ANDREW TULLOCH

STATISTICAL THEORY

TRINITY COLLEGE
THE UNIVERSITY OF CAMBRIDGE

Contents

1 *Introduction* 5

1.1 *Asymptotic Statistics* 6

1.2 *Statistical Inference* 6

2 *Stochastic Convergence Concepts* 7

2.1 *Uniform Laws of Large Numbers* 9

3 *Parametric Statistical Models* 11

3.1 *Consistency of M-Estimators* 12

3.2 *Verifying uniform convergence* 15

3.3 *Asymptotic Inference based on the MLE* 18

3.4 *Some Ideas from LeCam Theory* 21

4 *Bayesian Inference* 25

5 *Gaussian Linear Model* 27

5.1 *Over-fitting a linear model* 29

6 *High-Dimensional Statistics* 31

6.1 *Compressed Sensing and the Restricted Isometry Property* 37

6.2 *Inference with the LASSO* 45

7 *Conclusion* 47

7.1 *Outlook on Nonparametrics* 47

7.2 *Relevant previous Tripos questions* 47

8 *Bibliography* 49

1

Introduction

Consider observations $X_1, \ldots, X_n$ are copies of a random variable (r.v.) with distribution

$$F(t) = \mathbb{P}(X \leq t), t \in \mathbb{R}.$$

Definition 1.1. A statistical model is a family of candidate distributions

$$\mathbb{P}_\Theta = \{P_\theta | \theta \in \Theta\}$$

where Θ is a parameter space.

Example 1.2 (Linear Regression)**.** *Consider*

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\Theta} + \epsilon \tag{1.1}$$

$\mathbf{X}$ *is our design matrix,* $\mathbf{X}_i$ *are our explanatory variables,* $\mathbf{Y}$ *is our response,* ϵ *is our measurement error (e.g.* ϵ $N(0, \sigma^2)$*).*

Example 1.3 (Nonlinear regression)**.**

$$\mathbf{Y} = g(\mathbf{X}, \boldsymbol{\Theta}) + \epsilon \tag{1.2}$$

Example 1.4 (High dimensional linear model)**.**

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\Theta} + \epsilon \tag{1.3}$$

with $\boldsymbol{\Theta}$ *being sparse.*

If $\mathbf{X}$ *has some "properties" (restricted isometry property), then "miracle happens".*

1.1 *Asymptotic Statistics*

Investigate statistical problems where the sample size n is large ($n \to \infty$).

(i) Quick intuitions (about the complexity of Θ.

(ii) Large sample approximation can often proven to be good (concentration of measure).

(iii) Nice mathematics.

(iv) For n finite, Θ can be useless.

(v) Does not consider computation cost.

1.2 *Statistical Inference*

(i) Estimation: construct $\hat{\theta}_n - \hat{\theta}(X_1, \ldots, X_n)$, that estimates (approximates) θ well when $X_i \sim P_\theta$.

(ii) Hypothesis testing: $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$. Find test/decision rule $\psi_n = \psi(X_1, \ldots, X_n)$ such that $\psi_n = 0$ if H_0 is true.

(iii) Confidence Sets: Find $C_n = C(X_1, \ldots, X_n, \alpha) \subseteq \Theta, 0 \leq \alpha \leq 1$ such that $P_\theta(\theta \in C_n) = 1 - \alpha$ for all $n \in \mathbb{N}$. This is **uncertainty quantification**.

2

Stochastic Convergence Concepts

Definition 2.1 (Random Variable)**.** A random variable is a (measurable) mapping

$$X : (\Omega, \mathcal{A}, \mu) \to \mathbb{R}. \tag{2.1}$$

The **distribution function** is defined by

$$F(t) = \mathbb{P}(X \leq t) = \mu(w \in \Omega | X(w) \leq t), t \in \mathbb{R} \tag{2.2}$$

where $\mathbb{P} = \mu \circ X^{-1}$ is the law of X.

A random vector $\mathbf{X}$ is a vector of random variables with joint distribution

$$F(\mathbf{t}) = \mathbb{P}(\mathbf{X} \leq \mathbf{t}) = \mathbb{P}(\mathbf{X}_i \leq \mathbf{t}_i, 1 \leq i \leq n) \tag{2.3}$$

Definition 2.2 (Convergence almost surely)**.** A sequence $X_n, n \in \mathbb{N}$ of random variables converges almost surely to a random variable X if

$$\mathbb{P}(X_n \to X) = \mu(\omega \in \Omega | X_n(\omega) \to X(\omega)) = 1 \tag{2.4}$$

We say that $X_n \xrightarrow{as} X$.

Definition 2.3 (Convergence in probability)**.** $X_n \xrightarrow{p} X$ (in probability) if for all $\epsilon > 0$,

$$\mathbb{P}(|X_n - X) > \epsilon) \to 0 \tag{2.5}$$

as $n \to \infty$. For random vectors, we define analogously with taking the norm in $\mathbb{R}^n$.

Definition 2.4 (Convergence in distribution)**.** $X_n \xrightarrow{d} X$ or X_n con-

verges to X in distribution if

$$\mathbb{P}(X_n \leq t) \to \mathbb{P}(X \leq t) \tag{2.6}$$

whenever $t \mapsto \mathbb{P}(X \leq t)$ is continuous.

Proposition 2.5. *Let $(X_n, n \in \mathbb{N})$, X taking values in $\mathcal{X} \subseteq \mathbb{R}^d$.*

(i)

$$X_n \xrightarrow{as} X \Rightarrow X_n \xrightarrow{p} X \Rightarrow X_n \xrightarrow{d} X \tag{2.7}$$

(ii) If $X_n \to X$ in any mode, and if $g : \mathcal{X} \to \mathbb{R}^d$ is continuous, then $g(X_n) \to g(X)$ in the same mode.

(iii) ***Slutsky's lemma*** *If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c$ (a constant). Then*

(i)

$$Y_n \xrightarrow{p} c \tag{2.8}$$

(ii)

$$X_n + Y_n \xrightarrow{d} X + c \tag{2.9}$$

(iii)

$$X_n Y_n \xrightarrow{d} cX \tag{2.10}$$

where $Y_n \in \mathbb{R}$.

(iv)

$$X_n Y_n^{-1} \xrightarrow{d} c^{-1} X \tag{2.11}$$

where $Y_n \in \mathbb{R}, c \neq 0$.

(iv) If $(A_n, n \in \mathbb{N})$ are random matrices with $(A_n)_{ij} \xrightarrow{p} A_{ij}$ for all i, j and $X_n \xrightarrow{d} X$, then $A_n X_n \xrightarrow{d} AX$, and if A is invertible, $A_n^{-1} X_n \xrightarrow{d} A^{-1} X$, where $A = (A_{ij})$.

Proof. Exercise. □

Some key results from probability theory are statements about

$$\frac{1}{n} \sum_{i=1}^{n} X_i \tag{2.12}$$

where the $X_1, \ldots, X_n$ form an **infinite** sequence of IID copies of a fixed random variable $X \sim \mathbb{P}$. The $(X_1, X_2, \ldots)$ can be accommo-

dated as the coordinate projections of the product probability space

$$(\mathbb{R}^{\mathbb{N}}, \mathbb{B}^{\mathbb{N}}, \mathbb{P}^{\mathbb{N}}) \tag{2.13}$$

or, if the X_i's are random vectors in R^d, then

$$((\mathbb{R}^d)^{\mathbb{N}}, (\mathbb{B}^d)^{\mathbb{N}}, \mathbb{P}^{\mathbb{N}}) \tag{2.14}$$

where $Pr = \mathbb{P}^{\mathbb{N}}$ is the product space measure associated to the sequence $(X_1, \dots, X_n, \dots)$.

Theorem 2.6 (Law of Large Numbers)**.** *let $X_1, \dots, X_n$ be IID copies of X $\mathbb{P}$ such that $\mathbb{E}(|X_i|) < \infty$, then*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{as} \mathbb{E}(X) \tag{2.15}$$

Theorem 2.7 (Central limit theorem)**.** *Let $X_1, \dots, X_n$ be IID copies of $X \sim \mathbb{P}$ on $\mathbb{R}$ with $\mathbb{V}(X) = \sigma^2 < \infty$. Then*

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}(X)\right) \xrightarrow{d} N(0, \sigma^2) \tag{2.16}$$

In the multivariate case, where $X \sim \mathbb{P}$ on $\mathbb{R}^d$ with the covariance of X as Σ, then

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}(X)\right) \xrightarrow{d} N(0, \Sigma) \tag{2.17}$$

Assuming that the random variables X_i are bounded, say $|X_i| \leq 1$, the central limit theorem is in fact a non-asymptotic phenomena (at least for tail events), since by Hoeffding's inequality, for all $n \in \mathbb{N}$ and $u > 0$,

$$\mathbb{P}\left(\sqrt{n}\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}(X)\right| > u\right) \leq 2e^{-\frac{u^2}{2}} \tag{2.18}$$

which compares "well" to the Gaussian tail $\Phi(u) = \frac{1}{n}e^{-\frac{u^2}{2}}$.

2.1 *Uniform Laws of Large Numbers*

Consider $X_1, X_2, \dots, X_n$ IID from law $\mathbb{P}$ on T (e.g. $\mathbb{R}^d$), and let $h : T \to R$ such that $\mathbb{E}(|h(X)|) < \infty$. Then the $h(X_i)$'s are also IID, so by

the law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n} h(X_i) - \mathbb{E}(h(X)) \xrightarrow{as} 0 \tag{2.19}$$

For finitely many h's, say h_i, the exceptional set A_m such that the m'th LLN $\frac{1}{n}\sum_{i=1}^{n} h_m(X_i) - \mathbb{E}(h_m(X)) \xrightarrow{as} 0$ fails has probability zero, and clearly by the union bound

$$\mathbb{P}\left(\cup_{m=1}^{M} A_m\right) \leq \sum_{m=1}^{M} \mathbb{P}(A_m) = 0 \tag{2.20}$$

and so clearly

$$\max_{m=1,\ldots,M} \left| \sum_{1} n \sum_{i=1}^{n} h_m(X_i) - \mathbb{E}(h_m(X)) \right| \xrightarrow{as} 0 \tag{2.21}$$

as $n \to \infty$.

