Lasso and Related Methods

Prof. Myung Hwan Seo
Seoul National University

Spring 2023

1 Lasso

- Notation
  - For a given vector β, the ℓ_q-norm of β is denoted by $|\beta|_q := \left(\sum_{j=1}^n |\beta_j|^q\right)^{1/q}$.
  - Let $S(\beta) = \{j : \beta_j \neq 0, j = 1, ..., p\}$ denote the support of the vector β. We will also abbreviate $S(\beta^*)$ and $S\left(\hat{\beta}\right)$ by S_* and $\hat{S}$, respectively. In addition, $\hat{S}(\lambda) = S\left(\hat{\beta}(\lambda)\right)$, where $\hat{\beta}$ is a function of λ.
  - For a set S,the cardinality of the set S is denoted by $|S|$.

- The least absolute shrinkage and selection operator (lasso) proposed by Tibshirani (1996) solves the following ℓ_1-penalized least squares:

$$\min_{\beta} \mathcal{L}_n(\beta; \lambda_n) := \frac{1}{n}\sum_{i=1}^n (y_i - x_i'\beta)^2 + \lambda_n |\beta|_1, \tag{1}$$

  where β is a p-dimensional vector and each column of X is demeaned and normalized to have sample variance 1.

- (1) is a dual of the constrained minimization

$$\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - x_i'\beta\right)^2$$
$$s.t.\ \ |\beta|_1 \leq C_n,$$

  where there exists C_n corresponding to λ_n (1 to 1) to yield the same solution $\hat{\beta}$. (convex criterion function with convex constraint).

- There are many variants of the lasso method, such as the ridge or Bridge estimation or SCAD (smoothly clipped absolute deviation), where the penality function is respectively ℓ_2-, ℓ_q-norm with $0 < q < 1$, or a more general function $\sum_{j=1}^{p} g_{a,\lambda}\left(\beta_j\right)$, where

  $g_{a,\lambda}(u) = \lambda |u| 1\{|u| \leq \lambda\} - \left(u^2 - 2a\lambda |u| + \lambda^2\right) 1\{\lambda < |u| \leq a\lambda\} + (a+1)\left(\lambda^2/2\right) 1\{|u| > a\lambda\}$

  for some $a > 2$.

- When the penalty function is the ℓ_0-norm, i.e. $|\beta|_0^0 := \sum_{j=1}^{p} 1\{\beta_j \neq 0\}$, the problem corresponds to a model selection based on some information crite-

rion such as AIC and BIC. In this case the computation burden is $O\left(2^p\right)$. The ℓ_q-penalty with $0 \leq q < 1$ yields a non-convex criterion function.

- Efficient numerical algorithms to find the solution of (1) have been suggested, which exploits the convexity of the criterion function, such as the LARS algorithm (Efron et al 2004) and GLMnet (Friedman et al. 2007) and the codes are publically available in most statistical programs (e.g. Matlab, R, Stata, etc). This is an advantage of the lasso among many different penalized optimization estimators.

- In particular, the LARS algorithm computes all the solution paths along λ_n utilizing the fact that the solution paths are piecewise linear. GLMnet is faster as it computes the solution coordinate-wise (one-at-a-time) but for each fixed λ_n. This works because the lasso is separably convex and it works for many other similar problems. And it is particularly efficient due to the sparsity of the solution and the simplicity of the coordinate-wise solutions.

- One needs to pay attention to the (default) normalization of variables that takes place in the packages and differs across different packages. For instance, GLMnet in R standardizes all the variables y and x to have mean zero and standard deviation 1.

- Tuning parameter λ_n selection: there are theoretical results on the proper rate for $\lambda_n \to 0$ but in practice,

  1. some form of cross validation
  2. BIC criterion: no theoretical result yet and

$$BIC_\lambda = n \log\left(\frac{1}{n}\left|Y - \hat{Y}_\lambda\right|_2^2\right) + \log(n)\, df\left(\hat{Y}_\lambda\right),$$

  where $df\left(\hat{Y}_\lambda\right)$ is set as $\left|\hat{S}(\lambda)\right|$, typically.

