Supplementary Appendix: Weak Consistency of Lasso Penalized ℓ_1 Regression

October 8, 2007

Our proof is a straightforward adaptation of the arguments of Oberhofer (1983). He supposes that:

S1. The parameter vector θ is confined to a compact domain $K \subset \mathsf{R}^{p+1}$. The true parameter vector $\tilde{\theta}$ is an interior point of K.

S2. The random errors $e_i = y_i - \mu - x_i^t\beta$ are independent; e_i has distribution function $F_i(e)$ with $F_i(0) = \frac{1}{2}$.

S3. For every $c > 0$ there exists an $f > 0$ with

$$\inf_i \min\Big\{F_i(c) - \frac{1}{2}, \frac{1}{2} - F_i(-c)\Big\} \quad \geq \quad f.$$

S4. The predictor vectors $z_i^t = (1, x_i^t)$ satisfy $\|z_i\|_2 \leq B$ for some $B \geq 0$.

S5. For some $e > 0$ and $d > 0$, the predictors z_i satisfy

$$\inf_{\|v\|=1} \frac{1}{n}\sum_{i=1}^n 1_{\{|z_i^t v| \geq e\}} \quad \geq \quad d$$

for n sufficiently large.

Theorem 1 *Under the regularity conditions S1 through S5, the sequence of estimators minimizing the criterion, $f(\theta) = g(\theta) + \lambda\sum_{j=1}^p |\beta_j|$, is weakly consistent.*

Proof: Consider the random difference $d_n(\theta) = f_n(\theta) - f_n(\tilde{\theta})$, where

$$f_n(\theta) \quad = \quad \frac{1}{n}\sum_{i=1}^n \Big|y_i - z_i^t\theta\Big| + \frac{\lambda}{n}\sum_{j=1}^p |\beta_j|,$$

$z_i^t = (1, x_i^t)$, and $\tilde{\theta}$ is the true parameter vector. On one hand at an optimal point $\hat{\theta}_n$, we have $d_n(\hat{\theta}_n) \leq 0$. On the other hand according to Chebychev's inequality, the random variable $d_n(\theta)$ satisfies

$$\Pr\{d_n(\theta) \geq \mathrm{E}[d_n(\theta)] - \delta\} \quad \geq \quad 1 - \frac{\mathrm{Var}[d_n(\theta)]}{\delta^2} \tag{1}$$

for every positive δ. Our analysis hinges on three facts. Fact a) says that for all θ and θ^*

$$|d_n(\theta) - d_n(\theta^*)| \quad \leq \quad B\|\theta - \theta^*\|_2 + \frac{\lambda}{n}\sum_{j=1}^{p}|\beta_j - \beta_j^*|.$$

Facts b) and c) involve a compact subset $C \subset K$ excluding $\tilde{\theta}$. Fact b) says that $\inf_{\theta \in C} \mathrm{E}[d_n(\theta)]$ is bounded below by a positive constant η for all sufficiently large n, and fact c) says that $\lim_{n\to\infty} \sup_{\theta \in C} \mathrm{Var}[d_n(\theta)] = 0$. Before we prove these facts, let us demonstrate weak consistency.

If we take $\delta = \frac{1}{2}\eta$ in inequality (1), then fact b) entails

$$\Pr\left[d_n(\theta) \geq \frac{1}{2}\eta\right] \quad \geq \quad 1 - \frac{\mathrm{Var}[d_n(\theta)]}{\delta^2}.$$

The uniform continuity assertion a) implies that $d_n(\theta^*) - d_n(\theta) \geq -\frac{1}{4}\eta$ for all θ^* in some neighborhood N of θ. Hence,

$$\Pr\left[\inf_{\theta^* \in N} d_n(\theta^*) \geq \frac{1}{4}\eta\right] \quad \geq \quad 1 - \frac{\mathrm{Var}[d_n(\theta)]}{\delta^2}.$$