For a general class $\mathcal{H}$ of measurable functions $T \to \mathbb{R}$, we say that the brackets $\lfloor h_j, h_j \rceil, j = 1, \ldots, N$ **cover** $\mathcal{H}$ if for all $h \in \mathcal{H}$, there exists some j such that $\lfloor h_j(x) \leq h(x) \leq h_j(x) \rceil$ for all $x \in T$.

Proposition 2.8. *Suppose $\mathcal{H}$ is (for all $\epsilon > 0$), covered by brackets $\lfloor h_j, h_j \rceil, i = 1, \ldots, N_\epsilon$ such that*

$$\mathbb{E}(|h_j(X)\rceil|) < \infty, \mathbb{E}(|\lfloor h_j(X)|) < \infty, \tag{2.22}$$

and

$$\mathbb{E}(|\lfloor h_j(X) - h_j(X)\rceil|) < \epsilon. \tag{2.23}$$

Then

$$\sup_{h\in\mathcal{H}} \left| \frac{1}{n}\sum_{i=1}^{n} h(X_i) - \mathbb{E}(h(X)) \right| \xrightarrow{as} 0 \tag{2.24}$$

as $n \to \infty$.

Proof. Let $\epsilon > 0$ be given. By the law of large numbers for finitely many h_i's, we have that for all $n \geq N(\epsilon, \omega)$,

(2.25)

□

3

Parametric Statistical Models

Let $Y_1, \ldots, Y_n$ observations.

Example 3.1. *$Y_1 = (Z_i, X_i)$ where the Z_i's are response variables, and the covariates X_i are related to Z_i by the regression relationship $Z_i = g(X_i, \theta) + \epsilon_i$ for $\theta \in \Theta \subseteq \mathbb{R}^p$, ϵ_i IID with $\mathbb{E}(\epsilon_i) = 0$, and $g : X \times \Theta \to \mathbb{R}$.*

A regression function (possibly non-linear) and known — for example,

$$g(X_i, \theta) = X_i^T \theta, \tag{3.1}$$

a linear model.

A natural way to estimate θ is by nonlinear least squares (NLS) which finds $\hat{\theta}$ that minimizes

$$Q_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} (Z_i - g(X_i, \theta))^2 \tag{3.2}$$

Example 3.2. *We are given a model of PDF/PMF's $\{f(\cdot, \theta) : \theta \in \Theta\}, \Theta \subseteq \mathbb{R}^p$ for the distribution of a random variable Y. We view $Y_1, \ldots, Y_n$ as IID copies of Y.*

The likelihood function of the model is defined as

$$L_n(\theta) = \Pi_{i=1}^{n} f(Y_i, \theta) \tag{3.3}$$

*The log-likelihood function $l_n(\theta) = \log L_n(\theta)$. A **maximum likelihood estimator** (MLE) is any value $\hat{\theta} = \hat{\theta}_{MLE} \in \Theta$ that maximizes $L_n(\theta)$ over*

Θ. *Equivalently, we minimize*

$$Q_n(\theta) = -\frac{1}{n} l_n(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \log f(Y_i, \theta) \tag{3.4}$$

3.1 Consistency of M-Estimators

In both the examples, $\hat{\theta}_n$ is found by minimizing a random criterion function $Q_n(\theta)$ over Θ, and proved a "limiting function" $Q(\theta)$ exists, we expect these minimizers to converge to the minimizers of Q.

Theorem 3.3. *Let $\Theta \subseteq \mathbb{R}^p$ be compact. Let $Q : \Theta \to \mathbb{R}$ be a continuous, non-random function that has a unique minimizer $\theta_0 \in \Theta$.*

Let $Q_n : \Theta \to \mathbb{R}$ be any sequence of random functions such that

$$\sup_{\theta \in \Theta} |Q_n(\theta) - Q(\theta)| \xrightarrow{p} 0 \tag{3.5}$$

as $n \to \infty$.

*If θ_n is **any** sequence of minimizers of Q_n, then $\hat{\theta}_n \xrightarrow{p} \theta_0$ as $n \to \infty$.*

Proof. Let $\epsilon > 0$ be arbitrary. The set $\Theta_\epsilon = \{\theta \in \Theta : \|\theta - \theta_0\| \geq \epsilon\}$ is compact and Q is continuous on Θ_ϵ, so Q attains its infimum

$$c(\epsilon) = \inf_{\theta \in \Theta_\epsilon} Q(\theta) = Q(\bar{\theta}_\epsilon) \in \Theta_\epsilon > Q(\theta_0) \tag{3.6}$$

as θ_0 is the minimizer.

Pick $0 < \delta(\epsilon) < \frac{c(\epsilon) - Q(\theta_0)}{2}$, which implies

$$c(\epsilon) - \delta(\epsilon) > Q(\theta_0) + \delta(\epsilon) \tag{3.7}$$

Define the event

$$A_n(\epsilon) = \{\sup_{\theta \in \Theta} |Q_n(\theta) - Q(\theta)| < \delta(\epsilon)\}. \tag{3.8}$$

On this event we have

$$\begin{aligned}\inf_{\theta\in\Theta_\epsilon} Q_n(\theta) &= \inf_{\theta\in\Theta_\epsilon} [Q_n(\theta) - Q(\theta) + Q(\theta)] \\ &\geq \inf_{\theta\in\Theta_\epsilon} Q(\theta) - \sup_{\theta\in\Theta} |Q_n(\theta) - Q(\theta)| \\ &\geq C(\epsilon) - \delta(\epsilon) \\ &\geq Q(\theta_0) + \delta(\epsilon) \\ &\geq Q(\theta_0) + \delta(\epsilon) - |Q_n(\theta_0) - Q(\theta_0)| \\ &\geq Q_n(\theta_0)\end{aligned}$$

since on $A_n(\epsilon)$, in particular $|Q_n(\theta_0) - Q(\theta_0)| < \delta(\epsilon)$.

We conclude

$$\inf_{\theta:\|\theta-\theta_0\|\geq\epsilon} Q_n(\theta) > Q_n(\theta_0) \tag{3.9}$$

Now suppose $\hat{\theta}_n \in \Theta_\epsilon$, then $Q_n(\hat{\theta}_n) \geq \inf_{\theta\in\Theta_\epsilon} Q_n(\theta) > Q_n(\theta_0)$.

Hence, on $A_n(\epsilon)$, we have $\|\hat{\theta}_n - \theta_0\| < \epsilon$, $A_n(\epsilon) \subseteq \{\|\hat{\theta}_n - \theta_0\| < \epsilon\}$, so since by hypothesis $\mathbb{P}(A_n(\epsilon)) \to 1$ for all $\epsilon > 0$, we see $\mathbb{P}\big(\|\hat{\theta}_n - \theta_0\| < \epsilon\big) \to 1$, as $\mathbb{P}\big(\|\hat{\theta}_n - \theta_0\| \geq \epsilon\big) \to 0$ as $n \to \infty$. Since $\epsilon > 0$ was arbitrary, the result follows. □

Remark 3.4. *Uniform convergence of $Q_n \to Q$ is necessary. In fact, none of the conditions can be relaxed.*

Exercise 3.5. *(i) What is Q in Examples 3.1, 3.2?*

(ii) What is Θ_0?

(iii) When does uniform convergence occur?

Example 3.6. *Let $Y = (Z, X)$ such that $Z = g(X, \theta_0) + \epsilon$, where $\mathbb{E}(\epsilon|X) = 0$, θ_0 is the "true value", and based on* IID *observations $Y_1, \ldots, Y_n$, we minimize*

$$Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}(Z_i - g(X_i, \theta))^2 \tag{3.10}$$

over Θ. We expect

$$Q(\theta) = \mathbb{E}_{\theta_0}\Big((Z - g(X, \theta))^2\Big) \tag{3.11}$$

Inserting the model equation

$$Q(\theta) = \mathbb{E}_{\theta_0}\Big((g(X_1,\theta_0) - g(X,\theta) + \epsilon)^2\Big) = \mathbb{E}(g(X,\theta_0) - g(X,\theta))^2 + \mathbb{E}\Big(\epsilon^2\Big) \tag{3.12}$$

Hence $Q(\theta)$ is minimized at θ_0 if the regression parameterization is identifiable, that is

$$\theta = \theta' \iff g(\cdot,\theta) = g(\cdot,\theta') \tag{3.13}$$

$\mathbb{P}_X$ almost surely.

Example 3.7. *Let $Y_1,\dots,Y_n$ be* IID *copies of Y, and we maintain a parametric model*

$$\{f(\cdot,\theta) : \theta \in \Theta\} \tag{3.14}$$

of PDFs/PMFs and the MLE is found by minimizing

$$Q_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \log f(Y_i,\theta) \tag{3.15}$$

By the law of large numbers, assuming $f(y,\theta) > 0$ for all y,θ and

$$\mathbb{E}_{\theta_0}(|\log f(Y,\theta)|) < \infty \tag{3.16}$$

where Y is assumed to be distributed as $f(\cdot,\theta_0)$, then the limiting criterion function is

$$Q(\theta) = -\mathbb{E}_{\theta_0}(\log f(Y,\theta)) \tag{3.17}$$

Then

$$\begin{aligned}
Q(\theta_0) - Q(\theta) &= \mathbb{E}_{\theta_0} \log f(Y,\theta) - \mathbb{E}\theta_0 - \log f(Y,\theta_0) && (3.18)\\
&= \mathbb{E}_{\theta_0}\left(\log \frac{f(Y,\theta)}{f(Y,\theta_0)}\right) && (3.19)\\
&\le \log \mathbb{E}_{\theta_0}\left(\frac{f(Y,\theta)}{f(Y,\theta_0)}\right) && (3.20)\\
&= \log \int \frac{f(y,\theta)}{f(y,\theta_0)} f(y,\theta_0)\, dy && (3.21)\\
&= \log 1 && (3.22)\\
&= 0 && (3.23)
\end{aligned}$$

or in other words,

$$Q(\theta_0) \leq Q(\theta) \forall \theta \in \Theta \tag{3.24}$$

Equality in Jensen's inequality can only occur when

$$\frac{f(\cdot, \theta)}{f(\cdot, \theta_0)} = C \in \mathbb{R} \tag{3.25}$$

so since $\int f(y, \theta) dy = 1$, we see $C = 1$, and hence if the model is identifiable in the sense that $\theta = \theta' \iff f(\cdot, \theta) = f(\cdot, \theta')$ for all $\theta, \theta' \in \Theta$, then the value θ_0 that minimizes θ_0 is unique.

