=1}^{m} \beta_j x_{ij} + \epsilon_i, \quad \epsilon_i \sim N(0, 1), \quad i = 1, ..., n, \tag{1}$$

where $\epsilon_1, ..., \epsilon_n$ are independent and where $\{x_{ij}\}$ are orthogonal.

In nn-garrote the expression $\sum_{i=1}^{n} (y_i - \sum_{j=1}^{m} c_j \hat{\beta}_j x_{ij})^2$ is minimized under the constraints $c_j \geq 0$, for all j, and $\sum_{j=1}^{m} c_j = s$. The solution is easily obtained by $c_j = (1 - \frac{\lambda^2}{\hat{\beta}_j^2})^+$, where λ is determined from s by the condition $\sum_{j=1}^{m} c_j = s$ and where a^+ indicates the positive part of a. Therefore, in this case, nn-garrote coefficients are

$$\hat{\beta}_{j,G} = (1 - \frac{\lambda^2}{\hat{\beta}_j^2})^+ \hat{\beta}_j, \quad \lambda > 0.$$

Under the model (1), the prediction error (see Breiman (1995)) of any estimator $\tilde{\beta}$ is

$$PE(\tilde{\beta}) = n + \sum_{i=1}^{n} (\sum_{j=1}^{m} (\beta_j - \tilde{\beta}_j) x_{ij})^2 = n + MSE(\tilde{\beta}),$$

where $MSE(\tilde{\beta})$ means the mean square error of $\tilde{\beta}$. This component is the prediction error due to lack of fit to the underlying model and also called model error.

In this paper we consider the model (1) and study the behavior of the nn-garrote estimator when $\lambda = \lambda_n \to 0$. For example, we consider the case $\lambda = c\sqrt{\log n/n}$, $c > \sqrt{8}$. We eventually prove in the orthogonal case $X^T X = nI$, the nn-garrote estimator is consistent and its model error tends to zero as $n \to \infty$ and we obtain the rate of convergence.

2. Consistency of the nonnegative garrote estimator.

Convergence in probability - and hence consistency - can be defined in any metric space. Let $\{X_n\}$ be a sequence of random variables taking values in a metric space (M, d), where $d(x, y)$ denotes the distance between points x and y. Let x be a point in M, then we say that X_n converges in probability to x if $d(X_n, x)$ converges in probability to 0, that is if $P(d(X_n, x) > \epsilon) \to 0$ for every positive ϵ. In the simple case where (M, d) is two-dimensional Euclidean space $\mathbf{R^2}$ with the Euclidean distance, a sequence of random vectors $X_n = (Y_n, Z_n)$ converges to a point (y, z) in $\mathbf{R^2}$, if and only if Y_n converges in probability to y and Z_n converges in probability to z. Consistency of an estimator (in any metric space) means that the sequence of estimators converges in probability to the parameter it is supposed to be estimated. We shall write $\to^P$ for convergence of probability.

There is an old result that used to be called Slutsky's lemma, which says:

Slutsky's lemma. *If $X_n \to^P x$ in a metric space (M, d) and f is a function from (M, d) into another metric space (M', d') which is continuous at x, then $f(X_n) \to^P f(x)$ in (M', d').*

Proof. A function $f : (M, d) \to (M', d')$ is continuous at the point $x \in M$ if for any $\epsilon > 0$ there exists $\delta > 0$ such that $d'(f(x), f(y)) < \epsilon$ for every $y \in M$ with $d(x, y) < \delta$. So

$$0 \le P(d'(f(X_n), f(x)) \ge \epsilon) \le P(d(X_n, x) \ge \delta) \to 0$$

because $X_n \to^P x$.□

First consider the simple case

$$y_i = \beta x_i + \epsilon_i, \quad \epsilon_i \sim N(0, 1), \quad i = 1, ..., n, \quad \sum_{i=1}^{n} x_i^2 = n, \tag{2}$$

where $\epsilon_1, ..., \epsilon_n$ are independent. Now we have just a parameter β and the nn-garrote estimator for β is defined by

$$\hat{\beta}_G = (1 - \frac{\lambda^2}{\hat{\beta}^2})^+ \hat{\beta}, \quad \lambda > 0,$$

where $\hat{\beta}$ is the OLS estimator, i.e.,

$$\hat{\beta} = T_n = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} = \frac{1}{n}\sum_{i=1}^n x_i y_i.$$

The first lemma considers the case $\lambda = \lambda_n \to 0$ and proves the consistency of $\hat{\beta}_G$ when β is fixed.