1.1 Numerical property of the lasso solution

We can characterize the property of solutions $\hat{\beta}(\lambda_n)$ to (1) by the Karush-Kuhn-Tucker (KKT) conditions. Uniqueness requires more conditions.

- A subgradient vector z of $|\beta|_1$ is a vector s.t.

$$z_j = sign(\beta_j) \text{ if } \beta_j \neq 0, \text{ and } z_j \in [-1, 1], \text{ otherwise.}$$

- For a matrix X and an index set S, X_S is the $n \times |S|$ matrix that concatenates the columns $\{X_j : j \in S\}$.

Lemma 1 *Let $G(\beta) = -2X'(Y - X\beta)/n$. A vector $\hat{\beta}$ is a solution to (1) iff*

$$\begin{aligned} G_j\left(\hat{\beta}\right) &= -sign\left(\hat{\beta}_j\right)\lambda_n, \text{ if } \hat{\beta}_j \neq 0 \\ \left|G_j\left(\hat{\beta}\right)\right| &\leq \lambda_n, \text{ otherwise.} \end{aligned}$$

Moreover, if the solution is not unique and $\left|G_j\left(\hat{\beta}\right)\right| < \lambda_n$ for a solution $\hat{\beta}$, then $\hat{\beta}_j = 0$ for all solutions $\hat{\beta}$.

Proof See Lemma 2.1 in BG11.

Remarks

1. In other words, the FOC for the solution is

$$\frac{2}{n}X'X\hat{\beta} - \frac{2}{n}X'Y + \lambda_n \hat{z} = 0,$$

   where $\hat{z}$ is a subgradient vector of $\hat{\beta}$.

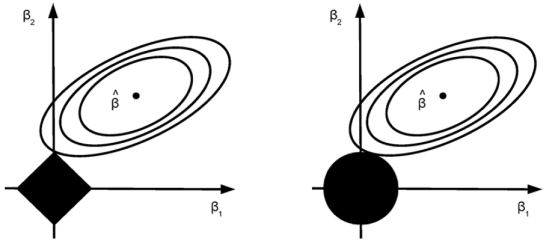
2. The set of all solutions are convex (convex opitimization under a convex constraint)

3. The zeros are unique among the solutions as long as a solution yields $\left|G_j\left(\hat{\beta}\right)\right| < \lambda_n$. To see this, note that $G(\beta)$ is continuous in β and for another solution $\tilde{\beta}$ s.t. $\tilde{\beta}_j \neq 0$, and for any $\rho \neq 1$, $\bar{\beta}_j = \rho\hat{\beta}_j + (1-\rho)\tilde{\beta}_j \neq 0$, which is also a solution. Thus, $\left|G_j\left(\rho\hat{\beta} + (1-\rho)\tilde{\beta}\right)\right| = \lambda_n$ while $\left|G_j\left(\hat{\beta}\right)\right| < \lambda_n$. This is a contradiction to the continuity of G.

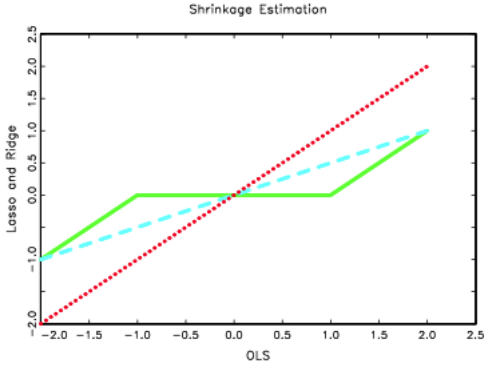
4. If, in addition, $X'_{S(\hat{\beta})}X_{S(\hat{\beta})}$ is invertible, then $\hat{\beta}$ is unique. For more complete discussion on the uniqueness of the solution, see Wainwright (2009).

5. The lasso can yield a unique solution even when $p > n$, while the OLS does not. Indeed, if $p \geq n$, $Y = X\hat{\beta}_{ols}$ for any solution.