3.2 *Verifying uniform convergence*

Proposition 3.8. *Let Θ be compact in $\mathbb{R}^p$, and let $\mathcal{X} \subseteq \mathbb{R}^d$ and consider observing $X_1, \ldots, X_n$ IID from $X \sim \mathbb{P}$ on X. Let $q : \mathcal{X} \times \Theta \to \mathbb{R}$ that is continuous in θ for all x and measurable in x for all $\theta \subseteq \Theta$.*

Assume

$$\mathbb{E}\left(\sup_{\theta \in \Theta} |q(X, \theta)| \right) < \infty \tag{3.26}$$

Then

$$\sup_{\theta \in \Theta} |\frac{1}{n} q(X_i, \theta) - \mathbb{E}(q(X, \theta))| \xrightarrow{as} 0 \tag{3.27}$$

as $n \to \infty$

Proof. We apply the uniform law of large numbers from Proposition 2.8 and we need to cover the set

$$\mathcal{H} = \{q(\cdot, \theta) : \theta \in \Theta\} \tag{3.28}$$

by suitable brackets.

Define open balls

$$B(\theta, m) = \{\theta' \in \Theta : \|\theta - \theta'\| < \eta\} \tag{3.29}$$

Construct "brute-force" brackets

$$\overline{q}(X, \theta, \eta) = \sup_{\theta' \in B(\theta, \epsilon)} q(X, \theta') \tag{3.30}$$

$$\underline{q}(X, \theta, \eta) = \inf_{\theta' \in B(\theta, \eta)} q(X, \theta') \tag{3.31}$$

which obviously cover all the $\{q(\cdot,\theta') : \theta' \in B(\theta,\eta)\}$.

Clearly,

$$\mathbb{E}\Big(\overline{q}(x,\theta,\eta)\Big) \leq \mathbb{E}\left(\sup_{\theta\in\Theta}|q(X,\theta)|\right) < \infty \tag{3.32}$$

by the domination condition.

By continuity and compactness, the supremum/infimum above are attained at $\overline{\theta},\underline{\theta} \in \Theta$ such that $\|\overline{\theta}-\theta\| \leq \eta$. So

$$|\overline{q}(X,\theta,\eta) - \underline{q}(X,\theta,\eta)| \leq |\overline{q}(X,\theta,\eta) - q(X,\theta)| + |q(X,\theta) - \underline{q}(X,\theta,\eta)| \tag{3.33}$$

which again by continuity tends to zero as $\eta \to 0$.

So $|\overline{q}(X,\theta,\eta) - \underline{q}(X,\theta,\eta)| \to 0$ as $\eta \to 0$.

By the dominated convergence theorem we can integrate this limit with respect to $\mathbb{E}$, (using the dominance condition). So,

$$\mathbb{E}\Big(|\overline{q}(X,\theta,\eta) - \underline{q}(X,\theta,\eta)\Big) \to 0 \tag{3.34}$$

as $\eta \to 0$.

Then for all $\epsilon > 0$, there exists $\eta = \eta(\epsilon,\theta)$ such that

$$\mathbb{E}\Big(|\overline{q}(X,\theta,\eta(\epsilon,\theta)) - \underline{q}(X,\theta,\eta(\epsilon,\theta))|\Big) < \epsilon \tag{3.35}$$

The balls $\{B(\theta,\eta(\epsilon,\theta)) : \theta \in \Theta\}$ form an open covering of Θ, so by compactness (Heine-Borel theorem in $\mathbb{R}^p$), there exists a finite subcover of θ, say with centers $\theta_1,\ldots,\theta_{N(\epsilon)}$. Then the corresponding brackets

$$[\underline{q}_i,\overline{q}_i] = [\underline{q}(\cdot,\theta_j,\eta(\epsilon,\theta_j)),\overline{q}(\cdot,\theta_j,\eta(\epsilon,\theta_j))] \tag{3.36}$$

cover $\mathcal{H}$ and satisfy the conditions of Proposition 2.8 □

Remark 3.9. *The above result is simply a law of large numbers in the Banach space of continuous functions on Θ, and*

$$\mathbb{E}\left(\sup_{\theta\in\Theta}|q(X,\Theta)|\right) = \mathbb{E}(\|Z\|) < \infty \tag{3.37}$$

which is necessary for the result to hold.

Fill in missing notes from previous lecture

Definition 3.10. A consistent estimator $\tilde{\theta}$ in a model $\{f(\cdot,\theta)|\theta \in \Theta\}$ is called **asymptotically efficient** if $\lim_n n\mathbb{V}(\tilde{\theta}) = I(\theta)^{-1}$ for all $\theta \in \text{int}(\Theta)$ where $I(\theta)$ is the Fisher information.

Theorem 3.11. *In a model satisfying Assumption B,*

$$\sqrt{n}(\hat{\theta}_{MLE} - \theta_0) \xrightarrow{d} N(0, I(\theta_0)^{-1}) \tag{3.38}$$

Proof. Let $\mathbb{P} = \mathbb{P}_{\theta_0}^N$, $\mathbb{E} = \mathbb{E}_\theta$.

For $\ell_n(\theta) = -Q_n(\theta) = \frac{1}{n}\log f(Y_i,\theta)$. When proving $Z_n \xrightarrow{d} Z$ we may restrict to events E_n such that $\mathbb{P}(E_n) \to 1$, since

$$\|\mathbb{P}(Z_n \leq t) - \mathbb{P}(Z_n \leq t, E_n)\| \leq \mathbb{P}(E_n^c) \to 0 \tag{3.39}$$

as $n \to \infty$. Since $\hat{\theta}_n \xrightarrow{p} \theta_0$ as $n \to \infty$, we can restrict to $E_n = \{\hat{\theta}_n\}$ where K is a closed ball centered at θ_0. By the assumptions, ln is C^2 on U, and $\hat{\theta}_n$ is a maximizer on the open set U, so necessarily,

$$0 = \frac{\partial}{\partial\theta} ln(\theta)_{|_{\theta=\hat{\theta}_n}} = \frac{\partial}{\partial\theta}\ln(\hat{\theta}_n) = \begin{bmatrix} \frac{\partial}{\partial\theta_1}\ln(\hat{\theta}_n) \\ \vdots \\ \frac{\partial}{\partial\theta_p}\ln(\hat{\theta}_n) \end{bmatrix} \tag{3.40}$$

For $h : K \to \mathbb{R}$ and $u, v \in K$ the line segment

$$tu + (1-t)v \tag{3.41}$$

for $0 < t < 1$ connection u, v does lie in the ball K by convexity, and the mean value theorem gives (for $h \in C^1(U)$),

$$h(u) = h(v) + \frac{\partial h}{\partial u}(\overline{v})^T(u - v) \tag{3.42}$$

where $\overline{v}$ is a mean-value on the line segment.

Applying this p-times to $\frac{\partial}{\partial\theta_i}\ell_n(\theta)$ we obtain

$$0 = \begin{bmatrix} \frac{\partial}{\partial\theta_1}\ell_n(\hat{\theta}_n) \\ \vdots \\ \frac{\partial}{\partial\theta_p}\ell_n(\hat{\theta}_n) \end{bmatrix} \tag{3.43}$$

o We have

Fill this in

$$(\overline{A_n})_{kj} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial^2}{\partial\theta_k\partial\theta_j}\log f(Y_i,\overline{\theta}_{(j)}) - \mathbb{E}\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\right)\right] \tag{3.44}$$

□ Fill in rest of proof

Remark 3.12 (Discussion of Theorem 3.11)**.** *(i) One can weaken the conditions to $\theta \mapsto f(\cdot,\theta)$ being "weakly C1", to model the Laplace family. For non-differentiable parameterizations, the asymptotics of the MLE may be non-normal. For example, consider $U[0,\theta]$ with $\theta \in \Theta = (0,\infty)$.*

(ii) When the "true" θ_0 is at the boundary of the parameter space, the asymptotics of the MLE are also non-normal. For example, $N(\theta,1), \theta \in \Theta = [0,\infty), \hat{\theta}_{MLE} = \max(\overline{X_n}, 0)$

(iii) Asymptotic efficiency is an optimality criterion that is meaningful only for "regular" estimators, that rules out the following super-efficient estimator e.g.

$$\tilde{\theta} = \begin{cases} \hat{\theta}_{MLE} & |\hat{\theta}_{MLE}| \geq n^{-\frac{1}{4}} \\ 0 & \text{otherwise} \end{cases} \tag{3.45}$$

One shows that under $P_\theta, \theta \neq 0$, that $\sqrt{n}(\tilde{\theta}-\theta) = \sqrt{n}(\hat{\theta}-\theta) \xrightarrow{d} N(0, I(\theta)^{-1})$ as $n \to \infty$. However, under P_0, one shows easily that $\sqrt{n}(\tilde{\theta}-\theta) \xrightarrow{d} 0 = N(0,0)$ which strictly beats the $N(0, I(0)^{-1})$-distribution (Hodges' estimator).

3.3 *Asymptotic Inference based on the MLE*

Suppose we want a confidence interval for θ_j, $j = 1,\dots,p$. We can write $\theta_j = e_j^T\theta$, $e_j = (0,\dots,0,\underbrace{1}_{j\text{-th position}},\dots,0)$. By the continuous mapping theorem, we have

$$\sqrt{n}(\hat{\theta}_j - \theta_j) = \sqrt{n}e_j^T(\hat{\theta}-\theta) \xrightarrow{d} N(0, e_j^T I(\theta)^{-1} e_j) = N(0, I^{-1}(\theta)_{jj}) \tag{3.46}$$

Suggesting that

$$C_n = \{v \in \mathbb{R} : |\hat{\theta}_{n,j} - v| \leq \frac{(I(\theta)^{-1})_{jj}^{\frac{1}{2}} Z_\alpha}{\sqrt{n}}\} \tag{3.47}$$

where Z_α are such that $\mathbb{P}(|Z| \leq Z_\alpha) = 1 - \alpha$ is a confidence interval for θ_j, since

$$\mathbb{P}_\theta^n(\theta_j \in C_n) = \P_\theta^n(\sqrt{n}(I(\theta)^{-1})_{jj}^{-\frac{1}{2}}|\hat{\theta}_{n,j} - \theta| \leq Z_\alpha) \to \mathbb{P}(|Z| \leq Z_\alpha) = 1 - \alpha. \tag{3.48}$$

This can only be used if $I(\theta)$ is known, otherwise $I(\theta)$ has to estimated consistently.

