Linear Methods for Regression: Theory

-Statistical Machine Learning-

Lecturer: Darren Homrighausen, PhD

Ridge theory

Set-up: assuming a linear model

Let $Y_i = X_i^\top \beta + \epsilon_i$, for $i = 1, \ldots, n$, where

- $X_i \in \mathbb{R}^p$
- $\mathbb{E}\epsilon_i = 0$ and $\mathbb{E}\epsilon\epsilon^\top = I_n$ (w.l.o.g. $\sigma^2 = 1$)
- $\mathbb{X}$ is the feature matrix, and $\text{rank}(\mathbb{X}) = p$

We'll consider various properties that may be of interest:

- Estimating β
- Doing good predictions

Estimating β in low dimensions

To get L^2 consistency, we need to show that
(writing for this section $\hat{\beta}_{\mathrm{ridge}}(\lambda) \equiv \hat{\beta}_\lambda$)

$$R(\hat{\beta}_\lambda) = \mathbb{E}_{\mathcal{D}}||\hat{\beta}_\lambda - \beta||_2^2$$

goes to zero.

Again, we can decompose this as ($\mathbb{E}_{\mathcal{D}} \equiv \mathbb{E}$)

$$R(\hat{\beta}) = \mathbb{E}||\hat{\beta} - \mathbb{E}\hat{\beta}||_2^2 + ||\mathbb{E}\hat{\beta} - \beta||_2^2 \quad (1)$$

$$= \mathrm{trace}\mathbb{V}\hat{\beta} + \sum_{j=1}^{p}(\mathbb{E}\hat{\beta}_j - \beta_j)^2 \quad (2)$$

Estimating β

Continuing from the previous slide

(Remember, $\hat{\beta}_\lambda = (\mathbb{X}^\top\mathbb{X} + \lambda I)^{-1}\mathbb{X}^\top Y$)

$$
\begin{aligned}
R(\hat{\beta}) &= \mathrm{trace}\mathbb{V}\hat{\beta}_\lambda + ||\mathbb{E}\hat{\beta}_\lambda - \beta||_2^2 && (3)\\
&= \mathrm{trace}(\mathbb{X}^\top\mathbb{X} + \lambda I)^{-1}\mathbb{X}^\top\mathbb{X}(\mathbb{X}^\top\mathbb{X} + \lambda I)^{-1} + && (4)\\
&\qquad + ||((\mathbb{X}^\top\mathbb{X} + \lambda I)^{-1}\mathbb{X}^\top\mathbb{X} - I)\beta||_2^2 && (5)\\
&= \mathrm{variance} + \mathrm{bias}^2 && (6)
\end{aligned}
$$

Let's address each of these terms separately

Estimating β: Bias

For the bias, let's use the Woodbury matrix inversion lemma

$$(A - BC^{-1}E)^{-1} = A^{-1} + A^{-1}B(C - EA^{-1}B)^{-1}EA^{-1}$$

(See Henderson, Searle (1980), equation (12) for a statement and discussion)

$$\begin{aligned}
\text{bias}^2 &= ||((\underbrace{\mathbb{X}^\top\mathbb{X}}_{E} + \underbrace{\lambda I}_{C})^{-1}\mathbb{X}^\top\mathbb{X} \underbrace{-I}_{A^{-1}})\beta||_2^2 && (7)\\
&= ||(I + (\mathbb{X}^\top\mathbb{X})\lambda^{-1})^{-1}\beta||_2^2 && (8)\\
&= \lambda^2||(\lambda I + \mathbb{X}^\top\mathbb{X})^{-1}\beta||_2^2 && (9)\\
&= \lambda^2||(\lambda I + VD^2V^\top)^{-1}\beta||_2^2 && (10)\\
&= \lambda^2||(V(\lambda V^\top V + D^2)V^\top)^{-1}\beta||_2^2 && (11)\\
&= \lambda^2||(\lambda I + D^2)^{-1}\theta||_2^2 \qquad (n>p,\ \theta=V^\top\beta) && (12)\\
&= \lambda^2\sum_{j=1}^{p}\frac{\theta_j^2}{(\lambda + d_j^2)^2} && (13)
\end{aligned}$$

Estimating β: Variance

Likewise,

$$\begin{aligned}
\text{variance} &= \text{trace}\left((\mathbb{X}^\top\mathbb{X} + \lambda I)^{-1}\mathbb{X}^\top\mathbb{X}(\mathbb{X}^\top\mathbb{X} + \lambda I)^{-1}\right) && (14)\\
&= \text{trace}\left(D^2(D^2 + \lambda I)^{-2}\right) && (15)\\
&= \sum_{j=1}^{p} \frac{d_j^2}{(d_j^2 + \lambda)^2} && (16)
\end{aligned}$$

Putting them together:

$$R(\hat{\beta}) = \sum_{j=1}^{p} \left(\frac{\lambda^2\theta_j^2 + d_j^2}{(\lambda + d_j^2)^2} \right)$$

What now?