6. Due to ℓ_1 penalty, the lasso performs variable selection in the sense that $\hat{\beta}_j$ takes value exact zero for some j. See the figure below, which compares the ℓ_1-penalty with ℓ_2-penalty.


β2
β̂
β1
β2
β̂
β1

7. The penalty term shrinks the OLS estimate, if available, thus resulting in **bias** toward zero. Illustration is given for the case with $p = 1,\ \lambda_n = 1,\ X'X/n = 1$ and $\hat{\beta}_{ols}$ is available and $\hat{\beta}_{ols} = \frac{1}{n}X'y$.



8. It is known that $\hat{\beta}(\lambda)$ is piecewise linear[1] along λ and that $\left|\hat{S}\right| \leq n$ in case

[1]This only holds for the linear regression.

$n \leq p$. (Efron et al's 2004 LARS). Heuristically, with $\lambda_n = 0$ $\left|\hat{S}\right| = n$ is sufficient to make SSR zero and the bigger λ_n, the sparser $\hat{\beta}$.

1.2 High-dimensional Asymptotics

The lasso is particularly useful when p is large. In fact, p can be even much larger than n as long as the sparsity of β^*–the number of non-zero elements in β^*–is maintained low. Here the setup is the fixed design, i.e., X is a fixed matrix with the scale normalization, with ε being a vector of *iid* $\mathcal{N}\left(0, \sigma^2\right)$. Main results in this area concern the prediction and parameter estimates' consistency and error bounds and the model selection consistency, which is also presented in a slightly stronger form as the sign-consistency.

1.2.1 Prediction Precision

We begin with the prediction accuracy of the lasso. (Indeed, many interesting economic issues involve good prediction).

Theorem 2 *A weak version of the results assumes that*

$$\lambda_n = 4\sigma\sqrt{\frac{t^2 + 2\log p}{n}},$$

to get, for $s = |S_*|$,

$$2\left|n^{-1}X\left(\hat{\beta}-\beta^*\right)\right|_2^2 \leq 3s\lambda_n,$$

with probability at least $1-2\exp\left(-t^2/2\right)$.

Proof. This requires not much more than the sparsity of β^*. Since

$$\left|Y-X\hat{\beta}\right|_2^2/n+\lambda_n\left|\hat{\beta}\right|_1 \leq |Y-X\beta|_2^2/n+\lambda_n|\beta|_1,$$

and

$$\begin{aligned}\left|\hat{\beta}-\beta^*\right|_1 &= \left|\hat{\beta}_{S_*}-\beta^*_{S_*}\right|_1+\left|\hat{\beta}_{S_*^c}\right|_1\\ |\beta^*|_1-\left|\hat{\beta}\right|_1 &= \left|\beta^*_{S_*}\right|_1-\left|\hat{\beta}_{S_*}\right|_1-\left|\hat{\beta}_{S_*^c}\right|_1 \leq \left|\hat{\beta}_{S_*}-\beta^*_{S_*}\right|_1-\left|\hat{\beta}_{S_*^c}\right|_1\end{aligned}$$

due to the triangle inequality, we get

$$\begin{aligned}
\left|X\left(\hat{\beta}-\beta^*\right)\right|_2^2/n &\leq 2\varepsilon' X\left(\hat{\beta}-\beta^*\right)/n+\lambda_n\left|\beta^*\right|_1-\lambda_n\left|\hat{\beta}\right|_1 \\
&\leq 2\max_j n^{-1}\left|\varepsilon' X_j\right|\cdot\left|\hat{\beta}-\beta^*\right|_1+\lambda_n\left(\left|\beta^*\right|_1-\left|\hat{\beta}\right|_1\right), \\
&\leq (3/2)\,\lambda_n\left(\left|\hat{\beta}_{S_*}-\beta^*_{S_*}\right|_1\right)-2^{-1}\lambda_n\left|\hat{\beta}_{S_*^c}\right|_1 \qquad (2) \\
&\leq (3/2)\,\lambda_n\left(\left|\hat{\beta}_{S_*}-\beta^*_{S_*}\right|_1\right)
\end{aligned}$$

conditional on

$$A=\left\{2\max_j n^{-1}\left|\varepsilon' X_j\right|\leq\lambda_n/2\right\}.$$

However, $\Pr\{A\}\geq 1-2\exp\left(-t^2/2\right)$ from elementary statistics.