Definition 3.13. The **observed** Fisher information is defined as

$$i_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(Y_i, \theta) \frac{\partial}{\partial\theta} \log f(Y_i, \theta)^T \tag{3.49}$$

One shows as in the proof of Theorem 3.11 that

$$\hat{i}_n = i_n(\hat{\theta}_{MLE}) \xrightarrow{p} I(\theta_0) \tag{3.50}$$

under P_{θ_0}.

Alternative, one can use

$$\hat{j}_n = j_n(\hat{\theta}_n) \tag{3.51}$$

where

$$j_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta\partial\theta^T} \log f(Y_i, \theta) \tag{3.52}$$

which does estimate $I(\theta_0)$ consistently.

To construct a confidence set for $\theta \in \Theta \subseteq \mathbb{R}^p$, it is consistent to consider the Wold-statistic

$$W_n(\theta) = n(\hat{\theta} - \theta)^T \hat{i}_n (\hat{\theta} - \theta) \tag{3.53}$$

which can be shown to have, under P_θ^N to have the χ_p^2 distribution. Thus

$$C_n = \{\theta \in \mathbb{R}^p | W_n(t) \leq \xi_\alpha\} \tag{3.54}$$

where ξ_α are the $1 - \alpha$ quartiles of the χ_p^2 distribution, is a confidence ellipsoid for θ of asymptotic coverage probability $1 - \alpha$.

To test $H_0 : \theta = \theta_0$ against $H_1 : \Theta \in \theta \backslash \{\theta_0\}$, we can refer $W_n(\theta_0)$ to the quartiles of the χ_p^2 distribution, since $W_n(\theta_0) \xrightarrow{d} \xi_p^2$ under H_0.

For such testing problems there exists an alternative approached based on the **likelihood ratio test statistic** for $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta \backslash \Theta_0$, with $\Theta_0 \subseteq \Theta$ as

$$\Lambda_n(\Theta, \Theta_0) = 2 \log \frac{\sup_{\theta \in \Theta} \prod_{i=1}^n f(Y_i, \theta)}{\sup_{\theta \in \Theta_0} \prod_{i=1}^n f(Y_i, \theta)} \tag{3.55}$$

$$= 2 \log \frac{\prod_{i=1}^n f(Y_i, \hat{\theta}_n)}{\prod_{i=1}^n f(Y_i, \hat{\theta}_{n,0})} \tag{3.56}$$

where $\hat{\theta}_n$ is the unrestricted MLE and $\hat{\theta}_{n,0}$ is the MLE restricted to H_0.

Theorem 3.14 (Wilks'). *If* $\dim(\theta_0) = p_0 < \dim(\Theta) = p$, *then*

$$\Lambda_n(\Theta, \Theta_0) \xrightarrow{d} \chi^2_{p-p_0} \tag{3.57}$$

as $n \to \infty$.

Proof. (Only for $H_0 = \{\theta_0\}, \dim \theta_0 = 0$).

Recall

$$Q_n(\theta) = -\frac{1}{n} \sum_{i=1}^n \log f(Y_i, \theta) = -l_n(\theta) \tag{3.58}$$

and so

$$\Delta_n(\theta, \theta_0) = 2nQ_n(\theta_0) - 2nQ_n(\hat{\theta}_n) \tag{3.59}$$

$$= 2n \frac{\partial}{\partial \theta} Q_n(\hat{\theta}_n)^T (\theta_0 - \hat{\theta}_n) + \frac{2n}{2} (\theta_0 - \hat{\theta}_n)^T \frac{\partial^2}{\partial \theta \partial \theta^T} Q_n(\bar{\theta})(\theta_0 - \hat{\theta}_n) \tag{3.60}$$

$$= \sqrt{n}(\hat{\theta}_n - \theta_0)^T \bar{A}_n \sqrt{n}(\theta_0 - \hat{\theta}_n) \tag{3.61}$$

$$= Z_n^T \bar{A}_n Z_n \tag{3.62}$$

where we then conclude that from Theorem 3 (in notes) that $Z_n \xrightarrow{d} Z \sim N(0, I(\theta_0)^{-1})$, and, as in the proof of Theorem 3, $\bar{A}_n \xrightarrow{p} I(\theta_0)$ as $n \to \infty$. Rewrite this as

$$Z_n^T I(\theta_0) Z_n + Z_n^T (\bar{A}_n - I(\theta_0)) Z_n \tag{3.63}$$

which by repeated applications of Slutsky's lemma.

The mapping $X \mapsto X^T I(\theta_0) X$ is continuous from $\mathbb{R}^p$ into $\mathbb{R}$, so by the continuous mapping theorem,

$$Z_n^T I(\theta_0) Z_n \xrightarrow{d} Z^T I(\theta_0) Z \tag{3.64}$$

and as $I(\theta_0)$ is positive semidefinite (and so has a square root), we can write this as

$$Z^T I(\theta_0) Z = Z_T I(\theta_0)^{\frac{1}{2}} I(\theta_0)^{\frac{1}{2}} Z = W^T W = \sum_{i=1}^{p} W_i^2 \sim \chi_p^2 \tag{3.65}$$

with $W \sim N(0, I)$. □

3.4 *Some Ideas from LeCam Theory*

Consider first a Gaussian shift experiment

$$N(g, I(\theta)^{-1}), g \in \mathbb{R}^p, I(\theta) \tag{3.66}$$

is the Fisher information of some statistical model

$$\{f(\cdot, \theta), \theta \in \Theta\} \tag{3.67}$$

The log-likelihood ratio

$$\log \frac{dN(h, I(\theta)^{-1})}{dN(0, I(\theta)^{-1})}(X) = h^T I(\theta) X - \frac{1}{2} h^T I(\theta) h \tag{3.68}$$

since the ratio is proportional to

$$\exp\left(-\frac{(X-h)^T I(\theta)(X-h)}{2} + \frac{X^T I(\theta) X}{2} \right) \tag{3.69}$$

Definition 3.15. A model $\{f(\cdot, \theta), \theta \in \Theta\}$ is called **locally asymptotically normal (LAN)** at $\theta_0 \in \int \Theta$ if for all $h \in \mathbb{R}^p$ (small enough),

$$\log \frac{\prod_{i=1}^n f(Y_i, \theta_0 + \frac{h}{\sqrt{n}})}{\prod_{i=1}^n f(Y_i, \theta_0)} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h^T \frac{\partial}{\partial \theta} \log f(Y_i, \theta)_{|_{\theta=\theta_0}} - \frac{1}{2} h^T I(\theta_0) h + Z_n \tag{3.70}$$

as $n \to \infty$, where $Z_n \xrightarrow{p} 0$ under $P_{\theta_0}^n$.

Remark 3.16. *The first term in the expansion (by the CLT) converges in distribution to $N(0, h^T I(\theta_0) h)$ as $n \to \infty$.*

Proposition 3.17. *Any statistical model that satisfies the conditions of Theorem 3 is also LAN.*

Proof. The LHS of (dagger) equals find reference

$$nl_n(\theta_0 + \frac{h}{\sqrt{n}}) - nl_n(\theta_0) = \sqrt{n}\frac{1}{n}\sum_{i=1}^{n} h^T \frac{\partial}{\partial\theta}\log f(Y_i, \theta_0) + \frac{n}{2}h^T \frac{\partial^2}{\partial\theta\partial\theta^T} l_n(\bar{\theta})h \tag{3.71}$$

$$= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} h^T \frac{\partial}{\partial\theta}\log f(Y_i, \theta_0) - h^T I(\theta_0)h + \underbrace{o_p(1)}_{Z_n} \tag{3.72}$$

□

Definition 3.18. Let $\mathbb{P}_n, \mathbb{Q}_n$ be sequences of probability measures. We say $\mathbb{Q}_n$ is **contiguous** with respect to $\mathbb{P}_n$ ($\mathbb{Q}_n \lhd \mathbb{P}_n$) if

$$\mathbb{P}_n(A_n) \to 0 \Rightarrow \mathbb{Q}_n(A_n) \to 0 \tag{3.73}$$

for any sequence of events A_n in the probability space. We say $\mathbb{P}_n, \mathbb{Q}_n$ are mutually contiguous if $\mathbb{P}_n \lhd \mathbb{Q}_n$ and $\mathbb{P}_n \rhd \mathbb{Q}_n$ and write $\mathbb{P}_n \lhd \rhd \mathbb{Q}_n$.

Lemma 3.19 (LeCam's 1st lemma)**.** *The following are equivalent:*

(i) $\mathbb{Q}_n \lhd \mathbb{P}_n$

(ii)

$$\frac{dQ_n}{dP_n}(X_n) \xrightarrow{d} U, X_n \sim P_n \tag{3.74}$$

along a subsequence, then $P(U > 0) = 1$.

(iii)

$$\frac{dP_n}{dQ_n}(X_n) \xrightarrow{d} V, X_n \sim Q_n \tag{3.75}$$

along a subsequence, $\mathbb{E}(V) = 1$.

(iv) For any sequence of statistics (measurable functions $T_n : \Omega_n \to \mathbb{R}$*), we have* $T_n \xrightarrow{p} 0$ *under* P_n *then* $T_n \xrightarrow{p} 0$ *under* Q_n *as* $n \to \infty$.

Remark 3.20. *For two probability measures* P, Q *that are absolutely continuous with respect to each other, the likelihood ratio is the random variable* $\frac{dP}{dQ}(X), X \sim Q$.

Corollary 3.21. *(i) If* $\frac{dQ_n}{dP_n} \xrightarrow{d} e^X$ *for* $X_n \sim P_n$*, and* $X \sim N(-\frac{\sigma^2}{2}), \sigma^2)$*,* $\sigma^2 > 0$*, then*

$$Q_n \lhd \rhd P_n \tag{3.76}$$

(ii) In any LAN model the product measures $P^n_{\theta_0+\frac{h}{\sqrt{n}}}, P^n_{\theta_0}$, corresponding to the joint distributions of a sample of size n fro the PDF/PMF $f(\theta_0 + \frac{h}{\sqrt{n}}), f(\theta_0)$ respectively, are mutually contiguous (for arbitrary $h \in \mathbb{R}^p$).

