Statistics 611
Lecture Notes

Lecturer: Harrison Zhou
Scribe: Addison J. Hu

Contents

CHAPTER 1

Introduction

This course is on selected topics in statistical decision theory. It is meant to prepare students to make serious theoretical contributions in their PhD thesis.

1. Overview

1.1. Nonparametric Regression: Gaussian Sequence Model.

1.1.1. *Frequentist Viewpoint.* Suppose you observe a sequence of univariate parameters perturbed by noise: $Y_i = \theta_i + \sigma Z_i, i \in [n]$. Customarily, but not always, we assume $Z_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$.

$$\Theta = \{\theta : \sum_{i=1}^{\infty} i^{2\alpha}\theta_i^2 \leq M\}$$

We assume that underlying the observations is a vector θ of true parameters, representing the sequence; this parameter vector is drawn from a model family represented by Θ.

Example 1.1. This model may seem arbitrarily, but a good example is in signal processing, in which case θ corresponds to wavelet or Fourier coefficient in denoising.

More generally, we have

$$\Theta = \{\theta : \sum_{i=1}^{\infty} \omega_i |\theta_i|^p \leq M\}$$

Example 1.2. Consider the $p = 0$ (ℓ_0 norm), and take $\omega_i = 1$. Then we have:

$$\Theta = \{\theta : \|\theta\|_0 \leq M\}$$

The goal is to estimate θ. We can, for example, use the ℓ_2 loss.

Definition.

$$\|\theta - \delta\|_2^2 = \sum_i (\theta_i - \delta_i)^2$$

But because the estimator is the function of data, we look at the expected loss (risk) instead:

Definition.

$$R(\delta) = \mathbf{E}_{Y|\theta} \left\|\delta - \theta\right\|^2$$

In minimax estimation, we look at the worst-case risk.

Definition. Minimax risk.

$$\inf_{\delta} \sup_{\theta \in \Theta} R(\delta)$$

Remark. Sometimes, our true θ may exist in a smaller parameter space, but if we consider a parameter space too big, then our worst case risk will be bigger too.

Definition. Adaptive minimaxity. Consider a sequence of parameter spaces:

$$\cdots \subset \Theta_1 \subset \Theta_2 \subset \Theta_3 \subset \cdots$$

Adaptive minimaxity gives an estimator that adapts to α and M.

Definition. Oracle inequality. This addresses the problem when we assume $\theta \in \Theta$ for minimax risk, but it actually is not (mis-specified model).

1.1.2. *Bayesian Viewpoint.* We assume a prior $\theta \sim \pi$ on the parameter, and try to calculate a posterior $\pi(\theta|Y)$, or sample from it.

Because Harry is a frequentist, he evaluates Bayesian methods using frequentist tools.

Definition. Posterior contraction. Assume θ_0 is the truth, and assume that the θ are truly concentrated about the truth. Consider a ball about the true θ:

$$\mathbf{E}_{Y|\theta_0} \pi(\|\theta - \theta_0\| \leq \gamma_\pi |Y) \to 1$$

Remark. We may also consider estimating $\sum \theta_i^2$ quadratic functionals, or linear functionals $\sum \alpha_i \theta_i$. There are still a lot of papers on this topic, especially for sparse models (ℓ_0 norm).

2. High Dimensional Linear Regression

Suppose we observe a vector

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \sigma Z_{n\times 1}$$

with customarily $Z_{n\times 1} \sim \mathcal{N}(0, \mathbf{I}_n)$.

If we assume $p = n$, then we have $XX^T = \mathbf{I}$. We observe that this is a simple extension of Gaussian sequence model (when the columns of X are orthogonal, it is identity subject to a rotation).

Remark. How to estimate when $p \gg n$? How to estimate θ under ℓ_2, ℓ_1, ℓ_0 loss?

We need to take some assumptions, such as ℓ_0 assumption. (Example 1.2) Then, we assume that M is unknown. If M is small, but we don't know it, what do we do? Adaptive model.

This can be evaluated from frequentist and Bayesian points of view. For Bayesian, we may use posterior contraction. For frequentist, we may find rate of convergence for minimax.

Remark. We can construct confidence intervals for both Gaussian sequence model (trivial) and for linear regression (harder when the columns are not orthogonal).

3. Matrix Estimation

Suppose $Y_{n\times m} = \Theta_{n\times m} + \sigma Z_{n\times m}$. We may consider the parameter space:

$$\Theta = \{\theta_{n\times m} : \text{rank}(\Theta_{n\times m}) \leq r\}$$

Here, we do not observe Y (e.g., Netflix prize). We observe $B_{ij}Y_{ij}$, where $B_{ij} \overset{\text{iid}}{\sim} Bern(p)$.

Remark. Trace regression. In trace regression, we consider

$$Y = X\Theta + \sigma Z$$

This is an example of high-dimensional linear regression.

4. Covariance Matrix Estimation

Consider $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu_{p\times 1}, \Sigma_{p\times p})$. We want to estimate the covariance matrix. The sample covariance is

$$\hat{\Sigma} = \frac{1}{n-1}\sum (Y_i - \bar{Y})(Y_i - \bar{Y})^T$$

This estimator is also unbiased.

Remark. Suppose we want to invert $\hat{\Sigma}$. The rank is at most $n-1$. So, we cannot invert when $p \gg n$. Sometimes, we can exploit structure in the sample covariance matrix to do something.

Possible structures include:

(1) Bandable: Along diagonal bands, they are not equal to zero, but outside the bands, they are equal to zero. This may result from time series structure. Sometimes, we desire this in the inverse covariance matrix.
(2) Sparse: Gaussian graphical models (linear regression), Ising model (logistic regression)

Remark. Principal Components Estimation: If we know some structure about the principal components, we can estimate them directly without estimating the covariance matrix. This is connected to factor models. We will also discuss computational barriers.

Remark. PCA may be viewed as a functional estimation problem for matrices. Another functional estimation problem is linear discriminant analysis.

5. Network Analysis

Graphon analysis, community detection.

6. Variational Bayes

Remark. For some statistical algorithms, we have good theoretical results, but they are intractable. How can we approximate it? Variational Bayes.

Remark. There is little theory for EM algorithm in high-dimensional statistics.

CHAPTER 2

Nonparametric Estimation

Consider the model: $Y_i = \theta_i + \sigma Z_i, Z_i \in [n]$. The parameter space is a ball with smoothness α (of f):

$$\Theta = \{\theta : \sum_{i=1}^{\infty} i^{2\alpha}\theta_i^2 \leq M\}$$

Example 0.1. Suppose we observe $Y_i = f(\frac{1}{n}) + Z_i, i \in [n]$. We may imagine this as observing a noisy version of a true signal. Oftentimes, we will take a discrete Fourier transformation:

$$A_{n\times n}\begin{bmatrix} Y_i \\ \vdots \\ Y_n \end{bmatrix} = \frac{1}{\sqrt{n}} A \begin{bmatrix} f(\frac{1}{n}) \\ \vdots \\ f(1) \end{bmatrix} + \frac{1}{\sqrt{n}} AZ$$

where $AA^T = \mathbf{I}_n$.

Remark. We may expand the function $f(x)$ into the Fourier basis $\sum_{i=1}^{\infty}\langle f, \varphi_i\rangle\phi_i(x) = \sum_{i=1}^{\infty}\theta_i\varphi_i(x)$, where φ is a sine or cosine, e.g., $\sin(cix)$. Therefore, the second derivative is given by $f''(x) = \sum_{i=1}^{\infty}\theta_i(ci)^2\varphi_i(x)$ when $\alpha = 2$.

The Sobolev ball assumes $\int (f''(x))^2 dx \leq M$, which implies that $\sum \left[\theta_i(ci)^2\right]^2 \leq M.$, where $\langle a, b\rangle = \int_0^1 a(x)b(x)dx$.

Remark. The Besov ball corresponds to wavelets, Sobolev ball corresponds to Fourier transformation.

1. Density Estimation Problem

Suppose we observe $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f$. How should we estimate f? Consider the parameter space

$$\Theta = \{f : \int (f''(x))^2 \leq M\}$$

If we expand

$$f(x) = \sum_{i=1}^{\infty}\langle f, \varphi_i\rangle\varphi_i$$

$$\theta_i = \int_0^1 f(x)\varphi_i(x)dx$$

We can construct the following unbiased estimator:

$$Y_i' = \frac{1}{n}\sum_{j=1}^{n}\varphi_i(Y_j)$$

such that, by central limit theory,

$$Y_i' = \theta_i + \frac{1}{n}\sigma_i Z_i, \qquad Z_i \approx \mathcal{N}(0,1)$$

An intuitive estimator for θ could be:

$$\hat{\theta}_i = \begin{cases} Y_i' & \text{for } i \leq I \\ 0 & \text{for } i > I \end{cases}$$

that is, we threshold the Fourier coefficients to zero beyond a certain point.

1.1. Model. Our goal is to estimate θ in the model $Y_i = \theta_i + \sigma Z_i$, with $\sigma = \frac{1}{\sqrt{n}}$. We take the squared loss. The parameter space is the Sobolev ball.

We may consider the estimator:

$$\hat{\theta}_i = \begin{cases} Y_i' & \text{for } i \leq I \\ 0 & \text{for } i > I \end{cases}$$

Observe that the risk when using Y_i is $\frac{1}{n}$, and the risk of zero is θ_i^2. So, we want to pick which one to use intelligently.

1.2. Risk Upper Bound. To evaluate our estimation technique, we may examine its risk. Can we place an upper bound on the risk (i.e., understand how poorly it might perform under a given parameter setting)?

$$\begin{aligned} \mathbf{E}\sum^{\infty}\left\|\hat{\theta} - \theta\right\|^2 &= \mathbf{E}\sum_{i=1}^{I}(Y_i - \theta_i)^2 + \sum_{i=I+1}^{\infty}\theta_i^2 \\ &\leq \frac{I}{n} + \frac{M}{I^{2\alpha}} \end{aligned}$$

This follows from:

$$\begin{aligned} \sum_{i\geq I+1} i^{2\alpha}\theta_i^2 &\leq M \\ \Rightarrow (I+1)^{2\alpha}\sum_{i\geq I+1}\theta_i^2 &\leq M \\ \Rightarrow \sum_{i\geq I+1}\theta_i^2 &\leq \frac{M}{(I+1)^{2\alpha}} \end{aligned}$$

We want to get the tightest upper bound, so we can try to pick I to minimize the upper bound. We need to pick an I that balances variance and bias.

We can choose $\frac{I}{n} = \frac{M}{I^{2\alpha}} \Rightarrow I^{1+2\alpha} = Mn$. Therefore:

$$\begin{aligned}\mathbf{E}\sum^{\infty} \left\|\hat{\theta} - \theta\right\|^2 &= \underbrace{\frac{I}{n}}_{\text{Variance}} + \underbrace{\frac{M}{I^{2\alpha}}}_{\text{Squared Bias}} \\ &= 2M^{\frac{1}{1+2\alpha}} n^{-\frac{2\alpha}{1+2\alpha}}\end{aligned}$$

It follows, for all θ, that:

$$\sup_{\Theta} \mathbf{E} \left\|\hat{\theta} - \theta\right\|^2 \leq 2M^{\frac{1}{1+2\alpha}} n^{-\frac{2\alpha}{1+2\alpha}}$$

Note that this is slower than we typically see for parametric models, where the rate is on the order of n^{-1}.

Having derived this upper bound, we should ask: is there a procedure whose rate of decay for the risk is lower? If so, we should use that procedure instead.

1.3. Risk Lower Bound (Le Cam's Method). It turns out that there does not exist an estimator with a faster rate of decay. We will show this by finding the best rate possible, a rate which matches the rate of decay for our own estimator. In fact, we will show that the fastest possible decay rate for the risk under this model is:

$$\inf_{\hat{\theta}} \sup_{\Theta} \mathbf{E} \left\|\hat{\theta} - \theta\right\|^2 \geq cn^{-\frac{2\alpha}{2\alpha+1}}; \qquad c > 0$$

Last semester, we proved this by taking a sub-parameter space $\Theta_0 \subset \Theta$. For any estimator,

$$\sup_{\Theta} \mathbf{E} \left\|\hat{\theta} - \theta\right\|^2 \geq \sup_{\Theta_0} \mathbf{E} \left\|\hat{\theta} - \theta\right\|^2$$

We want to use a sub-parameter space that admits a good lower bound. Suppose we choose

$$\Theta_0 = \left\{\theta : \begin{cases} \frac{1}{\sqrt{n}} \text{ or } 0 & \text{for } i \leq I \\ 0 & \text{for } i \geq I+1 \end{cases}\right\}$$

where $I = (Mn)^{\frac{1}{1+2\alpha}}$. We must confirm that $\Theta_0 \subset \Theta$:

$$\sum_{i=1}^{I} i^{2\alpha} \frac{1}{n} \leq \frac{1}{n} I I^{2\alpha} = M$$

Now, we want to show:

$$\sup_{\Theta_0} \mathbf{E} \left\|\hat{\theta} - \theta\right\|^2 \geq cn^{-\frac{2\alpha}{2\alpha+1}}; \qquad c > 0$$

Observe that:

$$\sup_{\Theta_0} \mathbf{E} \left\|\hat{\theta} - \theta\right\|^2 = \sup_{\Theta_0} \mathbf{E}_{X|\theta} \sum_{i=1}^{I} (\hat{\theta}_i - \theta_i)^2$$

We place a prior on θ_i; we have it equal $0, \frac{1}{\sqrt{n}}$ with equal probability. Recall that sup is bounded below by the average.

$$\begin{aligned}
\sup_{\Theta_0} \mathbf{E} \left\| \hat{\theta} - \theta \right\|^2 &\geq \sup_{\Theta_0} \mathbf{E}_{Y|\theta} \sum_{i=1}^{I} (\hat{\theta}_i - \theta_i)^2 \\
&\geq \mathbf{E}_\theta \, \mathbf{E}_{Y|\theta} \sum_{i=1}^{I} (\hat{\theta}_i - \theta_i)^2 \\
&\geq \mathbf{E}_\theta \, \mathbf{E}_{Y|\theta} \sum_{i=1}^{I} \left(\hat{\theta}_{i,Bayes} - \theta_i \right)^2
\end{aligned}$$

If the Bayes estimator is easy to derive, then we just show that this risk is lower bounded by $\frac{c}{\sqrt{n}}$. Alternatively, we can do the following, where $\varphi_{\alpha,\beta}$ is the normal density with mean α and variance β:

$$\begin{aligned}
\sup_{\Theta_0} \mathbf{E} \left\| \hat{\theta} - \theta \right\|^2 &\geq \mathbf{E}_\theta \, \mathbf{E}_{Y|\theta} \sum_{i=1}^{I} \left(\hat{\theta}_i - \theta_i \right)^2 \\
&\text{(Independence.)} \\
&= \sum_{i=1}^{I} \left[\frac{1}{2} \mathbf{E}_{Y_i|\theta_i=0} \left(\hat{\theta}_{i,Bayes} - 0 \right)^2 + \frac{1}{2} \mathbf{E}_{Y_i|\theta_i=\frac{1}{\sqrt{n}}} \left(\hat{\theta}_{i,Bayes} - \frac{1}{\sqrt{n}} \right)^2 \right] \\
&= \sum_{i=1}^{I} \frac{1}{2} \left[\int (\hat{\theta}_{i,Bayes})^2 \varphi_{0,\frac{1}{n}}(x) dx + \int (\hat{\theta}_{i,Bayes} - \frac{1}{\sqrt{n}})^2 \varphi_{\frac{1}{\sqrt{n}},\frac{1}{n}}(x) dx \right] \\
&\geq \sum_{i=1}^{I} \frac{1}{2} \int \left[(\hat{\theta}_{i,Bayes})^2 dx + (\hat{\theta}_{i,Bayes} - \frac{1}{\sqrt{n}})^2 \right] \min\{\varphi_{0,\frac{1}{n}}(x), \varphi_{\frac{1}{\sqrt{n}},\frac{1}{n}}(x)\} \\
&\left[\text{By } a^2 + b^2 \geq \frac{1}{2}(a-b)^2 \right] \\
&\geq \sum_{i=1}^{I} \frac{1}{2} \int \left[\frac{1}{2} \left(\frac{1}{\sqrt{n}} \right)^2 \right] \min\{\varphi_{0,\frac{1}{n}}(x), \varphi_{\frac{1}{\sqrt{n}},\frac{1}{n}}(x)\} dx \\
&\geq \frac{I}{n} \frac{1}{4} \int \min\{\varphi_{0,\frac{1}{n}}(x), \varphi_{\frac{1}{\sqrt{n}},\frac{1}{n}}(x)\} dx \\
&\geq c n^{-\frac{2\alpha}{2\alpha+1}} \\
&\geq c_1 M^{\frac{1}{1+2\alpha}} n^{-\frac{2\alpha}{2\alpha+1}}
\end{aligned}$$

Note that $a_n \asymp b_n$ implies $\exists c_1, c_2$ such that $c_1 \leq \frac{a_n}{b_n} \leq c_2$. Therefore, we have shown the minimax rate

$$\inf_{\hat{\theta}} \sup_{\Theta} \mathbf{E} \left\| \hat{\theta} - \theta \right\|^2 \asymp M^{\frac{1}{1+2\alpha}} n^{-\frac{2\alpha}{2\alpha+1}}$$

1.4. Questions. What if M is unknown? It must be known, otherwise this is not a procedure. We also must know α. To address these issues, we consider adaptive estimation; it is driven by data.

We could also replace the hard thresholding in $0, Y_i$ with a linear procedure $\hat{\theta}_i = c_i Y_i$, with c_i to be determined (to minimize the risk). Or just the best procedure in general.

What about general distributions other than i.i.d. Gaussian?

2. Best Linear Procedure

We want to find $c = (c_1, c_2, \cdots)$ to minimize

$$\sup_{\Theta} \mathbf{E} \sum_{i=1}^{\infty} (c_i Y_i - \theta_i)^2$$

Claim 1.

$$\inf_{c} \sup_{\Theta} \mathbf{E} \sum_{i=1}^{\infty} (c_i Y_i - \theta_i)^2 = (1 + o(1)) P_\alpha M^{\frac{1}{1+2\alpha}} n^{-\frac{2\alpha}{1+2\alpha}}$$

where P_α is the Pinsker constant.

We also know that for all procedures,

$$\inf_{\hat{\theta}} \sup_{\Theta} \mathbf{E} \sum_{i=1}^{\infty} (c_i Y_i - \theta_i)^2 \geq (1 + o(1)) P_\alpha M^{\frac{1}{1+2\alpha}} n^{-\frac{2\alpha}{1+2\alpha}}$$

The lower bound is harder to show; typically one constructs a parameter space and a prior, and use the fact that the lowest risk is bounded above by the average risk. But to get the optimal c, we need to know α and M.

Therefore we may conclude that linear procedure is best.

3. Adaptive Estimation and Pinsker Bound

What we want to do is mimic the best linear procedure. Recall that previously we studied a linear procedure: James-Stein Estimation, which is similar to Ridge Regression in that both procedures involve shrinkage.

3.1. James-Stein Estimation. Consider the Gaussian sequence model: $X_1 = \theta_1 + \sigma Z_1, \ldots, X_m = \theta_m + \sigma Z_m$, with $Z \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. Recall that we found the procedure:

$$\hat{\theta} = \left(1 - \frac{(m-2)\sigma^2}{\sum X_i^2}\right) X$$

and showed that:

$$\mathbf{E} \left\| \hat{\theta}_{\text{J-S}} - \theta \right\|^2 - \inf_{c} \mathbf{E} \left\| cX - \theta \right\|^2 \leq 2\sigma^2$$

where knowing the best c is not a valid procedure, since that requires knowledge of θ. Note that here c does not vary across i.

If we know θ, then the optimal c is given by $c = \frac{\|\theta\|^2}{\|\theta\|^2+m\sigma^2}$.

3.2. Derivation of the James-Stein Estimator. Recall that from geometric intuition, we try to shrink X as follows:

$$\frac{\langle\theta, X\rangle}{\|X\|^2}X \approx \frac{\|X\|^2 - m\sigma^2}{\|X\|^2}X$$

which is almost the form of the James-Stein estimator.

Claim 2.

$$\mathbf{E}\left\|\hat{\theta}_{J-S} - \theta\right\|^2 - \inf_c \mathbf{E}\,\|cX - \theta\|^2 \leq 2\sigma^2$$

Proof. Assume without loss of generality that $\sigma = 1$. Recall that $\|X\|^2$ is a mixture of χ^2_{m+2N} where $N \sim Pois\left(\frac{\|\theta\|^2}{2}\right)$ and that $\mathbf{E}\,\frac{1}{Y} \geq \frac{1}{\mathbf{E}\,Y}$ (Jensen's Inequality).

$$\begin{aligned}
\mathbf{E}\left\|\hat{\theta} - \theta\right\|^2 &= m - \mathbf{E}\,\frac{(m-2)^2}{\|X\|^2} \\
&= m - \mathbf{E}\left[\mathbf{E}\,\frac{(m-2)^2}{\chi^2_{m+2N}}|N\right] \\
&= m - \mathbf{E}\,\frac{(m-2)^2}{m+2N-2} \\
&\leq 2 + (m-2) - \frac{(m-2)^2}{(m-2)+\|\theta\|^2} \\
&= 2 + \frac{(m-2)\,\|\theta\|^2}{(m-2)+\|\theta\|^2} \\
&\leq 2 + \frac{m\,\|\theta\|^2}{m+\|\theta\|^2} \\
&= 2 + \inf_c \mathbf{E}\,\|cX - \theta\|^2
\end{aligned}$$

□

Remark. James-Stein Estimator is adaptive to ℓ_2 norm of θ.

4. Blockwise James-Stein

Consider the model $Y_i = \theta_i + \frac{1}{\sqrt{n}}Z_i$, with the the Sobolev ball as the parameter space. Our goal is to find $\hat{\theta}$ such that

$$\mathbf{E}\left\|\hat{\theta} - \theta\right\|^2 \lesssim M^{\frac{1}{1+2\alpha}} n^{-\frac{2\alpha}{1+2\alpha}}$$

without knowing M, α. James-Stein estimator assumes all θ_i are on about the same order.

Remark. Because we have n observations, the $\frac{1}{\sqrt{n}}$ is a normalization constant. It appears in linear regression as well.

To overcome the assumption that all θ_i are on the same order, we may apply shrinkage in a blockwise manner. We divide our observations into many blocks, for example $Y_1, \dots, Y_4$, $Y_5, \dots, Y_8$, $Y_9, \dots, Y_{16}$. We estimate $\hat{\theta}_i = 0$ for $i \geq n$. We have $\log_2 n$ blocks, and apply James-Stein estimation to each block.

Remark. We have $i \in [1, \infty)$ (recall that θ is in the Sobolev ball.

$$\begin{aligned}
\mathbf{E} \left\| \hat{\theta} - \theta \right\|^2 &= \mathbf{E} \sum_{i=1}^{n} (\hat{\theta}_i - \theta_i)^2 + \sum_{i=n+1}^{\infty} \theta_i^2 \\
&\leq \mathbf{E} \sum_{i=1}^{n} (\hat{\theta}_i - \theta_i)^2 + \frac{M}{n^{2\alpha}} \\
&= \mathbf{E} \sum_{i=1}^{n} (\hat{\theta}_i - \theta_i)^2 + o(n^{-\frac{2\alpha}{2\alpha+1}}
\end{aligned}$$

Consider the blocks $B_1, \dots, B_K$, $K \asymp \log_2 n$.

$$\begin{aligned}
\mathbf{E} \sum_{i=1}^{n} (\hat{\theta}_i - \theta_i)^2 &\leq \sum_{k=1}^{K} \frac{|B_k| \sigma^2 \left\| \theta_{B_k} \right\|^2}{|B_k| \sigma^2 + \left\| \theta_{B_k} \right\|^2} + 2\sigma^2 \\
&\leq \sum_{k=1}^{K} \frac{|B_k| n^{-1} \left\| \theta_{B_k} \right\|^2}{|B_k| n^{-1} + \left\| \theta_{B_k} \right\|^2} + 2\frac{1}{n}
\end{aligned}$$

Recall that last time we choose $\frac{I}{n} = \frac{M}{I^{2\alpha}}$. Now, we want to find K_0 such that the sum of block sizes is on the same order, i.e., $I \leq |B_1| + \cdots + |B_{K_0}| \leq 2I$. (In the proof, but not the procedure, we can assume knowledge of I.)

$$\begin{aligned}
\mathbf{E}\sum_{i=1}^{n}(\hat{\theta}_i-\theta_i)^2 &\leq \sum_{k=1}^{K}\frac{|B_k|\sigma^2\left\|\theta_{B_k}\right\|^2}{|B_k|\sigma^2+\left\|\theta_{B_k}\right\|^2}+2\sigma^2\\
&\leq \sum_{k=1}^{K}\frac{|B_k|n^{-1}\left\|\theta_{B_k}\right\|^2}{|B_k|n^{-1}+\left\|\theta_{B_k}\right\|^2}+2\frac{1}{n}\\
&\leq \sum_{k=1}^{K_0}|B_k|\frac{1}{n}+\sum_{k=K_0+1}K\left\|\theta_{B_k}\right\|^2+2\frac{1}{n}\log_2 n\\
&\leq \frac{1}{n}2I+\sum_{i>I}\theta_i^2\\
&= \frac{1}{n}2I+\frac{M}{I^{2\alpha}}\\
&\geq CM^{\frac{1}{1+2\alpha}}n^{-\frac{2\alpha}{2\alpha+1}}
\end{aligned}$$

For the last line, cf. Section 1.2.