Putting them together: Estimation risk

$$R(\hat{\beta}) = \sum_{j=1}^{p} \left(\frac{\lambda^2 \theta_j^2 + d_j^2}{(\lambda + d_j^2)^2} \right) \tag{17}$$

$$\Rightarrow \frac{\partial R(\hat{\beta})}{\partial \lambda} = \sum_{j=1}^{p} \frac{2 d_j^2 (\lambda \theta_j^2 - 1)}{(\lambda + d_j^2)^3} \tag{18}$$

Putting them together: Estimation risk

$$R(\hat{\beta}) = \sum_{j=1}^{p} \left(\frac{\lambda^2 \theta_j^2 + d_j^2}{(\lambda + d_j^2)^2} \right) \tag{17}$$

$$\Rightarrow \frac{\partial R(\hat{\beta})}{\partial \lambda} = \sum_{j=1}^{p} \frac{2 d_j^2 (\lambda \theta_j^2 - 1)}{(\lambda + d_j^2)^3} \tag{18}$$

This suggests taking $\hat{\lambda} = 1/\theta_{\max}^2$. Observe

$$R(\hat{\beta}_{\hat{\lambda}}) \leq \sum_{j=1}^{p} \left(\frac{1/\theta_{\max}^2 + d_j^2}{(1/\theta_{\max}^2 + d_j^2)^2} \right) = \sum_{j=1}^{p} \left(\frac{1}{\theta_{\max}^{-2} + d_j^2} \right) < \sum_{j=1}^{p} \left(\frac{1}{d_j^2} \right)$$

(As long as $0 < \theta_{\max} < \infty$)

High dimensional prediction

Let's now suppose $p > n$ and write

$$Y = \mathbb{X}\beta + \epsilon$$

We immediately run into a problem:

$$Y = \mathbb{X}(\beta + b) + \epsilon$$

for any b in the null space of $\mathbb{X}$.

This means that β is non-identified in high dimensions!

(Identifiable $\Rightarrow \mathbb{X}\beta = \mathbb{X}\beta'$ means $\beta = \beta'$)

Identifiability

If we let $\mathrm{rank}(\mathbb{X}) = r$ (and hence $r < p$), then

- $U \in \mathbb{R}^{n \times r}$
- $D \in \mathbb{R}^{r \times r}$
- $V \in \mathbb{R}^{p \times r}$
- $V_\perp \in \mathbb{R}^{p \times (p-r)}$ be orthonormal and $V^\top V_\perp = 0$

Let $\theta = \mathbb{X}^\top(\mathbb{X}\mathbb{X}^\top)^\dagger \mathbb{X}\beta = VV^\top \beta$

Then $\theta \in \mathbb{R}^p$ and $Y = \mathbb{X}\theta + \epsilon$ and hence estimating θ is enough for predictions.

Now, we form

$$\hat{\theta} = (\mathbb{X}^\top \mathbb{X} + \lambda I)^{-1} \mathbb{X}^\top Y$$

Bias

$$
\begin{aligned}
\mathrm{Bias}(\hat{\theta}) &= \mathbb{E}\hat{\theta} - \theta && (19)\\
&= -(\lambda^{-1}\mathbb{X}^\top\mathbb{X} + I)^{-1}\theta \qquad \text{(Woodbury)} && (20)\\
&= -\Gamma(\lambda^{-1}\Gamma^\top\mathbb{X}^\top\mathbb{X}\Gamma + I)^{-1}\Gamma^\top VV^\top\theta \quad (\Gamma=[V,V_\perp])^1 && (21)\\
&= -[V, V_\perp]\begin{bmatrix}(\lambda^{-1}D^2 + I_r)^{-1} & 0\\ 0 & I_{p-r}\end{bmatrix}\begin{bmatrix}V^\top\\ V_\perp^\top\end{bmatrix} VV^\top\theta && (22)\\
&= -[V(\lambda^{-1}D^2 + I_r)^{-1}, V_\perp]\begin{bmatrix}V^\top\theta\\ 0\end{bmatrix} && (23)\\
&= -V(\lambda^{-1}D^2 + I_r)^{-1}V^\top\theta && (24)
\end{aligned}
$$

(This derivation is a more general version of the previous)

[1] Γ is such that $\Gamma\Gamma^\top = \Gamma^\top\Gamma = I$

Variance

We can make a somewhat simple bound on the variance:

$$\mathbb{V}\hat{\theta} = (\mathbb{X}^\top \mathbb{X} + \lambda I)^{-1} \mathbb{X}^\top \mathbb{X} (\mathbb{X}^\top \mathbb{X} + \lambda I)^{-1} \tag{25}$$

$$\leq (\mathbb{X}^\top \mathbb{X} + \lambda I)^{-1} \tag{26}$$

$$\tag{27}$$

($A \leq B$ means $B - A$ is nonnegative definite)