Lemma 1. *Let β be fixed (independent of n) and $\lambda = \lambda_n \to 0$ as $n \to \infty$. Then $\hat{\beta}_G$ is consistent for β.*

Proof. If we view λ_n as a (degenerate) random variable, $\lambda_n \to 0$ means that $\lambda_n \to^P 0$. In view of the consistency of the least squares estimator we have $T_n \to^P \beta$. If $\beta \neq 0$, the function $f(t, \lambda) = (1 - (\lambda/t)^2)^+ \cdot t$ is continuous at the point $(t, \lambda) = (\beta, 0)$, so consistency follows from Slutsky's lemma. If $\beta = 0$, consistency follows from $T_n \to^P \beta$, because $|\hat{\beta}_G| \leq |T_n|$. □

It is easy to extend Lemma 1 immediately to the case where m and $\beta = (\beta_1, ..., \beta_m)$ are fixed (independent of n) and the columns of the matrix X are not necessarily orthogonal. In fact the orthogonality only plays a role in that it ensures that there is an explicit expression for $\hat{\beta}_G$ which involves a number $\lambda = \lambda_n$ and the lemma can then be formulated by requiring that $\lambda_n \to 0$ as $n \to \infty$. But for any fixed m and $\beta \neq 0$, this condition is equivalent to the assumption that $s = s_n \to m$ as $n \to \infty$. To prove this note that the function $g(u_1, ..., u_m) = \sum_{i=1}^n (y_i - \sum_{j=1}^m u_j \hat{\beta}_j x_{ij})^2$ is a quadratic function of u_j and if $s = \sum_{j=1}^m u_j$ is fixed, it has a unique minimizer (say $\mathbf{c} = (c_1, ..., c_m)$). The minimizer $\mathbf{c}$ is a continuous function of s. If $s = \sum_{j=1}^m u_j = m$, then $\mathbf{c} = (1, ..., 1)$, because (1,...,1) is the unique minimizer of g in $\mathbf{R}^m$. So $\mathbf{c} = \mathbf{c}_n \to^P (1, ..., 1)$ as $n \to \infty$.

Lemma 2. *Let m and β be fixed (independent of n) and $s = s_n \to m$ as $n \to \infty$. If $X^T X$ is nonsingular for every n and the diagonal elements of $(X^T X)^{-1}$ tend to infinity as $n \to \infty$, then $\hat{\beta}_G$ is consistent for β.*

Proof. As mentioned, $s = s_n \to m$ means $\mathbf{c}_n \to^P (1, ..., 1)$. The garrote estimator of β_j is obviously a continuous function of $c_{j,n}$ and the OLS estimator of β_j is consistent. So consistency will follow from Slutsky's lemma. □

We now show that $|\hat{\beta}_G - \beta| = O_p(n^{-q})$ where $n^q \lambda_n^2 \to 0$ and where $q \in [0, 1/2]$.

Lemma 3. *Let $\beta > 0$ be fixed and $n^q \lambda_n^2 \to 0$, where $q \in [0, 1/2]$. Then*

$$\lim_{L \to \infty} \limsup_{n \to \infty} P(n^q |\hat{\beta}_G - \beta| > L) = 0. \tag{3}$$

Proof. Let $L > 0$ be large and fixed and n be so large that $\epsilon = \frac{L}{n^q} < \beta$ and $0 < \lambda_n^2 < \beta\epsilon/4$. So

$$\lambda_n < \sqrt{\beta\epsilon}/2 < \beta/2.$$

Let $0 < |t - \beta| < \epsilon/2$. Then we have $\lambda_n < \beta/2 < \beta - \epsilon/2 < t$. Now

$$|(1 - (\lambda_n/t)^2)^+ \cdot t - \beta| = |t - \lambda_n^2/t - \beta| < |t - \beta| + \lambda_n^2/t < \epsilon,$$

because

$$\lambda_n^2 < \beta\epsilon/4 < (\beta - \epsilon/2)\epsilon/2 < t\epsilon/2.$$