Remark. We use K_0 in the proof, but we do not need to know it to use this procedure.

4.1. Alternative Blockwise Estimator. We may want more blocks. But we cannot have a fixed block size. Suppose we fix block size 4. Then we would have $\frac{n}{4}$ blocks, giving us risk $2\sigma^2\frac{n}{4}=2\frac{1}{n}\frac{n}{4}$. That risk is too big. How can we adapt the number of blocks to the sample size?

Consider:

$$|B_k|=\lfloor(1+a)^k\rfloor,\qquad a=\frac{1}{\log n}$$

so that the block size grows more slowly. Here, $K\asymp(\log n)^2$. Intuitively, this has better practical performance. It turns out that it has better theoretical performance as well. Let us consider the risk upper bound of the alternative blockwise estimator:

Claim 3.

$$\sup_{\Theta}\mathbf{E}\left\|\hat{\theta}_{ABE}-\theta\right\|^2\leq(1+o(1))P_\alpha m^{\frac{1}{1+2\alpha}}n^{-\frac{2\alpha}{2\alpha+1}}$$

so, by changing, the block size, we attain the best rate of convergence, and the best constant.

4.2. Pinsker Constant. To understand the Pinsker bound, we need to determine the best linear procedure. Here we have $c=(c_1,c_2,\cdots)$, and we consider:

$$\inf_{c}\sup_{\Theta}\mathbf{E}\sum_{i=1}^{\infty}(c_iY_i-\theta_i)^2$$

Claim 4. *Using the ABE procedure gives us the minimax rate up to* $(1+o(1))$.

We begin with bias-squared and variance decomposition.

$$\inf_c \sup_\Theta \mathbf{E} \sum_{i=1}^{\infty} (c_i Y_i - \theta_i)^2 = \inf_c \sup_\Theta \mathbf{E} \sum_{i=1}^{\infty} (c_i^2 \frac{1}{n} + (c_1 - 1)^2 \theta_i^2)$$

This optimization problem is convex in c and concave in $\beta_i = \theta_i^2$. Both sets are convex (c is vectors with elements between 0 and 1, θ is in the Sobolev ball). So, we can apply minimax (duality) theory and switch the order.

$$\begin{aligned} \inf_c \sup_\Theta \mathbf{E} \sum_{i=1}^{\infty} (c_i Y_i - \theta_i)^2 &= \inf_c \sup_\Theta \mathbf{E} \sum_{i=1}^{\infty} (c_i^2 \frac{1}{n} + (c_1 - 1)^2 \theta_i^2) \\ &= \sup_\Theta \inf_c \mathbf{E} \sum_{i=1}^{\infty} (c_i^2 \frac{1}{n} + (c_1 - 1)^2 \theta_i^2) \end{aligned}$$

We determine that $c_i = \frac{\theta_i^2}{\frac{1}{n} + \theta_i^2}$, giving us, in the $m = 1$ case (block size 1, which was our best linear procedure),

$$c_i^2 \frac{1}{n} + (c_i - 1)^2 \theta_i^2 = \frac{\frac{1}{n} \theta_i^2}{\frac{1}{n} + \theta_i^2}$$

$$\begin{aligned} \inf_c \sup_\Theta \mathbf{E} \sum_{i=1}^{\infty} (c_i Y_i - \theta_i)^2 &= \inf_c \sup_\Theta \mathbf{E} \sum_{i=1}^{\infty} (c_i^2 \frac{1}{n} + (c_1 - 1)^2 \theta_i^2) \\ &= \sup_\Theta \inf_c \mathbf{E} \sum_{i=1}^{\infty} (c_i^2 \frac{1}{n} + (c_1 - 1)^2 \theta_i^2) \\ &= \sup_{\Theta \in B_S} \sum_{i=1}^{\infty} \frac{\frac{1}{n} \theta_i^2}{\frac{1}{n} + \theta_i^2} \end{aligned}$$

Observe that our objective is now concave, and we can find the maximum. We now consider:

$$\begin{aligned} &\underset{\beta_i}{\text{maximize}} && \frac{\frac{1}{n} \beta_i}{\frac{1}{n} + \beta_i} - \lambda^{-2} \sum_{i=1}^{\infty} i^{2\alpha} \beta_i \\ &\text{subject to} && \beta_i > 0 \end{aligned}$$

We then get:

$$\begin{aligned}\frac{\frac{1}{n}\left(\frac{1}{n}+\beta_i\right)-\frac{1}{n}\beta_i}{\left(\frac{1}{n}+\beta_i\right)^2} &= \lambda^{-2}i^{2\alpha}\\ \frac{\frac{1}{n^2}}{\left(\frac{1}{n}+\beta_i\right)^2} &= \lambda^{-2}i^{2\alpha}\\ \frac{\frac{1}{n}}{\frac{1}{n}+\beta_i} &= \lambda^{-1}i^{\alpha}\\ \Rightarrow \beta_i &= \frac{1}{n}(\lambda i^{-\alpha}-1)\end{aligned}$$

This may only be used when $\lambda > i^\alpha$. Therefore, we take the positive part:

$$\beta_i = \left(\frac{1}{n}(\lambda i^{-\alpha}-1)\right)_+$$

Now we look for λ^*:

$$\lambda^* = \sum i^{2\alpha}\left[\frac{1}{n}\left(\frac{\lambda}{i^\alpha}-1\right)_+\right]^2 = M$$

and choose:

$$\beta_i^* = \left(\frac{1}{n}(\lambda^* i^{-\alpha}-1)\right)_+$$

and finally:

$$c_i^* = \frac{\beta_i^*}{\frac{1}{n}+\beta_i^2}$$

which is our best linear procedure.

Remark. If we can calculate λ^*, we know exactly what our risk is going to be. Therefore, the best linear risk is given by:

$$\sum \frac{\frac{1}{n}\beta_i^*}{\frac{1}{n}+\beta_i^*} = \frac{1}{n}\sum_{i=1}^{\infty} c_i^*$$

Remark. For large enough i, c_i^* will be zero. (Series will converge.) See the definition of c_i^*.

Exercise. For homework, try deriving λ^* and determining the form of the Pinsker constant.

CHAPTER 3

Frequentist Bayesian Investigation

We want to try to bridge the Frequentist and Bayesian worlds. Whether or not Bayesian techniques are valid, we may evaluate their performance from a Frequentist point of view.

1. Gaussian Sequence Model

Consider the:

(1) Model $Y_i = \theta_i + Z_i$, $Z \overset{\text{iid}}{\sim} \mathcal{N}(0, \frac{1}{n})$.
(2) Parameter space: $\{\theta : \sum_{i=1}^{\infty} i^{2\alpha}\theta_i^2 \leq M\}$

From the Bayesian point of view, we have a prior $\theta \sim \pi$, and the get the posterior $\pi(\theta|Y)$. At this point, a Bayesian would be satisfy and consider the job done. But as Frequentists, we want to evaluate this procedure, assuming there exists a true θ^*.

From the frequentist world, we have the result:

$$\mathbf{E}_{Y|\theta^*} \left\|\hat{\theta} - \theta\right\|^2 \lesssim n^{-\frac{2\alpha}{2\alpha+1}} = \epsilon_n^2$$

We assume that α, M are known. Our goal is to find a prior such that, for all θ^*:

$$\mathbf{E}_{Y|\theta^2}\, \pi(\|\theta - \theta^*\| < c\epsilon_n|Y) \to 1$$

That is, we want a prior such that the resultant posterior concentrates probability mass about the true parameter with high probability.

In fact, we will find that the LHS approaches 1 exponentially fast. That is, using this Bayesian method, we can get a very good estimator (from a frequentist point of view.)

We begin by crafting a prior, assuming α known:

$$\text{Prior} \quad \pi : \begin{cases} \theta_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1) & \text{for } 1 \leq i \leq n^{\frac{1}{1+2\alpha}} \\ \theta_i = 0 & \text{for } i > n^{\frac{1}{1+2\alpha}} \end{cases}$$

This is a very particular prior; it is easy to work with.

Remark. We can consider more general priors and get the same result (with more work; see Le Cam-Schwartz-Barron; Annals of Statistics, 1973; in this paper, Le Cam shows that testing and estimation is equally difficult, for this general class of priors.)

1.1. Conjugate Prior. We know that for $1 \leq i \leq n^{\frac{1}{1+2\alpha}}$, $Y_i \sim \mathcal{N}(\theta_i, \frac{1}{n})$. Then we have:

$$\theta_i | Y_i \sim \mathcal{N}\left(\frac{1}{1+\frac{1}{n}} Y_i, \frac{\frac{1}{n}}{1+\frac{1}{n}}\right)$$

Therefore, we have $\theta_i = \frac{1}{1+\frac{1}{n}} Y_i + \sqrt{\frac{\frac{1}{n}}{1+\frac{1}{n}}} w_i$

Then we have:

$$\begin{aligned}
\|\theta - \theta^*\|^2 &= \sum_{i=1}^{n^{\frac{1}{2\alpha+1}}} \left(\frac{1}{1+\frac{1}{n}} Y_i + \sqrt{\frac{\frac{1}{n}}{1+\frac{1}{n}}} w_i - \theta_i^*\right)^2 + \sum_{i > n^{\frac{1}{2\alpha+1}}} (\theta_i^*)^2 \\
&= \sum_{i=1}^{n^{\frac{1}{2\alpha+1}}} \left(\frac{n}{n+1}(\theta_i^* + Z_i) + \sqrt{\frac{\frac{1}{n}}{1+\frac{1}{n}}} w_i - \theta_i^*\right)^2 + \sum_{i > n^{\frac{1}{2\alpha+1}}} (\theta_i^*)^2 \\
&= \sum_{i=1}^{n^{\frac{1}{2\alpha+1}}} \left[\left(\frac{n}{n+1}\theta_i^* - \theta_i^*\right) + \left(\frac{n}{n+1} Z_i + \sqrt{\frac{\frac{1}{n}}{1+\frac{1}{n}}} w_i\right)\right]^2 + \sum_{i > n^{\frac{1}{2\alpha+1}}} (\theta_i^*)^2 \\
&= \sum_{i=1}^{n^{\frac{1}{2\alpha+1}}} \left(\frac{n}{n+1}\theta_i^* - \theta_i^*\right)^2 + 2\left(\frac{n}{n+1}\theta_i^* - \theta_i^*\right)\left(\frac{n}{n+1} Z_i + \sqrt{\frac{\frac{1}{n}}{1+\frac{1}{n}}} w_i\right) \\
&\quad + \left(\frac{n}{n+1} Z_i + \sqrt{\frac{\frac{1}{n}}{1+\frac{1}{n}}} w_i\right)^2 + \sum_{i > n^{\frac{1}{2\alpha+1}}} (\theta_i^*)^2 \\
&\leq \sum_{i=1}^{n^{\frac{1}{2\alpha+1}}} 2\left(\frac{n}{n+1}\theta_i^* - \theta_i^*\right)^2 + 2\left(\frac{n}{n+1} Z_i + \sqrt{\frac{\frac{1}{n}}{1+\frac{1}{n}}} w_i\right)^2 + \sum_{i > n^{\frac{1}{2\alpha+1}}} (\theta_i^*)^2
\end{aligned}$$

We want to show that every term here is bounded above by $c\epsilon_n^*$.

The fourth term is already bounded properly by $c_1 n^{-\frac{2\alpha}{2\alpha+1}}$, as known from a previous lecture. We know that the first term is bounded above by c_2/n. The third term is actually a χ^2 with n (?) degrees of freedom; and the second term is bounded by the sum of the first and third. Then we can show that the sum is bounded by above by a constant times ϵ with high probability.

For the χ^2 term, we know that:

$$\mathbf{P}\{\chi_I^2 \geq 2I\} \leq \exp\{-c'I\}$$

This means that:

$$\begin{aligned}\sum_{i=1}^{n^{\frac{1}{2\alpha+1}}}\left(\frac{n}{n+1}Z_i+\sqrt{\frac{\frac{1}{n}}{1+\frac{1}{n}}}w_i\right)^2 &= \left[\left(\frac{n}{n+1}\right)^2\frac{1}{n}+\frac{1}{1+n}\right]\chi^2_{n^{\frac{1}{2\alpha}}}\\ &\leq \frac{2}{n}I\\ &\leq 4n^{-\frac{2\alpha}{2\alpha+1}}\end{aligned}$$

Here, we chose a conjugate prior, which made the calculation of the posterior extremely easy. But in the general case, the analysis can be extremely difficult.

Remark. With this kind of result, the posterior mean is rate optimal from a frequentist point of view.

1.2. General Prior (Posterior Contraction). We may want to investigate for what priors we do, or do not, have rate optimal results. This is useful for providing guidance in practice of which priors do not yield terrible posteriors.

$$\mathbf{E}_{Y|\theta^*}\left[\frac{\int_A \mathbf{P}(Y|\theta)\,\mathrm{d}\,\pi(\theta)}{\int \mathbf{P}(Y|\theta)\,\mathrm{d}\,\pi(\theta)}\right]\to 1$$

where $A=\{\|\theta-\theta^*\|\leq L\epsilon_n\}$ and we want to show that the posterior distribution on the θ concentrates its mass near θ^* with high probability:

$$\pi(\|\theta-\theta^*\|\leq L\epsilon_n|Y)\geq \text{Something that approaches 1}$$

What conditions should our prior satisfy?

(1) $\pi(\|\theta-\theta^*\|)\leq\epsilon_n)\geq\exp\{-cn\epsilon_n^2\}$. We cannot explore an area that is too small; if it is too small, it may not contain what we are looking for. Note that this mass is exponentially small; we just need a little bit, and we will succeed eventually.

(2) $\pi(\mathcal{F}_n^c)\leq\exp\{-(c+4)n\epsilon_n^2\}$ We cannot explore an area that is too big, or we may have problems. $\mathcal{F}_n$ is the subparameter space we restrict ourselves to; for example, in our previous example, the Sobolev ball was the full parameter space, but we only considered nonzero θ_i up to $i=I$.

(3) Testing. There is a test as to whether the area we are examining contains the true parameter. For example, we could have:

$$\begin{aligned}&H_0:\theta=\theta^*\\ &H_1:\|\theta-\theta^*\|\geq L\epsilon_n\end{aligned}$$

for $\theta\in\mathcal{F}_n\cap\operatorname{supp}(\pi)$ such that $\mathbf{P}_{\theta^*}\phi+\sup_{\theta\in H_1}\mathbf{P}_\theta(1-\phi)\leq\exp\{-(c+4)\epsilon_n^2$ We define Φ to be the event that we reject the null hypothesis; that is,

$$\phi\triangleq\mathbf{1}\{\|\theta-\theta^*\|>L\epsilon_n\}$$

In so doing, the probability defined above is the sum of the Type I error and worst case Type II error. Note that here θ^* is not some true unknown parameter, but rather a null hypothesis that we set ourselves.

That is, if all three conditions are satisfied, then our estimator achieves the very good performance that we just discussed; i.e.,

$$\mathbf{E}_{Y|\theta^*}\,\pi(\|\theta-\theta^*\| \le L\epsilon_n|Y) \to 1$$

This essentially says that the posterior probability assigned by the Bayesian technique will put most of its mass on the θ near θ^* exponentially fast (with high probability). Posterior contraction!

Le Cam discussed (1) and (3). Barron added condition (2), so that in infinite parameter spaces (e.g., nonparametric estimation), we have useful conditions as well.

Suppose we instead considered

$$\mathbf{E}_{Y|\theta^*}\left[\frac{\int_{A^c}\mathbf{P}(Y|\theta)\,\mathrm{d}\,\pi(\theta)}{\int\mathbf{P}(Y|\theta)\,\mathrm{d}\,\pi(\theta)}\right] \to 0$$

We want an upper bound for the numerator, which we can get from condition 3, and a lower bound for the denominator, which we can get from condition 1.

Remark. We can check the satisfaction of conditions (1) and (3) for the normal prior with $\mathcal{F}_n = \{\theta : \theta_i = 0, i > n^{\frac{1}{1+2\alpha}}\}$.

1.2.1. *Lower bound for denominator.*

$$\begin{aligned}\mathbf{E}_{Y|\theta^*}\left[\frac{\int_{A^c}\mathbf{P}(Y|\theta)\,\mathrm{d}\,\pi(\theta)}{\int\mathbf{P}(Y|\theta)\,\mathrm{d}\,\pi(\theta)}\right] &= (\phi+1-\phi)\,\mathbf{E}_{Y|\theta^*}\left[\frac{\int_{A^c}\mathbf{P}(Y|\theta)\,\mathrm{d}\,\pi(\theta)}{\int\mathbf{P}(Y|\theta)\,\mathrm{d}\,\pi(\theta)}\right]\\ &\ge \mathbf{E}_{Y|\theta^*}\,\phi + \mathbf{E}_{Y|\theta^*}(1-\phi)\frac{\int_{A^c}\mathbf{P}(Y|\theta)\,\mathrm{d}\,\pi(\theta)}{\int\mathbf{P}(Y|\theta)\,\mathrm{d}\,\pi(\theta)}\end{aligned}$$

Note that condition (3) implies that the sum of Type I and Type II errors converge to zero exponentially fast, which implies that the Type I error ($\mathbf{E}_{\theta^*}\,\phi$ term) also converges to zero exponentially fast. Therefore, we may ignore it.

Denote $B = \{\theta : \|\theta-\theta^*\| \le \epsilon_n\}$. The intuition is that we want something like this:

$$\begin{aligned}\int\mathbf{P}(Y|\theta)\,\mathrm{d}\,\pi(\theta) &\ge \int_B p(y|\theta)\,\mathrm{d}\,\pi(\theta)\\ &\gtrsim p(y|\theta^*)\pi(\theta\in B)\\ &\ge p(y|\theta^*)\exp\{-cn\epsilon_n^2\}\end{aligned}$$

This may not necessarily be true (we are using some intuition about smoothness of the posterior to have this intuition.

Note that the integrals above are essentially over the posterior distribution. Observe that:

$$\begin{gathered}p(\theta|y) \propto p(p|\theta)p(\theta)\\ \Rightarrow \int p(\theta|y)\,\mathrm{d}\,\theta \propto \int p(y|\theta)\,\mathrm{d}\,\pi(\theta)\end{gathered}$$

Essentially, we are working with the total probability assigned by the posterior over some region of θ.

Instead, we can prove this proposition:

Lemma 1.

$$\mathbf{P}_{Y|\theta^*} \underbrace{\left\{ \int \frac{\mathbf{P}(Y|\theta)}{\mathbf{P}(Y|\theta^*)} \,\mathrm{d}\,\pi(\theta) \geq \exp\{-(c+2)n\epsilon_n^2 \right\}}_{D} \geq 1 - \exp\{-c' n\epsilon_n^2\}$$

Let

$$\begin{aligned} R &\triangleq \mathbf{E}_{Y|\theta^*}(1-\phi)\frac{\int_{A^c} \mathbf{P}(Y|\theta)\,\mathrm{d}\,\pi(\theta)}{\int \mathbf{P}(Y|\theta)\,\mathrm{d}\,\pi(\theta)} \\ &= \mathbf{E}_{Y|\theta^*}(\mathbf{1}_D + \mathbf{1}_{D^c})(1-\phi)\frac{\int_{A^c} \mathbf{P}(Y|\theta)\,\mathrm{d}\,\pi(\theta)}{\int \mathbf{P}(Y|\theta)\,\mathrm{d}\,\pi(\theta)} \\ &\triangleq R_1 + R_2 \end{aligned}$$

Note that $R_2 \leq \exp\{-c' n\epsilon_n^2\}$. Let us consider R_1 more closely:

$$\begin{aligned} R_1 &\leq \exp\{(c+2)n\epsilon_n^2\}\,\mathbf{E}_{Y|\theta^*}(1-\phi)\frac{\int_{A^c} \mathbf{P}(Y|\theta)\,\mathrm{d}\,\pi(\theta)}{\mathbf{P}(Y|\theta^*)} \\ &= \underbrace{\exp\{(c+2)n\epsilon_n^2\}\,\mathbf{E}_{Y|\theta^*}(1-\phi)\frac{\int_{A^c\cap\mathcal{F}_n} \mathbf{P}(Y|\theta)\,\mathrm{d}\,\pi(\theta)}{\mathbf{P}(Y|\theta^*)}}_{R_{11}} \\ &\quad + \underbrace{\exp\{(c+2)n\epsilon_n^2\}\,\mathbf{E}_{Y|\theta^*}(1-\phi)\frac{\int_{A^c\cap\mathcal{F}_n^c} \mathbf{P}(Y|\theta)\,\mathrm{d}\,\pi(\theta)}{\mathbf{P}(Y|\theta^*)}}_{R_{12}} \end{aligned}$$

Observe that $R_1 \leq R_{11} + R_{12}$. It follows that:

$$\begin{aligned} R_{12} &\leq \pi(\mathcal{F}_n^c) \\ &\leq \exp\{-(c+4)n\epsilon_n^2\}\exp\{(c+2)n\epsilon_n^2\} \\ &= \exp\{-4n\epsilon_n^2\} \end{aligned}$$

and that the type II error is:

$$\begin{aligned} R_{11} &\leq \int_{A^c\cap\mathcal{F}_n} \exp\{-(c+4)\epsilon_n^2\}\,\mathrm{d}\,\pi(\theta)\exp\{(c+2)n\epsilon_n^2\} \\ &= \exp\{-2n\epsilon_n^2\} \end{aligned}$$

Proof. Of Lemma 1

$$\begin{aligned}
&\mathbf{P}_{Y|\theta^*}\left\{\int \frac{\mathbf{P}(Y|\theta)}{\mathbf{P}(Y|\theta^*)}\,\mathrm{d}\,\pi(\theta) \le \exp\{-(c+2)n\epsilon_n^2\}\right\}\\
&\quad\le \mathbf{P}_{Y|\theta^*}\left\{\int_B \frac{\mathbf{P}(Y|\theta)}{\mathbf{P}(Y|\theta^*)}\,\mathrm{d}\,\pi(\theta) \le \exp\{-(c+2)n\epsilon_n^2\}\right\}\\
&\quad= \mathbf{P}_{Y|\theta^*}\left\{\int_B \frac{\mathbf{P}(Y|\theta)}{\mathbf{P}(Y|\theta^*)}\,\mathrm{d}\,\frac{\pi(\theta)}{\pi(B)} \le \exp\{-(c+2)n\epsilon_n^2\}\frac{1}{\pi(B)}\right\}\\
&\quad= \mathbf{P}_{Y|\theta^*}\left\{-\log\int_B \frac{\mathbf{P}(Y|\theta)}{\mathbf{P}(Y|\theta^*)}\,\mathrm{d}\,\frac{\pi(\theta)}{\pi(B)} \ge 2n\epsilon_n^2\right\}
\end{aligned}$$

Recall that $-\log X$ is convex. Therefore,

$$\begin{aligned}
&\mathbf{P}_{Y|\theta^*}\left\{-\log\int_B \frac{\mathbf{P}(Y|\theta)}{\mathbf{P}(Y|\theta^*)}\,\mathrm{d}\,\frac{\pi(\theta)}{\pi(B)} \ge 2n\epsilon_n^2\right\}\\
&\quad\le \mathbf{P}_{Y|\theta^*}\left\{\int_B -\log \frac{\mathbf{P}(Y|\theta)}{\mathbf{P}(Y|\theta^*)}\,\mathrm{d}\,\frac{\pi(\theta)}{\pi(B)} \ge 2n\epsilon_n^2\right\}\\
&\quad= \mathbf{P}_{Y|\theta^*}\left\{\int_B -\log \mathbf{P}(Y|\theta) + \log \mathbf{P}(Y|\theta^*)\,\mathrm{d}\,\frac{\pi(\theta)}{\pi(B)} \ge 2n\epsilon_n^2\right\}\\
&\quad= \mathbf{P}_{Y|\theta^*}\left\{\int_B \frac{n}{2}\left(\|Y-\theta\|^2 - \|Y-\theta^*\|^2\right)\mathrm{d}\,\frac{\pi(\theta)}{\pi(B)} \ge 2n\epsilon_n^2\right\}\\
&\quad= \mathbf{P}_{Y|\theta^*}\left\{\int_B \frac{n}{2}\left(\underbrace{2\langle Y-\theta^*, \theta^*-\theta\rangle}_{\mathcal{N}(0,\frac{4\|\theta^*-\theta\|^2}{n})} + \|\theta^*-\theta\|^2\right)\mathrm{d}\,\frac{\pi(\theta)}{\pi(B)} \ge 2n\epsilon_n^2\right\}
\end{aligned}$$

For all $\theta \in B$,

$$\begin{aligned}
\mathbf{P}_{Y|\theta^*}\{-\log\frac{\mathbf{P}(Y|\theta)}{\mathbf{P}(Y|\theta^*)} \ge 2n\epsilon_n^2\} &\le \mathbf{P}_{Y|\theta^*}\{\mathcal{N}(0, \frac{4\|\theta^*-\theta\|^2}{n}) \ge \epsilon_n^2\}\\
&\le \mathbf{P}_{Y|\theta^*}\{\frac{4\epsilon_n}{\sqrt{n}}\mathcal{N}(0,1) \ge \epsilon_n^2\}\\
&\le \mathbf{P}_{Y|\theta^*}\{\frac{4}{\sqrt{n}}\mathcal{N}(0,1) \ge \epsilon_n\}\\
&\lesssim \exp\{-c'n\epsilon_n^2\} \qquad \text{(Gaussian tail.)}
\end{aligned}$$

This bound is for every $\theta \in B$. Therefore, the average is also bounded. For more general distributions, the ball is defined by the Kullback-Leibler divergence. □

2. Remarks

These days, it is fashionable to consider Frequentist justifications of Bayesian approaches. Harry thinks there are problems more important than necessarily showing contraction for a prior:

(1) Negative results; show that for a big class of priors, we don't really have posterior contraction. It could be more important to give negative results, so theory can guide practice.

