A note on variable selection with concave penalty

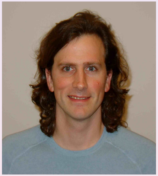
Sara van de Geer

January 2011

Workshop

Sparse Statistics, Optimization and Machine Learning

January 16-21, 2011

Linear model:

$$\mathbf{Y} = \mathbf{X}\beta^0 + \epsilon,$$

where $\mathbf{Y} \in \mathbb{R}^n$, $\mathbf{X} = (n \times p)$-matrix, $\beta \in \mathbb{R}^p$.
The Lasso [Tibshirani, 1995]

$$\hat{\beta} := \arg\min_{\beta} \left\{ \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 / n + \lambda \|\beta\|_1 \right\}.$$

Some notation

The columns of $\mathbf{X}$:

$$\psi_j := \begin{pmatrix} x_{1,j} \\ \vdots \\ x_{n,j} \end{pmatrix}, \; j = 1, \ldots, p,$$

i.e.,

$$\mathbf{X} = (\psi_1, \ldots, \psi_p).$$

We use the normalization

$$\|\psi_j\|_2^2 / n = 1.$$

The Gram matrix

$$\hat{\Sigma} := \mathbf{X}^T \mathbf{X} / n.$$

The "truth"

$$f^0 := \mathbf{X}\beta^0 = \sum_{j=1}^{p} \psi_j \beta_j^0.$$

The true active set

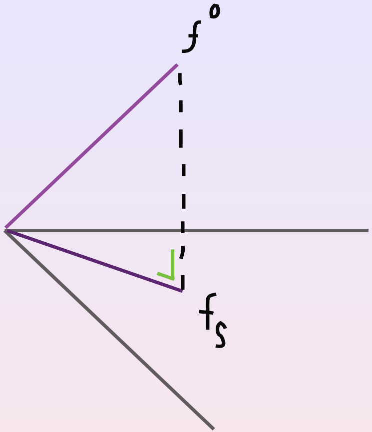
$$S_0 := \{j : \ \beta_j^0 \neq 0\}.$$

For an index set $S \subset \{1, \ldots, p\}$ and $\beta \in \mathbb{R}^p$, we set

$$\beta_{j,S} = \beta_j \mathrm{l}\{j \in S\},$$

i.e., $\beta_S \in \mathbb{R}^p$ has zeroes outside S.
The projection of f^0on the space spanned by $\{\psi_j\}_{j \in S}$:

$$\mathrm{f}_S := \mathbf{X}b^S, \ b^S := \min_{\beta} \|\mathbf{X}\beta_S - f^0\|_2.$$



The Projection on the Space Spanned by S

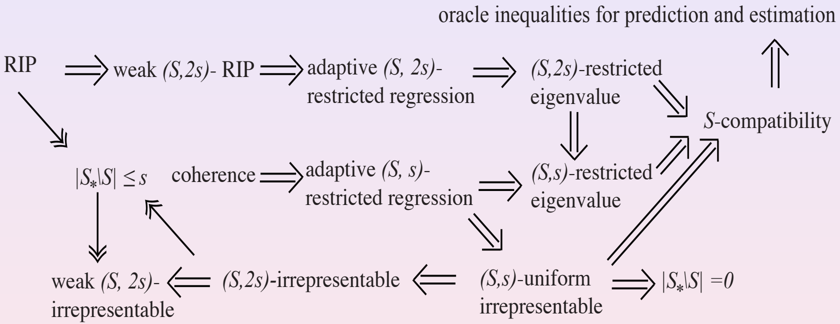
The ℓ_1-compatibility condition

Let $L > 0$ be some constant and S be an index set with cardinality $s = |S|$. We say that the ℓ_1-compatibility condition holds if

$$\phi^2(L,S) := \min\{\beta^T \hat{\Sigma} \beta s : \; \|\beta\|_1 = 1, \; \|\beta_{S^c}\|_1 \leq L\}$$

is strictly positive.

On the conditions used...



Handling of the noise

We assume throughout that

$$\max_{1 \le j \le p} 2|\epsilon^T \psi_j|/n \le \lambda_0,$$

and that

$$\lambda \ge 2\lambda_0.$$

Lemma

Suppose that $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. *Then for*

$$\lambda_0 = 2\sigma\sqrt{\frac{t^2 + 2\log p}{n}},$$

we have

$$\mathbb{P}\left(\max_{1 \le j \le p} 2|\epsilon^T \psi_j|/n > \lambda_0\right) \le 2\exp[-t^2/2].$$