Prediction risk

Theorem

There exists a constant C such that for n large enough

$$n^{-1}\mathbb{E}||\mathbb{X}(\hat{\theta}-\theta)||_2^2 \leq C\left(\frac{r}{n}+\lambda^2 n^{-(1+\eta-2\tau)}\right)$$

where $d_{\min}^{-2} \leq n^{-\eta}$ and $||\theta||_2 \leq n^{\tau}$

Note that

$$||\beta||_2^2 \geq ||VV^\top \beta||_2^2 = ||\theta||_2^2$$

Proof.

$$\mathbb{E}||\mathbb{X}(\hat{\theta} - \theta)||_2^2 = \mathrm{trace}(\mathbb{X}\mathbb{V}[\hat{\theta}]\mathbb{X}^\top) + ||\mathbb{X}\mathrm{bias}(\hat{\theta})||_2^2.$$

Using the variance bound

$$\mathbb{X}\mathbb{V}[\hat{\theta}]\mathbb{X}^\top \leq \mathbb{X}(\mathbb{X}^\top\mathbb{X} + \lambda I)^{-1}\mathbb{X}^\top \leq UU^\top$$

We get

$$\mathrm{trace}(\mathbb{X}\mathbb{V}[\hat{\theta}]\mathbb{X}^\top) \leq \mathrm{trace}(UU^\top) = r \quad (UU^\top \text{ is a rank } r \text{ projection})$$

$$\begin{aligned} ||\mathbb{X}\mathrm{bias}(\hat{\theta})||_2^2 &= ||UD(\lambda^{-1}D^2 + I_r)^{-1}V^\top\theta||_2^2 && (28)\\ &\leq ||D(\lambda^{-1}D^2 + I_r)^{-1}||_2^2||\theta||_2^2 && (29)\\ &\leq (\max_j \frac{\lambda^2}{d_j^2})^2||\theta||_2^2 = \lambda^2 d_{\min}^{-2}||\theta||_2^2 && (30) \end{aligned}$$

Prediction risk

So,

$$n^{-1}\mathbb{E}||\mathbb{X}(\hat{\theta}-\theta)||_2^2 \lesssim \frac{r}{n} + \frac{\lambda^2 d_{\min}^{-2}||\theta||_2^2}{n}$$

This shows the result:

Theorem

There exists a constant C such that for n large enough

$$n^{-1}\mathbb{E}||\mathbb{X}(\hat{\theta}-\theta)||_2^2 \leq C\left(\frac{r}{n} + \lambda^2 n^{-(1+\eta-2\tau)}\right)$$

where $d_{\min}^{-2} \leq n^{-\eta}$ and $||\theta||_2 \leq n^{\tau}$

Better result?

This result relies on a very crude bound of the variance, which I find unsatisfying.

Challenge: Can you tighten this bound? Is this as good as possible?

(As a reference, this material is in Shao, Deng (2012) in the Annals of Statistic. They introduce a thresholded ridge to get a faster rate.)

Normal means

A simpler model

Suppose that $Y \sim (\mu, 1)$ and let

$$L_q(\mu) = 2^{q-2}(Y - \mu)^2 + \lambda|\mu|^q$$

and

$$\hat{\mu}_q = \underset{\mu}{\text{argmin}}\, L_q(\mu)$$

Then,

- $q = 0 \Rightarrow \hat{\mu}_0 = Y$
- $q = 2 \Rightarrow \hat{\mu}_2 = Y/(\lambda + 1)$
- $q = 1$?

Subdifferential

To theoretically solve this optimization problem, we use the notion of a subderivative.

We call c a subderivative of f at x_0 provided

$$f(x) - f(x_0) \geq c(x - x_0)$$

A convex function can be optimized by setting the subderivative $= 0$

The subdifferential $\partial f|_{x_0}$ is the set of subderivatives.

x_0 minimizes a convex function f if and only if $0 \in \partial f|_{x_0}$.

Subdifferential in action

Reminder:

$$L_q(\mu) = 2^{q-2}(Y-\mu)^2 + \lambda|\mu|^q$$

For $\rho(\mu) = |\mu|$,

$$\partial\rho|_\mu = \begin{cases} \{-1\} & \text{if } \mu < 0 \\ [-1,1] & \text{if } \mu = 0 \\ \{1\} & \text{if } \mu > 0 \end{cases}$$

Therefore

$$\partial L_1|_\mu = \begin{cases} \{\mu - Y - \lambda\} & \text{if } \mu < 0 \\ \{\mu - Y + \lambda z : -1 \le z \le 1\} & \text{if } \mu = 0 \\ \{\mu - Y + \lambda\} & \text{if } \mu > 0 \end{cases}$$

ℓ_1 AND SOFT-THRESHOLDING

As L_1 is convex, $\hat{\mu}_1$ minimizes L_1 if and only if $0 \in \partial L_1$.