(2) Computational issues. Some people may have wonderful priors with great frequentist interpretation. The problem is often that we do not have to compute the posterior.

(3) Bernstein-von Mises Theorem: Last time, many people cared about the distribution of the parameter given data. Do we get asymptotic normality? In the parametric case, we get the posterior distribution, and see that it's very similar to the MLE distribution. In frequentist, we can get a confidence interval. We can get BvM and see if it is similar to frequentist results.

(4) Beyond KL Loss. Last time, we discussed the ℓ_2 norm, which is equivalent to KL Divergence. But, sometimes we care about other loss functions. Most of the theory we have these days is KL loss. We may want to try to go beyond that, for example, spectral norm for a covariance matrix. The KL divergence is equivalent to Frobenius norm.

   (a) Functional estimation: Rather than estimating θ, we may want to estimate a function of θ.

CHAPTER 4

Sparse Vector Estimation

Consider the model:

$$Y_i = \theta_i + \frac{1}{\sqrt{n}} Z_i, \qquad i = 1, \ldots, p$$

We have n samples, and p could be larger or smaller than n. We use $\sqrt{n}$ to make this model consistent with linear models.

Remark. Previously, we had an infinite number of parameters, and observed n of them once. Now, we have p parameters, and observe each of them n times.

Last time, we assumed that θ_i had an order; i.e., it decreases as i increases. Now, we have a different assumption on θ:

$$\Theta = \{\theta : \|\theta\|_0 \leq s\}, \qquad s \lesssim p$$

or, more generally:

$$\Theta = \{\theta : \|\theta\|_q \leq r\}$$

We will consider the simpler (sparse) case in this chapter. The concept of sparsity originated from the Gaussian sequence model, and has dominated statistics research in recent years.

1. Estimators

1.1. All Subsets Selection. We now construct our estimator. Suppose s is known. Consider the estimator:

$$\hat{\theta} = \arg \min_{\|\theta\|_0 \leq s} \|Y - \theta\|^2$$

That is, we consider the least squares estimator, searching through all the possible θ. Suppose we have $\|\theta\|_0 = s$. Then there are $\binom{p}{s}$ possible sparsity patterns. Intuitively, this makes sense. We are simply doing least squares for each sparsity pattern, and we choose the best one.

The problem is that we know how to calculate this in the sparse vector case, but we may not be able to in the sparse linear regression case. This is referred to as *all subsets selection.*

1.2. Dantzig Selector (2006). Introduced by Candes and Tao. We solve the problem:

$$\begin{aligned} &\underset{\theta}{\text{minimize}} && \|\theta\|_1 \\ &\text{subject to} && \sup_i |Y_i - \theta_i| \leq c\sqrt{\frac{2\log p}{n}} \qquad c \geq 1 \end{aligned}$$

This may be regarded as a convex relaxation of all subset selection.

1.3. LASSO (1996). We instead consider:

$$\hat{\theta} = \arg\min_{\theta} \|Y - \theta\|^2 + \lambda \|\theta\|_1$$

which is another convex relaxation of all subsets selection. $\frac{\lambda}{2}$ plays the same role as $c\sqrt{\frac{2\log p}{n}}$ in the Dantzig selector.

2. Risk Upper Bound

Risk upper bounds give a guarantee on the performance of an estimator.

2.1. All Subsets Selection. When we know the nonzero coordinates, we have a risk upper bound of $\frac{s}{n}$. We can get a similar upper bound when the coordinates are unknown, but we have to pay a price for not knowing the coordinates.

To start, we take advantage of the basic inequality, based on the definition of $\hat{\theta}$:

$$\left\|Y - \hat{\theta}\right\|^2 \leq \|Y - \theta^*\|^2$$

$$\begin{aligned} \left\|Y - \hat{\theta}\right\|^2 &= \left\|Y - \theta^* + \theta^* - \hat{\theta}\right\|^2 \\ &= \|Y - \theta^*\|^2 + 2\langle Y - \theta^*, \theta^* - \hat{\theta}\rangle + \left\|\theta^* - \hat{\theta}\right\|^2 \\ \Rightarrow \left\|Y - \hat{\theta}\right\|^2 - \|Y - \theta^*\|^2 &= 2\langle Y - \theta^*, \theta^* - \hat{\theta}\rangle + \left\|\theta^* - \hat{\theta}\right\|^2 \\ \Rightarrow \left\|\hat{\theta} - \theta^*\right\|^2 &= 2\langle Y - \theta^*, \hat{\theta} - \theta^*\rangle + \underbrace{\left\|Y - \hat{\theta}\right\|^2 - \|Y - \theta^*\|^2}_{\leq 0} \\ \Rightarrow \left\|\hat{\theta} - \theta^*\right\|^2 &\leq 2\left\langle Y - \theta^*, \hat{\theta} - \theta^*\right\rangle \\ &= \left\|\hat{\theta} - \theta^*\right\| 2\left\langle Y - \theta^*, \frac{\hat{\theta} - \theta^*}{\left\|\hat{\theta} - \theta^*\right\|}\right\rangle \end{aligned}$$

We normalize the second term of the inner product to make the following analysis simpler. We want to find an upper bound for that inner product. But this may be difficult as there is dependence between Y and $\hat{\theta}$.

We want to bound:

$$\left\langle Y - \theta^*, \frac{\hat{\theta} - \theta^*}{\left\|\hat{\theta} - \theta^*\right\|} \right\rangle$$

and so we can use:

$$\begin{aligned}\left\langle Y - \theta^*, \frac{\hat{\theta} - \theta^*}{\left\|\hat{\theta} - \theta^*\right\|} \right\rangle &\leq \sup_{c, \|c\|_2 = 1, \|c\|_0 \leq 2s} \langle Y - \theta^*, c\rangle \\ &\leq \sup_{c, \|c\|_2 = 1, \|c\|_0 \leq 2s} \frac{1}{\sqrt{n}} \langle \sqrt{n}(Y - \theta^*), c\rangle \\ &\leq \frac{1}{\sqrt{n}} \sup_{B : |B| = 2s} \sqrt{\sum_{i \in B} Z_i^2}\end{aligned}$$

We can see that now we just have standard normals. To maximize the inner product, we simply choose the elements of c proportional to the elements in Y, subject to the constraints; that is, adhere to sparsity, and normalize by the distance. But the distance is the square root of χ^2. This gives us:

$$\begin{aligned}\left\langle Y - \theta^*, \frac{\hat{\theta} - \theta^*}{\left\|\hat{\theta} - \theta\right\|} \right\rangle &\leq \sup_{c, \|c\|_2 = 1, \|c\|_0 \leq 2s} \langle Y - \theta^*, c\rangle \\ &\leq \sup_{c, \|c\|_2 = 1, \|c\|_0 \leq 2s} \frac{1}{\sqrt{n}} \langle \sqrt{n}(Y - \theta^*), c\rangle \\ &\leq \frac{1}{\sqrt{n}} \sup_{B : |B| = 2s} \sqrt{\sum_{i \in B} Z_i^2}\end{aligned}$$

Note that $Z_i = \sqrt{(}Y_i - \theta_i^*) \sim \mathcal{N}(0, 1)$, and the last line came from:

$$\frac{\sum_i Z_i^2}{\sqrt{\sum_i Z_i^2}}$$

Now we just need to find an upper bound for the sum of many random variables. We now take advantage of a fact for χ_d^2 distributions:

$$\mathbf{P}\left\{\sqrt{\chi_d^2} - \sqrt{d} > t\right\} \leq \exp\left\{-\frac{t^2}{2}\right\}$$

for $t > 0$. This tail bound does not depend on d.

Now we focus on this concentration bound:

$$\mathbf{P}\left\{\frac{1}{\sqrt{n}} \sup_{B : |B| = 2s} \sqrt{\sum_{i \in B} Z_i^2} > t\right\}$$

We may use Bonferroni (union bound):

$$\begin{aligned}\mathbf{P}\left\{\frac{1}{\sqrt{n}}\sup_{B:|B|=2s}\sqrt{\sum_{i\in B}Z_i^2}>t\right\} &\leq \binom{p}{2s}\mathbf{P}\left\{\sqrt{\chi^2_{2s}}>\sqrt{n}t\right\}\\ &=\binom{p}{2s}\mathbf{P}\left\{\sqrt{\chi^2_{2s}}-\sqrt{2s}>\sqrt{n}t-\sqrt{2s}\right\}\\ &\leq\binom{p}{2s}\exp\left\{-\frac{(\sqrt{n}t-\sqrt{2s})^2}{2}\right\}\end{aligned}$$

We want to find a t such that the upper bound on the inner product holds with high probability.

There is a fact that $\binom{p}{2s}$ can be bounded from above by $\left(\frac{ep}{2s}\right)^{2s}$.[1] Then, we get:

$$\exp\left\{2s\log\frac{ep}{2s}-\left(\frac{\sqrt{n}t-\sqrt{2s}}{2}\right)^2\right\}$$

Then, if we pick t such that:

$$\frac{\sqrt{n}t-\sqrt{2s}}{2}=2\sqrt{2s\log\frac{ep}{s}}$$

then we have this:

$$\exp\{-3\cdot 2s\log\frac{ep}{2s}\}$$

What really matters is that we are able to get a concentration bound such that $\sqrt{\chi^2_d}$ deviates far from $\sqrt{d}$ almost never. More specifically, we have t as follows:

$$t=\frac{4\sqrt{2s\log\frac{ep}{s}}}{\sqrt{n}}+\frac{\sqrt{2s}}{\sqrt{n}}$$

picking t so that the probability goes to zero. The proof is essentially looking for a tail bound on the χ^2 probability and then get a concentration bound.

Basically, we upper bounded the inner product by the sup over a number of χ^2 random variables, and then get a tail bound on its distribution.

In practice, t corresponds to the penalty constant that we choose. The produce for all subset selection is not practical. In practice, we don't have the ℓ_0 constraint, but instead we add a penalty term $c\left\|\theta\right\|_0\log\frac{ep}{\|\theta_0\|}$. With this kind of a penalty, we can get a risk upper bound:

$$\mathbf{E}\left\|\hat{\theta}-\theta\right\|^2\lesssim\frac{s\log\frac{ep}{n}}{n}$$

This is essentially regularizing with an ℓ_0 penalty; that is, BIC / AIC.

The following remarks may be done as homework practice.

[1] `http://math.stackexchange.com/questions/1352338/`

Remark. We are assuming $Z_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, but we can prove the same result for subgaussian variables (random variables with tails like Gaussian variables). For other distributions, it depends on the tails.

Remark. Although we are considering the ℓ_0 ball, we can extend it to the ℓ_p ball.

Remark. A sharper upper bound (Wu and Zhou) is possible:

$$(2+o(1))s\frac{\log eps}{n}$$

under some additional assumptions. In the proof of the upper bound, we showed the rate, but we don't know anything about the constant. This says that we can get a lower bound on the constant.

Harry is doubtful about calculating the exact constant because it may not always be helpful (e.g., the model is wrong).

Note that as before, this holds even for s unknown.

2.2. Dantzig Selector. Recall the formulation for the Dantzig Selector.

$$\begin{aligned} &\underset{\theta}{\text{minimize}} && \|\theta\|_1 \\ &\text{subject to} && \sup_i |Y_i - \theta_i| \le c\sqrt{\frac{2\log p}{n}} \qquad c \ge 1 \end{aligned}$$

Remark. Let us study the constraint more closely. There is a fact that says:

$$\mathbf{P}\left\{\sup_{1\le i\le p} Z_i > \sigma\sqrt{2\log p}\right\} \to 0$$

as $p \to \infty$.

Observe that in the constraint, $Y_i - \theta_i^*$ are essentially $\mathcal{N}(0,1)$ random variables. With the fact, this implies:

$$\mathbf{P}\left\{\theta^* : \sup_i |Y_i - \theta_i^*| \le \sqrt{\frac{2\log p}{n}}\right\} \to 1$$

where we get an extra factor of n because of the definition, which has $\frac{1}{\sqrt{n}}$. Then the basic inequality says (by definition of the objective):

$$\begin{aligned} \left\|\hat\theta\right\|_1 &\le \|\theta^*\|_1 \qquad \text{w.h.p.} \\ &= \|\theta_S^*\|_1 \end{aligned}$$

where we define:

$$(\theta_T)_i = \begin{cases} \theta_i & \text{for } i \in T \\ 0 & \text{otherwise} \end{cases}$$

with S being the true support of the θ^*. Then we may write:

$$\begin{aligned}
\left\|\hat{\theta}^*\right\|_1 &\geq \left\|\hat{\theta}\right\|_1 \\
&= \left\|\hat{\theta}_S\right\|_1 + \left\|\hat{\theta}_{S^c}\right\|_1 \\
\Rightarrow \left\|\hat{\theta}_{S^c}\right\|_1 &\leq \|\theta^*_S\|_1 - \left\|\hat{\theta}_S\right\|_1 \\
\left\|\hat{\theta}_{S^c}\right\|_1 &\leq \left\|(\theta^* - \hat{\theta})_S\right\|_1 \\
\Rightarrow \left\|(\hat{\theta} - \theta^*)_{S^c}\right\|_1 &\leq \left\|(\theta^* - \hat{\theta})_S\right\|_1
\end{aligned}$$

This says that the error outside the true support is bounded by the error inside, and we can bound the right hand side with high probability.

In our analysis, we need to know s, but we don't need it for our procedure. So, we can get a risk upper bound to estimate θ, and we are going to use the ℓ_2 loss:

$$\left\|\hat{\theta} - \theta^*\right\|_2^2 = \left\|(\hat{\theta} - \theta^*)_S\right\|_2^2 = \left\|(\hat{\theta} - \theta^*)_{S^c}\right\|_2^2$$

We will reduce the ℓ_2 loss to ℓ_1 to use the bound that we have found. Let us begin with the first term. Observe that:

$$|\hat{\theta}_i - \theta^*_i| \leq |\hat{\theta}_i - Y_i| + |Y_i - \theta^*_i|$$

by the triangle inequality. We have, with high probability, for all i:

$$\begin{aligned}
|\hat{\theta}_i - \theta^*_i| &\leq |\hat{\theta}_i - Y_i| + |Y_i - \theta^*_i| \\
&\leq 2\sqrt{\frac{2\log p}{n}}
\end{aligned}$$

as previously shown. That implies that:

$$\begin{aligned}
\left\|(\hat{\theta} - \theta^*)_S\right\|_2^2 &= \sum_{i \in S} (\hat{\theta}_i - \theta^*_i)^2 \\
&\leq 4\frac{2s\log p}{n}
\end{aligned}$$

Now, we can examine the second term. We cannot use the previous bound, because the size of S^c may be huge. What we can do is this. The value of every coordinate is bounded above by $2\sqrt{\frac{2\log p}{n}}$. Then, we have a trivial upper bound:

$$\begin{aligned}\left\|(\hat{\theta}-\theta^*)_{S^c}\right\|_2^2 &\leq 2\sqrt{\frac{2\log p}{n}}\left\|(\hat{\theta}-\theta^*)_{S^c}\right\|_1\\ &\leq 2\sqrt{\frac{2\log p}{n}}\left\|(\hat{\theta}-\theta^*)_{S}\right\|_1\\ &\leq 2\sqrt{\frac{2\log p}{n}}\cdot s\cdot\sqrt{\frac{2\log p}{n}}\\ &\leq 2s\frac{2\log p}{n}\end{aligned}$$

which is very close to the upper bound that we got from all subset selection.

2.3. LASSO.

The basic inequality we analyze here is:

$$\left\|Y-\hat{\theta}\right\|_2^2+\lambda\left\|\hat{\theta}\right\|_1\leq\|Y-\theta^*\|_2^2+\lambda\|\theta^*\|_1$$

Then we may rewrite:

$$\left\|Y-\hat{\theta}\right\|_2^2+\lambda\left\|\hat{\theta}\right\|_1=\left\|Y-\theta^*+\theta^*-\hat{\theta}\right\|_2^2+\lambda\left\|\hat{\theta}\right\|_1$$

which yields:

$$\left\|\theta^*-\hat{\theta}\right\|_2^2\leq-2\langle Y-\theta^*,\theta^*-\hat{\theta}\rangle+\lambda\|\theta^*_S\|_1-\lambda\left\|\hat{\theta}\right\|_1$$

With high probability (cf 2.2):

$$\leq 2\sqrt{\frac{2\log p}{n}}\left\|\hat{\theta}-\theta^*\right\|_1+\lambda\|\theta^*_S\|_1-\lambda\left\|\hat{\theta}_S\right\|_1-\lambda\left\|\hat{\theta}_{S^c}\right\|_1$$

$$\leq 2\sqrt{\frac{2\log p}{n}}\left[\left\|(\hat{\theta}-\theta^*)_S\right\|_1+\left\|\hat{\theta}_{S^c}\right\|_1\right]+\lambda\|\theta^*_S\|_1-\lambda\left\|\hat{\theta}_S\right\|_1-\lambda\left\|\hat{\theta}_{S^c}\right\|_1$$

Using triangle inequality:

$$\leq\left(2\sqrt{\frac{2\log p}{n}}+\lambda\right)\left\|(\hat{\theta}-\theta^*)_S\right\|_1$$

Suppose we let $\lambda=2\sqrt{\frac{2\log p}{n}}$, so have with high probability:

$$\left\|\theta^*-\hat{\theta}\right\|_2^2\leq\left(4\sqrt{\frac{2\log p}{n}}\right)\left\|(\hat{\theta}-\theta^*)_S\right\|_1$$

what are we going to do with ℓ_2 and ℓ_1 norms? Use this fact:

$$\left\|a\in\mathbf{R}^d\right\|_1\leq\sqrt{d}\,\|a\|_2$$

This gives us the upper bound:

$$\begin{aligned}\left\|\theta^* - \hat{\theta}\right\|_2^2 &\leq \left(4\sqrt{\frac{2\log p}{n}}\right)\left\|(\hat{\theta}-\theta^*)_S\right\|_1\\ &\leq \left(4\sqrt{\frac{2\log p}{n}}\right)\sqrt{s}\left\|(\hat{\theta}-\theta^*)_S\right\|_2\\ &\leq \left(4\sqrt{\frac{2\log p}{n}}\right)\sqrt{s}\left\|(\hat{\theta}-\theta^*)\right\|_2\end{aligned}$$

After putting this inequality back in and solving, we get essentially the same upper bound as for the Dantzig Selector.

3. Risk Lower Bound

A risk lower bound that matches the upper bound proven for our estimator will show that the guarantee on our estimator is good (cannot be improved).

3.1. All Subset Selection. Recall the model:

$$Y_i = \theta_i + \frac{1}{\sqrt{n}}Z_i, \qquad i = 1, \ldots, p$$

and we consider the parameter space:

$$\Theta_s = \{\theta : \|\theta\|_0 \leq s\}, \qquad s \lesssim p$$

We will show a lower bound:

$$\inf_{\hat{\theta}} \sup_{\theta\in\Theta_s} \mathbf{E}\left\|\hat{\theta}-\theta\right\|^2 \geq c_1\frac{s\log\frac{ep}{s}}{n}$$

which shows that the upper bound shown previously is the best rate of convergence. We will be sloppy in class, but if we are more careful, we can get the $2+o(1)$ lower bound for the constant.

We now begin our proof. Assume $p = n$ and $n^{a_1} \leq s \leq n^{a_2}$, with $0 < a_1 \leq a_2 < 1$. These assumptions are not necessary but makes the analysis easier.

Remark. In the first lecture, we found a trivial lower bound $c_1\frac{s}{n}$. We now want a log factor, which makes it slightly larger, but makes the proof much harder.

3.1.1. *Intuition.* Consider the case when $s = 1$. Then the risk bound is $\frac{\log ep}{n}$. Why is this intuitively true? Recall that:

$$\max_i \left|\frac{1}{\sqrt{n}}Z_i\right| = (1+o(1))\frac{1}{\sqrt{n}}\sqrt{2\log p}$$

If we want to be able to distinguish whether or not $\theta_i = 0$, then it has to be at least this order; otherwise it is indistinguishable from noise.

Intuitively, we would look at all Y_i, and see which one has the biggest magnitude, and say that θ_i is nonzero. But there is a risk there, as it could be the largest due to noise. If the true nonzero θ_i satisfies

$$|\theta_i| < (1+\epsilon)\sqrt{\frac{2\log p}{n}}$$

for $\epsilon \gg 0$, then with high probability we cannot identify the true support.

This is why the risk bound makes sense intuitively. The common wisdom is to take the i where Y_i is the largest in magnitude, but there is risk of choosing the wrong i.

Recall that for general s, $p = n$, we have $\binom{n}{s}$ ways to pick the support.

How would we recover the support, assuming s? We could take the the absolute values of the vector, and then rank the entries, and take the top s.

But this procedure is prone to mistakes. What if $\theta_i = 0$ for all i? Then we have the magnitudes:

$$|Y|_{[1]} = (1+o(1))\frac{1}{\sqrt{n}}\sqrt{2\log ep}$$

$$|Y|_{[s]} = (1+o(1))\frac{1}{\sqrt{n}}\sqrt{2\log\frac{ep}{s}}$$

So this procedure would fail in the case where:

$$\theta_i = \begin{cases} 0 \\ (1-\epsilon)\frac{1}{\sqrt{n}}\sqrt{2\log\frac{ep}{s}} \end{cases}$$

With high probability, we may estimate zero for entries that are truly nonzero; because the magnitude of the noise is greater than the magnitude of the signal. Then we expect a risk lower bound

$$s\left((1-\epsilon)\frac{\sqrt{2\log\frac{ep}{s}}}{\sqrt{n}}\right)^2 \approx 2\frac{s\log\frac{ep}{s}}{n}$$

3.1.2. *Proof.* Recall that previously, we used Le Cam's method: the maximum risk is bounded below by the average risk. We can also use Fano's Lemma. But both methods cannot give the proper constant.

What prior are we going to place on the parameter space?

(1) A first idea is to put a uniform prior over all possible supports, $\{T\}$, where the cardinality of each T is s.
(2) We can then say $\theta_i = (1-o(1))\sqrt{2\log\frac{ep}{s}}$ for $i \in T$ and zero otherwise, where the $o(1)$ corresponds to ϵ.

This prior would give us the desired lower bound, but is very hard to analyze. We will do this proof with a prior that is easier to work with; but gives us a weaker result.