Definition of the oracle

The active set of the oracle is

$$S_* := \min_S \left\{ \underbrace{\|\mathrm{f}_S - f_0\|_2^2/n}_{\text{approximation error}} + \underbrace{\frac{4\lambda^2 s}{\phi^2(3, S)}}_{\text{estimation error}} \right\}.$$

Here,

$$L = \frac{\lambda + \lambda_0}{\lambda - \lambda_0} = 3, \text{ because of our choice } \lambda = 2\lambda_0.$$

We write

$$\beta^* := b^{S_*},\ f^* := \mathrm{f}_{S_*} = \mathbf{X}\beta^*,\ s_* := |S_*|.$$

The prediction error of the Lasso

Theorem

We have

$$2\|\mathbf{X}\hat{\beta} - f^0\|_2^2/n + \lambda\|\hat{\beta} - \beta^*\|_1 \leq 3\left\{\|f^* - f^0\|_2^2 + \frac{4\lambda^2}{\phi^2(3, S_*)}\right\}.$$

The irrepresentable condition

[Meinshausen and Bühlmann, 2006] [Zhao and Yu, 2006]

Let

$$\hat{\Sigma}_{1,1}(S) := (\hat{\Sigma}_{j,k})_{j,k \in S}, \hat{\Sigma}_{1,2}(S) := (\hat{\Sigma}_{j,k})_{j \in S, k \notin S},$$

E.g., when $S = \{1, \ldots, s\}$,

$$\hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_{1,1}(S) & \hat{\Sigma}_{1,2}(S) \\ \hat{\Sigma}_{2,1}(S) & \hat{\Sigma}_{2,2}(S) \end{pmatrix},$$

where $\hat{\Sigma}_{2,2}(S) = \hat{\Sigma}_{1,1}(S^c)$.

Write $\hat{S} := \{j : \hat{\beta}_j \neq 0\}$.

Lemma

Suppose the irrepresentable condition

$$\sup_{\|\tau_{S_0}\|_\infty \leq 1} \|\hat{\Sigma}_{2,1}(S_0)\hat{\Sigma}_{1,1}(S_0)\tau_{S_0}\|_\infty \leq \theta < \frac{\lambda - \lambda_0}{\lambda + \lambda_0}.$$

Then there are no false positives:

$$\hat{S} \subset S_0.$$

Lemma

Suppose the irrepresentable condition

$$\sup_{\|\tau_{S_0}\|_\infty \le 1} \|\hat\Sigma_{2,1}(S_0)\hat\Sigma_{1,1}(S_0)\tau_{S_0}\|_\infty \le \theta < \frac{\lambda - \lambda_0}{\lambda + \lambda_0}.$$

Then the compatibility condition holds for $L\theta < 1$,

$$\phi^2(L, S_0) \ge (1 - L\theta)^2 \Lambda^2_{\min}(\hat\Sigma_{1,1}(S_0)),$$

where $\Lambda^2_{\min}(\hat\Sigma_{1,1}(S_0))$ is the smallest eigenvalue of $\hat\Sigma_{1,1}(S_0)$.

Recall we applied

$$L = \frac{\lambda + \lambda_0}{\lambda - \lambda_0}.$$

Benchmark: the ℓ_0-penalty

Let $\epsilon \sim \mathcal{N}(0, \sigma)$, $\lambda \asymp \sqrt{\log p/n}$ and

$$\hat{\beta}_{\text{ideal}} := \arg\min_{\beta} \Big\{ \|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda^2 \underbrace{\|\beta\|_0^0}_{:= \#\{\beta_j \neq 0\}} \Big\}.$$

Then one can show that with large probability

$$\|\mathbf{X}\hat{\beta}_{\text{ideal}} - f^0\|_2^2/n + \lambda^2 \hat{s}_{\text{ideal}} \leq \text{const.} \Big\{ \|f^* - f^0\|_2^2/n + \lambda^2 s_* \Big\},$$

and hence

$$\hat{s}_{\text{ideal}} = \mathcal{O}(s_*).$$

[Barron et al. 1999]

The number of false positives of the Lasso

Recall that S_* is the oracle active set. Generally

$$S_* \subset S_0.$$

Lemma

We have

$$|\hat{S} \backslash S_*| \leq \left[\frac{\Lambda_{\max}^2}{\phi^2(3, S_*)} \right] \mathcal{O}(s_*),$$

where $\Lambda_{\max}^2$ is the largest eigenvalue of $\hat{\Sigma}$.