■

In case of random design, the result can be stated as

$$\mathrm{E}\left[\left(x'_{n+1}\left(\hat{\beta}-\beta^*\right)\right)^2\right]=O_p\left(s\sqrt{\frac{\log p}{n}}\right),$$

where the expectation is taken with respect to x_{n+1}.

1.2.2 Oracle Inequality

A stronger version, a.k.a. *oracle inequality*, can be obtained under the compatibility (or restricted eigenvalue) assumption between the design matrix X (or the gram matrix $\hat{M} = n^{-1}X'X$) and the ℓ_1-norm of β. Refer to Bickel et al. (2009), BV (Ch 6.13).

Condition 3 *For an index set $S \subset \{1, ..., p\}$ and a constant $\mu > 1$, there exists a positive constant ϕ such that*

$$\min_{\delta:|\delta_{S^c}|_1 \leq \mu|\delta_S|_1} \frac{\delta'\hat{M}\delta}{|\delta_S|_1^2} |S| \geq \phi^2.$$

We call ϕ^2 a *compatibility constant.*

- Note that the denominator is based on the ℓ_1-norm.
- Restricted eigenvalues,

$$\phi_{\min}(s) = \min_{\beta:S(\beta)\leq s} \frac{\beta' M \beta}{\beta'\beta},$$

can provides a certain lower bound for the compatibility constant ϕ^2.

- Due to the condition $|\delta_{S^c}|_1 \leq \mu |\delta_S|_1$, this is essentially a condition on restricted eigenvalue. We do not need $\hat{M}$ to be p.d. as in the OLS.

- As a consequence, any s-dimensional square submatrix of $\hat{M}$ is positive definite if the compatibility condition holds for all S with $|S| \leq s$.

- For two symmetric design matrixes s.t. $|M_1 - M_2|_\infty \leq \lambda$, and for all δ s.t. $|\delta_{S^c}|_1 \leq \mu |\delta_S|_1$, we can show that

$$\left| \frac{\delta' M_2 \delta}{\delta' M_1 \delta} - 1 \right| \leq (1+\mu)^2 \frac{\lambda s}{\phi_1^2},$$

when the design matrix M_1 satisfies the compatibility condition with ϕ_1^2. This follows simply by noting that

$$|\delta' (M_2 - M_1) \delta| = \left| \sum_{i,j} \delta_i (M_2 - M_1)_{ij} \delta_j \right| \leq \lambda |\delta|_1^2 .$$

As a corollary of this, if $\lambda \leq \phi_1^2 \left(2\left(1+\mu\right)^2 s\right)^{-1}$ in addition, the design matrix M_2 satisfies the compatibility condition with some $\phi_2^2 \geq \phi_1^2/2$.

Then,

Theorem 4 *Suppose the compatibility condition holds for S_* with $\mu = 3$. Then, with probability approaching to 1 as $n \to \infty$,*

$$n^{-1}\left|X\left(\hat{\beta}-\beta^*\right)\right|_2^2 + \lambda_n\left|\hat{\beta}-\beta^*\right|_1 \leq O\left(\frac{\lambda_n^2 s_*}{\phi_*^2}\right),$$

where $s_* = |S_*|$ and ϕ_*^2 is the compatibility constant for S_*.

Proof. 1. Rewrite (2) after multiplying 2 both sides to get

$$2n^{-1}\left|X\left(\hat{\beta}-\beta^*\right)\right|_2^2 + \lambda_n\left|\hat{\beta}_{S_*^c}\right|_1 \leq 3\lambda_n\left|\hat{\beta}_{S_*}-\beta^*_{S_*}\right|_1, \tag{3}$$

and note that $\left|\hat{\beta}_{S_*^c}\right|_1 \leq 3\left|\hat{\beta}_{S_*}-\beta^*_{S_*}\right|_1$, which implies we can apply the compatibility condition with $\delta = \hat{\beta}-\beta^*$.