Also $U_n = n^q(T_n - \beta) \sim N(0, n^{2q-1})$. Then

$$\begin{aligned} P(|T_n - \beta| > \epsilon/2) &= P(|T_n - \beta| > \frac{L}{2n^q}) \\ &= P(|U_n| > \frac{L}{2}) \\ &= \int_{|y|>\frac{L}{2}} \frac{1}{\sqrt{2\pi n^{2q-1}}} \exp(-\frac{1}{2n^{2q-1}} y^2) dy \\ &\quad \to 0 \quad \text{as } n \to \infty \ \& \ L \to \infty. \end{aligned}$$

Now

$$P(n^q|\hat{\beta}_G - \beta| > L) \le P(n^q|T_n - \beta| > L/2) \to 0 \quad \text{as } n \to \infty \ \& \ L \to \infty.$$

□

The condition $\lambda = \lambda_n \to 0$ is sufficient to have the consistency of $\hat{\beta}_G$. Now suppose $\lambda_n \not\to 0$. Note that for n large, $|T_n - \beta| \approx \frac{1}{\sqrt{n}}$. Then when $T_n > \lambda_n$ and $\beta \neq 0$,

$$\hat{\beta}_G = T_n - \frac{\lambda^2}{T_n} \approx \beta - \frac{\lambda^2}{\beta}$$

and $\hat{\beta}_G - \beta \approx -\frac{\lambda^2}{\beta} \neq 0$. Hence intuitively, the consistency of $\hat{\beta}_G$ needs $\lambda_n \to 0$ for $\beta \neq 0$. The next lemma ensures this.

Lemma 4. *Let $\lambda_n \not\to 0$. Then $\hat{\beta}_G$ is not consistent for all β.*

Proof. There are $\eta > 0$ and a subsequence λ_{n_k} such that $\lambda_{n_k} > \eta$ for all k. Let $\beta \in (0, \eta)$. Then

$$\begin{aligned} \frac{1}{2} &= P(T_n < \beta) \\ &= P(T_{n_k} < \beta) \\ &\le P(T_{n_k} < \eta) \\ &\le P(T_{n_k} < \lambda_{n_k}). \end{aligned}$$

We also have

$$P(T_{n_k} < -\lambda_{n_k}) \leq P(T_{n_k} < -\eta) = \int_{-\infty}^{-\sqrt{n}\eta} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}y^2)dy \to 0.$$

So $P(|T_n| < \lambda_n) \not\to 0$. For $\epsilon < \beta$,

$$P(|\hat{\beta}_G - \beta| > \epsilon) \geq P(|T_n| < \lambda_n).$$

Therefore $\hat{\beta}_G$ is not consistent for $\beta < \eta$. □

Remark 1. Note that $\hat{\beta}_G$ is consistent for $\beta = 0$ even if $\lambda_n \not\to 0$ (see the proof of Lemma 1 for $\beta = 0$).

Although we need $\lambda_n \to 0$, this convergency does not need to be very fast. A famous example for λ_n is $c\sqrt{\log n/n}$, $c > 0$ (see van de Geer (2000) and Donoho and Johnstone (1994)). In the rest of the paper we apply this case to λ_n.

Corollary 1. If $\lambda = c\sqrt{\log n/n}$ and $c > 0$, then $n^q\lambda_n^2 \to 0$, for $q \in [0, 1/2]$. Then by Lemma 3

$$\lim_{L\to\infty} \limsup_{n\to\infty} P(n^q|\hat{\beta}_G - \beta| > L) = 0.$$

In particular it is true when $q = 1/2$.

Lemma 5 is the first step to prove the mean square error of $\hat{\beta}_G$ tends to zero as $n \to \infty$.

Lemma 5. *Let $c > \sqrt{8}$ and $\lambda = c\sqrt{\log n/n}$. Then the functions*

$$f_1(n) = n \int_{\sqrt{n}\lambda/2}^{\infty} x^2 \exp(-x^2/2)dx$$

and

$$f_2(n) = \frac{n\sqrt{n}}{\log n} \int_{-\lambda}^{\lambda} \exp(-\frac{n}{2}(x-\beta)^2)dx$$

are bounded.