An idealized example: equal correlation

Let

$$\hat{\Sigma} := \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \rho \\ \rho & \rho & \cdots & 1 \end{pmatrix}$$

$$= (1-\rho)I + \rho\tau\tau^T,$$

where $0 < \rho < 1$ and $\tau = (1, \ldots, 1)^T$. Then

$$\Lambda^2_{\max} = (1-\rho) + \rho p,$$

and

$$\phi^2(L, S) = 1 - \rho.$$

We take

$$\Delta := \sup_{\|\tau_{S_0}\|_\infty \le 1} \|\hat{\Sigma}_{2,1}(S_0)\hat{\Sigma}_{1,1}(S_0)\tau_{S_0}\|_\infty - \frac{\lambda - \lambda_0}{\lambda + \lambda_0}.$$

which holds for $\rho \gg 1/s_0$.

We assume

$$2\epsilon^T \psi_j / n = \begin{cases} -\lambda_0 & j \in S_0 \\ +\lambda_0 & j \notin S_0 \end{cases},$$

and

$$\beta_j^0 = b_0 \; \forall \, j \in S_0,$$

where

$$b_0 > \frac{\lambda + \lambda_0}{2} \left(\frac{1}{1 - \rho + \rho s_0} + \frac{\rho(p - s_0)\Delta}{(1-\rho)(1 - \rho + \rho p)} \right).$$

Then for $j \in S_0$,

$$\hat{\beta}_j = b_0 - \frac{\lambda + \lambda_0}{2} \left(\frac{1}{1 - \rho + \rho s_0} + \frac{\rho(p - s_0)\Delta}{(1-\rho)(1 - \rho + \rho p)} \right),$$

and for $j \notin S_0$,

$$\hat{\beta}_j = \frac{\lambda + \lambda_0}{2} \frac{\Delta(1 - \rho + \rho s_0)}{(1-\rho)(1 - \rho + \rho p)}.$$

Thus the Lasso selects **all** variables, so that

$$|\hat{S} \backslash S_0| = p - s_0!$$

The ℓ_r-"norm" penalty, $0 < r < 1$.

We let

$$\hat{\beta} := \arg\min_{\beta} \left\{ \|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda^{2-r}\|\beta\|_r^r \right\}.$$

[Zhang, 2010]

The ℓ_r-compatibility condition

We say that the ℓ_r-compatibility condition holds if

$$\phi_r^2(L,S) := \min\{\beta^T \hat{\Sigma}\beta|S|^{\frac{2-r}{2}} : \ \|\beta_S\|_r = 1, \ \|\beta_{S^c}\|_r \leq L\}$$

is strictly positive.

Handling the noise

We assume that

$$\sup_{\beta} \frac{2|\epsilon^T \mathbf{X}\beta|/n}{\|\beta\|_r^{\frac{2}{2-r}} (\|\mathbf{X}\beta\|_n^2/n)^{\frac{1-r}{2-r}}} \leq \lambda_0,$$

and that

$$\lambda^{2-r} \geq 5\lambda_0^{2-r} 4^{1-r}.$$

Lemma

Suppose $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. Then for a constant c_r, and for

$$\lambda_0 = c_r \sigma \sqrt{\frac{\log(2p) + t^2}{n}},$$

we have

$$\mathbb{P}\left(\sup_{\beta} \frac{2|\epsilon^T \mathbf{X}\beta|/n}{\|\beta\|_r^{\frac{2}{2-r}} (\|\mathbf{X}\beta\|_n^2/n)^{\frac{1-r}{2-r}}} > \lambda_0 \right) \leq 2\exp[-t^2/2].$$

Prediction error of the ℓ_r-norm penalized estimator

Definition of the oracle

$$S_* := \arg\min_S \left\{ \|\mathrm{f}_S - f^0\|_2^2/n + \frac{3(9\lambda)^2 s^2}{\phi_r^{\frac{2r}{2-r}}(S)} \right\},$$

and $f^* := \mathrm{f}_{S_*}$, $s_* := |S_*|$.

Theorem

It holds that

$$\|\mathbf{X}\hat{\beta} - f^0\|_2^2/n + 4\lambda^{2-r}\|\hat{\beta} - \beta^*\|_r^r/5 \leq 4\left\{ \|f^* - f^0\|_2^2/n + \frac{3(9\lambda)^2 s_*}{\phi_r^{\frac{2r}{2-r}}(S_*)} \right\}.$$

Definition sparse eigenvalue

$$\Lambda_{\rm sparse}^2(s) := \max_{S:\ |S|=s} \Lambda_{\rm max}(\hat\Sigma_{1,1}(S)).$$

Example: equal correlation Let

$$\hat\Sigma := (1-\rho)I + \rho\tau\tau^T.$$

Then

$$\Lambda_{\rm sparse}^2(s) = (1-\rho) + \rho s.$$

Variable selection with ℓ_r-penalty

Theorem

$$|\hat{S} \backslash S_*| = \left[\frac{\Lambda_{\mathrm{sparse}}(s_*)}{\phi_r(3, S_*)}\right]^{\frac{r}{1-r}} \left[\frac{1}{\phi_r(3, S_*)}\right]^{\frac{r}{1-r}} \mathcal{O}(s_*)$$

$$\wedge \left[\frac{1}{\phi_r(3, S_*)}\right]^{\frac{r}{1-r}} \mathcal{O}(s_*^{1+\frac{r}{2(1-r)}}).$$