Here is an alternative idea, where $p = n$.

$$\theta_i \overset{\text{iid}}{\sim} \begin{cases} \frac{1}{\sqrt{n}}\sqrt{\log \frac{ep}{s}} & \text{with probability } q = \frac{s}{2n} \\ 0 & \text{with probability } 1-q \end{cases}$$

The expected number of nonzero θ_i is $\frac{s}{2}$. However, the support of this prior is not a subset of the parameter space. But, because of the expectation, with high probability, the support of this prior *will* be in the parameter space; we will show that the probability that we choose a parameter outside our parameter space is negligible. This is a very important idea. We may calculate the probability of this binomial random variable:

$$\mathbf{P}\{\|\theta\|_0 > s\}$$

by using this tail bound for $X \sim Bin\left(n, \frac{s}{2n}\right)$:

$$\mathbf{P}\left(\left|\sqrt{X} - \sqrt{\frac{s}{2}}\right| > t\right) \leq 2\exp\{-t^2\}$$

which implies:

$$\mathbf{P}\left(X > s\right) \leq c_2 \exp\{-c_3 s\}$$

$c_2, c_3 > 0$; that is, away from the center, it is like a Gaussian tail. This is true for any s.

So, although the support of our prior could be outside the parameter space, it is nearly right.

Typically, we want to show something about:

$$\sup_{\theta\in\Theta_s} \mathbf{E}\left\|\hat{\theta} - \theta\right\|^2 \geq \mathbf{E}_\theta\,\mathbf{E}_{Y|\theta}\left\|\hat{\theta}_{\text{Bayes}} - \theta\right\|^2$$

but this is not right, because our support is invalid. But it's nearly right, so let's examine the right hand side more closely.

$$\mathbf{E}_\theta\,\mathbf{E}_{Y|\theta}\left\|\hat{\theta}_{\text{Bayes}} - \theta\right\|^2 = \underbrace{\int_{\theta\in\Theta_s} \mathbf{E}_{Y|\theta}\left\|\hat{\theta}_{\text{Bayes}} - \theta\right\|^2 \pi(\theta)}_{\text{A}} + \underbrace{\int_{\theta\in\Theta_s^c} \mathbf{E}_{Y|\theta}\left\|\hat{\theta}_{\text{Bayes}} - \theta\right\|^2 \pi(\theta)}_{\text{B}}$$

where we can show that B is very small, because the probability is small (subgaussian) and because the norm is small (definition of the point masses). Let's examine A more closely. We may show that:

$$\int_{\theta\in\Theta_s} \mathbf{E}_{Y|\theta}\left\|\hat{\theta}_{\text{Bayes}} - \theta\right\|^2 \pi(\theta) \leq \left(1 - \mathbf{P}(\Theta_s^c)\right)\frac{1}{1-\mathbf{P}(\Theta_s^c)}\int_{\theta\in\Theta_s} \mathbf{E}_{Y|\theta}\left\|\hat{\theta}_{\text{Bayes}} - \theta\right\|^2 \pi(\theta) + n^{-100}$$

Observe that $(1 - \mathbf{P}(\Theta_s^c))$ is nearly one. If we move the inverse inside the integral, we normalize the measure of integration so that we turn the prior not supported on Θ_s into a prior supported on Θ_s.

Let's try again. Observe that:

$$\begin{aligned}
\sup_{\theta \in \Theta_s} \mathbf{E}_{Y|\theta} \left\|\hat{\theta} - \theta\right\|^2 &\geq \int_{\Theta_s} \frac{\pi(\theta)}{\mathbf{P}(\Theta_s)} \mathbf{E}_{Y|\theta} \left\|\hat{\theta} - \theta\right\|^2 \mathrm{d}\,\theta \\
&= \frac{1}{\mathbf{P}(\Theta_s)} \left[\int \pi(\theta)\, \mathbf{E}_{Y|\theta} \left\|\hat{\theta} - \theta\right\|^2 - \int_{\Theta_s^c} \cdots \right] \\
&\geq \frac{1}{\mathbf{P}(\Theta_s)} \left[\int \pi(\theta)\, \mathbf{E}_{Y|\theta} \left\|\hat{\theta}_{\text{Bayes}} - \theta\right\|^2 - \int_{\Theta_s^c} \cdots \right] \\
&\geq \int \pi(\theta)\, \mathbf{E}_{Y|\theta} \left\|\hat{\theta}_{\text{Bayes}} - \theta\right\|^2 - \int_{\Theta_s^c} \cdots
\end{aligned}$$

where the integral over the complement is a lower order term. We originally include the $\mathbf{P}(\Theta_s)$ to make the integral a valid average.

All we need to show now is that:

$$\int \pi(\theta)\, \mathbf{E}_{Y|\theta} \left\|\hat{\theta}_{\text{Bayes}} - \theta\right\|^2 \geq c_3 \frac{s \log \frac{ep}{s}}{n}$$

where $p = n$. Because we have independent observations, we may calculate the Bayes estimator for the ith coordinate, letting $a = \frac{1}{\sqrt{n}}\sqrt{\log \frac{ep}{n}}$.

$$\begin{aligned}
\hat{\theta}_{\text{Bayes},i} &= \mathbf{E}[\theta_i | Y_i] \\
&= a \cdot \frac{q \frac{1}{\sqrt{2\pi}\frac{1}{\sqrt{n}}} \exp\left\{-\frac{(Y_i - a)^2}{2\frac{1}{n}}\right\}}{q \frac{1}{\sqrt{2\pi}\frac{1}{\sqrt{n}}} \exp\left\{-\frac{(Y_i - a)^2}{2\frac{1}{n}}\right\} + (1-q) \frac{1}{\sqrt{2\pi}\frac{1}{\sqrt{n}}} \exp\left\{-\frac{(Y_i - 0)^2}{2\frac{1}{n}}\right\}} \\
&= a \cdot \frac{q \exp\left\{-\frac{n(Y_i - a)^2}{2}\right\}}{q \exp\left\{-\frac{n(Y_i - a)^2}{2}\right\} + (1-q) \exp\left\{-\frac{nY_i^2}{2}\right\}} \\
&= a \cdot \frac{q \exp\left\{-\frac{n(-2Y_i a + a^2)}{2}\right\}}{q \exp\left\{-\frac{n(-2Y_i a + a^2)}{2}\right\} + (1-q)}
\end{aligned}$$

Therefore, we have:

$$\begin{aligned}\int \pi(\theta)\,\mathbf{E}_{Y|\theta}\left\|\hat{\theta}_{\text{Bayes}}-\theta\right\|^2 &= \sum_{i=1}^{n}\int_{\pi}(\theta_i)\,\mathbf{E}_{Y_i|\theta_i}\left\|\hat{\theta}_{\text{Bayes},i}-\theta_i\right\|^2\\ &= \sum_{i=1}^{n}\int_{\pi}(\theta_i)q\,\mathbf{E}_{Y_i|\theta_i=a}\left\|\hat{\theta}_{\text{Bayes},i}-a\right\|^2+\underbrace{(1-q)\,\mathbf{E}_{Y_i|\theta_i=0}\left\|\hat{\theta}_{\text{Bayes},i}\right\|^2}_{\geq 0}\\ &= nq\,\mathbf{E}_{Y_1|\theta_1=a}\left\|\hat{\theta}_{\text{Bayes},1}-1\right\|^2\\ &\geq \frac{s}{2}\,\mathbf{E}_{Y_1|\theta_1=a}\left\|\hat{\theta}_{\text{Bayes},1}-1\right\|^2\end{aligned}$$

Now, we want to get a lower bound for this. Let's examine the Bayes estimator:

$$\begin{aligned}\left(\hat{\theta}_{\text{Bayes},1}-a\right)^2 &= a^2\left(\frac{1-q}{q\exp\left\{-\frac{n}{2}(-2Y_ia)\right\}\exp\left\{-\frac{n}{2}a^2\right\}+1-q}\right)^2\\ \mathbf{E}_{Y_1|\theta_1=1}\left(\hat{\theta}_{\text{Bayes},1}-a\right)^2 &= a^2\underbrace{\mathbf{E}_{Y_1|\theta_1=1}\left(\frac{1-q}{q\exp\left\{-\frac{n}{2}(-2Y_ia)\right\}\exp\left\{-\frac{n}{2}a^2\right\}+1-q}\right)^2}_{\text{D}}\end{aligned}$$

Now, we recall that $a^2=\frac{1}{n}\log\frac{ep}{s}$. So, we just need to show that D is bounded away by a constant. One way to do is is using Jensen's inequality:

$$\begin{aligned}a^2\,\mathbf{E}_{Y_1|\theta_1=1}&\left(\frac{1-q}{q\exp\left\{-\frac{n}{2}(-2Y_ia)\right\}\exp\left\{-\frac{n}{2}a^2\right\}+1-q}\right)^2\\ &\geq \frac{(1-q)^2}{\mathbf{E}\left(q\exp\left\{-\frac{n}{2}(-2Y_ia)\right\}\exp\left\{-\frac{n}{2}a^2\right\}+1-q\right)^2}\end{aligned}$$

which gives a nice, determinist bound. Another way is as follows. We observe that the hard part is to analyze

$$\begin{aligned}A &= q\exp\left\{2Y_ia\right\}\exp\left\{-\frac{n}{2}a^2\right\}\\ &= q\exp\left\{n\left(a+\frac{1}{\sqrt{n}}Z_i\right)a\right\}\exp\left\{-\frac{n}{2}a^2\right\}\\ &= q\exp\{\sqrt{n}Z_ia\}\exp\left\{\frac{na^2}{2}\right\}\\ &= \frac{1}{2n}\exp\left\{Z_i\sqrt{\log\frac{ep}{s}}\right\}\exp\left\{\frac{1}{2}\log\frac{en}{s}\right\}\end{aligned}$$

So, as $n \to \infty$, $A \to 0$, so the limit of the term preceding a^2 goes to one by dominated convergence theory.

4. Bayesian Posterior Contraction

4.1. All Subset Selection.

We will construct a prior π such that

$$\mathbf{E}_{Y|\theta^*}\, \pi \left(\|\theta - \theta^*\|^2 \geq c_2 \frac{s \log \frac{ep}{s}}{n} |Y \right) \to 0$$

for some $c_2 > 0$.

4.2. Contraction Rate.

4.3. Contraction Rate Matches Frequentist Risk.

4.4. Failure on Some Priors (Lasso).

CHAPTER 5

High-Dimensional Linear Regression

In this chapter, we will review the theoretical foundations of high-dimensional linear regression and give some justfications.

Suppose the model:

$$Y_{n\times 1} = \mathbf{X}_{n\times p}\beta_{p\times 1} + \varepsilon_{n\times 1}$$

with noise:

$$\varepsilon_{n\times 1} \sim \mathcal{N}(0, \sigma\mathbf{I}_p)$$

we typically further assume:

$$\lambda_{\max}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right) \leq M$$

which is a normalization of sorts. Sometimes we take additional assumptions; for example, if we want to perform inference on β rather than $\mathbf{X}\beta$, we would have to take some assumptions to prevent multicollinearity.

Suppose our goal is to measure the loss in this way:

$$\mathbf{E}\left\|\mathbf{X}\hat{\beta} - \mathbf{X}\beta\right\|^2$$

which is prediction loss, or

$$\mathbf{E}\left\|\hat{\beta} - \beta\right\|^2$$

We can measure the risk in each of these ways, and hope that the risk is small, or "optimal" in some sense. There are a few ways we could do this:

(1) All subsets selection, in which we assume $\|\beta\|_0 \leq s$, with $s \ll p$. Then we seek:

$$\hat{\beta} = \arg\min_{\|\beta\|_0 \leq s} \|Y - \mathbf{X}\beta\|^2$$

If s is known, then this is just maximum likelihood estimation over this parameter space. When s is unknown, we may use the estimator:

$$\hat{\beta} = \arg\min \|Y - \mathbf{X}\beta\|^2 + \lambda\|\beta\|_0$$

We will not be discussing this procedure, as it is intractable when p, s is large, and the problem is not convex.

(2) Dantzig Selector

(3) LASSO

These ideas should be familiar; they were previously explored in the context of sparse vector estimation.

1. All subsets selection

We analyze the problem where s is known, and the design is identity. The analysis can be generalized. First, we apply the basic inequality:

$$\begin{aligned}
\left\|Y - \mathbf{X}\hat{\beta}\right\|^2 &\leq \|Y - \mathbf{X}\beta\|^2 \\
\Rightarrow \left\|Y - \mathbf{X}\beta + \mathbf{X}\beta - \mathbf{X}\hat{\beta}\right\|^2 &\leq \|Y - \mathbf{X}\beta\|^2 \\
\Rightarrow \|Y - \mathbf{X}\beta\|^2 + 2\langle Y - \mathbf{X}\beta, \mathbf{X}\beta - \mathbf{X}\hat{\beta}\rangle + \left\|\mathbf{X}\beta - \mathbf{X}\hat{\beta}\right\|^2 &\leq \|Y - \mathbf{X}\beta\|^2 \\
\Rightarrow \left\|\mathbf{X}\beta - \mathbf{X}\hat{\beta}\right\|^2 &\leq 2\langle Y - \mathbf{X}\beta, \mathbf{X}\hat{\beta} - \mathbf{X}\beta\rangle \\
&= \left\|\mathbf{X}\hat{\beta} - \mathbf{X}\beta\right\| 2 \left\langle Y - \mathbf{X}\beta, \frac{\mathbf{X}(\hat{\beta} - \beta)}{\left\|\mathbf{X}(\hat{\beta} - \beta)\right\|} \right\rangle \\
&\leq \left\|\mathbf{X}\hat{\beta} - \mathbf{X}\beta\right\| 2 \sup_{\|u\|_0 \leq 2s} \left\langle Y - \mathbf{X}\beta, \frac{\mathbf{X}u}{\|\mathbf{X}u\|} \right\rangle
\end{aligned}$$

We now want to show:

$$\sup_{\|u\|_0 \leq 2s} \langle Y - \mathbf{X}\beta, \frac{\mathbf{X}u}{\|\mathbf{X}u\|} \rangle \leq C\sqrt{s \log \frac{ep}{s}}$$

with high probability, which would imply

$$\left\|\mathbf{X}\hat{\beta} - \mathbf{X}\beta\right\|^2 \leq c^2 s \log \frac{ep}{s}$$

with high probability.

If we want to further perform inference on β, we need to take on the additional assumption (RIP):

$$\inf_{\|u\|_0 \leq ks} \frac{\|\mathbf{X}u\|_2^2}{n\,\|u\|_2^2} \geq \gamma > 0$$

where $k = 2$; i.e., we take at most $2s$ subcolumns of XX, and then do classical linear regression, which then gives us the result:

$$\left\|\hat{\beta} - \beta\right\|_2^2 \leq \frac{c^2}{\gamma} \frac{s \log \frac{ep}{s}}{n}$$

with high probability. Equivalently, this assumption states that for the subsets of columns chosen $\mathbf{X}_S$, we have, with high probability,

$$\lambda_{\min}\left(\frac{1}{n}\mathbf{X}_S^\top \mathbf{X}_S\right) \geq \gamma$$

This smallest eigenvalue helps us give the lower bound on RIP assumption so that we can achieve the rate for the parameter estimation.

Observe that we need this assumption for inference on β, but it is not necessary for prediction (i.e., estimating $\mathbf{X}\beta$).

What does a $\frac{1}{n}$ rate mean? Consider:

$$\hat{\beta} - \beta \sim \mathcal{N}\left(0, \frac{1}{n}\left(\frac{\mathbf{X}^\top \mathbf{X}}{n}\right)^{-1}\right)$$

note that each entry of $\mathbf{X}^\top \mathbf{X}$ is like a length $\sqrt{n}$ vector times a length $\sqrt{n}$ vector, so we normalize it.

2. Dantzig Selector

Recall that last time, we wanted to:

$$\begin{aligned} &\underset{\beta}{\text{minimize}} && \|\beta\|_1 \\ &\text{subject to} && \|Y - \beta\|_\infty \leq \rho \end{aligned}$$

when $\mathbf{X}$ was orthogonal (or equivalently, the identity). So now, we instead consider:

$$\begin{aligned} &\underset{\beta}{\text{minimize}} && \|\beta\|_1 \\ &\text{subject to} && \left\|\mathbf{X}^\top (Y - \mathbf{X}\beta)\right\|_\infty \leq \rho \end{aligned}$$

Suppose we find a minimizer $\hat{\beta}$ that satisfies the constraints. Then by the basic inequality:

$$\left\|\hat{\beta}\right\|_1 \leq \|\beta\|_1 = \|\beta_S\|_1$$

where S is the true support of β, i.e., $S = \{i : \beta_i \neq 0\}$. Recall that last time, we did:

$$\begin{aligned} \left\|\hat{\beta}_S\right\|_1 + \left\|\hat{\beta}_{S^c}\right\|_1 &\leq \|\beta_S\|_1 \\ \Rightarrow \left\|\hat{\beta}_{S^c}\right\|_1 &\leq \|\beta_S\|_1 - \left\|\hat{\beta}_S\right\|_1 \leq \left\|(\beta - \hat{\beta})_S\right\|_1 \end{aligned}$$

by the Triangle Inequality. This will be the key for the proof today as well.

As we have done for all subsets selection, we want to an upper bound for:

$$\begin{aligned} \left\|\mathbf{X}\beta - \mathbf{X}\hat{\beta}\right\|^2 &= \langle \mathbf{X}(\hat{\beta} - \beta), \mathbf{X}(\hat{\beta} - \beta)\rangle \\ &= \langle (\hat{\beta} - \beta), \underbrace{\mathbf{X}^\top \mathbf{X}(\hat{\beta} - \beta)}_{\text{A}}\rangle \end{aligned}$$

Assuming we want to perform prediction, this is the most natural loss function. If we want to use a similar trick as last time, we want to get an upper bound for term A.

By assumption, we have:

$$\left\|\mathbf{X}^\top (Y - \mathbf{X}\hat{\beta})\right\|_\infty \leq \rho$$

which implies:

$$\begin{aligned}\left\|\mathbf{X}^\top\varepsilon_n + \mathbf{X}^\top(\mathbf{X}\beta - \mathbf{X}\hat{\beta})\right\|_\infty &\leq \rho \\ \Rightarrow \left\|\mathbf{X}^\top\mathbf{X}(\hat{\beta} - \beta)\right\|_\infty &\leq \rho + \left\|\mathbf{X}^\top\varepsilon_n\right\|_\infty \\ &\leq 2\rho\end{aligned}$$

if we assume $\left\|\mathbf{X}^\top\varepsilon_n\right\|_\infty$. The magnitude of ρ is therefore $c\sqrt{n\log p}^{[\text{why?}]}$with high probability. If $\mathbf{X}$ is the identity then the magnitude is (??).

Remark. Observe that $\mathbf{X}^\top\varepsilon_n$ is normally distributed with mean zero and covariance $\mathbf{X}^\top\mathbf{X}$.

We now have:

$$\left\|\mathbf{X}\hat{\beta} - \mathbf{X}\beta\right\|^2 \leq 2\rho\left\|\hat{\beta} - \beta\right\|_1 \qquad \text{w.h.p.}$$

Remark.

$$|\langle a, b\rangle| \leq \|a\|_1 \cdot \|b\|_\infty$$

We now split the ℓ_1 norm on the true support and the complement.

$$\begin{aligned}\left\|\mathbf{X}\hat{\beta} - \mathbf{X}\beta\right\|^2 &\leq 2\rho\left\|\hat{\beta} - \beta\right\|_1 \qquad \text{w.h.p.} \\ &= 2\rho\left[\left\|(\hat{\beta} - \beta)_{S^c}\right\|_1 + \left\|(\hat{\beta} - \beta)_S\right\|_1\right] \\ &\leq 4\rho\left\|(\hat{\beta} - \beta)_S\right\|_1\end{aligned}$$

We want to transform the ℓ_1 norm to an ℓ_2 norm on the right-hand side. On the left-hand side, we want to change it to a statement about the distance between β and $\hat{\beta}$.

Remark.

$$\|a\|_1 \leq \sqrt{m}\,\|a\|_2$$

where m is the length of a. It then follows that:

$$\left\|\mathbf{X}\hat{\beta} - \mathbf{X}\beta\right\|^2 \leq 4\rho\sqrt{s}\left\|(\hat{\beta} - \beta)_S\right\|_2$$

with high probability.

Suppose we want a bound on the risk for inference on β. We take on a restricted eigenvalue condition (REC), which states:

$$\inf_{\|u_{S^c}\|_1 \leq \|u_S\|_1} \frac{\|\mathbf{X}u\|^2}{n\,\|u\|^2} \geq \gamma_1 > 0$$

this is a restriction on the design.

Recall that:

$$\left\|(\hat{\beta} - \beta)_{S^c}\right\|_1 \leq \left\|(\hat{\beta} - \beta)_S\right\|_1$$

It thus follows that:

$$\gamma_1 n \left\|\hat{\beta}-\beta\right\|^2 \leq 4\rho\sqrt{s}\left\|(\hat{\beta}-\beta)_S\right\|$$

with high probability, which implies:

$$\left\|\hat{\beta}-\beta\right\| \leq \frac{4\rho\sqrt{s}}{\gamma_1 n}$$

with high probability. We thus conclude that:

$$\left\|\hat{\beta}-\beta\right\|^2 \leq C_m \frac{s\log p}{n}$$

with high probability, by substituting in the order of ρ ($\rho = C\sqrt{n\log p}$ we assumed earlier to get the desired bound on $\left\|\mathbf{X}^\top \varepsilon_n\right\|_\infty$).

Remark. Restricted Isometry Property "implies" Restricted Eigenvalue Condition, depending on parameters. That is, RIP is stronger.

3. LASSO

We consider the problem:

$$\hat{\beta} = \arg\min \|Y-\mathbf{X}\beta\|^2 + \lambda\|\beta\|_1$$

Observe that the λ here depends on the M in $\lambda_{\max}\left(\operatorname{diag}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)\right) \leq M$. Suppose we want to get risk bounds on prediction or inference. Let us now apply the basic inequality:

$$\left\|Y-\mathbf{X}\hat{\beta}\right\|^2 + \lambda\left\|\hat{\beta}\right\|_1 \leq \|Y-\mathbf{X}\beta\|^2 + \lambda\|\beta_S\|_1$$

Applying the same trick as before, we have:

$$\begin{aligned}\|Y-\mathbf{X}\beta\|^2 + 2\langle Y-\mathbf{X}\beta, \mathbf{X}\beta-\mathbf{X}\hat{\beta}\rangle + \left\|\mathbf{X}\beta-\mathbf{X}\hat{\beta}\right\|^2 + \lambda\left\|\hat{\beta}_S\right\|_1 + \lambda\left\|\hat{\beta}_{S^c}\right\|_1 &\leq \|Y-\mathbf{X}\beta\|^2 + \lambda\|\beta_S\|_1 \\ \Rightarrow 2\langle Y-\mathbf{X}\beta, \mathbf{X}\beta-\mathbf{X}\hat{\beta}\rangle + \left\|\mathbf{X}\beta-\mathbf{X}\hat{\beta}\right\|^2 + \lambda\left\|\hat{\beta}_S\right\|_1 + \lambda\left\|\hat{\beta}_{S^c}\right\|_1 &\leq \lambda\|\beta_S\|_1\end{aligned}$$

We then have:

$$\begin{aligned}\left\|\mathbf{X}\beta-\mathbf{X}\hat{\beta}\right\|^2 &\leq 2\langle Y-\mathbf{X}\beta, \mathbf{X}(\hat{\beta}-\beta)\rangle - \lambda\left\|\hat{\beta}_S\right\| + \lambda\|\beta_S\|_1 - \lambda\left\|\hat{\beta}_{S^c}\right\| \\ &= 2\langle\underbrace{\mathbf{X}^\top(Y-\mathbf{X}\beta)}_{\text{A}}, (\hat{\beta}-\beta)\rangle - \lambda\left\|\hat{\beta}_S\right\| + \lambda\|\beta_S\|_1 - \lambda\left\|\hat{\beta}_{S^c}\right\|\end{aligned}$$

How do we then get an upper bound on the right-hand side? We get an upper bound on A. As before, we assume $\left\|\mathbf{X}^\top\varepsilon_n\right\|_\infty \leq \rho$ with high probability.