Proof. Note that

$$\begin{aligned}
\lim_{n\to\infty} f_1(n) &= \lim_{n\to\infty} \frac{\int_{c\sqrt{\log n}/2}^{\infty} x^2 \exp(-x^2/2)dx}{1/n} \\
&= \lim_{n\to\infty} \frac{-(c^2/4)\log n \exp(-c^2 \log n/8)\frac{1/n}{c\sqrt{\log n}}}{-1/n^2} \\
&= \lim_{n\to\infty} \frac{c\sqrt{\log n}}{4n^{c^2/8-1}} = 0.
\end{aligned}$$

Since f_1 is continuous and it tends to zero, then it is bounded. Also,

$$f_2(n) = \frac{n}{\log n} \int_{-c\sqrt{\log n}-\beta\sqrt{n}}^{c\sqrt{\log n}-\beta\sqrt{n}} \exp(-\frac{1}{2}x^2)dx$$

and

$$\begin{aligned} f_2'(n) &= \frac{\log n - 1}{\log^2 n} \int_{-c\sqrt{\log n}-\beta\sqrt{n}}^{c\sqrt{\log n}-\beta\sqrt{n}} \exp(-\frac{1}{2}x^2)dx \\ &\quad + \frac{n}{\log n}[(\frac{c}{2n\sqrt{\log n}} - \frac{\beta}{2\sqrt{n}}) \exp(-\frac{1}{2}(c\sqrt{\log n} - \beta\sqrt{n})^2) \\ &\quad + (\frac{c}{2n\sqrt{\log n}} + \frac{\beta}{2\sqrt{n}}) \exp(-\frac{1}{2}(c\sqrt{\log n} + \beta\sqrt{n})^2)] \\ &= \frac{\sqrt{n}}{\log n}[\frac{\log n - 1}{\sqrt{n}\log n} \int_{-c\sqrt{\log n}-\beta\sqrt{n}}^{c\sqrt{\log n}-\beta\sqrt{n}} \exp(-\frac{1}{2}x^2)dx \\ &\quad + (\frac{c}{2\sqrt{n\log n}} - \frac{\beta}{2}) \exp(-\frac{1}{2}(c\sqrt{\log n} - \beta\sqrt{n})^2) \\ &\quad + (\frac{c}{2\sqrt{n\log n}} + \frac{\beta}{2}) \exp(-\frac{1}{2}(c\sqrt{\log n} + \beta\sqrt{n})^2)] \\ &= \frac{\sqrt{n}}{\log n}[I_1 + I_2 + I_3], \end{aligned}$$

where

$$I_1 = \frac{\log n - 1}{\sqrt{n}\log n} \int_{-c\sqrt{\log n}-\beta\sqrt{n}}^{c\sqrt{\log n}-\beta\sqrt{n}} \exp(-\frac{1}{2}x^2)dx,$$

$$I_2 = (\frac{c}{2\sqrt{n\log n}} - \frac{\beta}{2}) \exp(-\frac{1}{2}(c\sqrt{\log n} - \beta\sqrt{n})^2)$$

and

$$I_3 = (\frac{c}{2\sqrt{n\log n}} + \frac{\beta}{2}) \exp(-\frac{1}{2}(c\sqrt{\log n} + \beta\sqrt{n})^2).$$

$I_1 \to 0$ when $n \to \infty$. Also $\frac{c}{2\sqrt{n\log n}} \to 0$ when $n \to \infty$. If $\beta > 0$, then

$$-\frac{1}{2}(c\sqrt{\log n} + \beta\sqrt{n})^2 < -\frac{1}{2}(c\sqrt{\log n} - \beta\sqrt{n})^2$$

and therefore, $I_2 + I_3 < 0$, when $n \to \infty$. If $\beta < 0$, then

$$-\frac{1}{2}(c\sqrt{\log n} + \beta\sqrt{n})^2 > -\frac{1}{2}(c\sqrt{\log n} - \beta\sqrt{n})^2$$

and again, $I_2 + I_3 < 0$, when $n \to \infty$. Consequently, $\lim_{n\to\infty} f_2'(n) < 0$. Since f_2 is continuous in n, $f_2(2) < \infty$ and it is decreasing when $n \to \infty$, then it is bounded. □

In Lemma 6 we show that the MSE is bounded by a multiple of λ^2 or $\frac{1}{n^2}$. Therefore it tends to zero.