So..

$$\hat{\mu}_1 = \begin{cases} Y + \lambda & \text{if } Y < -\lambda \\ 0 & \text{if } -\lambda \leq Y \leq \lambda \\ Y - \lambda & \text{if } Y > \lambda \end{cases}$$

This can be written

$$\hat{\mu}_1 = \operatorname{sgn}(Y)(|Y| - \lambda)_+$$

This is known as soft thresholding

Orthogonal design: Example

Suppose now that $(p \leq n)$

$$Y = \mathbb{X}\beta + \epsilon,$$

where $\mathbb{X}^\top \mathbb{X}/n = I$.

Let's solve

$$\hat{\beta}_\lambda = \underset{\beta}{\operatorname{argmin}} \frac{1}{2n}||\mathbb{X}\beta - Y||_2^2 + \lambda||\beta||_1$$

$$\frac{1}{2n}||\mathbb{X}\beta - Y||_2^2 \propto \frac{\beta^\top \mathbb{X}^\top \mathbb{X}\beta}{2n} - \frac{\beta^\top \mathbb{X}^\top Y}{n} = \frac{\beta^\top \beta}{2} - \beta^\top \hat{\beta}_{LS}$$

Now,

$$\frac{1}{2n}||\mathbb{X}\beta - Y||_2^2 + \lambda||\beta||_1 = \sum_{j=1}^{p} \left(\beta_j^2/2 - \beta_j \hat{\beta}_{LS,j} + \lambda|\beta_j|\right)$$

Orthogonal design

We can minimize this component wise:

$$L(\beta) = \beta^2/2 - \beta\hat{\beta}_{LS} + \lambda|\beta| \qquad \text{(dropping the } j\text{)}$$

This can be optimized using subdifferentials [Exercise]
This results in soft-thresholding the least squares solution.

This rationale can be extended to make the lasso coordinate descent explicit

From orthogonal to non-orthogonal

An iterative algorithm for finding $\hat{\beta} \equiv \hat{\beta}_\lambda$ is:

Set $\hat{\beta} = (0, \ldots, 0)^\top$. Then for $j = 1, \ldots, p$:

1. Define $R_i = \sum_{k \neq j} \hat{\beta}_k X_{ik}$
2. Form $\hat{\beta}_j$ by simple linear regression of $(R_i)_{i=1}^n$ on x_j
3. Soft-threshold these coefficients:
   $\hat{\beta}_j = \mathrm{sgn}(\hat{\beta}_j)(|\hat{\beta}_j| - \lambda/||x_j||_2^2)_+$

In words, to implement coordinate descent for lasso, soft-threshold the least squares coordinate descent

(This same insight was extended to sparse additive models as well Ravikumar, Lafferty, Liu, Wasserman (2009))

Normal means

Note that the orthogonal design linear model is an example of a normal means problem:

Let $\epsilon \sim N(0, I)$, then

$$Y = \mathbb{X}\beta + \epsilon \Leftrightarrow W \overset{D}{=} \beta + \frac{1}{\sqrt{n}}\epsilon$$

This turns out to be an even more powerful idea..

Normal means

Let

- $\mathcal{H}$ be a real, separable Hilbert space with inner product $\langle \cdot, \cdot \rangle$
- (ϕ_i) be an orthonormal basis for $\mathcal{H}$

Then we can imagine a signal h being observed with a white noise Gaussian process

$$Y(t) = h(t) + \epsilon(t)$$

(Technically, this doesn't exist. Rather we can observe functionals $Y(t)dt = h(t)dt + d\epsilon(t)$)

We make observations of this signal via inner products:

$$y_i = \langle Y, \phi_i \rangle = \langle h + \epsilon, \phi_i \rangle = h_i + \epsilon_i$$

As linear operations of Gaussians are Gaussians, $\epsilon_i \sim N(0, 1)$

Assumptions in high dimension

Most theoretical papers on high-dimensional regression have several components:

- The linear model is correct.
- The variance is constant.
- The errors have a Normal distribution (or related distribution)
- The parameter vector is sparse.
- The feature matrix has very weak collinearity.

  (E.g. incoherence, eigenvalue restrictions, or incompatibility assumptions. We'll return to this later)

These assumptions are not testable when $p > n$

In fact, high collinearity is the rule, rather than exception

Low assumption prediction: the lasso

Remember: Prediction risk

$$R(\beta) = \mathbb{E}_Z\left[\left(Y - X^\top \beta\right)^2\right] = \mathbb{E}_Z\left[\left(Y - X^\top \beta\right)^2 |\mathcal{D}\right]$$

Define the oracle estimator

$$\beta_t^* = \underset{\{\beta: ||\beta||_1 \leq t\}}{\operatorname{argmin}} R(\beta)$$

(Important: This does not assume that $\mathbb{E}Y|X$ is linear in X!)

The excess risk is

$$\mathcal{E}(\hat{\beta}_t, \beta_t^*) = R(\hat{\beta}_t) - R(\beta_t^*)$$

Persistence

A procedure is persistent for a set of measures $\mathcal{P}$ if

$$\forall \mathbb{P} \in \mathcal{P}, \qquad \mathcal{E}(\hat{\beta}_t, \beta_t^*) \xrightarrow{p} 0$$

(This is convergence in probability. What is random?)