Remark. For $M = 1$, pick $\rho = \sqrt{n}\sqrt{2\log p}$ We thus have:

$$\begin{aligned}\left\|\mathbf{X}\beta - \mathbf{X}\hat{\beta}\right\|^2 &\le 2\rho\left\|\hat{\beta} - \beta\right\|_1 + \lambda\left\|(\hat{\beta} - \beta)_S\right\|_1 - \lambda\left\|\hat{\beta}_{S^c}\right\|_1 \\ &\le 2\rho\left[\left\|(\hat{\beta} - \beta)_S\right\|_1 + \left\|\hat{\beta}_{S^c}\right\|_1\right] + \lambda\left\|(\hat{\beta} - \beta)_S\right\|_1 - \lambda\left\|\hat{\beta}_{S^c}\right\|_1\end{aligned}$$

We want to pick λ such that we get a good bound. We want to kill off the first term. We can choose $\lambda = 4\rho$. We then have:

$$\begin{aligned}\left\|\mathbf{X}\beta - \mathbf{X}\hat{\beta}\right\|^2 &\le 2\rho\left[\left\|(\hat{\beta} - \beta)_S\right\|_1 + \left\|\hat{\beta}_{S^c}\right\|_1\right] + \lambda\left\|(\hat{\beta} - \beta)_S\right\|_1 - \lambda\left\|\hat{\beta}_{S^c}\right\|_1 \\ &= 2\rho\left[3\left\|(\hat{\beta} - \beta)_S\right\|_1 - \left\|\hat{\beta}\right\|_1\right]\end{aligned}$$

Remark. For Dantzig (cf 2.2), we know that $\left\|\hat{\beta}_{S^c}\right\| \le \left\|\hat{\beta}_S\right\| = 1$. We want something similar for the LASSO. Note that because norms are non-negative, we have:

$$\left\|(\hat{\beta} - \beta)_{S^c}\right\|_1 \le 3\left\|(\hat{\beta} - \beta)_S\right\|_1$$

which gives us the control we need. Now, if we take a restricted eigenvalue condition:

$$\inf_{u:\|u_{S^c}\|_1 \le 3\|u_S\|_1} \frac{\|\mathbf{X}u\|^2}{n\|u\|^2} \ge \gamma_2 > 0$$

under which, we have:

$$\gamma_2 n\left\|\hat{\beta} - \beta\right\|^2 \le 2\rho\left[3\left\|(\hat{\beta} - \beta)_S\right\|_1 - \left\|\hat{\beta}\right\|_1\right]$$

with high probability, which implies:

$$\begin{aligned}\gamma_2 n\left\|\hat{\beta} - \beta\right\|^2 &\le 2\rho\left[3\left\|(\hat{\beta} - \beta)_S\right\|_1 - \left\|\hat{\beta}\right\|_1\right] \\ &\le 2\rho\left[3\left\|(\hat{\beta} - \beta)_S\right\|_1\right] \\ &\le 2\rho 3\sqrt{s}\left\|(\hat{\beta} - \beta)_S\right\|_2 \\ \Rightarrow \gamma_2 n\left\|\hat{\beta} - \beta\right\| &\le 6\rho\sqrt{s}\end{aligned}$$

Recall that we can bound the ℓ_1 norm from above from the ℓ_2 norm times the square root of the length.

CHAPTER 6

Bayesion Posterior Contraction: Positive Results

1. Background

We will focus on linear regression, but this can be extended to other models.

Consider the model:

$$Y_{n\times 1} = \mathbf{X}_{n\times p}\beta_{p\times 1} + \varepsilon_{n\times 1}$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_{n\times 1})$. This analysis can be extended to the subgaussian case.

Assumption: Let $\mathcal{S}^* = \operatorname{supp}(\beta^*)$, where β^* is the truth. This is the dimension or cardinality of β^*. We define $S^* = \mathcal{S}^*$. Further assume $\det(X_{\mathcal{S}}^\top X_{\mathcal{S}}) \neq 0$.

Our goal for today is to take a prior π and calculate the posterior. Then we consider the ball:

$$\pi(\|X\beta - \mathbf{X}\beta^*\|_2^2)$$

Recall that previously, we got the rate:

$$\mathbf{E}\left\|X\hat{\beta} - X\beta^*\right\| \leq CS^* \log \frac{ep}{S^*}$$

Our hope is that we can get the same result in a Bayesian setting:

$$\mathbf{E}_{Y|\beta^*}\, \pi\left(\|X\beta - \mathbf{X}\beta^*\|_2^2 \geq M^{\sigma^2} S^* \log \frac{ep}{S^*} |Y\right) \to 0$$

The rate depends on M, S^*, and p.

This implies that this is a good prior; we can get the same result as the Frequentist approach.

Remark. The result today is part of Gao, et al. (2016). The also consider graphon estimation, wavelet estimation, fourier basis estimation, etc. is a unified framework. This paper is an extention of Carstillo et al (2015, aos). Gao's paper uses a different prior, which allows for weaker conditions and a general framework.

Remark. Computational issues. For both papers, they can provide theoretical results, but they cannot compute. But Yang, Wainwright, Jordan (2016 aos). They propose a different prior, with MCMC convergence at a linear rate, but takes stronger assumptions.

Remark. Lasso prior:

$$\pi(\theta) \propto \exp\{-\lambda \|\theta_1\|\}$$

We can show that the posterior does not converge. This means that the LASSO prior fails. We should not use it (we do not have a theoretical guarantee).

If we look at MAP, then LASSO prior works. From a Bayesian point of view, though, people do not look at MAP (which only indicates a maximum, ignoring the rest of the distribution).

Harry thinks that there may be a general framework specifying which priors fail.

Remark. Prior mass and testing approach (Le Cam, Swartz, Barron). This approach often fails to get the optimal rate.

Consider $Y \sim \mathcal{N}(\theta, \frac{1}{n})$. What prior are we going to use to estimate θ under the ℓ_2 loss.

(1) Suppose we take $\pi : \theta \sim \mathcal{N}(0,1)$. Then we have $\mathbf{E}[\theta|X] = \left(1 + \frac{1}{n}\right)^{-1} Y$. If we use the posterior mean to estimate, it is always biased.

We want to estimate all β^*. Posterior contraction says that the mass outside the ball is exponentially small. But we have a bias which is:

$$\frac{\frac{1}{n}}{1+\frac{1}{n}}\theta^* = \frac{1}{1+n}\theta^*$$

There is no limit on the size of θ^*, which means the bias could be infinite. No matter how large Y is, the bias is always there. It is okay when we have an assumption on the size of θ^*.

Note that this bias is essentially the center of the random variable $\|\mathbf{X}\beta - \mathbf{X}\beta^*\|_2^2$.

(2) We can try a different prior (Laplace):

$$\pi : \pi(\theta) = \frac{1}{2}\exp\{-|theta|\}$$

If we use this prior, it can be shown that:

$$\mathbf{E}_{Y|\theta^*}\,\pi\left(|\theta - \theta^*|^2 \geq M\frac{1}{n}|Y\right) \to 0$$

But prior mass and testing fails to prove result above. Recall the first condition is (taking $\gamma_n^2 = \frac{1}{n}$.

$$\pi(|\theta - \theta^*| \leq c\gamma_n^2) \geq \exp\left\{-c_1 n\gamma_n^2\right\}$$

We want the integral over $\frac{1}{n}$ to be greater than a positive constant, which is impossible as $n \to \infty$.

Remark. David: For what kind priors does the bias go to zero as Y grows? (1980s).

2. Warm-up: Non-Sparse

Assume $S^* = p < n$. Then we are considering classical linear regression. Then we want to prove:

$$\mathbf{E}_{Y|\beta^*}\,\pi\left(\|\mathbf{X}\beta - \mathbf{X}\beta^*\|_2^2 \geq M^{\sigma^2}S^*Mp|Y\right) \to 0$$

How would we extend the Laplace prior? We may consider $\pi(\theta) \propto \exp\{-\|\theta\|_1\}$. This is what Carstillo uses, but needs to take on some ugly assumptions. We consider instead:

$$\pi(\beta) \propto \exp\{-\|\mathbf{X}\beta\|_2\}$$

Note that we don't square the ℓ_2 norm, which would just be Gaussian. This is called elliptical Laplace prior. What's the normalization constant?

$$\int \exp\left\{-\left\|\mathbf{X}\beta\right\|_2\right\} \mathrm{d}\,\beta = \frac{1}{\sqrt{\det(X^\top X)}} \int \exp\left\{-\left\|\theta\right\|_2\right\} \mathrm{d}\,\theta$$

Note that now $\theta \in \mathbf{R}^p$. We have reduced the dimension. θ lives in a lower dimensional space anyways, so we just rotate it down to a p dimensional space, and we have the same length.

Now, we take a polar transformation. Then we have:

$$\begin{aligned}\int \exp\left\{-\left\|\mathbf{X}\beta\right\|_2\right\} \mathrm{d}\,\beta &= \frac{1}{\sqrt{\det(X^\top X)}} \int \exp\left\{-\left\|\theta\right\|_2\right\} \mathrm{d}\,\theta \\ &= \frac{1}{\sqrt{\det \mathbf{X}^\top \mathbf{X}}} \int \exp\{-r\} r^{p-1} 2 \frac{\pi^{\frac{p}{2}}}{\Gamma(\frac{p}{2})} \mathrm{d}\,r\end{aligned}$$

Therefore, for all β^*, we have the integral:

$$\frac{1}{\sqrt{\det \mathbf{X}^\top \mathbf{X}}} \frac{2\pi^{\frac{p}{2}}\Gamma(p)}{\Gamma(\frac{p}{2})}$$

Assume without loss of generality that $\sigma = 1$.

Claim 5.

$$\mathbf{E}_{Y|\beta^*}\, \pi\left(\left\|\mathbf{X}\beta - \mathbf{X}\beta^*\right\|_2^2 \geq M^{\sigma^2} S^* M p | Y\right) \to 0$$

Proof. Observe that the LHS is equal to:

$$\mathbf{E}_{Y|\beta^*} \frac{C_N \int_\beta \exp\left\{-\frac{\|Y-\mathbf{X}\beta\|^2}{2}\right\} \exp\left\{-\left\|\mathbf{X}\beta\right\|_2\right\} \mathrm{d}\,\beta}{C_N \int \exp\left\{-\frac{\|Y-\mathbf{X}\beta\|^2}{2}\right\} \exp\left\{-\left\|\mathbf{X}\beta\right\|_2\right\} \mathrm{d}\,\beta} = \mathbf{E}_{Y|\beta^*} \frac{\int_\beta \exp\left\{-\frac{\|Y-\mathbf{X}\beta\|^2}{2}\right\} \exp\left\{-\left\|\mathbf{X}\beta\right\|_2\right\} \mathrm{d}\,\beta}{\int \exp\left\{-\frac{\|Y-\mathbf{X}\beta\|^2}{2}\right\} \exp\left\{-\left\|\mathbf{X}\beta\right\|_2\right\} \mathrm{d}\,\beta}$$

As Frequentists, we assume that $Y = \mathbf{X}\beta^* + \varepsilon$. This implies that:

$$\begin{aligned}&\mathbf{E}_{Y|\beta^*} \frac{\int_\beta \exp\left\{-\frac{\|Y-\mathbf{X}\beta\|^2}{2}\right\} \exp\left\{-\left\|\mathbf{X}\beta\right\|_2\right\} \mathrm{d}\,\beta}{\int \exp\left\{-\frac{\|Y-\mathbf{X}\beta\|^2}{2}\right\} \exp\left\{-\left\|\mathbf{X}\beta\right\|_2\right\} \mathrm{d}\,\beta} \\ &= \mathbf{E}_{\varepsilon|\beta^*} \frac{\int_\beta \exp\left\{-\frac{\|\mathbf{X}\beta^*+\varepsilon-\mathbf{X}\beta\|^2}{2}\right\} \exp\left\{-\left\|\mathbf{X}\beta\right\|_2\right\} \mathrm{d}\,\beta}{\int \exp\left\{-\frac{\|\mathbf{X}\beta^*+\varepsilon-\mathbf{X}\beta\|^2}{2}\right\} \exp\left\{-\left\|\mathbf{X}\beta\right\|_2\right\} \mathrm{d}\,\beta} \\ &= \mathbf{E}_{\varepsilon|\beta^*} \frac{\int_\beta \exp\left\{-\frac{\|\mathbf{X}\beta^*-\mathbf{X}\beta\|^2}{2} - \langle \mathbf{X}\beta^* - \mathbf{X}\beta, \varepsilon\rangle - \|\mathbf{X}\beta\|\right\}}{\int_\beta \exp\left\{-\frac{\|\mathbf{X}\beta^*-\mathbf{X}\beta\|^2}{2} - \langle \mathbf{X}\beta^* - \mathbf{X}\beta, \varepsilon\rangle - \|\mathbf{X}\beta\|\right\}}\end{aligned}$$

We note the denimator is lower bounded by (Triangle Inequality):

$$\begin{aligned}
&\int_\beta \exp\left\{-\frac{\|\mathbf{X}\beta^* - \mathbf{X}\beta\|^2}{2} - \langle \mathbf{X}\beta^* - \mathbf{X}\beta, \varepsilon\rangle - \|\mathbf{X}\beta\|\right\} \\
&\quad \geq \exp\left\{-\|\mathbf{X}\beta\|\right\} \int_\beta \exp\left\{-\frac{\|\mathbf{X}\beta^* - \mathbf{X}\beta\|^2}{2} - \langle \mathbf{X}\beta^* - \mathbf{X}\beta, \varepsilon\rangle - \|\mathbf{X}\beta - \mathbf{X}\beta^*\|\right\} \mathrm{d}\,\beta \\
&\quad = \exp\left\{-\|\mathbf{X}\beta^*\|\right\} \frac{1}{\sqrt{\det \mathbf{X}^\top \mathbf{X}}} \int \exp\left\{-\frac{\|\theta\|^2}{2} - \|\theta\| - \langle \theta, \varepsilon\rangle\right\} \mathrm{d}\,\theta \\
&\quad \geq \exp\left\{-\|\mathbf{X}\beta^*\|\right\} \frac{1}{\sqrt{\det \mathbf{X}^\top \mathbf{X}}} \int \exp\left\{-\frac{\|\theta\|^2}{2} - \|\theta\|\right\} \mathrm{d}\,\theta
\end{aligned}$$

How do we get the last line? Let us denote:

$$A = \int \exp\left\{-\frac{\|\theta\|^2}{2} - \|\theta\|\right\} \mathrm{d}\,\theta$$

Then we have:

$$\begin{aligned}
&\int \exp\left\{-\frac{\|\theta\|^2}{2} - \|\theta\| - \langle \theta, \varepsilon\rangle\right\} \mathrm{d}\,\theta \\
&\quad = A \int \exp\left\{-\langle \theta, \varepsilon\rangle\right\} \underbrace{\frac{\exp\left\{-\frac{\|\theta\|^2}{2} - \|\theta\|\right\}}{A} \mathrm{d}\,\theta}_{\text{Measure of Integration.}} \\
&\quad = A\, \mathbf{E}_{\theta'} \exp\left\{-\langle \theta, \varepsilon\rangle\right\} \\
&\quad \geq A \exp\left\{\mathbf{E}_{\theta'} -\langle \theta, \varepsilon\rangle\right\} \\
&\quad = A
\end{aligned}$$

And we see that our lower bound on the denominator is independent of the randomness in the problem.

Now, we consider the numerator:

$$\mathbf{E}_{\varepsilon|\beta^*} \int_\beta \exp\left\{-\frac{\|\mathbf{X}\beta - \mathbf{X}\beta^*\|^2}{2} - \langle \mathbf{X}\beta^* - \mathbf{X}\beta, \varepsilon\rangle - \|\mathbf{X}\beta\|\right\} \mathrm{d}\,\beta$$

The analysis here will be very similar to the analysis for all subsets selection.

$$\mathbf{E}_{\varepsilon|\beta^*}\int_\beta \exp\left\{-\frac{\|\mathbf{X}\beta-\mathbf{X}\beta^*\|^2}{2} - \langle \mathbf{X}\beta^* - \mathbf{X}\beta, \varepsilon\rangle - \|\mathbf{X}\beta\|\right\} \mathrm{d}\,\beta$$

$$= \mathbf{E}_{\varepsilon|\beta^*}\int_\beta \exp\left\{-\frac{\|\mathbf{X}\beta-\mathbf{X}\beta^*\|^2}{2} - \|\mathbf{X}\beta^*-\mathbf{X}\beta\| \underbrace{\langle \frac{\mathbf{X}\beta^*-\mathbf{X}\beta}{\|\mathbf{X}\beta^*-\mathbf{X}\beta\|}, \varepsilon\rangle}_{L} - \|\mathbf{X}\beta\|\right\}\mathrm{d}\,\beta$$

Note that $|L| \leq \|\varepsilon\| \leq 2n$ This has connections to χ^2. We can get a better bound if $n \gg p$. Then we can project onto a lower dimensional space and get the bound:

$$|L| \leq 2p$$

with high probability $\geq 1 - c_1 \exp\{-c_2 p\}$. This result can be proved with the singular value decomposition. Let us denote $H = \{|L| \leq 2p\}$. Then we expand the indicator over H and H^c. We return our attention to the original ratio. We have that:

$$\text{LHS } \leq \mathbf{E}_{\varepsilon|\beta^*}\,\mathbf{1}_{H^c} + \mathbf{E}_{\varepsilon|\beta^*}\,\frac{\mathbf{1}_H \text{ Numerator}}{\text{Denominator}}$$

We observe that because the fraction is at most 1, then we see that $\mathbf{E}\,\mathbf{1}_{H^c}$ is a valid upper bound. Now we consider the numerator.

$$\mathbf{E}_{\varepsilon|\beta^*}\,\mathbf{1}_H\int_B \exp\left\{-\frac{\|\mathbf{X}\beta-\mathbf{X}\beta^*\|^2}{2} - \|\mathbf{X}\beta^*-\mathbf{X}\beta\| \underbrace{\langle \frac{\mathbf{X}\beta^*-\mathbf{X}\beta}{\|\mathbf{X}\beta^*-\mathbf{X}\beta\|}, \varepsilon\rangle}_{L} - \|\mathbf{X}\beta\|\right\}\mathrm{d}\,\beta$$

$$\leq \mathbf{E}_{\varepsilon|\beta^*}\exp\{-\|\mathbf{X}\beta^*\|\}\int_B \exp\left\{-\frac{\|\mathbf{X}\beta-\mathbf{X}\beta^*\|^2}{2} - \sqrt{2p}\,\|\mathbf{X}\beta^*-\mathbf{X}\beta\| - \|\mathbf{X}\beta-\mathbf{X}\beta^*\|\right\}\mathrm{d}\,\beta$$

$$\leq \exp\{-\|\mathbf{X}\beta^*\|\}\int_B \exp\left\{-\frac{\|\mathbf{X}\beta-\mathbf{X}\beta^*\|^2}{2} - \sqrt{2p}\,\|\mathbf{X}\beta^*-\mathbf{X}\beta\| - \|\mathbf{X}\beta-\mathbf{X}\beta^*\|\right\}\mathrm{d}\,\beta$$

The expectation disappears because we replace ε terms with upper bound.

$$\exp\{-\|\mathbf{X}\beta^*\|\}\int_B \exp\left\{-\frac{\|\mathbf{X}\beta-\mathbf{X}\beta^*\|^2}{2} - \sqrt{2p}\,\|\mathbf{X}\beta^*-\mathbf{X}\beta\| - \|\mathbf{X}\beta-\mathbf{X}\beta^*\|\right\}\mathrm{d}\,\beta$$

Note that on the set B, $-\frac{\|\mathbf{X}\beta-\mathbf{X}\beta^*\|^2}{4} \leq -\frac{1}{4}MS^*\log\frac{ep}{S^*}$. Therefore, we have

$$\leq \exp\{-\|\mathbf{X}\beta^*\|\}\exp\left\{-\frac{M}{4}S^*\log\frac{ep}{S^*}\right\}\int \exp\left\{-\frac{\|\mathbf{X}\beta-\mathbf{X}\beta^*\|^2}{4} - (\sqrt{2p}-1)\,\|\mathbf{X}\beta^*-\mathbf{X}\beta\|\right\}\mathrm{d}\,\beta$$

$$= \exp\{-\|\mathbf{X}\beta^*\|\}\exp\left\{-\frac{M}{4}S^*\log\frac{ep}{S^*}\right\}\frac{1}{\sqrt{\det\mathbf{X}^\top\mathbf{X}}}\int \exp\left\{-\frac{\|\theta\|^2}{4} - (\sqrt{2p}-1)\,\|\theta\|^2\right\}\mathrm{d}\,\theta$$

Then, if we can show that for a large enough M, we have:

$$\frac{\int \exp\left\{-\frac{\|\theta\|^2}{4} - (\sqrt{2p}-1)\|\theta\|^2\right\} \mathrm{d}\,\theta}{\int \exp\left\{-\frac{\|\theta\|^2}{2} - \|\theta\|\right\} \mathrm{d}\,\theta} \leq \exp\left\{-\frac{M}{8}p\right\}$$

Observe that because the terms with β^* cancel out, we do not need to put any assumptions on β^*. □

3. Sparse Case

What prior should we use?

(1) First, we put a prior on the size of the model. We observe that we could have $\binom{p}{s}$ possible models. We should put a very small weight when the model size is large:

$$\pi(s) = \exp\left\{-Ds\log\frac{ep}{s}\right\}$$

for $D > 2$. This will let us obtain a union bound later on (over the $\binom{p}{s}$ models). Note that $\binom{p}{s} \leq \exp\left\{s\log\frac{ep}{s}\right\}$.

Harry believes a uniform prior would be better here, but that makes the analysis more difficult.

(2) For every given model size, we put a uniform prior $d(s)$ over the (at most) $\binom{p}{s}$ models.

(3) Finally, given $\mathcal{S}$, we use the elliptical Laplace prior:

$$\pi(\beta_{\mathcal{S}}) \propto \exp\{-\|\mathbf{X}_{\mathcal{S}}\beta_{\mathcal{S}}\|\}$$

We have

$$\pi(\beta_{\mathcal{S}}) = C_{N,\beta}\exp\{-\|\mathbf{X}_{\mathcal{S}}\beta_{\mathcal{S}}\|\}$$

so we choose:

$$\pi(s) = (2\pi)^{\frac{s}{2}}C_{N,\beta}^{-1}\exp\left\{-Ds\log\frac{ep}{s}\right\}$$

to tidy it up (and the Gaussian normalization). Finally, to ensure the third step is meaningful, we put the uniform mass only over $\mathcal{S}$ such that $\det \mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}} \neq 0$.

3.1. Step One.

Prove that:

$$\mathbf{E}_{Y|\beta^*}\,\pi(s > A_i S^*|Y) \leq C_1\exp\left\{-C_2 S^*\log\frac{ep}{S^*}\right\}$$

That is; indeed, this prior is doing model selection. With high probability, the model size will be less that A_1S^*.

What is this conditional probability?

$$\text{LHS } = \mathbf{E}_{Y|\beta^*}\frac{\sum_{s>AS^*}\pi(s)d(s)\sum_{\mathcal{S}:|\mathcal{S}|=s}\int\exp\left\{-\frac{\|Y-\mathbf{X}\beta_{\mathcal{S}}\|^2}{2} - \|\mathbf{X}\beta_{\mathcal{S}}\|\right\}\frac{1}{\sqrt{\det\mathbf{X}_{\mathcal{S}}^\top\mathbf{X}_{\mathcal{S}}}}}{\sum_s\pi(s)d(s)\sum_{\mathcal{S}:|\mathcal{S}|=s}\int\exp\left\{-\frac{\|Y-\mathbf{X}\beta_{\mathcal{S}}\|^2}{2} - \|\mathbf{X}\beta_{\mathcal{S}}\|\right\}\frac{1}{\sqrt{\det\mathbf{X}_{\mathcal{S}}^\top\mathbf{X}_{\mathcal{S}}}}}$$

How can we feasibly analyze this? Let's try bounding the denominator below by one term; that of the true model:

$$\text{Denominator } \geq \exp\left\{-DS^* \log \frac{ep}{S^*}\right\} d(S^*) \int \exp\left\{-\frac{\|Y - \mathbf{X}\beta_{S^*}\|^2}{2} - \|\mathbf{X}\beta_{\mathcal{S}^*}\|\right\} \frac{1}{\sqrt{\det \mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}}}}$$

This is a terrible lower bound, but it works because of the prior we have chosen.

3.2. Step Two. After we have done model selection, then what? We are now back in the classical model. The analysis is then nearly identical to the warmup step.

CHAPTER 7

Bayesian Posterior Contraction: Negative Results

In the previous chapter, we got positive results about how Bayesian posterior contraction achieves Frequestist rates. In this chapter, we consider negative results. Can we show that an entire class of priors does not produce good results? This would provide some theoretical structure and guide our search for best practices.