2. Furthermore, by adding $\lambda_n \left|\hat{\beta}_{S_*} - \beta^*_{S_*}\right|_1$ both sides of (3),

$$\begin{aligned} 2n^{-1}\left|X\left(\hat{\beta}-\beta^*\right)\right|_2^2 + \lambda_n\left|\hat{\beta}-\beta^*\right|_1 &\leq 4\lambda_n\left|\hat{\beta}_{S_*} - \beta^*_{S_*}\right|_1, \\ &\leq 4\sqrt{s_*}\lambda_n\left(n\phi_*^2\right)^{-1/2}\left|X\left(\hat{\beta}-\beta^*\right)\right|_2 \\ &\leq n^{-1}\left|X\left(\hat{\beta}-\beta^*\right)\right|_2^2 + 4\lambda_n^2 s_*/\phi_*^2, \end{aligned}$$

where the second inequality is the compatibility condition and the last inequality is due to the inequality $2ab \leq a^2 + b^2$.

3. Finally, bring $n^{-1}\left|X\left(\hat{\beta}-\beta^*\right)\right|_2^2$ to the left to conclude. ■

The plug-in of λ_n formula yields that

$$\left|n^{-1}X\left(\hat{\beta}-\beta^*\right)\right|_2^2 = O\left(\frac{s_*\log p}{n\phi_*^2}\right), \tag{4}$$

which is called an *oracle inequality*.

- oracle? The rate achieved in (4) for the mean-squared prediction error is the same order–up to the $\log p$ term–that can be reached if the OLS is

employed with only the variables in S_* as the regressors. That is, the oracle knows which variables are relevant.

- The prediction and estimation error bounds get smaller as λ_n becomes smaller. But at the same time, the probability of the conditioning event A decreases. The rate for λ_n provided here is, thus, the minimal value that guarantees the probability goes to 1.

- Theorem 4 also yields that

$$\left|\hat{\beta} - \beta^*\right|_1 = O\left(\frac{s_*}{\phi^2}\sqrt{\frac{\log p}{n}}\right).$$

This has an important implication for the variable selection, i.e.,

$$\Pr\left\{S_* \subset \hat{S}\left(\lambda_n\right)\right\} \to 1,$$

under the so-called *beta-min* condition

$$\min\left\{\left|\beta_j^*\right| : j \in S_*\right\} >> O\left(s_*\sqrt{n^{-1}\log p}\right).$$

However, in general,

$$\Pr\left\{S_* = \hat{S}\left(\lambda_n\right)\right\} < 1.$$

In other words, the lasso select all the relevant variables if it overfits the model.

1.2.3 Model Selection

The next theorem (Wainwright 2009) gives a sufficient condition for the sign-consistency of the lasso estimate $\hat{\beta}$.

Theorem 5 *Write* $S_* = S\left(\beta^*\right)$ *and assume that there exists some* incoherence parameter $\gamma \in (0, 1]$ *s.t.*

$$\left|X'_{S_*^c} X_{S_*} \left(X'_{S_*} X_{S_*}\right)^{-1}\right|_\infty \leq (1 - \gamma), \tag{5}$$

and some $C_{\min} > 0$ *s.t.*

$$\Lambda_{\min}\left(\frac{1}{n} X'_{S_*} X_{S_*}\right) \geq C_{\min},$$

where $\Lambda_{\min}$ *denotes the minimal eigenvalue. Furthermore, assume that each columns of* X *are normalized so that* $\|X_j\| \leq 1$ *for all* $j \in S_*^c$ *and*