Proof. (i) By LeCam's lemma, $P(e^X > 0) = 1$ for any normal random variable X, and $\mathbb{E}(e^X) = e$.

□ Complete proof

IN a LAN model, the product measures $\P^n_\Theta = \otimes^n_{i=1}\P_\Theta$ and $P_{\theta+\frac{h}{\sqrt{n}}}$ are mutually contiguous.

Example 3.22. *Recall the Hodges' estimator*

$$\tilde{\theta}_n = \hat{\theta}_n \mathbb{I}\left(|\hat{\theta}_n| \geq n^{-\frac{1}{4}}\right) \tag{3.77}$$

in a regular parametric model, $\Theta = \mathbb{R}$, and where $\hat{\theta}_n$ is the MLE. One shows under $P_\theta, \theta \neq 0$, we have

$$\sqrt{n}(\tilde{\theta} - \theta) \xrightarrow{d} N(0, I(\theta)^{-1}) \tag{3.78}$$

as $n \to \infty$. But when sampling from P_0, then $P^n_0(\tilde{\theta}(X_1, \ldots, X_n) \neq 0) = P^n_0(|\hat{\theta}_n| \geq n^{-\frac{1}{4}}) = P^n_0(\sqrt{n}|\theta_n - \theta| \geq n^{-\frac{1}{4}}) \to 0$. This follows as $X_n \xrightarrow{d} X \Rightarrow (X_n, n \in \mathbb{N})$ is stochastically bounded, that is, there exists $M(\epsilon)$ such that $P(|X_n| > M(\epsilon)) < \epsilon$). Hence, under P_0, $\sqrt{n}(\tilde{\theta} - \theta) \xrightarrow{d} N(0, 0)$ which outperforms the Cramer-Rao lower bound at $\theta = 0$.

Consider now the minimax quadratic risk of $\tilde{\theta}$, equal to (for $n \in \mathbb{N}$ fixed),

$$\sup_{\theta \in \Theta} \mathbb{E}^h_\theta (\sqrt{n}(\tilde{\theta} - \theta))^2 \tag{3.79}$$

Consider the local alternative $0 + \frac{h}{\sqrt{n}}, h \in \mathbb{R}$ arbitrary. Then the minimax risk exceeds

$$\geq \mathbb{E}^n_{\frac{h}{\sqrt{n}}} n(\tilde{\theta} - \frac{h}{\sqrt{n}})^2 \mathbb{I}(\tilde{\theta} = 0) \tag{3.80}$$

$$= h^2 P^n_{\frac{h}{\sqrt{n}}}(\tilde{\theta} = 0) \tag{3.81}$$

$$= h^2 (1 - P^n_{\frac{h}{\sqrt{n}}(\tilde{\theta} \neq 0)}) \tag{3.82}$$

$$\geq \frac{h^2}{2} \tag{3.83}$$

by contiguity of $P_0^n \lhd \rhd P^n_{\frac{h}{\sqrt{n}}}$.

Conclude that

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} \mathbb{E}^h_\Theta (\sqrt{n}(\tilde{\theta} - \theta))^2 \to \infty \tag{3.84}$$

whereas

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} \mathbb{E}^n_\theta (\sqrt{n}(\hat{\theta} - \theta))^2 \leq \sup_{\theta\in\Theta} I(\theta)^{-1} < \infty \tag{3.85}$$

4

Bayesian Inference

In any parametric model $\{f(\cdot,\theta), \theta \in \Theta\}$, we can consider a prior distribution Π on Θ, and model the observations $X_1, \ldots, X_n$ as IID copies of the random variable $X|\theta \sim f(\cdot,\theta)$, where $\theta \sim \Pi$. The posterior distribution is the law of $\theta|X_1, \ldots, X_n$.

Formally, if $\mathcal{X}$ is the sample space that X takes values in, consider on $\mathcal{X} \times \Theta$ the probability distribution Q with pdf/pmf by $dQ(x,\theta) = f(x,\theta)\Pi(\theta)dxd\theta$ by the laws/definition of conditional probability,

$$X|\theta \sim \frac{f(x,\theta)\Pi(\theta)dx}{\int_{\mathcal{X}} f(x,\theta)dx\Pi(\theta)} = f(x,\theta)dx \tag{4.1}$$

and conversely

$$\theta|X \sim \frac{f(x,\theta)\Pi(\theta)d\theta}{\int_{\Theta} f(x,\theta)\Pi(\theta)d\theta} = \Pi(\theta|X). \tag{4.2}$$

In particular, for $(X_i, i = 1, \ldots, n)$ IID copies of $X|\theta$, the posterior distribution equals

$$\theta|X_1, \ldots, X_n \sim \frac{\prod_{i=1}^n f(x_i,\theta)\Pi(\theta)}{\int_{\Theta} \prod_{i=1}^n f(X_i,\theta)\Pi(\theta)d\theta} \tag{4.3}$$

The posterior distribution can be used for all purposes of statistical inference on θ:

(i)

$$\bar{\theta}(X_1, \ldots, X_n) = \mathbb{E}(\theta|X_1, \ldots, X_n) \tag{4.4}$$

estimates θ,

(ii)

$$C_n = \{\theta \in \Theta : \|\theta - \bar{\theta}\| \leq R_n\} \tag{4.5}$$

where R_n is such that $\Pi(C_n|X_1, \ldots, X_n) = 1 - \alpha$, giving a credible set for Θ.

Theorem 4.1 (Bernstein-von Mises theorem)**.** *The Bernstein-von Mises theorem states that in LAN-models $\{f(\cdot, \theta), \theta \in \Theta\}$ and for **any** prior that has a positive continuous density at θ_0, we have*

$$\Pi\cdot, X_1, \ldots, X_n \approx N(\hat{\theta}_{MLE}, \frac{1}{n} I(\theta_0)^{-1}) \tag{4.6}$$

under $P^n_{\theta_0}$ (in total variation distance), which in particular implies that any credible set C_n such that $\Pi(C_n|X_1, \ldots, X_n) = 1 - \alpha$ satisfies $P^n_{\theta_0}(\theta_0 \in C_n) \to 1 - \alpha$ as $n \to \infty$. In particular, asymptotic Bayesian inference coincides with asymptotic inference based on the MLE.

5

Gaussian Linear Model

Let $(Y_i, i = 1, \ldots, n)$ be the response variable, with $X_{ij}, i = 1, \ldots, n, j = 1, \ldots, p$ **covariates**. We have **influence parameters** θ_j, related in a way that

$$Y_i = \sum_{i=1}^{p} \theta_j X_{ij} + \text{error}_i, i = 1, \ldots, n \tag{5.1}$$

Assume that error_i is a sum of **many small independent** "measurement" errors,

$$\text{error}_i = \sum_{m=1}^{m} \epsilon_{im} \tag{5.2}$$

which is approximately normal.

Assume that $\epsilon_i \sim N(0, \sigma^2)$. Gauss proceeded to compute the maximum likelihood estimate.

Proposition 5.1 (Approach I — MLE Interpretation)**.** *Joint distribution of the Y_i's is*

$$f(y_i, \ldots, y_n; \theta) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \sum_{j=1}^{p} \theta_j X_{ij})^2\} \tag{5.3}$$

so the log likelihood is

$$l_n(\theta, \sigma^2) = -\frac{n}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \sum_{i=1}^{p} \theta_j X_{ij})^2 \tag{5.4}$$

$$\frac{\partial}{\partial \theta_j} l_n(\theta, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} 2(Y_i - \sum_{j=1}^{p} \theta_j X_{ij})(-X_{ij}) \tag{5.5}$$

$$\frac{\partial}{\partial (\sigma^2)} l_n(\theta, \sigma^2) = -\frac{n}{2} \frac{1}{\sigma^2} + \frac{1}{\sigma^4} \sum_{i=1}^{n} \sum_{i=1}^{n} (Y_i - \sum_{i=1}^{p} \theta_j X_{ij})^2 \tag{5.6}$$

Solving these for zero, we obtain $X^TY = X^TX\hat{\theta}$, *and* $\hat{\sigma}^2 = \|Y - X\hat{\theta}\|^2$, *where we rewrite the linear model in matrix form as* $Y = X\theta + \epsilon$, *and note that* $\frac{\partial^2}{\partial\theta_j\partial\theta_k} l_n(\theta) = -\frac{1}{\sigma^2}(X^TX)_{jk}$. *If* X *has full column rank (the column vectors* X_j *are linearly independent), then*

$$\hat{\theta} = \hat{\theta}_{MLE} = (X^TX)^{-1}X^TY \tag{5.7}$$

Proposition 5.2 (Approach II — Geometric Interpretation)**.** *The model can be written as*

$$Y = \sum_{j=1}^{p} \theta_j X_j + \epsilon \tag{5.8}$$

Heuristically, we can see to find the ***best approximation*** *of* $Y \in \mathbb{R}^n$ *from the p-dimensional subspace (*$p \leq n$*) of* $\mathbb{R}^n$ *spanned by* $X_j, j = 1, \ldots, p$, *assumed to be linearly independent. Setting*

$$\langle Y - X\theta, X\theta \rangle = 0 \tag{5.9}$$

gives the projection.

The projection matrix P *onto* $\langle X \rangle = \text{span}(X_j, j = 1, \ldots, p)$ *is given by* $P = X(X^TX)^{-1}X^T$ *since*

(i) $PX = X(X^TX)^{-1}X^TX = X$,

(ii) $P = P^T$,

(iii) $P^2 = P$,

and so the best approximation of Y *from* $\langle X \rangle$ *is* $PY = X(X^TX)^{-1}X^TY = X\hat{\theta}$.