1. Lasso Prior Fails

If we use double-exponential prior, we don't get a good posterior contraction. We know that Lasso prior works with posterior mode. But Bayesians care about the entire posterior distribution. Saying a Bayesian approach works means saying that the posterior is concentrated about the truth. This is what we mean by "Lasso prior fails". Many papers in the past few years look at the mode, which is the Frequentist side, but Bayesians are interested in the whole distribution.

Let us consider the simplest model:

$$Y_i = \theta_i + Z_i$$

with $i = 1, \ldots, p$. We assume the parameter space:

$$\{\Theta = \{\theta : \|\theta\|_0 \leq s\}$$

Recall the Lasso (double exponential) prior:

$$\exp\{-\lambda \|\theta\|_1\}$$

typically with the choice of $\lambda = \sqrt{2 \log p}$. If we used lasso, we would pick this λ. We have discussed this before.

We claim that the Lasso prior fails. In what sense? Suppose we assume the truth $\theta^* = 0$. For this truth, the ideal estimator estimates every θ_i as zero.

We also know that for this parameter space, we can get the risk bound:

$$s \log p$$

(if we assume risk is scaled by $n^{-\frac{1}{2}}$, then this would be $\frac{s \log p}{n}$.

We can consider:

$$\pi\left(\|\theta - \theta^*\|^2 \leq c \frac{p}{\log p} \Big| Y\right)$$

Note that this is a huge ball. This corresponds to the risk upper bound for $s = \frac{p}{(\log p)^2}$. We can then evaluate at $\theta^* = 0$:

$$\mathbf{E}_{X|\theta^*=0}\, \pi\left(\|\theta - \theta^*\|^2 \leq c\frac{p}{\log p} \middle| Y \right)$$

If this is a good prior, this probability should go to 1 (this ball is huge!). It turns out that this probability goes to zero – even a huge ball cannot cover the truth.

Claim 6.

$$\mathbf{E}_{X|\theta^*=0}\, \pi\left(\|\theta - \theta^*\|^2 \leq c\frac{p}{\log p} \middle| Y \right) \to 0$$

We will eventually show this for a small $c \approx \exp\{-6\}$

Proof. Let us denote the set $B = \left\{ \theta : \|\theta - \theta^*\|^2 \leq c\frac{p}{\log p} \right\}$. We now consider, with $Y = Z \sim \mathcal{N}(0, \mathbf{I}_p)$[1]:

$$\begin{aligned}
&\mathbf{E}_{X|\theta^*=0}\pi\left(\|\theta - \theta^*\|^2 \leq c\frac{p}{\log p} \middle| Y \right) \\
&\quad= \mathbf{E}_{Y|\theta^*=0} \frac{\int_B \exp\left\{-\frac{\|Y-\theta\|^2}{2}\right\} \exp\left\{-\lambda \|\theta\|_1\right\} \mathrm{d}\,\theta}{\int \exp\left\{-\frac{\|Y-\theta\|^2}{2}\right\} \exp\left\{-\lambda \|\theta\|_1\right\} \mathrm{d}\,\theta} \\
&\quad= \mathbf{E}_Z \frac{\int_B \exp\left\{\langle Y, \theta\rangle - \frac{\|\theta\|^2}{2} - \lambda \|\theta\|_1\right\} \mathrm{d}\,\theta}{\int \exp\left\{\langle Y, \theta\rangle - \frac{\|\theta\|^2}{2} - \lambda \|\theta\|_1\right\} \mathrm{d}\,\theta} \\
&\quad\leq \mathbf{E}_Z \frac{\int_B \exp\left\{\langle Z, \theta\rangle - \frac{\|\theta\|^2}{2}\right\} \exp\left\{-\lambda \|\theta\|_1\right\} \mathrm{d}\,\theta}{\int \exp\left\{- \frac{\|\theta\|^2}{2} - \lambda \|\theta\|_1\right\} \mathrm{d}\,\theta} \qquad \text{(Jensen, see prev. Chapt., n.b. density symm)} \\
&\quad\leq \frac{\int_B \mathbf{E}_Z \exp\left\{\langle Z, \theta\rangle - \frac{\|\theta\|^2}{2}\right\} \exp\left\{-\lambda \|\theta\|_1\right\} \mathrm{d}\,\theta}{\int \exp\left\{- \frac{\|\theta\|^2}{2} - \lambda \|\theta\|_1\right\} \mathrm{d}\,\theta}
\end{aligned}$$

Note that $\mathbf{E}_Z \exp\left\{\langle Z, \theta\rangle - \frac{\|\theta\|^2}{2}\right\} = 1$. We can get this by completing the square (write out the integral, Z is standard normal). We therefore get:

$$\frac{\int_B \mathbf{E}_Z \exp\left\{\langle Z, \theta\rangle - \frac{\|\theta\|^2}{2}\right\} \exp\left\{-\lambda \|\theta\|_1\right\} \mathrm{d}\,\theta}{\int \exp\left\{- \frac{\|\theta\|^2}{2} - \lambda \|\theta\|_1\right\} \mathrm{d}\,\theta} = \frac{\int_{\|\theta\|^2 \leq c\frac{p}{\log p}} \exp\left\{-\lambda \|\theta\|_1\right\} \mathrm{d}\,\theta}{\int \exp\left\{- \frac{\|\theta\|^2}{2} - \lambda \|\theta\|_1\right\} \mathrm{d}\,\theta}$$

[1] This is the proof for $\theta^* = 0$. The analysis can of course be extended.

We want to show that this goes to zero. The intuition is that this is very close to:

$$\frac{\int_{\|\theta\|^2\le c\frac{p}{\log p}} \exp\left\{-\lambda\left\|\theta\right\|_1\right\}\mathrm{d}\,\theta}{\int \exp\left\{-\frac{\|\theta\|^2}{2}-\lambda\left\|\theta\right\|_1\right\}\mathrm{d}\,\theta} \approx \frac{\int_{\|\theta\|^2\le c\frac{p}{\log p}} \exp\left\{-\lambda\left\|\theta\right\|_1\right\}\mathrm{d}\,\theta}{\int \exp\left\{-\lambda\left\|\theta\right\|_1\right\}\mathrm{d}\,\theta}$$

Therefore we have θ_i distributed $\exp\{-\lambda|\theta_i|\}$. Then we have:

$$\begin{aligned}\mathbf{E}\,\theta_i &\propto \frac{1}{\lambda}\\ \mathbf{E}\left\|\theta\right\|^2 &\propto \frac{p}{\lambda^2}\\ &\propto \frac{p}{\log p}\end{aligned}$$

It's like exponential, and there isn't enough mass. That's the intuition; now let's do it more rigorously:

$$\begin{aligned}&\frac{\int_{\|\theta\|^2\le c\frac{p}{\log p}} \exp\left\{-\lambda\left\|\theta\right\|_1\right\}\mathrm{d}\,\theta}{\int \exp\left\{-\frac{\|\theta\|^2}{2}-\lambda\left\|\theta\right\|_1\right\}\mathrm{d}\,\theta}\\ &=\frac{\int_{\|\theta\|^2\le c\frac{p}{\log p}} \exp\left\{-\lambda\left\|\theta\right\|_1\right\}\mathrm{d}\,\theta}{\int_{\|\theta\|^2\le\frac{p}{\log p}} \exp\left\{-\frac{\|\theta\|^2}{2}-\lambda\left\|\theta\right\|_1\right\}\mathrm{d}\,\theta}\\ &\le\frac{\int_{\|\theta\|^2\le c\frac{p}{\log p}}\mathrm{d}\,\theta}{\exp\left\{-\frac{p}{2\log p}-\lambda\sqrt{p}\sqrt{\frac{p}{\log p}}\right\}\int_{\|\theta\|^2\le\frac{p}{\log p}}\mathrm{d}\,\theta} \qquad (\left\|\theta\right\|_1\le\sqrt{p}\left\|\theta\right\|_2)\\ &\le\exp\left\{\frac{p}{2\log p}+\sqrt{2}p\right\}\cdot\sqrt{c}^p \qquad (p\text{-dim ball}).\\ &\to 0\end{aligned}$$

Harry believes that we should be able to build from the intuition to give a unified framework with necessary and sufficient conditions for when a prior succeeds or fails. □

CHAPTER 8

High-Dimensional Inference

Let us consider the high-dimensional linear model:

$$Y_{n\times 1} = \mathbf{X}_{n\times p}\beta_{p\times 1} + \sigma\varepsilon_{n\times 1}$$

Oftentimes, we want to determine whether β_j is zero.

Remark. Under restricted eigenvalue condition, we may estimate:

$$\mathbf{E}\left\|\hat{\beta} - \beta\right\|_2^2 \leq c\frac{s\log p}{n}$$

Homework: prove:

$$\mathbf{E}\left\|\hat{\beta} - \beta\right\|_1 \leq cs\sqrt{\frac{\log p}{n}}$$

with the lasso procedure.

What we are considering today is not estimating the whole β; we are only estimating β_j. Without loss of generality, we take $j = 1$. We want to determine, is this false discovery or not, and how much confidence do we have in this statement?

For $p < n$, we assume $\det(\mathbf{X}^\top\mathbf{X}) \neq 0$. Then we can get an ordinary least squares estimator for the whole vector with:

$$\hat{\beta}_{OLS} - \beta \sim \mathcal{N}(0, \frac{1}{n}(n^{-1}\mathbf{X}^\top\mathbf{X})^{-1})$$

What about $p > n$? If $X_1 \perp X_j$ for all $j \neq 1$, then we have:

$$X_1^\top Y = X_1^\top X_1\beta_1 + X_1^\top\varepsilon$$

from which it follows:

$$\hat{\beta}_1 = \frac{X_1^\top Y}{\|X_1\|^2}$$

(this is again ordinary least squares).

But what if this were not true (we do not have this orthogonality)? Today we present an idea Cun-Hui Zhang from five-years ago. We define:

$$Z_1 = X_1 - P_{X_{-1}}X_1$$

where X_{-1} is the space formed by $X_2, \ldots, X_p$. That is, we are projecting X_1 the space spanned by all other problems. We define Z_1 to be the residual, which *is* orthogonal to all other columns. If Z_1 is not equal to zero, then we can use the original idea:

$$Z_1^\top Y = Z_1^\top X_1\beta_1 + Z_1^\top\varepsilon$$

We can then estimate β_1 by:

$$\hat{\beta}_1 = \frac{Z_1^\top Y}{Z_1^\top X_1}$$

This, and X_1 orthogonal, are the two ideal cases.

Homework: show that this is an ordinary least squares estimator for β_1.

But what if $Z_1 = 0$? We cannot just drop X_1, because we want to make sure to keep the smallest subset of representative columns. We set aside this issue for now.

Consider instead the nearly orthogonal case. Suppose $X_{ij} \sim \mathcal{N}(0,1)$ [1]. Then

$$\mathbf{E}\, X_i^\top X_j = 0$$

i.e., they are orthogonal in the population sense. Moreover,

$$\sqrt{\operatorname{var}(X_i^\top X_j)} \asymp \sqrt{n}$$

Now, we have:

$$X_1^\top Y = X_1^\top X_1 \beta_1 + \sum_{j=2}^{p} X_1^\top X_j \beta_j + X_1^\top \varepsilon$$

Then a natural estimator is:

$$\hat{\beta}_1 = \frac{X_1^\top Y}{X_1^\top X_1} = \frac{X_1^\top Y}{\|X_1\|^2}$$

But how large is the error:

$$\sum_{j=2}^{p} X_1^\top X_j \beta_j$$

We can look at the difference:

$$\hat{\beta}_1 - \beta_1 = \underbrace{\frac{\sum_{j=2}^{p} X_1 X_j^\top \beta_j}{\|X_1\|^2}}_{\text{(Could be huge!)}} + \underbrace{\frac{X_1^\top \varepsilon}{\|X_1\|^2}}_{(\mathcal{N}(0, \frac{1}{\|X_1\|^2})^{-1})}$$

Because the errors could be very large, this may be bad for inference.

Remark. By central limit theory,

$$\frac{X_1 X_j^\top}{\|X_1\|^2} \approx \mathcal{N}(0, \frac{C_j}{n})$$

We can obtain $\hat{\beta}^{\text{initial}}$ by lasso or another approach. Suppose we want to do inference for β_1. Then we will do:

$$Y - \sum_{j=2}^{p} X_j \hat{\beta}_j^{\text{initial}} = X_1 \beta_1 + \sum_{j=2}^{p} X_j (\beta_j - \hat{\beta}_j^{\text{initial}}) + \varepsilon$$

[1] This is very good design; you can prove restricted eigenvalue condition, etc., all with high probability.

Now, we do:

$$X_1^\top \left(Y - \sum_{j=2}^{p} X_j \hat{\beta}_j^{\text{initial}} \right) = X_1^\top X_1 \beta_1 + \sum_{j=2}^{p} X_1^\top X_j (\beta_j - \hat{\beta}_j^{\text{initial}} + X_1^\top \varepsilon$$

$$\frac{X_1^\top \left(Y - \sum_{j=2}^{p} X_j \hat{\beta}_j^{\text{initial}} \right)}{\|X_1\|^2} = \beta_1 + \frac{\sum_{j=2}^{p} X_1^\top X_j (\beta_j - \hat{\beta}_j^{\text{initial}})}{\|X_1\|^2} + \frac{X_1^\top \varepsilon}{\|X_1\|^2}$$

which gives us the new estimator:

$$\hat{\beta}^{\text{new}} = \frac{X_1^\top \left(Y - \sum_{j=2}^{p} X_j \hat{\beta}_j^{\text{initial}} \right)}{\|X_1\|^2}$$

We can then see the original procedure as a special case where $\hat{\beta}^{\text{initial}} = 0$. We effectively shifted the mass of the "error" onto the estimator itself (?).

Now, let us consider the error term:

$$\hat{\beta}_1 - \beta_1 = \underbrace{\sum_{j=2}^{p} \frac{X_1 X_j^\top}{\|X_1\|^2} (\beta_j - \hat{\beta}_j^{\text{initial}})}_{R} + \mathcal{N}(0, \frac{1}{\|X_1\|^2})$$

We want to control R. We claim that:

$$\begin{aligned} R &\leq \max_j \left| \frac{X_1 X_j^\top}{\|X_1\|^2} \right| \sum_{j=2}^{p} |\beta_j - \hat{\beta}_j^{\text{initial}}| \\ &\leq \underbrace{\max_{2 \leq j \leq p} \frac{|X_1 X_j^\top|}{\|X_1\|^2}}_{\leq C\sqrt{\frac{\log p}{n}}} \underbrace{\left\| \hat{\beta} - \beta \right\|_1}_{\leq s\sqrt{\frac{\log p}{n}}} \\ &\leq \frac{C_1 s \log p}{n} \\ &= o(n^{-\frac{1}{2}}) \end{aligned}$$

with high probability when $s = o\left(\frac{\sqrt{n}}{\log p} \right)$; that is, if our model is not too large (relative to n), then we will not fail. Ren et al (?) shows that this condition cannot be avoided.

But this is not the end of the story. The design is not necessarily $X_{ij} \sim \mathcal{N}(0, 1)$. What about the general design? That means we have to replace X_1 with something else. Our procedure can be:

(1) Start with good (bounded distance from truth) $\hat{\beta}^{\text{initial}}$
(2)

$$Z_1^\top (Y - \sum_{j=2}^{p} X_j \hat{\beta}_j^{\text{initial}}) = Z_1^\top X_1 \beta_1 + \sum_{j=2}^{p} Z_1^\top X_j (\beta_j - \hat{\beta}_j^{\text{initial}}) + Z_1^\top \varepsilon$$

and we divide through by $Z_1^\top X_1$ to get our estimator.

Note that we want to find Z_1 such that $Z_1^\top X_1$ is sufficiently large, and $Z_1^\top X_j$ small, to make the errors small enough. We can do this by projection (regression on the rest of the columns):

$$\frac{\left\|X_1 - \sum_{j=2}^{p} X_j \gamma_j\right\|^2}{2} + \lambda \left\|\gamma\right\|_1$$

We want $\hat{\gamma}$ to minimize the above quantity. Then we have:

$$Z_1 = X_1 - \sum_{j=2}^{p} X_j \hat{\gamma}_j$$

that is, we use the residual as Z_1.

The KKT conditions then give the following bound by taking the derivative:

$$|X_j^\top Z_1| \leq \lambda$$

CHAPTER 9

Large Covariance & Precision Matrix Estimation: Upper Bounds

In classical parametric models, the number of parameters is finite. But what if the number of parameters is growing, like in sparse linear models, and nonparametric function estimation?

Up until now, we have been talking about mean estimation. Now we will discuss variance estimation.

1. Motivation

Suppose we observe $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{I}_p)$. We can calculate the sample covariance $\hat{\Sigma}$. (Note that p can be large – a function of n, or as large of n). Then the sample covariance is:

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top$$

In the classical setting, this would be a good estimator of the population covariance. It turns out that this may not be a good estimator.

Remark. Some may say this is a good estimator because it is unbiased. But it has some problems.

One problem is that the eigenvalues lie in:

$$\left(1 - \sqrt{\frac{p}{n}}\right)^2 \leq \lambda_i \leq \left(1 + \sqrt{\frac{p}{n}}\right)^2$$

when really they are all one.

Theorem 1.1 (Marchenko-Pastur). *Suppose $\frac{p}{n} \to \lambda \leq 1$. We can find the empirical eigenvalues $\hat{\lambda}_i$. They studied the empirical distribution of the eigenvalues and found that:*

$$\frac{1}{p}\#\{i : \hat{\lambda}_i \in A\} \to \nu(A) = \int_A \nu(x)\,\mathrm{d}\,x$$

where

$$\nu(x) = \frac{1}{2\pi}\frac{\sqrt{(\lambda_+ - x)(x - \lambda_-)}}{\lambda x}$$

where $\lambda_+ = (1 + \sqrt{\lambda})^2$ and $\lambda_- = (1 - \sqrt{\lambda})^2$.

We see that the distribution of the empirical eigenvalues is very different from the true eigenvalues when $\frac{p}{n}$ is very large.

2. Eigenvectors

Consider the special case where the population Σ_p is identity with the first entry $l > 1$.[1] If we do principal components analysis, then we find that the first eigenvector has the largest eigenvalue.

We could also calculate the first principal component with $\hat{\Sigma}_p$ by way of the singular value decomposition. This may be expressed as:

$$\hat{\Sigma}_p = \sum_{i=1}^{p} \hat{\lambda}_i \hat{u}_i \hat{u}_i^\top$$

Suppose we have $\hat{\lambda}_i > \hat{\lambda}_j$, $i < j$. Then $\hat{\lambda}_1$ is the largest eigenvalue.

It turns out that when p is large, $\langle \hat{u}_1, u_1 \rangle \to 0$ almost surely when $l - 1$ is small $(l - 1 \leq \sqrt{\lambda})$. This result can be found in some papers by Iain Johnstone.

This suggests we have to go beyond the sample covariance. There's a lot of research on how to improve $\hat{\Sigma}$.

Remark. Wall Street often care about low-rank covariance matrix analysis, which suggests that there are latent factors. This is very popular, and Donoho, Johnstone, have produced some research regarding low-rank matrix analysis.

Because the eigenvalues are very large, they suggest shrinkage of the eigenvalues.

3. Covariance Matrix Estimation

Suppose we assume the covariance is identity (due to data normalization) plus a low-rank perturbation:

$$\Sigma_p = \mathbf{I}_p + \sum_{i=1}^{k} \lambda_k u_i u_i^\top$$

In this case, one possible estimator is:

$$\hat{\Sigma}_p^{\text{new}} = \omega \mathbf{I}_p + (1 - \omega)\hat{\Sigma}_p$$

This kind of shrinkage is very popular (the eigenvalues $\hat{\lambda}_i, \ldots, \hat{\lambda}_k$ are subject to shrinkage).

4. Structured Covariance Matrices

Let $\Sigma_{p \times p} = (\sigma_{ij})_{p \times p}$. One possible structure is diagonal.

[1] In the Gaussian case, we can always make this diagonal subject to a rotation.

4.1. Bandable. Another possible structure, motivated by time series, is bandable. We assume that σ_{ij} decays to zero as we move away from the diagonal:

$$|\sigma_{ij}| \leq M(1 + |i - j|)^{-\alpha-1}, \alpha > 0$$

For the purposes of this exposition, we take the Gaussian assumption. This assumes we have a few big entries, and the rest are close to zero.

4.2. Sparse. Sparse assumption can be seen as a relaxation of bandable. One possible assumption (row-sparse) may be:

$$\#\{i : \sigma_{ij} \neq 0\} \leq s$$

for every j.

5. Structured Precision Matrices

We define $\Omega_{p\times p} = \Sigma_{p\times p}^{-1}$.

5.1. Sparse. We may assume:

$$\#\{i : \omega_{ij} \neq 0\} \leq s$$

for all j. under the Gaussian assumption, the nonzero entries give us a Gaussian Graphical Model.

5.2. Bandable. The bandable assumption can also be applied to the precision matrix.

6. Estimation

Suppose we take on these extra assumptions. How would we estimate the covariance matrix?

We start with the sample covariance:[2]

$$\hat{\Sigma}_{\text{sample}} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top$$

where each entry is given by:

$$\begin{aligned}\hat{\sigma}_{ij} &= \frac{1}{n}\sum_{k=1}^{n} x_{ki}x_{kj} \\ &\approx \mathcal{N}(\sigma_{ij}, \frac{1}{n}(\sigma_{ii}\sigma_{jj} + \sigma_{ij}^2))\end{aligned}$$

which suggests that looking at eigenvalues, the sample covariance is not great, but it does well as an entrywise estimator.

[2] The unbiased estimator scaled by $n-1$ stems from not taking the assumption of $\mu = 0$, which requires a projection of data and a loss of a degree of freedom.

We tend to use the operator norm to measure the performance of an estimator:

$$\|A\|_{\text{op}} = \sup_{v \in \mathbf{R}^p} \frac{\|A_{p\times p}\nu\|_2}{\|\nu\|_2}$$

which gives us the largest absolute eigenvalue when A is symmetric. We can then define the loss function as:

$$\left\|\hat{\Sigma} - \Sigma\right\|_{\text{op}}^2$$

we can also use the Frobenius norm. We tend to use the operator norm due to Davis-Kahan theory, which says that for an estimator:

$$|\hat{\lambda}_i - \lambda_i| \leq \left\|\hat{\Sigma} - \Sigma\right\|_{\text{op}}$$

where the $\hat{\lambda}_i, \lambda_i$ are ranked decreasing. We then have a corresponding eigenspace $\nu_i, \hat{\nu}_i$, assuming that the λ_i are distinct. It turns out that:

$$\min\{\|-\nu_i - \hat{\nu}_i\|_2, \|\nu_i - \hat{\nu}_i\|\} \leq \frac{2^3 \left\|\hat{\Sigma} - \Sigma\right\|_{\text{op}}}{\min\{\lambda_i - \lambda_{i-1}, \lambda_{i+1} - \lambda_i\}}$$

which says that we have a good control of the angle of the principal components, if there is a good eigengap, and if the operator norm goes to zero.

7. Bandable Covariance Estimation

This is similar to what we did at the beginning of the semester; we have a function, expand it into Fourier basis[3], what did we do to the coefficients as the index becomes larger? We cut off at some index (here, a band) and estimate the rest as zero. We can call this a banding estimator with a band of k.

7.1. Analysis.

Let us analyze $\left\|\hat{\Sigma}_k - \Sigma\right\|_{\text{op}}$:

$$\begin{aligned}\left\|\hat{\Sigma}_k - \Sigma\right\|_{\text{op}}^2 &= \left\|\hat{\Sigma}_k - \Sigma_k + \Sigma_k - \Sigma\right\|_{\text{op}}^2 \\ &\leq 2\left(\left\|\hat{\Sigma}_k - \Sigma_k\right\|_{\text{op}}^2 + \|\Sigma_k - \Sigma\|_{\text{op}}^2\right)\end{aligned}$$

How do we bound this? We can use the fact that, with high probability,

$$\left\|\hat{\Sigma}_k - \Sigma\right\|_{\text{op}}^2 \leq C\left(\frac{k}{n} + \frac{\log p}{n}\right)$$

The proof of the variance relies on packing in some space with random matrices. This is similar to the bound we had before on vector estimation where we pick k entries.

[3] Function in a Sobolev ball.

Why is there a $\log p$ term? If $k = 1$, then the operator norm is the maximum absolute difference. Then the difference is on the order of $\sqrt{\frac{\log p}{n}}$, and then we square it.