By compactness, a finite number of such neighborhoods $N_1, \ldots, N_m$ cover C. It follows that

$$\Pr\left[\inf_{\theta^* \in C} d_n(\theta^*) < \frac{1}{4}\eta\right] \;\leq\; \sum_{i=1}^{m}\Pr\left[\inf_{\theta^* \in N_i} d_n(\theta^*) < \frac{1}{4}\eta\right] \;\leq\; \sum_{i=1}^{m}\frac{\mathrm{Var}[d_n(\theta)]}{\delta^2}.$$

According to assertion c), the scaled sum of variances in the second of these inequalities can be made smaller than any $\epsilon > 0$ by taking n sufficiently large. Hence, $\hat{\theta}_n \in C$ with probability at most ϵ for large n. Taking C to be the complement in K of a small open ball around $\tilde{\theta}$ then implies that $\hat{\theta}_n$ converges in probability to $\tilde{\theta}$.

Let us tackle facts a) through c) in reverse order. Because

$$d_n(\theta) \quad = \quad \frac{1}{n}\sum_{i=1}^{n}\Big[|e_i + z_i^t(\theta - \tilde{\theta})| - |e_i|\Big] + \frac{\lambda}{n}\sum_{j=1}^{p}\Big[|\beta_j| - |\tilde{\beta}_j|\Big],$$

we have $\mathrm{Var}[d_n(\theta)] = n^{-2}\sum_{i=1}^n \mathrm{Var}[|e_i + z_i^t(\theta-\tilde{\theta})| - |e_i|]$. In view of the inequality

$$\Big||e_i + z_i^t(\theta-\tilde{\theta})| - |e_i|\Big| \;\le\; |z_i^t(\theta-\tilde{\theta})| \;\le\; \|z_i\|_2 \cdot \|\theta-\tilde{\theta}\|_2$$

and assumptions S1, S2, and S4, we conclude that

$$\mathrm{Var}[d_n(\theta)] \;\le\; \frac{1}{n^2}\sum_{i=1}^n \|z_i\|_2^2 \cdot \|\theta-\tilde{\theta}\|_2^2 \;\le\; \frac{B^2}{n} 4 \sup_{\phi \in K} \|\phi\|_2^2.$$

This clearly proves assertion c).

To deal with assertion b), Oberhofer considers the objective function $g_n(\theta) = \frac{1}{n}\sum_{i=1}^n |y_i - z_i^t\theta|$ without the penalty and proves the lower bound

$$\mathrm{E}\Big[g_n(\theta) - g_n(\tilde{\theta})\Big] \;\ge\; \frac{2}{n}\sum_{i=1}^n |h_i(\theta)| \min\Big\{F_i[|h_i(\theta)|] - \frac{1}{2}, \frac{1}{2} - F_i[-|h_i(\theta)|]\Big\},$$

where $h_i(\theta) = \frac{1}{2} z_i^t(\theta-\tilde{\theta})$. If we set $c = \frac{e}{2}\min_{\theta\in C}\|\theta-\tilde{\theta}\|$, then assumptions S3 and S5 imply that

$$\begin{aligned}
\inf_{\theta\in C} \mathrm{E}\Big[g_n(\theta) - g_n(\tilde{\theta})\Big] \;&\ge\; \inf_{\theta\in C}\frac{2}{n}\sum_{i=1}^n 1_{\{h_i(\theta)\ge\frac{e}{2}\|\theta-\tilde{\theta}\|_2\}} \frac{e}{2}\|\theta-\tilde{\theta}\|_2 \\
&\quad\; \cdot \min\Big\{F_i[|h_i(\theta)|] - \frac{1}{2}, \frac{1}{2} - F_i[-|h_i(\theta)|]\Big\} \\
&\ge\; \inf_{\theta\in C}\frac{1}{n}\sum_{i=1}^n 1_{\{h_i(\theta)\ge\frac{e}{2}\|\theta-\tilde{\theta}\|_2\}} e\|\theta-\tilde{\theta}\|_2 \\
&\quad\; \cdot \min\Big\{F_i(c) - \frac{1}{2}, \frac{1}{2} - F_i(c)\Big\} \\
&\ge\; \inf_{\theta\in C} def \cdot \|\theta-\tilde{\theta}\|_2 \\
&>\; 0
\end{aligned}$$

for n large and appropriate constants d and f. Because $\frac{\lambda}{n}\sum_{j=1}^p[|\beta_j| - |\tilde{\beta}_j|]$ tends to 0 uniformly on the compact set C, assertion b) now follows.