Lemma 6. *If $\lambda = c\sqrt{\log n/n}$, $c > \sqrt{8}$, then*

$$E((\hat{\beta}_G - \beta)^2) \leq \begin{cases} c_1\lambda^2 & |\beta| > \lambda/2 \\ c_2\beta^2 + c_3\frac{1}{n^2} & |\beta| \leq \lambda/2, \end{cases} \tag{4}$$

for some finite and positive c_1, c_2 and c_3.

Proof. Let $T = T_n$. Then

$$E((\hat{\beta}_G - \beta)^2) = \beta^2 P(|T| < \lambda) + E((T - \frac{\lambda^2}{T} - \beta)^2 1(|T| \geq \lambda)). \tag{5}$$

Let $|\beta| \leq \lambda/2$. If $T \geq \lambda$, then

$$(T - \frac{\lambda^2}{T} - \beta)^2 \leq (T-\beta)^2 1(T - \frac{\lambda^2}{T} - \beta > 0) + \beta^2 1(T - \frac{\lambda^2}{T} - \beta \leq 0).$$

Similarly if $T \leq -\lambda$, then

$$(T - \frac{\lambda^2}{T} - \beta)^2 \leq (T-\beta)^2 1(T - \frac{\lambda^2}{T} - \beta \leq 0) + \beta^2 1(T - \frac{\lambda^2}{T} - \beta > 0)$$

and consequently

$$\begin{aligned} E(\hat{\beta}_G - \beta)^2 &\leq c_2\beta^2 + E((T-\beta)^2 1(|T| \geq \lambda)) \\ &\leq c_2\beta^2 + \frac{1}{n^2}(nE(X^2 1(|X| \geq \sqrt{n}\lambda/2))) \end{aligned}$$

where $X = \sqrt{n}(T - \beta) \sim N(0,1)$. Now it is sufficient to show

$$nE(X^2 1(|X| \geq \sqrt{n}\lambda/2)) < c_3.$$

for some $c_3 \in (0, \infty)$. It follows from Lemma 5. Lemma 5 also implies that

$$\frac{1}{\lambda^2} P(|T| < \lambda) < c_4, \tag{6}$$

for some $0 < c_4 < \infty$. Note also that if $|\beta| > \lambda/2$, then

$$\begin{aligned} & E((T - \frac{\lambda^2}{T} - \beta)^2 1(|T| \geq \lambda)) \\ \leq\ & 2E((T-\beta)^2 1(|T| \geq \lambda)) + 2E(\frac{\lambda^4}{T^2} 1(|T| \geq \lambda)) \\ \leq\ & 2E((T-\beta)^2) + 2\lambda^2 \\ =\ & 2/n + 2\lambda^2 \leq c_5\lambda^2, \quad \forall n \geq 2 \end{aligned}$$

for some $0 < c_5 < \infty$. Using this, (4) and (5), (6) holds. □

Note that OLS $\hat{\beta}$ is a linear estimator while $\hat{\beta}_G$ is non-linear. There are certain limits on the extent to which nonlinear estimators can improve on linear ones in the worst case. For example, Donoho and Johnstone (1994) considered the model $y_i = \beta_i + \epsilon_i$ with $\sum_{i=1}^n |\beta_i|^\rho \leq 1$, where $\rho > 0$.

In fact they assume the parameter space $\Theta_\rho = \{\beta : \sum_{i=1}^n |\beta_i|^\rho \leq 1\}$ rather than $\Theta = \mathbf{R}^n$.

They proved that in some function estimation problems of a linear nature, the optimal rate of convergence over certain convex function classes is not attained by any linear estimate. Loubes and van de Geer (2000) considered the limitation Θ_ρ in adaptive estimation using soft thresholding type penalties. We shall also assume the model (1) with $\sum_{i=1}^n |\beta_i|^\rho \leq 1$ in the following theorem.