Variable selection with ℓ_r-penalty

Theorem

$$|\hat{S} \backslash S_*| = \left[\frac{\Lambda_{\mathrm{sparse}}(s_*)}{\phi_r(3, S_*)}\right]^{\frac{r}{1-r}} \left[\frac{1}{\phi_r(3, S_*)}\right]^{\frac{r}{1-r}} \mathcal{O}(s_*)$$

$$\wedge \left[\frac{1}{\phi_r(3, S_*)}\right]^{\frac{r}{1-r}} \mathcal{O}(s_*^{1+\frac{r}{2(1-r)}}).$$

The adaptive Lasso

Let

$$\hat{\beta}_{\mathrm{init}} = \arg\min_{\beta} \left\{ \|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda_{\mathrm{init}} \sum_j |\beta_j| \right\},$$

and

$$\hat{\beta}_{\mathrm{adap}} = \arg\min_{\beta} \left\{ \|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda_{\mathrm{init}}\lambda_{\mathrm{adap}} \sum_j |\beta_j|/|\hat{\beta}_{\mathrm{init},j}| \right\}.$$

[Zou, 2006]

Theorem

Take

$$\lambda_{\text{adap}} \asymp \left[\frac{\Lambda_{\text{sparse}}(s_*)}{\phi_*^3}\right]\lambda_{\text{init}}.$$

Then

$$\|\mathbf{X}\hat{\beta}_{\text{adap}} - f^0\|_2^2 = \left[\frac{\Lambda^2_{\text{sparse}}(s_*)}{\phi_*^2}\right]\mathcal{O}\left(\frac{\lambda^2_{\text{init}} s_*}{\phi_*^2}\right),$$

and

$$|\hat{S}_{\text{adap}} \backslash S_*| = \left[\frac{\Lambda^2_{\text{sparse}}(s_*)}{\phi_*^2}\right]\mathcal{O}(s_*).$$

Define for $\lambda_{\text{thres}} > 0$,

$$S_*^{\text{thres}} := \{j : \ |\beta_j^*| > 4\lambda_{\text{thres}}\},$$

and

$$f_{\text{thres}}^* = \mathrm{f}_{S_*^{\text{thres}}}.$$

Let

$$|\beta^*|_{\text{trim}}^2 := \left(\frac{1}{s_*} \sum_{|\beta_j^*| > 2\lambda_{\text{thres}}} \frac{1}{|\beta_j^*|^2} \right)^{-1}.$$

Note that

$$|\beta^*|_{\text{trim}} > 2\lambda_{\text{thres}}.$$

Theorem

Suppose

$$\|\hat{\beta}_{\rm init} - \beta^*\|_\infty \le \lambda_{\rm thres}.$$

Take

$$\lambda_{\rm adap} \asymp \left(1 + \frac{\|f^*_{\rm thres} - f^0\|_2^2/n}{\lambda_{\rm init}^2 s_*/\phi_*^2}\right) |\beta^*|^2_{\rm trim}.$$

Then

$$\|\mathbf{X}\hat{\beta}_{\rm adap} - f^0\|_2^2/n = \left[\frac{\lambda^2_{\rm adap}}{|\beta^*|^2_{\rm trim}}\right] \mathcal{O}(\frac{\lambda^2_{\rm init} s_*}{\phi_*^2}),$$

and

$$|\hat{S}_{\rm adap} \backslash S_*| = \left[\frac{\lambda^2_{\rm init}}{|\beta^*|^2_{\rm trim}}\right] \mathcal{O}\left(\frac{s_*^2}{\phi_*^6}\right).$$

Conclusion

- When $\Lambda_{\text{sparse}}(s_*) \asymp 1$ the adaptive Lasso mimics the ℓ_r-Lasso
- When $\Lambda_{\text{sparse}}(s_*)$ is very large the ℓ_r-Lasso still has good prediction error. Under beta-min conditions, that is, conditions which require $|\beta_j^*|$ to be sufficiently large for all $j \notin S_*$, the prediction error of the adaptive Lasso is also good. Otherwise, it may be problematic. (The same holds for the thresholded Lasso.)

Reference

P. Bühlmann and S.A. van de Geer. *Statistics for high-dimensional data. Methods, Theory and Applications* Springer, to appear (2011).

THANK YOU!