Inserting the model equation $Y = X\theta + \epsilon$, *we see*

$$\begin{aligned}
\hat{\theta} &= (X^TX)^{-1}X^T(X\theta + \epsilon) && (5.10)\\
&= (X^TX)^{-1}(X^TX)\theta + (X^TX)^{-1}X^T\epsilon && (5.11)\\
&\sim N(\theta, \sigma^2(X^TX)^{-1}) && (5.12)\\
&= N(\theta, I_n(\theta)^{-1}) && (5.13)
\end{aligned}$$

so we achieve the Cramer-Rao lower bound. So $\hat{\theta}$ *is efficient among all unbiased estimators.*

5.1 Over-fitting a linear model

Consider adding X_{p+1} to the model, then fit this model. Then

$$\|Y - P_{p+1}Y\|_2^2 \leq \|Y - P_pY\|_2^2 \tag{5.14}$$

since we are projection to a lower dimensional subspace. If $p = n$, then

$$\|Y - P_nY\|_2^2 = 0 \tag{5.15}$$

and so the fit is perfect.

In contrast, the prediction error increases with the dimensionality of the model p, since

$$\begin{aligned}
\mathbb{E}\Big(\|X\hat{\theta} - X\theta\|^2\Big) &= \mathbb{E}\Big(\|PY - X\theta\|^2\Big) && (5.16)\\
&= \mathbb{E}\Big(\|P\epsilon\|^2\Big) && (5.17)\\
&= \mathbb{E}\Big(\epsilon^T P^T P\epsilon\Big) && (5.18)\\
&= \mathbb{E}\Big(\operatorname{tr}(\epsilon^T P\epsilon)\Big) && (5.19)\\
&= \mathbb{E}\Big(\operatorname{tr}(P\epsilon\epsilon^T)\Big) && (5.20)\\
&= \operatorname{tr}(P\mathbb{E}\Big(\epsilon\epsilon^T\Big)) && (5.21)\\
&= \sigma^2 \operatorname{tr}(P) && (5.22)\\
&= \sigma^2 \operatorname{tr}(X(X^TX)^{-1}X^T) && (5.23)\\
&= \sigma^2 p && (5.24)
\end{aligned}$$

since $PY = PX\theta + P\epsilon = X\theta + P\epsilon$, and the trace of a projection matrix is the sum of the eigenvalues, which are either 0 or 1, and the number of multiplicity of the latter eigenvalues is equal to the dimension of the subspace.

This tradeoff between the fit and predictive accuracy leads to the problem of model selection. For typical matrices X, we have

$$\|\hat{\theta} - \theta\|^2 \sim \|X(\hat{\theta} - \theta)\|^2 \sigma^2 \frac{p}{n} \tag{5.25}$$

When p is moderate compared to N, we can use model selection

criteria to choose the MLE of low-dimensional model. When $p > n$, the vectors $(X_j, j = 1, \ldots, n)$ can **never** be linearly independent, and $\hat{\theta}$ cannot be used at all ($X^T X$ is not invertible).

6

High-Dimensional Statistics

Consider a functional form

$$Y_i = \sum_{j=1}^{p} X_{ij}\theta_j^0 + \epsilon_i, i = 1, \ldots, n \tag{6.1}$$

with $p \geq n$ or $p >> n$. We believe that θ^0 is **sparse** in the sense that most if it's p coefficients are zero. Formally, assume that

$$\theta^0 \in B_0(k) = \{\theta \in \mathbb{R}^p | \text{at most } k \text{ non-zero entries}\} \tag{6.2}$$

with $k \leq n$, or $k << n$.

We call $S_0 = \S(\theta^0) = \{j : \theta_j^0 \neq 0\}$ the **active set** of θ^0, satisfying $|S_0| \leq k$. We would like to fit LS in the "true" submodel

$$Y_i = \sum_{j \in S_0} X_{ij}\theta_j^0 + \epsilon_i \tag{6.3}$$

with prediction risk

$$\mathbb{E}_\theta \|\hat{\theta}(S^0) - \theta\| \approx \frac{1}{n}\mathbb{E}_\theta \|X(\hat{\theta}(S_0) - \theta)\|_2^2 \approx \frac{|S_0|}{n} \leq \frac{k}{n} \tag{6.4}$$

In practice both the position of S_0 and k are unknown.

Question 6.1. *Can we minimize the oracle risk?*

Consider first $p \leq n$, and a fixed submodel M of $Y = X\theta + \epsilon$ given by $Y = X^M\theta^M + \epsilon$, $X = (X^m, X^{\bar{m}}), \dim(M) = k$.

In the full model, we use the LS estimates to obtain prediction risk

$$\mathbb{E}\Big(\|PY - X\Theta\|^2\Big) = \sigma^2 p \tag{6.5}$$

where P projects onto $\langle X \rangle$.

For the restricted model, the LS-fit for P_M, the projection onto $\langle X^M \rangle$, the prediction risk is

$$\begin{aligned}
\mathbb{E}\Big(\|P_M Y - X\theta\|^2\Big) &= \mathbb{E}\Big(\|P_M Y - P_M X\Theta + P_M X\theta - X\theta\|^2\Big) && (6.6)\\
&= \mathbb{E}\Big(\|P_M(Y - X\theta)\|^2 + \|(I - P_M)X\theta\|^2\Big) && (6.7)\\
&= \mathbb{E}\Big(\|P_M\epsilon\|^2\Big) + \theta^T X^T (I - P_M) X\theta && (6.8)\\
&= \sigma^2 k + \theta^T X^T (I - P_M) X\theta && (6.9)
\end{aligned}$$

If we estimate the second term by $\hat{\theta}^T X^T (I - P_M) X\hat{\theta}$, which has expectation ($\hat{\theta} = \hat{\theta}_{full}$)

$$\begin{aligned}
\mathbb{E}\Big(Y^T P(I - P_M)PY\Big) &= \mathbb{E}\Big((X\theta + \epsilon)^T (P - P_M)(X\theta + \epsilon)\Big) && (6.10)\\
&= \theta^T X^T (I - P_M) X\theta + 2\mathbb{E}\Big(\epsilon^T (P - P_M) X\theta\Big) + \mathbb{E}\Big(\epsilon^T (P - P_M)\epsilon\Big) && (6.11)\\
&= \theta^T X^T (I - P_M) X\theta + 0 + \sigma^2(p - k) && (6.12)
\end{aligned}$$

(as in the previous lecture).

So if we take

$$\hat{MSPE}(M) = \hat{\theta}^T X^T (I - P_M) X\hat{\theta} + 2\sigma^2 k - \sigma^2 p \tag{6.13}$$

is an unbiased "estimate" of

$$MSPE(M) = \mathbb{E}\Big(\|P_M Y - X\theta\|^2\Big) \tag{6.14}$$

Now replace σ^2 by it's unbiased estimate $\hat{\sigma}^2 = \frac{1}{n-p}\|Y - PY\|^2$ to obtain the estimated predictive risk of model M,

$$\begin{aligned}
\underset{C_p}{\text{crit}}(M) &= \hat{\theta}^T X^T (I - P_M) X\hat{\theta} + 2\hat{\sigma}^2 k - \hat{\sigma}^2 p && (6.15)\\
&= \|Y - P_M Y\|^2 + 2\hat{\sigma}^2 k - n\sigma^2 && (6.16)
\end{aligned}$$

Comparing all submodels M of R^p by computing $\text{crit}_{C_p}(M)$, we fit the LS estimator in the model $\hat{M}$ that minimizes $\text{crit}_{C_p}(M)$.

This is called Mallow's C_p.

Several such model selection criterion exist, all of the form

$$\text{crit}(M) = \|Y - P_M Y\|^2 + \lambda \dim(M) + \text{const} \tag{6.17}$$

The "derivation" of each criterion suggests a choice of λ.

For $p \geq n$ this approach cannot be used directly, since $\|Y - P_M Y\| = 0$ for $\dim(M) \geq n$, but we can adapt it by minimizing $\|Y - X\theta\|^2 + \lambda \#\{\theta_j \neq 0\}$ over $\mathbb{R}^p$. This is a combinatorially hard optimization problem (to find k, we need to compute $\binom{p}{k}$ solutions).

Let us try to find a **convex relaxation** of this optimization problem, and note that the penalty can be "written" as

$$\lambda \|\theta\|_0 = \lambda \sum_{i=1}^{p} |\theta_j|^0 \tag{6.18}$$

Note further that the l_q means

$$\|\theta\|_q^q = \sum_{i=1}^{p} |\theta_j|^q \downarrow \|\theta\|_0 \tag{6.19}$$

These "norms" are convex when $q \geq 1$, and the program

$$\min_{\theta \in \Theta} \|Y - X\theta\|^2 + \lambda \|\theta\|_q \tag{6.20}$$

is convex. So $q = 1$ arises as a natural compromise and we define $\tilde{\theta}$ to be any solution of

$$\min_{\theta \in \mathbb{R}^p} \frac{\|Y - X\theta\|^2}{n} + \lambda \|\theta\|_1 \tag{6.21}$$

known as the LASSO estimator.

If we consider

$$\sup_{\theta \in B_0(h)} \frac{\mathbb{E}_\theta \|X\tilde{\theta}_{LASSO} - X\theta\|^2}{n} \leq \frac{k}{n} \log p \tag{6.22}$$

Any solution to the minimization problem

$$\min_{\theta \in \mathbb{R}^p} \frac{\|Y - X\Theta\|_2^2}{n} + \lambda \|\theta\|_1 \tag{6.23}$$

for λ a tuning parameter, is called a LASSO solution, denoted by $\tilde{\theta} = \tilde{\theta}_{LASSO}$. $\tilde{\theta}$ is generally not unique, but $X\tilde{\theta}$ and also $\|\theta\|_1$ give the same numerical values for all LASSO solutions.[1]

[1] Exercise sheet

Theorem 6.2. *Let $Y = X\theta^0 + \epsilon$ with $\epsilon \sim N(0, I_n)$ and θ^0 is k-sparse with* ***active set*** *S_0. Let $\tilde{\theta}$ be any LASSO solution, with*

$$\lambda = 4\overline{\sigma}\sqrt{\frac{t^2 + \log p}{n}} \tag{6.24}$$

$$\overline{\sigma}^2 = \max_{j=1,\dots,p} \hat{\sum}_{jj} \tag{6.25}$$

$$\hat{\sum} = \frac{X^T X}{n} \tag{6.26}$$

the Gram matrix.