How can we control the operator norm of the bias? We can use another fact, where for A symmetric:

$$\|A\|_{\text{op}} \le \|A\|_1 \triangleq \max_j \sum_i |a_{ij}|$$

Then, we have:

$$\begin{aligned}\|\Sigma_k - \Sigma\|_{\text{op}} &\le \max_j \sum_{i:|i-j|\ge k} |\sigma_{ij}| \\ &\le 2M \sum_{m\ge k} (1+m)^{-\alpha-1}\end{aligned}$$

If we square everything, and apply the geometric sum[4], then we have:

$$\|\Sigma_k - \Sigma\|_{\text{op}}^2 \le C_2 k^{-2\alpha}$$

This is exactly the same result of the smooth function estimation with smoothness α where we only estimate the first k terms. We get the same results for the variance and the bias. The analysis is nearly identical! Here, as in there, we pick a k to minimize the bias and variance. We pick on the order of $k = n^{\frac{1}{1+2\alpha}}$. We have known how to pick k for nonparametric function estimation for a long time; now we see that we do the same for high-dimensional covariance matrix estimation.

Remark. Does this produce the optimal k under cross validation? We do not necessarily known M, α. A positive or negative result would both be interesting.

8. Sparse Covariance Estimation

Suppose we have a sample covariance. We know that asymptotically:

$$\hat{\sigma}_{ij}^{\text{sample}} \approx \mathcal{N}(\sigma_{ij}, \frac{1}{n}(\sigma_{ii}\sigma_{jj} + \sigma_{ij}^2))$$

We could use:

$$\hat{\sigma}_{ij} = \begin{cases} \hat{\sigma}_{ij}^{\text{sample}} & \text{for } \left|\frac{\hat{\sigma}_{ij}^{\text{s}}}{\sqrt{\frac{1}{n}(\hat{\sigma}_{ii}^{\text{s}}\hat{\sigma}_{jj}^{\text{s}} + \hat{\sigma}_{ij}^{\text{s}2}}}\right| \ge \lambda \\ 0 & \text{otherwise}\end{cases}$$

Here, we have an unbiased estimator to start with.

9. Sparse Precision Matrix

In the precision matrix case, we do not have an unbiased estimator to start with. When $p > n$, the sample covariance is not even invertible.

[4] $\sum_{m\ge k}^{\infty} \le C_\beta k^{-\beta+1}$

CHAPTER 10

Large Covariance & Precision Matrix Estimation: Lower Bounds

Recall that we discussed three kinds of matrices:

(1) Bandable covariance
(2) Sparse covariance
(3) Sparse precision

We will discuss the procedures and state the results.

1. Bandable Covariance

We have already discussed the bandable case and proposed a banding procedure in the previous lecture. Recall that the parameter space was:

$$\Sigma = (\sigma_{ij})_{p\times p}, \qquad |\sigma_{ij}| \le M(1+|i-j|)^{-\alpha-1}$$

where the observations are i.i.d. $\mathcal{N}(0,\Sigma)$. For bandable matrices, we estimated under the spectral norm, and with very high probability, we attained the upper bound:

$$\left\|\hat{\Sigma}-\Sigma\right\|_2^2 \le C\left(n^{-\frac{2\alpha}{2\alpha+1}} + \frac{\log p}{n}\right)$$

This improves upon a bound given by Bickel and Levina.

2. Sparse Covariance

We assume the number of nonzero entries for each row and column is small:

$$\max_j |\{i : \sigma_{ij} \neq 0, 1 \le i \le p\}| \le s \ll p$$

Typically, we assume the largest eigenvalue is bounded above: $\lambda_{\max}(\Sigma) \le M$. What estimator would we use? Start with the sample covariance:

$$\hat{\Sigma}_S = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top$$

Recall that the sample covariance is unbiased. What is the variance of each entry?

$$\operatorname{var}(\hat{\sigma}_{ij}^S) = \frac{\sigma_{ii}\sigma_{jj}+\sigma_{ij}^2}{n}$$

One possible estimator is to threshold. We can set:

$$\hat{\sigma}_{ij} = \begin{cases} |\hat{\sigma}^S_{ij}| & \text{for } |\hat{\sigma}^S_{ij}| > \lambda \\ 0 & \text{otherwise} \end{cases}$$

We can choose λ on the order of the standard deviation.

Another idea is to rank:

$$\frac{|\hat{\sigma}^S_{ij}|}{\sqrt{\frac{1}{n}\left(\hat{\sigma}^S_{ii}\hat{\sigma}^S_{jj} + \hat{\sigma}^{S^2}_{ij}\right)}}$$

and pick the s largest. Theoretically, we can prove that:

$$\frac{\hat{\sigma}^S_{ij} - \sigma_{ij}}{\sqrt{\frac{1}{n}\left(\hat{\sigma}^S_{ii}\hat{\sigma}^S_{jj} + \hat{\sigma}^{S^2}_{ij}\right)}}$$

behaves like a $\mathcal{N}(0,1)$ random variable (we can find a random variable that is very similar).

If all $\sigma_{ij} = 0$ for $i \neq j$, then we have:

$$\max_{i \neq j} \frac{\hat{\sigma}^S_{ij}}{\sqrt{\frac{1}{n}\left(\hat{\sigma}^S_{ii}\hat{\sigma}^S_{jj} + \hat{\sigma}^{S^2}_{ij}\right)}} \approx \sqrt{2 \log \binom{p}{2}} \approx 2\sqrt{\log p}$$

We can then formulate a test:

$$Y_{ij} = \frac{|\hat{\sigma}^S_{ij}|}{\sqrt{\frac{1}{n}\left(\hat{\sigma}^S_{ii}\hat{\sigma}^S_{jj} + \hat{\sigma}^{S^2}_{ij}\right)}} \geq 2\sqrt{\log p}$$

We could try estimating $\mu_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj} + \sigma^2_{ij}}}$ with penalized least squares, as in the Gaussian sequence model:

$$\sum_{i \neq j} (Y_{ij} - \mu_{ij})^2 + 2\,\|\mu\|_0 \log \frac{e\binom{p}{2}}{\|\mu\|_0}$$

But this is harder to analyze, because the squared penalty is harder to connect to the spectral norm, and because there is dependency between Y_{ij}, whereas the observations in the Gaussian sequence model were assumed to be independent.

Fact 2.1. With high probability:

$$|\hat{\sigma}_{ij} - \sigma_{ij}| \leq \begin{cases} C\sqrt{\frac{\log p}{n}} & \text{for } \sigma_{ij} \neq 0 \\ 0 & \text{w.h.p. if } \sigma_{ij} \neq 0 \end{cases}$$

The sample covariance has better error for case 1, but worse performance for case 2, which is why we use this (thresholding) procedure. It performs better when the true value is zero, which is more common than nonzero. Suppose we can prove this result (with C very close to 2). How can we control the spectral norm? Recall that for symmetric A, we may bound:

$$\|A\|_{\text{op}} \leq \max_j \sum_i |A_{ij}|$$

then we may bound:

$$\begin{aligned}\left\|\hat{\Sigma} - \Sigma\right\|_2^2 &\leq \left(\max_j \sum_i |\hat{\sigma}_{ij} - \sigma_{ij}|\right)^2 \\ &\leq C^2 \cdot \frac{s^2 \log p}{n}\end{aligned}$$

The matching lower bound proof is very difficult. A more elegant proof may be possible through random matrix theory.

3. Sparse Precision Matrix

Recall the precision matrix $\Omega = \Sigma^{-1}$. We assume:

$$\Omega = (\omega_{ij})_{1 \leq i,j \leq p}$$

where we assume:

$$\max_j |\{i : \omega_{ij} \neq 0\}| \leq s$$

If we have $Y \sim \mathcal{N}(0, \Omega^{-1})$, then the joint density for Y is given by:

$$f(y) \propto \exp\left\{-\frac{1}{2}\sum_{i,j} y_i y_j \omega_{ij}\right\}$$

With this formula, it is easy to show that if $\omega_{12} = 0$, then $Y_1 \perp\!\!\!\perp Y_2 | Y_3, \ldots, Y_p$. Moreover, we know that $Y_1, Y_2 | Y_3, \ldots, Y_p$ is distributed:

$$\mathcal{N}\left(-\begin{bmatrix}\omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22}\end{bmatrix}^{-1}\begin{bmatrix}\omega_{13} & \ldots & \omega_{1p} \\ \omega_{23} & \ldots & \omega_{2p}\end{bmatrix}\begin{bmatrix}Y_3 \\ \vdots \\ Y_p\end{bmatrix}, \begin{bmatrix}\omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22}\end{bmatrix}^{-1}\right)$$

How would we estimate Ω? A natural proposal is to do:

$$\hat{\Omega}^S = \left(\hat{\Sigma}^S\right)^{-1}$$

and then propose a thresholding procedure. But sparsity in a matrix does not imply sparsity in its inverse. We probably want to estimate the precision directly. Other issues:

(1) Not invertible when $p > n$.
(2) Claim: $\mathbf{E}\,\hat{\Omega}^S \neq \Omega$.
(3) Claim: $\mathbf{E}\,\hat{\Omega}^S = \frac{n}{n-p-1}\Omega$ (very biased when p is close to n).

What can we do? One way is the graphical lasso. We formulate a penalized estimation procedure based on the log-likelihood:

$$-\log \ell(\Omega) + \lambda \sum_{i \neq j} |\omega_{ij}|$$

It's not a very good approach though, because we are treating the whole matrix as a vector. To justify theoretically, we typically need very strong conditions for Ω. What can we do instead? We can exploit the conditional likelihood. Suppose we observe $X_1, \ldots, X_n \in \mathbf{R}^p$. The conditional distribution of X_{i1}, X_{i2} given the rest is normal. Therefore, we can formulate a least squares problem:

$$(X_{i1}, X_{i2}) = \underbrace{(X_{i3}, \ldots, X_{ip})}_{X_{-12} = \tilde{X}} \beta + Z_i$$

where $\beta = -\begin{bmatrix}\omega_{13} & \omega_{23} \\ \vdots & \vdots\end{bmatrix} \begin{bmatrix}\omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22}\end{bmatrix}^{-1}$ and where the noise is: $\epsilon_i \sim \mathcal{N}\left(0, \begin{bmatrix}\omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22}\end{bmatrix}^{-1}\right)$. Hypothetically, if we observe ϵ_i, then we could estimate $\Omega^{-1}_{1,2}$ with the sample covariance:

$$\widehat{\begin{bmatrix}\omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22}\end{bmatrix}}^{\text{oracle}} = \frac{1}{n} \sum_{i=1}^{n} \epsilon_i^\top \epsilon_i$$

If we observe the noise, then its a trivial two-dimensional sample covariance estimation. We can prove asymptotic normality for this estimator.

But we do not know the ϵ_i directly. But we can run regresion, knowing that β is up to $2s$-sparse. We can run lasso, and show that with high probability[1]:

$$\left\|\tilde{X}\hat{\beta} - \tilde{X}\beta\right\|_2^2 \leq c\frac{s \log p}{n}$$

With $\hat{\beta}$, we can find $\hat{\epsilon}_i$. With $\epsilon \in \mathbf{R}^{p \times 2}$, we can then get:

$$\|\hat{\epsilon} - \epsilon\|_2^2 \leq c\frac{s \log p}{n}$$

And use the estimator:

$$\widehat{\begin{bmatrix}\omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22}\end{bmatrix}}^{-1} = \frac{1}{n} \sum_{i=1}^{n} \hat{\epsilon}_i^\top \hat{\epsilon}_i$$

Note that the sparsity assumption is very important. We can also prove asymptotic normality when $\frac{s \log p}{n} = o(n^{-\frac{1}{2}})$.

$$\frac{\hat{\omega}_{ij} - \omega_{ij}}{\sqrt{\frac{1}{n}\underbrace{(\omega_{ii}\omega_{jj} + \omega_{ij}^2)}_{\hat{I}_{ij}}}} \approx \mathcal{N}(0, 1)$$

[1] Restricted eigenvalue condition, etc.

which is the same as what we got for covariance estimation. Therefore, we can formulate a thresholding procedure:

$$\hat{\omega}_{ij}^{\text{thre}} = \begin{cases} \hat{\omega}_{ij} & \text{for } |\hat{\omega}_{ij}| > \lambda_{ij} \\ 0 & \text{otherwise} \end{cases}$$

We can set a signal-to-noise ratio:

$$2\sqrt{\log p}\sqrt{\frac{1}{n}\hat{I}_{ij}}$$

We can then prove a similar property as before. With high probability:

$$|\hat{\omega}_{ij}^{\text{thre}} - \omega_{ij}| \leq \begin{cases} C\sqrt{\frac{\log p}{n}} & \text{for } \omega_{ij} \neq 0 \\ 0 & \text{for } \omega_{ij} = 0 \end{cases}$$

And it is possible to prove, as in the covariance estimation case:

$$\left\|\hat{\Omega}^{\text{thre}} - \Omega\right\|_{\text{op}}^2 \leq \left(\max_j \sum_i |\hat{\omega}_{ij}^{\text{thre}} - \omega_{ij}|\right)^2 \leq C^2 \frac{s \log p}{n}$$

Why do we need the assumption that $\frac{s\log p}{n} = o(\frac{1}{\sqrt{n}})$. Recall that we need to estimate the residuals. We need s to be small for the estimated residuals to be truly close to the true residuals. Note that we can prove that it is necessary to take this assumption.

CHAPTER 11

Sparse PCA & the Hidden Clique Problem

Consider the ideal model $X_1, \ldots, X_n$, drawn i.i.d. from a $\mathcal{N}(0, \Sigma)$ distribution, where $\Sigma_{p\times p} = \mathbf{I}_{p\times p} + \lambda u u^\top$, where $u \in \mathbf{R}^{p\times 1}$. That is, the covariance matrix is identity, plus a rank-1 matrix. Asuume that $\lambda > 0$. The goal for us is to estimate u, subject to it being in the ℓ_2 ball.

Remark. This idea can easily be extended to a more general case:

$$\Sigma_{p\times p} = \Sigma_0 + \sum_{r=1}^{R} \lambda_r u_r u_r^\top + \sum_{r=R+1}^{p} \lambda_r u_r u_r^\top$$

where $\lambda_R - \lambda_{R+1} \geq \gamma > 0$, and the λ_r are ordered. Here we want to recover a rank-R principal subspace.

Because p is high-dimensional, it may be imposible to estimate u. If we look at the sample covariance, we can calculate the principal components, and it's possible realize that the principal components are orthogonal to the truth. We may have to assume some conditions to estimate u, e.g., $\|u\|_0 \leq s \ll p$, i.e., a sparsity assumption.

We will discuss three topics:

(1) Minimax rate. Assume $\hat{u}$ has a positive angle with u; otherwise, flip it.

$$\inf_{\hat{u}} \sup_{\|u\|_0 \leq s} \|\hat{u} - u\|_2^2 \asymp \frac{s \log \frac{ep}{s}}{n\lambda^2}$$

If λ is constant, this is equivalent to the rate for sparse linear regression, sparse vector model. This also tells us that it is easier to recover u for large λ, and nearly impossible when λ is too small. The lower bound will be very similar to the lower bound for the sparse vector model. The algorithm for the upper bound is not computable (in polynomial time).

(2) Under the assumption $\frac{s^2 \log \frac{ep}{s}}{n\lambda^2} \to 0$ we can find a polynomial time algorithm to attain the rate $\frac{s \log \frac{ep}{s}}{n\lambda^2}$. Suppose $\lambda = 1$; we then need s^2 far smaller than n (very tough). Can we relax this assumption?

(3) $\frac{s^2 \log \frac{ep}{s}}{n\lambda^2} \to 0$ is "essentially" necessary to estimate u consistently among polynomial time algorithms.

Remark. This problem will be connected to the hidden clique problem; if we want to discover hidden cliques, the clique cannot be too small; otherwise, we will not be able to discover it.

Remark. In computer science, if we want to show that a problem is not solvable in polynomial time, we usually reduce it to hidden clique, or k-SAT.

1. Minimax Rate

1.1. Johnstone, Lu (JASA 200?) We start with the sample covariance $\hat{\Sigma}$. Then we know:

$$\mathbf{E}\left(\hat{\Sigma} - \mathbf{I}\right) = (\lambda u_i u_j)$$

Then, we may look at the diagonals, and we can get the estimator, assuming λ is known:

$$\lambda \hat{u}_i = \hat{\sigma}_{ii} - 1 \approx \mathcal{N}\left(0, \frac{\sigma_{ii}\sigma jj + \sigma_{ij}^2}{n}\right)$$

Then we have the accuracy:

$$\hat{\sigma}_{ii} - 1 - \hat{u}_i^2 \approx \mathcal{N}\left(0, \frac{C_{ii}}{n\lambda^2}\right)$$

We can then threshold, and set $u_i = \sqrt{\hat{\sigma}_{ii} - 1}$ if $\hat{\sigma}_{ii} - 1 - \hat{u}_i^2 \geq c\sqrt{\frac{\log p}{n\lambda^2}}$, and 0 otherwise. We then get:

$$\mathbf{E}\left(\hat{u}_i^{2\text{thres}} - u_i^2\right)^2 \lesssim \frac{\log p}{n\lambda^2}$$

But how does it perform for u_i itself (not squared)? The derivative of u_i^2 is $2u_i$, when u_i is very small, then the Fisher information is also very small.

Remark. If we have $T - \theta^2 \sim \mathcal{N}\left(0, \frac{1}{n}\right)$. Then we have $\sqrt{T} - \theta \sim \mathcal{N}\left(0, \frac{1}{4n\theta^2}\right)$. So, when θ is very small, then the accuracy is very bad (for estimating θ).

Therefore, if we use the thresholding procedure, then we get:

$$\mathbf{E}\left(\hat{u}_i^{\text{thres}} - u_i\right)^2 \lesssim \sqrt{\frac{\log p}{n\lambda^2}}$$

which is a slower rate.

Then, if we look at all the errors:

$$\mathbf{E}\sum_{i=1}^{p}\left(\hat{u}_i^{\text{thres}} - u_i\right)^2 \lesssim s\sqrt{\frac{\log p}{n\lambda^2}}$$

If we want good rates, we need s to be on a lower order than $\sqrt{n}$ (assuming everything else is constant).

All of these results may be considered exercises.

1.2. All Subset Selection. We can try to find u with $\|u\|_0 \leq s$ and $\|u\|_2 = 1$ to minimize:

$$\left\| \hat{\Sigma}_{\text{sample}} - \underbrace{(\mathbf{I} + \lambda u u^\top)}_{\Sigma} \right\|_F^2$$

This is a minimax optimal procedure. To prove this result, we can apply the basic inequality to obtain (denoting the minimizers by $\hat{\lambda}$ and $\hat{u}$):

$$\begin{aligned} \left\| \hat{\Sigma}_{\text{sample}} - \mathbf{I} - \hat{\lambda}\hat{u}\hat{u}^\top \right\|_F^2 &\leq \left\| \hat{\Sigma}_{\text{sample}} - \mathbf{I} - \lambda u u^\top \right\|_F^2 \\ \Rightarrow \left\| \hat{\Sigma}_{\text{sample}} - \Sigma + \Sigma - \underbrace{(\mathbf{I} + \hat{\lambda}\hat{u}\hat{u}^\top)}_{\hat{\Sigma}} \right\|_F^2 &\leq \left\| \hat{\Sigma}_{\text{sample}} - \mathbf{I} - \lambda u u^\top \right\|_F^2 \\ \Rightarrow \left\| \hat{\Sigma}_{\text{sample}} - \Sigma + \Sigma - \hat{\Sigma} \right\|_F^2 &\leq \left\| \hat{\Sigma}_{\text{sample}} - \Sigma \right\|_F^2 \end{aligned}$$

Expanding the quadratic, we then get:

$$\left\| \Sigma - \hat{\Sigma} \right\|_F^2 \leq 2 \left\langle \hat{\Sigma}_{\text{sample}} - \Sigma, \hat{\Sigma} - \Sigma \right\rangle$$

Every entry is $\hat{\Sigma}_{\text{sample}} - \Sigma$ behaves like a normal random variable, with standard deviation like $\frac{1}{n}$. The problem with bounding this guy is that $\hat{\Sigma}$ depends on $\hat{\Sigma}_{\text{sample}}$.

$$\begin{aligned} \left\| \Sigma - \hat{\Sigma} \right\|_F^2 &\leq 2 \left\langle \hat{\Sigma}_{\text{sample}} - \Sigma, \frac{\hat{\Sigma} - \Sigma}{\left\| \hat{\Sigma} - \Sigma \right\|_F^2} \right\rangle \left\| \hat{\Sigma} - \Sigma \right\|_F^2 \\ &\leq \left\| \hat{\Sigma} - \Sigma \right\|_F^2 \sup_{v : \|v\|_2 = 1, \|v\|_0 = s} 2 \left\langle \hat{\Sigma}_{\text{sample}} - \Sigma, \frac{\mathbf{I} + vv^\top - \Sigma}{\|\mathbf{I} + vv^\top - \Sigma\|_F^2} \right\rangle \end{aligned}$$

The rest of this proof is similar to all subsets selection for sparse linear regression. This is like looking like the max over $\binom{p}{s}$ $\mathcal{N}(0,1)$ random variables. We may show that this is bounded above with high probability:

$$\begin{aligned} \text{above } &\lesssim \sqrt{\frac{1}{n} \log \binom{p}{s}} \left\| \Sigma - \hat{\Sigma} \right\|_F \\ \left\| \Sigma - \hat{\Sigma} \right\|_F^2 &\leq \frac{s}{n} \log \frac{ep}{s} \end{aligned}$$

This implies that:

$$\left\| \hat{\lambda}\hat{u}\hat{u}^\top - \lambda u u^\top \right\|_F^2 \lesssim \frac{s \log \frac{ep}{s}}{n}$$

with high probability. Now, we want to get u. By sine-theta theorem, or Davis-Kahan, we have:

$$\min\left\{\|\hat{u}-u\|_2^2, \|\hat{u}+u\|_2^2\right\} \lesssim \frac{\frac{s\log\frac{ep}{s}}{n}}{\lambda^2}$$

Contrast this with Johnstone's original procedure:

$$\min\left\{\|\hat{u}-u\|_2^2, \|\hat{u}+u\|_2^2\right\} \leq \sqrt{\frac{s^2\log p}{n\lambda^2}}$$

If this goes to zero, we have a consistent estimator. Is this procedure rate-optimal?

2. Relaxing the Assumption

Suppose Johnstone, the truth, is in New Haven, and we, an initial or current estimate, are in New Haven as well. Can we find him? That is what it means to be rate-optimal. The idea here is very much same analogy as Bayesian Posterior Contraction.

Basically, the idea is like this: can we improve the consistent estimator to have it achieve the minimax rate?

Remark. Recall the proof of efficiency of MLE. First, show MLE is consistent. Then, show that it is asymptotically efficient.

We already have a consistent estimator (assume $\frac{s^2\log p}{n} \to 0$). Can we make it rate-optimal? The current procedure ignores all off-diagonal entries, so in that sense, it cannot be optimal.

Recall the power-method. We can use this. We assume that what we want to recover is rank-1. Then we only need one iteration. Assume $\langle u_0, u\rangle \to 1$ (this is equivalent to saying that the distance goes to zero).

What we do is this. From the $\hat{u}$, we get a u_0. We know that:

$$\begin{aligned} \mathbf{E}\,\hat{\Sigma} &= \mathbf{I} + \lambda u u^\top \\ \Rightarrow \mathbf{E}\,\hat{\Sigma}u_0 &= u_0 + \lambda u u^\top u_0 \\ \Rightarrow \lambda u \underbrace{\hat{u^\top} u_0}_{c\approx 1}&\hat{\Sigma}u_0 - u_0 \end{aligned}$$

We are not worried about the constant factor. Now, we are estimating $\hat{u}$ directly, instead of $\hat{u}^2$. We then get:

$$\lambda u^\top \hat{u_0} \cdot u_i \approx \mathcal{N}\left(0, \frac{C_i}{n}\right)$$

So we estimate u_i up to a constant scaling. Now, we may apply a thresholding procedure. This gives us:

$$\left\| \underbrace{\widehat{\lambda u^\top u_0}}_{\text{constant}} \cdot u - \lambda u^\top u_0 \cdot u \right\|_2^2 \lesssim \frac{s \log \frac{ep}{s}}{n}$$

$$\Rightarrow \|\hat{u} - u\|_2^2 \lesssim \frac{s \log \frac{ep}{s}}{n\lambda^2}$$

Remark. Typically, we can just take any u_0 where u_0 is not orthogonal to the eigenvector u, and take a few (or even just one) iteration. But, in the high dimensional case, a random vector has probability close to zero of being orthogonal to the eigenvector. That is why we must initialize, for example, with Johnstone's procedure.