$$\lambda_n > \frac{2}{\gamma}\sqrt{\frac{2\sigma^2 \log p}{n}}.$$

Then, with probability greater than $1 - 4\exp\left(-c_1 n\lambda_n^2\right)$ *for some* $c_1 > 0$*, the lasso has a unique solution with* $S\left(\hat{\beta}\right) \subset S\left(\beta^*\right)$ *and*

$$\left\|\hat{\beta}_{S_*} - \beta^*_{S_*}\right\|_\infty \leq g\left(\lambda_n\right),$$

where g *is a linear function. In addition, if*

$$\min\left\{\beta_j^* : j \in S_*\right\} > g\left(\lambda_n\right), \tag{6}$$

then,

$$sign\left(\hat{\beta}\right) = sign\left(\beta^*\right).$$

The condition (5) is called *mutual incoherence* condition and also known as *neighborhood stability* or *irrepresentable* condition. And the other condition (6)

is referred to as the *beta-min* condition. It is known that the mutual incoherence condition is **essentially a necessary and sufficient** condition, which fails to hold if the design matrix X exhibits too much linear dependence within smaller submatrices of X. For instance, if $n^{-1}X_j'X_k = \rho^{|j-k|}$, then $|\rho| < 1$ is required.

- Comparing the conditions in this subsection to the previous one, we can see that the model selection demands stronger conditions than estimation precision.
- relation between restrictions on design matrix X : mutual incoherence $\Rightarrow$ restricted eigenvalue $\Rightarrow$ compatibility.
- The minimal value of λ_n for the model selection consistency is bigger than the one for the best prediction. Thus, we note that the prediction benefits from a slightly bigger model.
- We need larger λ for smaller γ, that is, higher correlation between relevant and irrelevant variables demands larger penalty and larger signal in β.

- See also Meinshausen and Bühlmann (2006) for a slightly different exposition.

1.2.4 Maximal Inequality

To begin with, we review the relation between moments and tail bounds. The two are linked to each other by means of Markov inequality and the formula:

$$E|X| = \int_0^\infty \Pr\{|X| > x\}\, dx,$$

which follows from the integral-by-parts.

- Maximal inequality: Let

$$\mathrm{E}G(|X_j|) \leq K < \infty, \quad \forall j \tag{7}$$

where G on $\mathbb{R}_+$ is convex and increasing. Then,

$$\max_{1\leq j\leq p} |X_j| = O_p(r_p), \tag{8}$$

with $r_p = G^{-1}(p)$.

– By the union bound, for any ε there exists $M < \infty$ such that

$$\begin{aligned}\Pr\left\{r_p^{-1}\max_{j\leq p}|X_j| > M\right\} &\leq \left(p\max_{j\leq p}\Pr\left\{|X_j| > r_pM\right\}\wedge 1\right)\\ &\leq \frac{p\mathrm{E}G\left(|X_j|\right)}{G\left(r_pM\right)} \quad \text{or} \quad pF\left(r_pM\right)\\ &\leq \frac{K}{M} < \varepsilon,\end{aligned}$$

because the increasing convexity means that

$$G\left(r_pM\right) \geq G\left(r_p\right)M = pM,$$

for $M > 1$.

- Moment and tail bounds:

1. Hoeffding inequality: Assume X_i's are centered, independent, and **bounded** s.t. $|X_i| \leq c_i$ a.s. Then for any $L > 0$,

$$E\exp\left(S_n/L\right) \leq \exp\left(\frac{\sum_{i=1}^n c_i^2}{2L^2}\right).$$

2. Bernstein inequality: Let $\{z_i\}$ be a sequence of independent random elements on $\mathcal{L}$ and $\{g_j\}$ functions on $\mathcal{L}$. Assume that

$$\mathrm{E}g_j(z_i) = 0, \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left|g_j(z_i)\right|^m \leq \frac{m!}{2}K^{m-2},\ m = 2,3,... \tag{9}$$

Then

$$\mathrm{E}\exp\left(\sum_{i=1}^{n}\frac{g_j(z_i)}{L}\right) \leq \exp\left(\frac{n}{2L(L-K)}\right)$$

for any $L > K$.