Define the following set of distributions on $\mathbb{R} \times \mathbb{R}^p$: Let $C_{\mathcal{P}} < \infty$ and

$$\mathcal{P} = \{\mathbb{P} : \mathbb{P} Y^2 < C_{\mathcal{P}}, \text{ and } |X_j| < C_{\mathcal{P}} \text{ almost surely}, j = 1, \ldots, p\}$$

We'd like to know how fast t can grow while still maintaining persistency.

Useful results and observations

Let

- $Z = (Y, X_1, \ldots, X_p)^\top \in \mathbb{R}^{p+1}$
- $\gamma = (-1, \beta_1, \ldots, \beta_p)^\top$

Then, for $\ell_\beta(Z) = (Y - X^\top \beta)^2$,

$$\mathbb{P}\ell_\beta = \mathbb{P}(Y - X^\top \beta)^2 = \gamma^\top \Sigma \gamma,$$

where $\Sigma_{jk} = \mathbb{P} Z_j Z_k$ for $0 \leq j, k \leq p$

Likewise,

$$\hat{\mathbb{P}}\ell_\beta = \gamma^\top \hat{\Sigma} \gamma,$$

where $\hat{\Sigma}_{jk} = n^{-1} \sum_{i=1}^n Z_{ij} Z_{ik}$ for $0 \leq j, k \leq p$

(These can be written: $\Sigma = \mathbb{P} Z Z^\top$ and $\hat{\Sigma} = \hat{\mathbb{P}} Z Z^\top$)

Persistence theorem

Theorem

Over any $\mathbb{P}$ in $\mathcal{P}$, the procedure

$$\underset{\beta \in \{\beta : ||\beta||_1 \leq t\}}{\text{argmin}} \quad \hat{\mathbb{P}} \ell_\beta$$

is persistent provided $\log p = o(n)$ *and*

$$t = t_n = o\left(\left(\frac{n}{\log p}\right)^{1/4}\right)$$

(This theorem appears in Greenshtein, Ritov (2004), though with a substantially different proof)

Deterministic asymptotic notation

We write $a_n = O(b_n)$ (and say big ohh) provided

$$\frac{a_n}{b_n} = O(1),$$

where

$$c_n = O(1)$$

means

- There exists a C
- Such that for sufficiently large N
- For all $n \geq N$
- $c_n \leq C$

Deterministic asymptotic notation

We write $a_n = o(b_n)$ (and say little ohh) provided

$$\frac{a_n}{b_n} = o(1),$$

where

$$c_n = o(1)$$

means

- For all $\epsilon > 0$
- There exists an N
- Such that for all $n \geq N$
- $c_n \leq \epsilon$

Stochastic asymptotic notation

We write $a_n = O_p(b_n)$ (and say big ohh p) provided

$$\frac{a_n}{b_n} = O_p(1),$$

where

$$c_n = O_p(1)$$

means

- For all δ
- There exists a C
- Such that for sufficiently large N
- For all $n \geq N$
- $\mathbb{P}(|c_n| \geq C) \leq \delta$

(This is also called bounded in probability, and is related to convergence in distribution)

Stochastic asymptotic notation

We write $a_n = o_p(b_n)$ (and say little ohh p) provided

$$\frac{a_n}{b_n} = o_p(1),$$

where

$$c_n = o_p(1)$$

means

- For all $\epsilon > 0, \delta > 0$
- There exists an N
- Such that for all $n \geq N$
- $\mathbb{P}(|c_n| \geq \epsilon) \leq \delta$

Stochastic asymptotic notation

Note that if we have random variables (X_n) and X, then

$$X_n \to X \text{ in probability} \Leftrightarrow X_n - X = o_p(1)$$

We can also express Slutsky's theorem$^\sharp$

- $o_p(1) + O_p(1) = ?$
- $o_p(1)O_p(1) = ?$
- $o_p(1) + o_p(1)O_p(1) = ?$

Persistence proof

Note that, $\sup_{\beta \in \{b : ||b||_1 \leq t\}} ||\beta||_1 \leq t$. Also,

Lemma

Suppose $a \in \mathbb{R}^p$ and $A \in \mathbb{R}^{p \times p}$. Then

$$a^\top A a \leq ||a||_1^2 \, ||A||_\infty \, ,$$

where $||A||_\infty := \max_{i,j} |A_{ij}|$ is the entry-wise max norm.

Proof.

$$a^\top A a \underbrace{\leq}_{\text{Hölder's}^\sharp} ||a||_1 \, ||Aa||_\infty \leq ||a||_1 \max_{ij} |A_{ij}| \, ||a||_1 = ||a||_1^2 \, ||A||_\infty \, ,$$

$\square$

These facts imply..