Now, assume, for some r_0

$$\|\tilde{\theta}_{S_0} - \theta^0\|_1 \leq r_0 k (\tilde{\theta}_{S_0} - \theta^0)^T \hat{\sigma} (\tilde{\theta}_{S_0} - \theta^0) \tag{6.27}$$

with probability $\geq 1 - \beta$. Then with probability at least

$$1 - \beta - e^{-\frac{t^2}{2}}, \tag{6.28}$$

we have

$$\frac{1}{n}\|X(\tilde{\theta} - \theta^0)\|_2^2 + \lambda \|\tilde{\theta} - \theta^0\|_1 \leq 4 k r_0 \lambda^2 \lesssim \frac{k}{n} \log p \tag{6.29}$$

Proof. Note that $\theta^0_{S_0} = \theta^0$. By definition of $\tilde{\theta}$, we have

$$\frac{1}{n}\|Y - X\tilde{\theta}\|_2^2 + \lambda \|\tilde{\theta}\|_1 \leq \frac{1}{n}\|Y - X\theta^0\|_2 + \lambda \|\theta^0\|_1 \tag{6.30}$$

Inserting the model equation $Y = X\theta^0 + \epsilon$, we get

$$\frac{1}{n}\|X(\theta^0 - \tilde{\theta}) + \epsilon\|_2^2 + \lambda \|\tilde{\theta}\|_1 \leq \frac{1}{n}\|\epsilon\|_2^2 + \lambda \|\theta^0\|_1 \tag{6.31}$$

$$= \|X(\theta^0 - \tilde{\theta})\|_2^2 + \frac{2}{n}\epsilon^T X(\theta^0 - \tilde{\theta}) + \frac{1}{n}\|\epsilon\|_2^2 \tag{6.32}$$

and so we obtain

$$\frac{1}{n}\|X(\theta^0-\tilde{\theta})\|_2^2+\lambda\|\tilde{\theta}\|_1 \leq \frac{2}{n}\epsilon^T X(\tilde{\theta}-\theta^0)+\lambda\|\theta^0\|_1 \tag{6.33}$$

Lemma 6.3.

$$\mathbb{P}\left(\max_{j=1,\ldots,p}\frac{2}{n}|(\epsilon^T X)_j| \leq \frac{\lambda}{2}\right) \geq 1-e^{-\frac{t^2}{2}} \tag{6.34}$$

Proof.

$$\frac{1}{n}\epsilon^T X \sim N(0,\frac{1}{n}X^T X) = N(0,\hat{\sigma}) \tag{6.35}$$

and so $\frac{1}{\sqrt{n}}(\epsilon^T X)_j \sim N(0,\hat{\sigma}_{jj})$, and so

$$\mathbb{P}\left(\max_{j=1,\ldots,p}\frac{2}{n}|(\epsilon^T X)_j| > \frac{\lambda}{2}\right) \leq \sum_{j=1}^{p} \tag{6.36}$$

$$\tag{6.37}$$

□

Fill in proof from lecture notes

Now, can conclude. We have

$$\frac{1}{n}\|X(\theta^0-\tilde{\theta})\|_2^2+\lambda\|\tilde{\theta}\|_1 \leq \frac{\lambda}{2}\|\tilde{\theta}-\theta^0\|_1+\lambda\|\theta^0\|_1 \tag{6.38}$$

$$\|\tilde{\theta}\|_1 = \sum_{i=1}^{p}|\tilde{\theta}_i| = \|\tilde{\theta}_{S_0}\|_1+\|\theta_{S_0^c}\|_1 \geq \|\theta^0\|_1-\|\tilde{\theta}_{S_0}-\theta^0\|+\|\tilde{\theta}_{S_0^c}\|_1 \tag{6.39}$$

which gives

$$\|\tilde{\theta}_{S_0^c}\|_1 \leq \|\tilde{\theta}\|_1+\|\tilde{\theta}_{S_0}-\theta^0\|_1 \tag{6.40}$$

and so

$$\frac{2}{n}\|X(\tilde{\theta}-\theta^0)\|_2^2+2\lambda\|\tilde{\theta}_{S_0^c}\|_1 \tag{6.41}$$

$$\leq \frac{2}{n}\|X(\tilde{\theta}-\theta^0)\|_2^2+2\lambda\|\tilde{\theta}\|_1-2\lambda\|\theta^0\|_1+2\lambda\|\tilde{\theta}_{S_0}-\theta^0\|_1 \tag{6.42}$$

$$\leq \lambda\|\tilde{\theta}-\theta^0_{S_0}\|_1+2\lambda\|\theta^0\|_1-2\lambda\|\theta^0\|_1+2\lambda\|\tilde{\theta}_{S_0}-\theta^0_{S_0}\|_1 \tag{6.43}$$

$$= 3\lambda\|\tilde{\theta}_{S_0}-\theta^0_{S_0}\|_1+\lambda\|\tilde{\theta}_{S_0^c}\|_1 \tag{6.44}$$

and so we obtain

$$\frac{2}{n}\|X(\tilde{\theta}-\theta^0)\|_2^2+\lambda\|\tilde{\theta}_{S_0^c}\|_1 \leq 3\lambda\|\tilde{\theta}_{S_0}-\theta^0_{S_0}\|_1 \tag{6.45}$$

and then

$$\begin{aligned}\frac{2}{n}\|X(\tilde{\theta}-\theta^0)\|_2^2+\lambda\|\tilde{\theta}-\theta^0_{S_0}\|_1 &= \frac{2}{n}\|X(\tilde{\theta}-\theta^0)\|_2^2+\lambda\|\tilde{\theta}_{S_0^c}\|_1+\lambda\|\tilde{\theta}_{S_0}-\theta^0_{S_0}\|_1 && (6.46)\\ &\leq 4\lambda\|\tilde{\theta}_{S_0}-\theta^0_{S_0}\|_1 && (6.47)\\ &\leq 4\lambda\sqrt{r_0 k}(\tilde{\theta}_{S_0}-\theta^0)^T\hat{\sigma}(\tilde{\theta}_{S_0}-\theta^0) && (6.48)\\ &= 4\lambda\sqrt{r_0 k}\frac{1}{n}\|X(\tilde{\theta}_{S_0}-\theta^0) && (6.49)\\ &\leq 4\lambda^2 r_0 k+\frac{1}{n}\|X(\tilde{\theta}_{S_0}-\theta^0)\|_2^2 && (6.50)\end{aligned}$$

Thus

$$\frac{1}{n}\|X(\tilde{\theta}-\theta^0)\|_2^2+\lambda\|\tilde{\theta}-\theta^0\|_1 \leq 4\lambda^2 r_0 k \tag{6.51}$$

□

Remark 6.4 (Remarks on Theorem 6 in Lecture Notes). *(i) Using* $\mathbb{E}(X)=\int_0^\infty P(X>u)du$ *and if* $\beta=0$ *the result of Theorem 6 be "integrated" to give a risk bound*

$$\sup_{\theta\in B_n(k)}\mathbb{E}_\theta\frac{1}{n}\|X(\tilde{X}-\theta^0)\|_2^2 \lesssim \frac{k}{n}\log p \tag{6.52}$$

(ii) One can show (exercise sheet) that with high probability,

$$\|\tilde{\theta}-\theta^0\|_2^2 \lesssim \frac{k}{n}\log p \tag{6.53}$$

(iii) If th error variance is σ^2 *is not known, we may approximate it by* $\hat{\sigma}^2=\frac{1}{n}Y^TY$*, and multiplying* λ *by* $\hat{\sigma}$*.*

About condition (6.27):

$$\begin{aligned}\|\tilde{\theta}_{S_0}-\theta^0\|_1^2 &\leq k r_0(\tilde{\theta}-\theta^0)^T\hat{\Sigma}(\tilde{\theta}-\theta^0) && (6.54)\\ &= k r_0\frac{1}{n}\|X(\tilde{\theta}-\theta^0)\|_2^2 && (6.55)\\ \hat{\Sigma} &= \frac{X^TX}{n} && (6.56)\end{aligned}$$

It suffices to check (6.27) with $\tilde{\theta}$ replaced with $\forall \theta \in U$, with $P(\tilde{\theta} \in U) = 1$.

In the proof of Theorem 6, we have shown

$$\|\tilde{\theta}_{S_0^c} \leq 3\|\tilde{\theta}_{S_0} - \theta^0\|_1 \tag{6.57}$$

holds with probability 1.

Corollary 6.5. *Theorem 6 still holds if* (6.27) *is replaced by the condition*

$$\forall \theta \in U = \{\theta \in R^p : \|\theta_{S_0^c}\|_1 \leq 3\|\theta_{S_0} - \theta^0\|_1\} \tag{6.58}$$

we have

$$\|\theta_{S_0} - \theta\|_1^2 \leq k r_0 (\theta - \theta^0)^T \hat{\Sigma} (\theta - \theta^0) \tag{6.59}$$

Note that $\|\theta_{S_0} - \theta^0\|_1^2 \leq k\|\theta_{S_0} - \theta^0\|_2^2$ *since there are at most k non-zero entries of* $\theta_{S_0} - \theta_0$.

So it remains to check that

$$\|\theta_{S_0} - \theta^0\|_2^2 \leq r_0 (\theta - \theta^0)\hat{\Sigma}(\theta - \theta^0) = r_0 \frac{1}{n}\|X(\theta - \theta^0)\|_2^2 \tag{6.60}$$

6.1 Compressed Sensing and the Restricted Isometry Property

Consider a signal $Y \in \mathbb{R}^n$ and a "sensing"/design matrix $X_j \in \mathbb{R}^n, j = 1, \ldots, p$ such that $Y = X\theta = \sum_{i=1}^p X_j \theta_j$, without noise.

If $p > n$, the representation is under-determined, but we may aim to find the "most sparse" solution. Formally, if $\|\theta_0\|_0 = |\{\theta_j \neq 0\}|$, we want to find the solution

$$\min_{\theta \in \mathbb{R}^p} \|\theta\|_0 \tag{6.61}$$

such that $Y = X\theta$.

If the sensing matrix satisfies the restricted isometry property (RIP)

$$(1 - \epsilon)\|\theta\|_2 \leq \frac{1}{n}\|X\theta\|_2^2 \leq (1 + \epsilon)\|\theta\|_2^2 \quad \forall \theta \in B_0(k) \tag{6.62}$$

for some/all $\epsilon (= \epsilon_k)$, and if there exists a k-sparse solution θ^0 such

that $Y = X\theta^0$, and if $k < n$, then the solution of the convex relaxation of (6.61) ($\min_{\theta \in \mathbb{R}^p} \|\theta\|_1$ such that $Y = X\theta$) is exactly equal to the solution of (6.61).