3. Self-normalized sums: Let X_i be a sequence of independent r.v.'s such that $EX_i = 0, 0 < E\left|X_i\right|^3 \leq K < \infty$ for all $i = 1,...$ Then,

$$\Pr\left\{\frac{\sum_{i=1}^{n}X_i}{\sqrt{\sum_{i=1}^{n}X_i^2}} \leq -x\right\} = \Phi(-x)\left(1 + O(1)\left(\frac{1+x}{d_n}\right)^3\right),$$

uniformly in $x \in [0, d_n]$, where Φ is the cdf of standard normal and

$$d_n = \frac{\left(\sum_{i=1}^{n}EX_i^2\right)^{1/2}}{\left(\sum_{i=1}^{n}E\left|X_i\right|^3\right)^{1/3}}.$$

Lemma 6 *Let $\{z_i\}$ be a sequence of independent random elements on $\mathcal{L}$ and $\{g_j\}$ functions on $\mathcal{L}$. Assume that*

$$\mathrm{E}g_j(z_i) = 0, \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left|g_j(z_i)\right|^m \leq \frac{m!}{2}K^{m-2},\ m = 2, 3, ... \tag{10}$$

Then,

$$\mathrm{E}\left(\max_{j\leq p}\left|\frac{1}{n}\sum_{i=1}^{n}g_j(z_i)\right|^m\right) \leq \left(\frac{K\log\left(p+e^{m-1}\right)}{n} + \left(\frac{2\log\left(p+e^{m-1}\right)}{n}\right)^{1/2}\right)^m.$$

See Lemma 14.12 in Bühlman and van der Geer's Ch 14 for a proof.

1.3 Approximation

Typically, the linear regression is an approximation. That is,

$$Y = f_0 + \varepsilon,$$

and

$$\|f_0 - X\beta\| \to 0,$$

as n, $p \to \infty$ at a certain sense.

For instance, consider the series estimation of the conditional mean function. See Hansen's note for introduction and Chen's (2007) handbook of econometrics chapter.

- We may define the oracle β^* over a collection $\mathcal{S}$ as

$$\beta^* = arg \min_{\beta : S(\beta) \in \mathcal{S}} \left\{ n^{-1} \left| f_0 - X\beta \right|_2^2 + \frac{4\lambda^2 s_\beta}{\phi^2 \left(S\left(\beta\right)\right)} \right\},$$

  where ϕ is the compatibility constant. Accordingly, $S^* = S\left(\beta^*\right)$.

Then,

Theorem 7 *Assume the compatibility condition holds for all $S \in \mathcal{S}$. For some $t > 0$, let*

$$\lambda = 8\hat{\sigma}\sqrt{\frac{t^2 + 2\log p}{n}},$$

where $\hat{\sigma}$ is an estimator of σ. Then,

$$2n^{-1} \left| X\hat{\beta} - f_0 \right|_2^2 + \lambda \left| \hat{\beta} - \beta^* \right|_1 \leq 6n^{-1} \left| X\beta^* - f_0 \right|_2^2 + \frac{24\lambda^2 s_*}{\phi_*^2}$$

with probability at least

$$1 - \left(2 \exp\left(\frac{-t^2}{2} \right) + \Pr\left\{ \hat{\sigma} < \sigma \right\} \right).$$

Proof. Similar to the exact case. The basic inequality now becomes

$$\left| f_0 - X\hat{\beta} \right|_2^2 / n + \lambda_n \left| \hat{\beta} \right|_1 \leq 2\varepsilon' X \left(\hat{\beta} - \beta^* \right) / n + \lambda_n \left| \beta^* \right|_1 + \left| f_0 - X\beta^* \right|_2^2 / n,$$

See BV ch 6.2.3. for further details. ■

- Here we are free to choose $\mathcal{S}$ but the compatibility needs to hold for $S \in \mathcal{S}$. The larger $\mathcal{S}$ the smaller approximation error but the larger requirement in terms of the compatibility condition.

- Now we can see that the oracle β^* is chosen to minimize the prediction error.