Persistence proof

$$
\begin{aligned}
\mathcal{E}(\hat{\beta}_t, \beta_t^*) &= \underbrace{R(\hat{\beta}_t)}_{\hat{\gamma}_t^\top \Sigma \hat{\gamma}_t} - \underbrace{R(\beta_t^*)}_{(\gamma_t^*)^\top \Sigma (\gamma_t^*)} && (31)\\
&= \hat{\gamma}_t^\top \Sigma \hat{\gamma}_t - \hat{\gamma}_t^\top \hat{\Sigma} \hat{\gamma}_t + \hat{\gamma}_t^\top \hat{\Sigma} \hat{\gamma}_t - (\gamma_t^*)^\top \Sigma (\gamma_t^*) && (32)\\
&\leq \hat{\gamma}_t^\top \Sigma \hat{\gamma}_t - \hat{\gamma}_t^\top \hat{\Sigma} \hat{\gamma}_t + (\gamma_t^*)^\top \hat{\Sigma} \gamma_t^* - (\gamma_t^*)^\top \Sigma \gamma_t^* && (33)\\
&= \hat{\gamma}_t^\top (\Sigma - \hat{\Sigma}) \hat{\gamma}_t + (\gamma_t^*)^\top (\hat{\Sigma} - \Sigma)(\gamma_t^*) && (34)\\
&\leq 2 \sup_{\beta \in \{b : ||b||_1 \leq t\}} \gamma_t^\top (\Sigma - \hat{\Sigma}) \gamma_t \quad \text{(2}\epsilon\text{ trick)} && (35)\\
&\leq 2 \sup_{\beta \in \{b : ||b||_1 \leq t\}} ||\gamma_t||_1^2 \left|\left|\Sigma - \hat{\Sigma}\right|\right|_\infty \quad \textit{(Lemma)} && (36)\\
&\leq 2(t+1)^2 \left|\left|\Sigma - \hat{\Sigma}\right|\right|_\infty && (37)
\end{aligned}
$$

Can we control the sup-norm part?

Persistence proof

Nemirovski's inequality: Let $\xi_i \in \mathbb{R}^p$, $i = 1, \ldots, n$ be independent, zero mean, finite variance random variables with $p \geq 3$. Define $S_n = \sum_{i=1}^n \xi_i$. Then for every $q \in [2, \infty]$

$$\mathbb{E}||S_n||_q^2 \leq e(2\log(p) - 1)\min\{q, \log(p)\} \sum_{i=1}^n \mathbb{E}||\xi_i||_q^2$$

(Juditsky, Nemirovski (2000), Dümbgen, et al. (2010))

This should be compared with the naïve bound:

$$\mathbb{E}||S_n||_q^2 \leq \sum_{i=1}^n \sum_{i'=1}^n \mathbb{E}||\xi_i||_q ||\xi_{i'}||_q$$

Persistence proof: Nemirovski's inequality

Motivation: Under Nemirovski's assumptions,

- Suppose $p = 1$, then $\mathbb{E}S_n^2 = \sum_{i=1}^n \mathbb{E}\xi_i^2$
- In a Hilbert space with inner product $\langle \cdot, \cdot \rangle$

$$\mathbb{E}||S_n||^2 = \sum_{i,i'}^n \mathbb{E}\langle \xi_i, \xi_{i'} \rangle = \sum_{i=1}^n \mathbb{E}||\xi_i||^2$$

- What about a Banach space (e.g. $|| \cdot ||_q, q \neq 2$)?

Persistence proof

Nemirovski's inequality: Let $\xi_i \in \mathbb{R}^p$, $i = 1, \ldots, n$ be independent, zero mean, finite variance random variables with $p \geq 3$. Define $S_n = \sum_{i=1}^n \xi_i$. Then for every $q \in [2, \infty]$

$$\mathbb{E}||S_n||_q^2 \leq e(2\log(p) - 1)\min\{q, \log(p)\} \sum_{i=1}^n \mathbb{E}||\xi_i||_q^2$$

Let $\xi_i = \mathrm{vec}\left(\frac{1}{n}\left(Z_{ij}Z_{ik} - \mathbb{E}Z_jZ_k\right)\right) \in \mathbb{R}^{(p+1)^2}$ be the vectorized difference of empirical covariance and the true covariance

Then

$$\left|\left|\Sigma - \hat{\Sigma}\right|\right|_\infty = \left|\left|\sum_{i=1}^n \xi_i\right|\right|_\infty$$

Persistence proof

$$\mathbb{E}||S_n||_q^2 \leq e(2\log(p)-1)\min\{q,\log(p)\}\sum_{i=1}^{n}\mathbb{E}||\xi_i||_q^2$$