2.1. Alternative Procedure to Pick Initializer.

The matrix Σ is sparse. Therefore, we using previously discussed techniques, we may have:

$$\left\| \hat{\Sigma} - \Sigma \right\|_{\text{op}} \lesssim \frac{s^2 \log p}{n}$$

with high probability, using thresholding for each entry. Then, apply Davis-Kahan, which gives us:

$$\min \left\{ \|\hat{u} - u\|_2^2 , \|\hat{u} + u\|_2^2 \right\} \lesssim \frac{s^2 \log p}{n\lambda^2}$$

This procedure tends to be more robust than Johnstone's original procedure. The leading s^2 term means this procedure performs much worse than the power method and thresholding procedure.

3. Computational Barrier

Claim 7. *If*

$$s > \left(n\lambda^2\right)^{\frac{1}{2}+\delta}$$

with $\delta > 0$, then we may pick a λ such that there exists no polynomial time algorithm (in average sense) to perform consistent estimation if the Hidden Clique Hypothesis is true.

Therefore, if this assumption does not hold (very similar but not equivalent to assuming $\frac{s^2 \log p}{n\lambda^2} \to 0$), then we cannot find a polynomial time algorithm.

3.1. Hidden Clique Problem.

Suppose we observe an adjacency matrix $\mathbf{R}^{n\times n} \ni A = (a_{ij}) \overset{\text{iid}}{\sim} \text{Bern}(1/2)$. A does not have a hidden clique (denote this H_0).

Now, we will implant a hidden clique to create H_1. This is just a clique, but we do not know the indices of the vertices.

We can find the hidden clique with high probabiltiy when $k \geq n^{\frac{1}{2}+\delta}$. But when $k \leq n^{\frac{1}{2}-\delta}$, there does not exist a polynomial time algorithm to find the Hidden Clique.

It turns out that this is very similar to $\Sigma = \mathbf{I} + \lambda u u^\top$, where u is k-sparse, and $\lambda \asymp \frac{k^2}{n}$, which comes from the largest eigenvalue.

It turns out that this is PCA with this particular setting of λ. This means that we need the condition that $\frac{s^2 \log p}{n\lambda^2} \to 0$. In principle, the Johnstone procedure needs $s < \left(n\lambda^2\right)^{\frac{1}{2}-\delta}$. This condition can be shown to be equivalent to $k \leq n^{\frac{1}{2}-\delta}$.

CHAPTER 12

Stochastic Block Model

The stochastic block model is a model for network analysis, which is an increasingly popular topic in statistics.

In SBM, we observe an adjacency matrix $A = (A_{ij})_{n\times n}$ of binary entries. One simple model is:

$$A_{ij} \overset{\text{iid}}{\sim} \text{Ber}(\theta_{ij}), \qquad A_{ii} = 0$$

Remark. In mathematical physics, they often consider this model, where $\theta_{ij} \in \{p, q\}$.

Suppose we assume our n nodes fall into k groups with mapping: $\sigma : \{1, \dots, n\} \mapsto \{1, \dots, k\}$.

Remark. There are k^n possible assignments σ.

Then, we can assume the model:

$$\theta_{ij} = B_{\sigma(i)\sigma(j)}, \qquad B_{k\times k} = (B_{ij})_{1\leq i,j\leq k}$$

Assume A, B symmetric.

There are many goals for this problem. In this chapter, we will discuss how to estimate $\Theta_{n\times n} = (\theta_{ij})_{1\leq i,j\leq n}$. In the following chapter, we will discuss how to estimate σ. If we have a good estimator of σ, we can understand the clustering structure. This is often called *community detection*.

Remark. Assume k is known.

1. Intractable Algorithm: Least Squares

1.1. Algorithm. We first consider a simple, intractable method. First, we find $\hat{\Theta}$, an $n \times n$ matrix to minimize $\|A - \Theta\|_F^2$ among all Θ determined by σ and B.

Remark. This is an alternative definition of Θ:

$$\Theta : \Theta_{n\times n} = Z_{n\times k} B_{k\times k} Z_{n\times k}^\top$$

where $Z_{n\times k} \in \{0, 1\}^{n\times k}$ and $\|Z_{i.}\|_0 = 1$. We may see that Z parameterizes σ.

1.2. Theoretical Justification. We apply the basic inequality:

$$\left\|A - \hat{\Theta}\right\|_F^2 \leq \|A - \Theta\|_F^2$$

Now, we write:

$$\begin{aligned}\left\|A-\Theta+\Theta-\hat{\Theta}\right\|_F^2 &\leq \|A-\Theta\|_F^2\\ \|A-\Theta\|_F^2+\left\|\Theta-\hat{\Theta}\right\|_F^2+2\langle A-\Theta,\Theta-\hat{\Theta}\rangle &\leq \|A-\Theta\|_F^2\\ \left\|\Theta-\hat{\Theta}\right\|_F^2+2\langle A-\Theta,\Theta-\hat{\Theta}\rangle &\leq 0\\ \left\|\Theta-\hat{\Theta}\right\|_F^2 &\leq 2\langle A-\Theta,\hat{\Theta}-\Theta\rangle\end{aligned}$$

Remark. $\mathbf{E}\,A=\Theta$ Therefore, we have:

$$\left\|\hat{\Theta}-\Theta\right\|_F^2 \leq 2\left\|\hat{\Theta}-\Theta\right\|_F\cdot\left\langle A-\Theta,\frac{\hat{\Theta}-\Theta}{\left\|\hat{\Theta}-\Theta\right\|_F}\right\rangle$$

Claim 8. *With high probability,*

$$\left|\left\langle A-\Theta,\frac{\hat{\Theta}-\Theta}{\left\|\hat{\Theta}-\Theta\right\|_F}\right\rangle\right| \leq C\sqrt{k^2+n\log k}$$

Remark. The k^2 term is from the estimation of the parameters, and $n\log k$ is from discovery of σ.

If we trust this claim, then we know:

$$\left\|\hat{\Theta}-\Theta\right\|_F \leq 2C\sqrt{k^2+n\log k}$$

with high probability. This is equivalent to:

$$\left\|\hat{\Theta}-\Theta\right\|_F^2 \leq 4C^2\left(k^2+n\log k\right)$$

with high probability. Many people like to divide by n^2, because we have n^2 parameters, which gives us the convergence rate:

$$\frac{1}{n}\left\|\hat{\Theta}-\Theta\right\|_F^2 \leq 4C^2\left(\frac{k^2}{n^2}+\frac{\log k}{n}\right)$$

Remark.

$$\frac{k^2}{n^2}+\frac{\log k}{n}\asymp\begin{cases}\frac{1}{n^2} & \text{for } k=1\\ \frac{1}{n} & \text{for } n=2\\ \frac{\log k}{n} & \text{for } k\leq\sqrt{n\log n}\\ \frac{k^2}{n^2} & \text{for } k\geq\sqrt{n\log n}\end{cases}$$

Observe that we recover binomial rate, and then there is a phase transition. Also observe that the two terms are equal when $k^2=n\log k$.

Remark. Low-rank approximation is another way to estimate Θ. If we use a low-rank approximation, then we get the rate:

$$\frac{1}{n^2}\left\|\hat{\Theta}-\Theta\right\|_F^2 \lesssim \frac{k}{n}$$

Remark. Harry says that $\frac{\log k}{n}$ is ambitious, but it is also interesting to consider whether there is a computationally tractable algorithm for beating low-rank.

Proof. Of Claim 8. To prove the claim, we use the Hanson-Wright Inequality (Hsu et al (2013)).

Let $\xi \in \mathbf{R}^n$ be a σ-subgaussian random vector. A special case could be: $\xi \in \mathcal{N}(0, \sigma^2\mathbf{I}_{n\times n})$. Let $A \in \mathbf{R}^{n\times n}$ be an $n \times n$ matrix. Set $\Sigma = A^\top A$. Then:

$$\mathbf{P}\left\{\|A\xi\|^2 \geq \sigma^2\left(\operatorname{tr}(\Sigma) + 2\sqrt{\operatorname{tr}(\Sigma^2)t} + 2\lambda_{\max}(\Sigma)t\right)\right\} \leq \exp\{-t\}$$

for any $t > 0$. So we may treat $A - \Theta$ from the Claim as our ξ; every entry in $A - \Theta$ is subgaussian.

Now, we give a corollary to the Hanson-Wright Inequality. Let $u \in \mathbf{R}^n$ be a vector. Suppose $u \in W$, where W is a space with dimension d. Then, we have:

$$\mathbf{P}\left\{\sup_{\|u\|_2\leq 1, u\in W} |u^\top\xi|^2 \geq \sigma^2\left(d + 2\sqrt{dt} + 2t\right)\right\} \leq \exp\{-t\}$$

Remark. If ξ is Gaussian, this inequality gives a “nice” tail for χ^2.

How do we apply this inequality to our Claim? First, we note that:

$$u^\top\xi = u^\top P_W\xi$$

In the inner product, only the projection of ξ onto W matters. Then, we have:

$$\sup_{\|u\|_2\neq 1, u\in W} = \sup_{\|u\|_2=1, u\in W} |u^\top P_W\xi|^2$$

We may rewrite $P_W\xi = \|P_W\xi\| \cdot u'$, where $\|u'\| = 1$. And, if we want to maximize:

$$u^\top \|P_W\xi\| \, u'$$

We choose $u = u'$, which gives us:

$$\begin{aligned}\sup_{\|u\|_2\neq 1, u\in W} &= \sup_{\|u\|_2=1, u\in W} |u^\top P_W\xi|^2 \\ &= \|P_W\xi\|_2^2\end{aligned}$$

We may rewrite $P_W = V\Lambda V^\top$, where V is orthogonal and Λ is diagonal with d ones and the rest zeros. Therefore, we have $\operatorname{tr}(P_W) = d = \operatorname{tr}(P_W^2)$, and $\lambda_{\max}(P_W) = 1$.

We may now use this corollary to prove the claim. Note that we may always write:

$$\hat{\Theta} = \hat{Z}_{n\times k}\hat{B}_{k\times k}(\hat{Z}_{n\times k})^\top$$

We may then write the inner product:

$$\left|\left\langle A-\Theta, \frac{\hat{\Theta}-\Theta}{\left\|\hat{\Theta}-\Theta\right\|_F}\right\rangle\right| \leq \sup_Z \sup_B \left|\left\langle A-\Theta, \frac{ZBZ^\top-\Theta}{\|ZBZ^\top-\Theta\|_F}\right\rangle\right|$$

Now, we apply the corollary. From the corollary:

$$\mathbf{P}\left\{\sup_B \left|\left\langle A-\Theta, \frac{ZBZ^\top-\Theta}{\|ZBZ^\top-\Theta\|_F}\right\rangle\right|^2 \geq \sigma^2\left(d+2\sqrt{dt}+2t\right)\right\} \leq \exp\{-t\}$$

for $t>0$, where $d=\dim(B)\leq k^2$.[1] This is true for every given Z. Then we use union bound to get the probability for the supremum over Z:

$$\mathbf{P}\left\{\sup_Z \sup_B \left|\left\langle A-\Theta, \frac{ZBZ^\top-\Theta}{\|ZBZ^\top-\Theta\|_F}\right\rangle\right|^2 \geq \sigma^2\left(d+2\sqrt{dt}+2t\right)\right\} \leq k^n\exp\{-t\}$$

We want this to be true with very high probability, so what t should we pick? If we pick $t=2n\log k$, then we get the bound is exceeded with probability less than k^{-n}. That is, with very high probability, we have:

$$\begin{aligned}\sup_Z \sup_B \left|\left\langle A-\Theta, \frac{ZBZ^\top-\Theta}{\|ZBZ^\top-\Theta\|_F}\right\rangle\right|^2 &\leq \sigma^2\left(d+2\sqrt{dt}+2t\right)\\ &\leq \sigma^2\left(d+(d+t)+2t\right)\\ &\leq C\left(k^2+n\log k\right)\end{aligned}$$

□

2. Application to Sparse Linear Regression

Remark. For sparse linear regression, we discussed All Subset Selection, and with basic inequality, we can get a good bound.

Suppose

$$Y_{n\times 1}=\mathbf{X}_{n\times p}\beta_{p\times 1}+\xi, \qquad \xi\sim\mathcal{N}(0,\sigma^2\mathbf{I})$$

or more generally, the noise is σ-subgaussian. In All Subsets Selection, we may obtain:

$$\hat{\beta}=\arg\min_{\|\beta\|_0=s}\|Y-X\beta\|^2$$

From the basic inequality, we obtain:

$$\left\|Y-\mathbf{X}\hat{\beta}\right\|^2\leq\|Y-\mathbf{X}\beta\|^2$$

[1] In this case, "dimension" is in the sense of unknown parameters. Here, the inner product is a vector inner product, so the latent space within which they exist is W, with dimensionality at most k^2.

Therefore:

$$\begin{aligned}\left\|\mathbf{X}\beta - \mathbf{X}\hat{\beta}\right\|^2 &\leq 2\langle Y - \mathbf{X}\beta, \mathbf{X}\hat{\beta} - \mathbf{X}\beta\rangle \\ &= 2\left\|\mathbf{X}\hat{\beta} - \mathbf{X}\beta\right\| \left\langle \underbrace{Y - \mathbf{X}\beta}_{\xi}, \underbrace{\frac{\mathbf{X}\hat{\beta} - \mathbf{X}\beta}{\left\|\mathbf{X}\hat{\beta} - \mathbf{X}\beta\right\|}}_{u} \right\rangle\end{aligned}$$

Given the support S of $\hat{\beta}$,

$$\mathbf{P}\left\{\sup_{\tilde{\beta}, \text{supp}(\tilde{\beta})=s} \left|\left\langle Y - \mathbf{X}\beta, \frac{\mathbf{X}\tilde{\beta} - \mathbf{X}\beta}{\left\|\mathbf{X}\tilde{\beta} - \mathbf{X}\beta\right\|}\right\rangle\right|^2\right\} \leq \exp\{-t\}$$

$$\mathbf{P}\left\{\sup_{S} \sup_{\tilde{\beta}, \text{supp}(\tilde{\beta})=s} \left|\left\langle Y - \mathbf{X}\beta, \frac{\mathbf{X}\tilde{\beta} - \mathbf{X}\beta}{\left\|\mathbf{X}\tilde{\beta} - \mathbf{X}\beta\right\|}\right\rangle\right|^2\right\} \leq \binom{p}{s}\exp\{-t\}$$

We could pick t such that $\exp\{-t\} = \left(\frac{1}{\binom{p}{s}}\right)^2$, which turns out to give $t \leq 2s\log\frac{ep}{s}$. We can then apply this bound to bound the prediction error:

$$\left\|\mathbf{X}\hat{\beta} - \mathbf{X}\beta\right\| \leq 2\sqrt{s + 2\sqrt{st} + t}$$

with high probability. With an appropriate choice of t, we can get:

$$\leq Cs\log\frac{ep}{s}$$

Both of these rates that we have proven today are optimal (the proof of the lower bound is possible).

3. Beyond

3.1. Mixture Models. We may extend this analysis to mixture models. Typically, we observe $Y_1, \ldots, Y_n$ independent, with:

$$Y_i \sim \mathcal{N}(\mu_{\sigma(i)}, \sigma^2\mathbf{I}_{p\times p})$$

where $\sigma : \{1, \ldots, n\} \mapsto \{1, \ldots, k\}$. This model may be rewritten as:

$$\begin{aligned}Y_{p\times n} &= (Y_1, \ldots, Y_n) \\ &= \Theta_{p\times n} + \xi_{p\times n}\end{aligned}$$

We can then decompose Θ:

$$\Theta = \mu_{p\times k}Z_{k\times n}$$

The Z in this equation is binary, and $\|Z_{.i}\|_0 = 1$.

3.2. Sparse Dictionary Learning. Suppose we have a regression model:

$$Y_{n\times 1} = \mathbf{X}_{n\times p}\beta_{p\times 1} + \xi_{n\times 1}$$

Suppose this is just one image, and we have many, many similar images. Then we instead consider:

$$Y_{n\times k} = \mathbf{X}_{n\times p}\beta_{p\times k} + \xi_{n\times k}$$

We want to find $\mathbf{X}$, a sparse dictionary, or basis.

3.3. Other Models. Other models of note include:

1. Mixed Membership Models
2. Bi-clustering

It is possible to analyze these models with the same technique.

4. Minimaxity

Minimaxity in the context of community detection was established in an Annals paper by Harry and Anderson.

The upper bound is derived via the MLE, which is an NP-Hard algorithm. So, once we establish the minimax rate, we consider whether it is possible for us to derive a tractable algorithm that is still rate-optimal.

A rate-optimal and tractable algorithm was proposed in a second paper, using spectral clustering with refinement. We may also want to consider other ways to formulate this problem, such as:

1. Bayesian Framework
2. Semidefinite Programming

We may also want to consider models more complex than the stochastic block model, such as:

1. Degree-corrected Block Model (also in a paper)
2. Overlapping SBM
3. Dynamic Model

Consider the 2-community case. Suppose a vector $Z \in \{1,2\}^n$ which maps each node to its community. Then, we may create an adjacency matrix by: $(\mathbf{1}\{Z_i = Z_j\})_{ij}$. We may generate the entries $A_{ij} \sim \text{Ber}(p_{ij})$. We further assume that A is symmetric and $A_{ii} = 0$ for all i. Let $p_{ij} = p$ for within-community connection and $p_{ij} = q$ for between community connection.

Our goal is to estimate Z, where our loss function is the Hamming distance $H(Z, \hat{Z})$.

4.1. Minimax Result. Suppose the communities are relatively balanced; that is:

$$Z = \left\{ Z \in \{1,2\}^n : c_1 n \le \sum \mathbf{1}\{Z_i = 1\} \le C_2 n \right\}$$

We assume p, q, c_1, C_2 are known.

The minimax rate has a nice form:

$$\min_{\hat{Z}} \max_{z \in \mathcal{Z}} \ell(Z, \hat{Z}) = n \exp\left\{ -\frac{(1+o(1))n(p-q)^2}{2p} \right\}$$

4.2. Algorithm. The intuition is as follows. Suppose there are $n+1$ nodes, and we know the label for all but one node. How do we estimate the label of the last node? This is essentially a hypothesis-testing problem:

$$\begin{aligned} H_0 &: Z_{n+1} = 1 \\ H_1 &: Z_{n+1} = 2 \end{aligned}$$

We know that for this, we have a minimax upper bound:

$$\min_{\hat{Z}(n+1)} \max_{Z(n+1) \in \{1,2\}} \mathbf{E}\,\mathbf{1}\{\hat{Z}(n+1) \neq Z(n+1)\} \leq \min \left[\mathbf{P}_{H_0}(\hat{Z}(n+1) = 1) + \mathbf{P}_{H_1}(\hat{Z}(n+1) = 2) \right]$$

We also have a lower bound based on the Bayes risk:

$$\geq \min_{\hat{Z}(n+1)} \left[\frac{1}{2} \mathbf{P}_{H_0}(\hat{Z}(n+1) = 1) \frac{1}{2} + \mathbf{P}_{H_1}(\hat{Z}(n+1) = 2) \right]$$

The optimal procedure for this two-point hypothesis test is the likelihood ratio test. This is the most powerful test; that is, the Type I + Type II error is minimal:

$$\tilde{Z} = \begin{cases} 1 & \text{for } L_1(A_{n+1}) > L_2(A_{n+1}) \\ 2 & \text{for } L_2(A_{n+1}) > L_1(A_{n+1}) \end{cases}$$

$$\Rightarrow \tilde{Z} = \begin{cases} 1 & \text{for } \sum_{j=1}^{\frac{n}{2}} A_{n+1,j} > \sum_{j=\frac{n}{2}+1}^{n} A_{n+1,j} \\ 2 & \text{for } \sum_{j=1}^{\frac{n}{2}} A_{n+1,j} < \sum_{j=\frac{n}{2}+1}^{n} A_{n+1,j} \end{cases}$$

This is a simple majority voting algorithm. This reduces to finding the probability that one binomial is greater than another binomial.

Then we may apply the Chernoff Bound and the Chernoff-Cramer Theorem:

$$\begin{aligned} \mathbf{P}\left\{ \sum_{i=1}^{\frac{n}{2}} X_i > \sum_{i=\frac{n}{2}=1}^{n} Y_i | X_i \sim Ber(q), Y_i \sim Ber(p) \right\} &\leq \exp\left\{ -\frac{n(p-q)^2}{2p} \right\} \\ &\leq \exp\left\{ -\frac{(1-o(1))n(p-q)^2}{2p} \right\} \end{aligned}$$

We may generalize to the unbalanced case using penalized majority voting.

$$\tilde{Z} = \begin{cases} 1 & \text{for } \sum_{j=1}^{m} A_{n+1,j} - \lambda m > \sum_{j=m}^{n} A_{n+1,j} - \lambda(n-m) \\ 2 & \text{for } \sum_{j=1}^{m} A_{n+1,j} - \lambda m < \sum_{j=m}^{n} A_{n+1,j} - \lambda(n-m) \end{cases}$$

where λ is a function of p, q.

4.3. Upper & Lower Bound. Generally, the lower bound is an information-theoretical result, whereas the upper bound requires a functional algorithm. Here we use the MLE.

$$\hat{Z} = \arg\max_{Z \in \mathcal{Z}} \log\left(L(A; Z)\right)$$

where $L(A; Z)$ is just a product of binomials. It turns out that:

$$\hat{Z} = \arg\max \sum_{i<j} A_{ij} \mathbf{1}\{Z(i) = Z(j)\} - \lambda \sum_{i<j} \mathbf{1}\{Z(i) = Z(j)\}$$

By the "local to global" connection in network science, the minimax rate is very similar for the global case. So, if we can solve the problem for estimating for one node, we can generalize to the network.

Note that although the MLE is rate-optimal, it is not computationally feasible (the parameter space is too huge).

4.4. Computationally Feasible Algorithms. There are many methods that are common in practice. The most popular one is spectral clustering.

4.4.1. *Spectral Clustering.* Input: A. Output: $\hat{Z}$

(1) SVD on $A = U\Lambda U^\top$
(2) V $= [U_1, U_2]$ keep the top two eigenvectors
(3) k-means on the data projected onto the two top eigenvectors (rows of V).

Lemma 2. *Lei and Rindd, 2014.*

$$\frac{1}{n}\ell(\hat{Z}, Z) \leq \frac{cp}{n(p-q)^2}$$

Recall our minimax rate:

$$\min\max \frac{1}{n}\ell(\hat{Z}, Z) = \exp\left\{-\frac{(1+o(1))n(p-q)^2}{p}\right\}$$

So if we desire $\frac{1}{n}\ell(\hat{Z}, Z) < 1$, i.e., consistency, we need $\frac{n(p-q)^2}{p} \to \infty$. The necessary condition for consistency is the same in spectral clustering, but the rate there is polynomial, not exponential.

4.4.2. *Spectral Clustering with Refinement.*

(1) Obtain $\hat{Z}^{\mathrm{sc}}$ by spectral clustering.
(2) Refine estimates using penalized majority voting.

We may show that as long as we have consistency, we have the exponential rate:

$$\frac{1}{n}\ell(\hat{Z}, Z) \leq \exp\left\{-\frac{n(p-q)^2}{2p}\right\}$$

if $\frac{n(p-q)^2}{y} \to \infty$. That is, as long as we have good initial estimates, the refinement will give us the minimax rate.

4.5. Degree-Corrected Block Model. In this model, we still have p, q, Z, but we have additional parameters θ_i, then we have:

$$\begin{aligned} A_{ij} &\sim Ber(p_{ij}) \\ p_{ij} &= \theta_i \theta_j p \mathbf{1}\{Z(i) = Z(j)\} \\ &= \theta_i \theta_j q \mathbf{1}\{Z(i) \neq Z(j)\} \end{aligned}$$

Then, the expected degree is proportional to θ_i:

$$\mathbf{E}\, d_i = \mathbf{E} \sum p_{ij} \asymp \theta_i \sum_j \theta_j p_{ij}$$

Recall that last time we encoded the Stochastic Block Model:

$$\mathbf{E}\, A = \Pi B \Pi^\top$$

We can then generalize to DCBM:

$$\mathbf{E}\, A = \Theta \Pi B \Pi^\top \Theta$$

where Θ is a diagonal matrix consisting of $\theta_1, \ldots, \theta_n$.

We can then generalize the loss:

$$\ell(\hat{\Pi}, \Pi) = \left\| \hat{\Pi} \hat{\Pi}^\top - \Pi \Pi^\top \right\|_F$$

We do not need to worry about identifiability in this case.