To prove assertion a), note that

$$\begin{aligned}
|d_n(\theta) - d_n(\theta^*)| \;&\le\; \frac{1}{n}\sum_{i=1}^n \Big||y - z_i^t\theta| - |y_i - z_i^t\theta^*|\Big| + \frac{\lambda}{n}\sum_{j=1}^p \Big||\beta_j| - |\beta_j^*|\Big| \\
&\le\; \frac{1}{n}\sum_{i=1}^n |z_i^t(\theta-\theta^*)| + \frac{\lambda}{n}\sum_{j=1}^p |\beta_j - \beta_j^*|.
\end{aligned}$$

Application of assumption S4 and the Cauchy-Schwarz inequality finishes the proof of assertion a) and the theorem. ■

Assumption S5 is unusual and deserves comment. Suppose the x_i constitute a random sample from a bounded random vector X with positive definite covariance matrix Σ. Consider the random vector $Z^t = (1, X^t)$ and the continuous function $U(z, v) = \min\{|z^t v|, 1\}$. The mean $\mu(v)$ of $U(Z, v)$ is a continuous function of v owing to the dominated convergence theorem. On the unit sphere $\|v\|_2 = 1$, $\mu(v)$ attains its minimum. If the minimum is 0, then $Z^t v = u + \sum_{j=1}^p X_j w_j$ is identically 0 for the pertinent vector $v^t = (u, w^t)$. This implies that $X^t w$ is constant. Because $\mathrm{Var}(X^t w) = w^t \Sigma w$, we must have $w = \mathbf{0}$. However, this can only occur if $u = 0$ as well, contradicting the condition $\|v\|_2 = 1$. We conclude that there exists a $\delta > 0$ such that $\mu(v) \geq \delta$ for all v on the unit sphere.

According to the uniform strong law of large numbers (Ferguson 1996),

$$\lim_{n\to\infty} \inf_{\|v\|=1} \frac{1}{n} \sum_{i=1}^n U(z_i, v) \quad \geq \quad \frac{1}{2} \min_{\|v\|=1} \mu(v) \quad = \quad \frac{\delta}{2} \tag{2}$$

with probability 1. We now claim that $inf_{\|v\|=1} \frac{1}{n} \sum_{i=1}^n 1_{\{|z_i^t v| \geq \frac{\delta}{6}\}} > \frac{\delta}{6}$ for large n. If this condition fails, then for infinitely many n there exists a unit vector v with

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n U(z_i, v) &= \frac{1}{n} \sum_{i=1}^n U(z_i, v) 1_{\{|z_i^t v| \geq \frac{\delta}{6}\}} + \frac{1}{n} \sum_{i=1}^n U(z_i, v) 1_{\{|z_i^t v| < \frac{\delta}{6}\}} \\
&\leq \frac{1}{n} \sum_{i=1}^n 1_{\{|z_i^t v| \geq \frac{\delta}{6}\}} + \frac{1}{n} \sum_{i=1}^n \frac{\delta}{6} \\
&\leq \frac{\delta}{6} + \frac{\delta}{6},
\end{aligned}$$

an evident contradiction to inequality (2). Thus, assumption S5 follows.

References

Ferguson, T. S. (1996), *A Course in Large Sample Theory*, London: Chapman & Hall.

Oberhofer, W. (1983), "The consistency of nonlinear regression minimizing the ℓ_1-norm," *Ann Stat*, 10, 316–319.