(Nemirovski's inequality)

$$\begin{aligned}
\left(\mathbb{E}\left|\left|\Sigma-\hat{\Sigma}\right|\right|_\infty\right)^2 &\leq \mathbb{E}\left|\left|\Sigma-\hat{\Sigma}\right|\right|_\infty^2 && \text{(Jensen's inequality)}\\
&= \mathbb{E}\left|\left|\sum_{i=1}^{n}\xi_i\right|\right|_\infty^2\\
&\leq C\log((p+1)^2)\sum_{i=1}^{n}\mathbb{E}||\xi_i||_\infty^2\\
&\leq 4CC_{\mathcal{P}}^2\log(p+1)\frac{1}{n} && (\mathbb{P}\in\mathcal{P})\\
&\lesssim \frac{\log(p)}{n}
\end{aligned}$$

Persistence proof: Conclusion

$$\mathbb{P}\left(\mathcal{E}(\hat{\beta}_t, \beta_t^*) > \delta\right) \leq \mathbb{E}[\mathcal{E}(\hat{\beta}_t, \beta_t^*)]\delta^{-1} \quad \text{(Markov's inequality)} \tag{38}$$

$$\leq 2\delta^{-1}(t+1)^2 \mathbb{E}\left\|\Sigma - \hat{\Sigma}\right\|_\infty \tag{39}$$

$$\lesssim 2\delta^{-1}(t+1)^2 \sqrt{\frac{\log p}{n}} \tag{40}$$

Therefore, we have persistence provided $\log p = o(n)$ and

$$t_n = o\left(\left(\frac{n}{\log p}\right)^{1/4}\right)$$

Alternatively

$$\mathcal{E}(\hat{\beta}_t, \beta_t^*) = O_p\left(t^2\sqrt{\frac{\log(p)}{n}}\right)$$

Probably approximately correct (PAC)

Probability bound $\Leftrightarrow$ high probability upper bound:

$$\mathbb{P}(\text{error} > \delta) \leq \epsilon$$

gets converted to: with probability $1 - \epsilon$

$$\text{error} \leq \delta$$

(This is known as a PAC bound)

Example:

$$\mathbb{P}(|\overline{X} - \mu| > \delta) \leq \frac{\mathbb{E}(\overline{X} - \mu)^2}{\delta^2}$$

Hence, with probability at least $1 - \frac{\sigma^2}{n\delta^2}$

$$|\overline{X} - \mu| \leq \delta$$

Low assumption lasso: Summary

It is important to note that we do not assume...

- a linear model
- an additive stochastic component (let alone, Gaussian errors)
- that the design is 'almost' uncorrelated

and we get that, with probability at least $1 - C\delta^{-1}t^2\sqrt{\frac{\log(p)}{n}}$,

$$R(\hat{\beta}_t) \leq R(\beta_t^*) + \delta$$

Not low assumption lasso

Compare this to a classic result about the lasso that says more, but under much stronger assumptions

Assume that $Y = \mathbb{X}\beta^* + \epsilon$

Then we have the basic inequality[♯] for the lasso ($\hat{\beta} \equiv \hat{\beta}_\lambda$)

$$\left|\left|\mathbb{X}(\hat{\beta} - \beta^*)\right|\right|_2^2 / n + \lambda \left|\left|\hat{\beta}\right|\right|_1 \leq 2\epsilon^\top \mathbb{X}(\hat{\beta} - \beta^*)/n + \lambda \left|\left|\beta^*\right|\right|_1$$

- The $2\epsilon^\top \mathbb{X}(\hat{\beta} - \beta^*)/n$ term is the empirical process
- The $\lambda \left|\left|\beta^*\right|\right|_1$ is the deterministic part

Goal: choose λ so that deterministic $>>$ empirical process

Not low assumption lasso

Observe that the empirical process can be bounded

$$2\epsilon^\top \mathbb{X}(\hat{\beta} - \beta^*)/n \leq \left(\max_{1 \leq j \leq p} 2|\epsilon^\top x_j|/n \right) \left\| \hat{\beta} - \beta^* \right\|_1$$

Set

$$\mathcal{T} = \left\{ \max_{1 \leq j \leq p} 2|\epsilon^\top x_j|/n \leq \lambda_0 \right\}$$

Assume $\mathbb{X}$ is standardized: $\hat{\Sigma} = \mathbb{X}^\top \mathbb{X}/n$ has 1's on diagonal.

Theorem

If $\lambda_0 = 2\sigma\sqrt{(t^2 + 2\log p)/n}$, *and* $\epsilon \sim N(0, \sigma^2 I)$, *then*

$$\mathbb{P}(\mathcal{T}) \geq 1 - 2e^{-t^2/2}$$

Hence, $\mathcal{T}$ *is 'large'*

Proof: [Exercise]

Not low assumption lasso

Let

$$\beta_{j,S} = \begin{cases} \beta_j & \text{if } j \in S \\ 0 & \text{if } j \notin S \end{cases}$$

(Hence, $\beta = \beta_S + \beta_{S^c}$)