Theorem 1. *Consider the regression model*

$$y_i = \sum_{j=1}^m \beta_j x_{ij} + \epsilon_i, \quad \epsilon_i \sim N(0,1), \quad i = 1, ..., n, \quad X^T X = nI, \quad \sum_{i=1}^m |\beta_i|^\rho \leq 1,$$

where $\epsilon_1, ..., \epsilon_n$ are independent and $\rho \in (1,2)$. If $\lambda = c\sqrt{\log n/n}$, $c > \sqrt{8}$, then there exists $k < \infty$, such that

$$E(\sum_{i=1}^m |\hat{\beta}_{i,G} - \beta_i|^2) \leq k\lambda^{2-\rho},$$

where $\hat{\beta}_{i,G}$ is the nn-garrote estimator of β_i.

Proof. By Lemma 6,

$$E(\sum_{i=1}^m |\hat{\beta}_{i,G} - \beta_i|^2) \leq c_1 \sum_{i:|\beta_i|>\lambda/2} \lambda^2 + c_2 \sum_{i:|\beta_i|\leq\lambda/2} |\beta_i|^2 + c_3 \sum_{i:|\beta_i|\leq\lambda/2} \frac{1}{n^2}.$$

for some finite and positive c_1, c_2 and c_3. But

$$\lambda^2 \sum_{i:|\beta_i|>\lambda/2} 1 = \lambda^{2-\rho} \sum_{i:|\beta_i|>\lambda/2} \lambda^\rho \leq \lambda^{2-\rho} 2^\rho \sum_{i:|\beta_i|>\lambda/2} |\beta_i|^\rho \leq \lambda^{2-\rho} 2^\rho$$

and also

$$\sum_{i:|\beta_i|\leq\frac{\lambda}{2}} |\beta_i|^2 = \frac{\lambda^2}{4} \sum_{i:|\beta_i|\leq\frac{\lambda}{2}} |\frac{\beta_i}{\frac{\lambda}{2}}|^2 \leq \frac{\lambda^2}{4} \sum_{i:|\beta_i|\leq\frac{\lambda}{2}} |\frac{\beta_i}{\frac{\lambda}{2}}|^\rho = \frac{\lambda^{2-\rho}}{2^{2-\rho}} \sum_{i:|\beta_i|\leq\frac{\lambda}{2}} |\beta_i|^\rho \leq (\lambda/2)^{2-\rho}.$$

If $l \in (2-\rho, 1)$, then there exists N such that for any $n \geq N$,

$$\sum_{i:|\beta_i|\leq\frac{\lambda}{2}} \frac{1}{n^2} \leq \frac{1}{n} \leq \lambda^l \leq \lambda^{2-\rho}.$$

Therefore,

$$E(\sum_{i=1}^{m} |\hat{\beta}_{i,G} - \beta_i|^2) \leq k\lambda^{2-\rho},$$

for some finite k. □

Corollary 2. By Theorem 1, we have

$$MSE(\hat{\beta}_{i,G}) = E(|\hat{\beta}_{i,G} - \beta_i|^2) \leq E(\sum_{i=1}^{m} |\hat{\beta}_{i,G} - \beta_i|^2) \leq k\lambda^{2-\rho}$$

This not only shows $MSE(\hat{\beta}_{i,G}) \to 0$ when $n \to \infty$, but also shows the rate of convergence. If $\rho \to 1$, then this convergency is fast. Now by Markov's inequality for all $\epsilon > 0$

$$P(|\hat{\beta}_{i,G} - \beta_i| > \epsilon) \leq \frac{E(|\hat{\beta}_{i,G} - \beta_i|)}{\epsilon} \leq \frac{E^{1/2}(|\hat{\beta}_{i,G} - \beta_i|^2)}{\epsilon} \to 0$$

when $n \to \infty$. Therefore $\hat{\beta}_{i,G}$ is consistent for β_i.

References

Breiman, L. (1995), Better subset regression using the nonnegative garrote. *Technometrics,* November, **37**, N. 4, 373-384.

Donoho, D. L. and Johnstone, I. M. (1994), Minimax risk for l_q losses over l_p-balls. *Probab. Theory and Related Fields,* **99**, 277–303.

Loubes J. M. and van de Geer S. A. (2000), Adaptive estimation in regression using soft thresholding type penalties. *Statistica Neerlandica.* Unpublished

Van de Geer S. A. (2000), Empirical processes in M-estimation. Cambridge University Press.