Theorem

On $\mathcal{T}$, with $\lambda \geq 2\lambda_0$ and $S^* = \{j : \beta_j^* \neq 0\}$

$$2\left|\left|\mathbb{X}(\hat{\beta} - \beta^*)\right|\right|_2^2 / n + \lambda \left|\left|\hat{\beta}_{S_*^c}\right|\right|_1 \leq 3\lambda \left|\left|\hat{\beta}_{S_*} - \beta_{S_*}^*\right|\right|_1$$

Proof sketch: Use the basic inequality along with the triangle inequality

$$\left|\left|\hat{\beta}\right|\right|_1 \geq ||\beta_{S_*}||_1 - \left|\left|\hat{\beta}_{S_*} - \beta_{S_*}\right|\right|_1 + \left|\left|\hat{\beta}_{S_*^c}\right|\right|_1$$

[Exercise]

Not low assumption lasso

Here is where structural assumptions come in

We need to get the term and the term 'together'

$$2\left|\left|\mathbb{X}(\hat{\beta} - \beta^*)\right|\right|_2^2/n + \lambda\left|\left|\hat{\beta}_{S_*^c}\right|\right|_1 \leq 3\lambda\left|\left|\hat{\beta}_{S_*} - \beta^*_{S_*}\right|\right|_1$$

This occurs in two steps ($s_* = |S_*|$):

1. By Cauchy-Schwarz: $\left|\left|\hat{\beta}_{S_*} - \beta^*_{S_*}\right|\right|_1 \leq \sqrt{s_*}\left|\left|\hat{\beta}_{S_*} - \beta^*_{S_*}\right|\right|_2$
2. Next convert $||b||_2$ into Mahalanobis distance $||\mathbb{X}b||_2$

Not low assumption lasso

To accomplish step 2., if $d_{\min} > 0$, then

$$\left|\left|\mathbb{X}(\hat{\beta} - \beta)\right|\right|_2^2 \geq d_{\min}^2 \left|\left|\hat{\beta} - \beta\right|\right|_2^2$$

and we can continue the chain of inequalities

However...

$d_{\min} > 0 \Leftrightarrow \mathbb{X}$ is full rank!

This is too strong as it gives a guarantee for all β

Not low assumption lasso

Observe that on $\mathcal{T}$:

$$\left|\left|\hat{\beta}_{S_*^c}\right|\right|_1 \leq 3 \left|\left|\hat{\beta}_{S_*} - \beta^*_{S_*}\right|\right|_1$$

(This follows from previous theorem)

We can restrict ourselves to only those β that satisfy this constraint

This gives us the compatibility condition for a set S and constant $\phi > 0$:

$$\forall \beta \text{ such that } ||\beta_{S^c}||_1 \leq 3\, ||\beta_S||_1 \Rightarrow ||\beta_S||_1^2 \leq \left(\beta^\top \hat{\Sigma} \beta\right) |S| / \phi^2$$

($||\beta_S||_1^2 \leq |S|\, ||\beta_S||_2^2 = \left(\beta^\top \hat{\Sigma}\hat{\Sigma}^{-1}\beta\right)|S| \leq \left(\beta^\top \hat{\Sigma}\beta\right) \frac{n|S|}{d^2_{\min}}$ provided $\hat{\Sigma}$ is invertible)

Not low assumption lasso

Related notions are:

- Restricted eigenvalue: Check the compatibility inequality $||\beta_{S^c}||_1 \leq 3\,||\beta_S||_1$ for all sets S of a given cardinality.
- Restricted isometry: An isometry U doesn't deform angles: $||Ux||_2 = ||x||_2$. A restricted isometry doesn't deform the angles too much over relevant parts of the space: that is, $\exists \delta > 0$ such that for all interesting β

$$(1-\delta)\,||\beta||_2^2 \leq ||\mathbb{X}\beta||_2^2 \leq (1+\delta)\,||\beta||_2^2$$

(See Larry Wasserman's blog post for an interesting discussion on this topic:

`http://normaldeviate.wordpress.com/2012/08/07/rip-rip-restricted-isometry-property-rest-in-peace`)

These are known as structural assumptions

Not low assumption lasso

Suppose that the compatibility condition holds for S_* with constant ϕ_*. Then on $\mathcal{T}$ and for $\lambda \geq 2\lambda_0$

$$n^{-1} \left|\left|\mathbb{X}(\hat{\beta} - \beta_*)\right|\right|_2^2 + \lambda \left|\left|\hat{\beta} - \beta_*\right|\right|_1 \leq 4\lambda^2 \frac{|S_*|}{\phi_*}$$

[Exercise] Use the compatibility condition on the previous theorem and use the useful inequality $4ab \leq a^2 + 4b^2$

This of course implies that:

- $\left|\left|\mathbb{X}(\hat{\beta} - \beta_*)\right|\right|_2^2 \leq 4n\lambda^2 |S_*|/\phi_*$
- $\left|\left|\hat{\beta} - \beta_*\right|\right|_1 \leq 4\lambda |S_*|/\phi_*$

(Write this as a PAC bound$^\sharp$)