We note that in a Gaussian random matrix, we have $\frac{X^T X}{n}$ satisfies RIP with probability greater than $1 - e^{-k \log p}$.

Remark 6.6. *An intuition for Theorem 6 can thus be given as follows.*

(i) In the noise model $Y = X\theta + \epsilon$, with $p > n$, we can detect sparse submodels from model selection criterion of the form

$$\|Y - X\theta\|_2^2 + \|\theta\|_0 \tag{6.63}$$

.

What was the second term

(ii) A convex relaxation of the l_0 penalty is also possible, along the CRT-ideas, using condition (6.27), which can be shown to be implied by (RIP).

(iii) Candes and Tao (Annals of Statistics) showed further that a "dual" estimator of the LASSO is obtained from

$$\min_{\theta \in \mathbb{R}^p} \|\theta\|_1 \tag{6.64}$$

such that

$$\frac{1}{n}\|Y - X\theta\|_2^2 \leq t \tag{6.65}$$

*where t is the analogue of λ in the definition of the LASSO. This estimator is called the **Danzig selector** $\tilde{\theta}_{DANTZIG}$. One then shows that the "exact recovery" of the most sparse selector θ^0 still holds approximately, with large probability.*

Recall the restricted isometry property,

$$(1 - \epsilon)\|\theta\|_2^2 \leq \|\hat{\Sigma}\theta\|_2^2 \leq (1 + \epsilon)\|\theta\|_2^2 \, \forall \theta \in B_0(k) \tag{6.66}$$

Recall that

$$\hat{\Sigma} = \frac{X^T X}{n} \tag{6.67}$$

for $p > n$ is not invertible, so that difficult part

$$\inf_{\theta \in \mathbb{R}^p, \|\theta\|_2 \leq 1} \theta^T \hat{\Sigma} \theta = |(\lambda_{min}(\Sigma))| = 0 \tag{6.68}$$

Consider $(X_{ij}) \sim N(0,1)$ giving rise to X. Then

$$(\hat{\Sigma})_{jj} = \frac{1}{n} \sum_{i=1}^{n} X_{ij}^2 (\hat{\Sigma})_{jk} = \frac{1}{n} \sum_{i=1}^{n} X_{ij} X_{ik} \tag{6.69}$$

and we see $\mathbb{E}\left(\hat{\Sigma}_{jj}\right) = 1$ and $\mathbb{E}\left(\hat{\Sigma}_{jk}\right) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left(X_{ij} X_{ik}\right) = 0$. Thus $\mathbb{E}\left(\hat{\Sigma}\right) = I_p$, the identity matrix.

Theorem 6.7. *For X_{ij} as above, and $\frac{n}{\log p} \to \infty$ (or $\min(p,n) \to \infty$) is large enough. Then for all $k \in N$, there exists a constant $0 < c < \infty$ such that for all n large enough,*

$$\mathbb{P}\left(\theta^T \hat{\Sigma} \theta \geq \frac{1}{2} \theta^T \theta \, \forall \theta \in B_0(k)\right) \geq 1 - 2e^{-Ck \log p} \tag{6.70}$$

Proof. It suffices to bound

$$\mathbb{P}\left(\frac{\theta^T \hat{\Sigma} \theta}{\theta^T \theta} - 1 \geq -\frac{1}{2} \, \forall \theta \in B_0(k) \backslash \{0\}\right) = \mathbb{P}\left(1 - \frac{\theta^T \hat{\Sigma} \theta}{\theta^T \theta} \leq \frac{1}{2} \, \forall \theta \in B_0(k) \backslash \{0\}\right) \tag{6.71}$$

$$\geq \mathbb{P}\left(\sup_{\theta \in B_0(k), \theta \neq 0} \left| \frac{\theta^T \hat{\Sigma} \theta}{\theta^T \theta} - 1 \right| \leq \frac{1}{2}\right) \tag{6.72}$$

Let $S \subseteq \{1, \ldots, p\}$ of cardinality $|S| = k$, then we can bound

$$\mathbb{P}\left(\sup_{\theta \in B_0(k), \theta \neq 0} \left| \frac{\theta^T \hat{\Sigma} \theta}{\theta^T \theta} - 1 \right| > \frac{1}{2}\right) = \mathbb{P}\left(\max_S \sup_{\theta \in R_S^p} \left| \frac{\theta^T \hat{\Sigma} \theta}{\theta^T \theta} - 1 \right| > \frac{1}{2}\right) \tag{6.73}$$

where R_S^p is the subspace of $\mathbb{R}^p$ with $\theta_j = 0$ for $\theta \in S^c$.

This is bounded by

$$\leq \sum_{S \in \{1, \ldots, p\}} \mathbb{P}\left(\sup_{\theta \in R_S^p, \theta \neq 0} \left| \frac{\theta^T \hat{\Sigma} \theta}{\theta^T \theta} - 1 \right| > \frac{1}{2}\right) \tag{6.74}$$

where the sum extends over $\binom{p}{k \leq p^k}$ elements. Thus the last sum is

$$\leq^{(\star)} \sum_{S \subseteq \{1,\dots,p\}} 2e^{-(c+1)k\log p} \leq 2e^{-Ck\log p} \underbrace{p^k e^{-k\log p}}_{\leq 1} = 2e^{-Ck\log p} \quad (6.75)$$

where we used

$$\mathbb{P}\left(\sup_{\theta \in R_S^p, \theta \neq 0} \left| \frac{\theta^T \hat{\Sigma} \theta}{\theta^T \theta} - 1 \right| > \frac{1}{2} \right) \leq 2e^{-(C+1)k\log p} \qquad (\star)$$

Lemma 6.8.

$$\mathbb{P}\left(\left| \frac{\theta^T \hat{\Sigma} \theta}{\theta^T \theta} - 1 \right| > 18(\sqrt{\frac{t + c_0 k}{n}} + \frac{t + c_0 k}{n}) \right) \leq 2e^{-t} \; \forall t > 0 \text{ and some } c_0 < \infty \quad (6.76)$$

This lemma implies $(\star)$ *since for* $t = (c+1)k\log p$ *we have*

$$18(\sqrt{\frac{(c+1)k\log p + c_0 k}{n}} + \frac{(c+1)k\log p + c_0 k}{n}) \to 0 \qquad (6.77)$$

since we had assumed $\frac{n}{\log p} \to 0$.

Proof (Proof of Lemma).

$$\sup_{\theta \in R_S^p, \theta \neq 0} \left| \frac{\theta^T \hat{\Sigma} \theta}{\theta^T \theta} - 1 \right| = \sup_{\dots} \left| \frac{\theta^T (\hat{\Sigma} - I) \theta}{\theta^T \theta} \right| \qquad (6.78)$$

$$\leq \underbrace{\sup_{\theta \in \mathbb{R}_S^p, \|\theta\|_2 \leq 1}}_{B(S)} \left| \theta^T (\underbrace{\hat{\Sigma} - I}_{\Phi}) \theta \right| \qquad (6.79)$$

$$= \sup_{\theta \in B(S)} |\theta^T \Phi \theta| \qquad (6.80)$$

Since the unit ball $B(S)$ of R_S^p is compact, we can find (for all $\delta > 0$) a "net" of points $\theta^l, l = 1, \dots, N(\delta)$ such that every $\theta \in B(S)$ is at distance at most $\|\theta - \theta^l\| < \delta$ away from some θ^l. We may take $\theta^l \in B(S)$.

Writing

$$\theta^T\Phi\theta = (\theta - \theta^l + \theta^l)^T\Phi(\theta - \theta^l + \theta^l) \tag{6.81}$$

$$= (\theta^l)^T\Phi\theta^l + \underbrace{(\theta - \theta^l)\Phi(\theta - \theta^l)}_{(I)} + \underbrace{2(\theta - \theta^l)\Phi\theta^l}_{(II)} \tag{6.82}$$

$$|(I)| = \|\theta - \theta^l\|_2^2 \frac{(\theta - \theta^l)^T}{\|\theta - \theta^l\|_2}\Phi\frac{(\theta - \theta^l)^T}{\|\theta - \theta^l\|_2} \tag{6.83}$$

$$\leq \delta^2 \sup_{v\in B(S)} |v^T\Phi v| \tag{6.84}$$

$$|(II)| = 2\|\theta - \theta^l\|_2 \left|\frac{(\theta - \theta^l)^T}{\|\theta - \theta^l\|_2}\Phi\theta^l\right| \tag{6.85}$$

$$\leq^{CS} 2\delta\|\Phi\theta^l\|_2 \tag{6.86}$$

$$\leq 2\delta\|\Phi\|_{op}\|\theta^l\|_2 \tag{6.87}$$

$$= 2\delta \sup_{v\in B(S)} |v^T\Phi v| \tag{6.88}$$

So we proved

$$\sup_{\theta\in B(S)} |\theta^T\Phi\theta| \leq \max_{l=1,\ldots,N(\delta)} |\theta^l\Phi\theta^l| + (\delta^2 + 2\delta) \sup_{v\in B(S)} |v^T\Phi v| \tag{6.89}$$

Set $\delta = \frac{1}{3}$.

$$(1 - \frac{7}{9}) \sup_{\theta\in B(S)} |\theta^T\Phi\theta| \leq \max_{l=1,\ldots,N(\delta)} |\theta^l\Phi\theta^l| \tag{6.90}$$

showing

$$\sup_{\theta\in B(S)} |\theta^T\Phi\theta| \leq \frac{9}{2} \max_{l=1,\ldots,N(\delta)} |\theta^l\Phi\theta^l| \tag{6.91}$$

$$\theta^T\Phi\theta = \theta^T(\hat{\Sigma} - I)\theta = \frac{1}{n}\sum_{i=1}^n (X\theta)_i^2 - \mathbb{E}\Big((X\theta)_i^2\Big) \tag{6.92}$$

since

$$\theta^T\hat{\Sigma}\theta = \frac{1}{n}(X\theta)^T X\theta = \frac{1}{n}\sum_{i=1}^n (X\theta)_i^2 \tag{6.93}$$

and

$$\mathbb{E}\Big((X\theta)_i^2\Big) = \mathbb{E}\left(\left(\sum_{m=1}^{p} X_{im}\theta_m\right)^2\right) \tag{6.94}$$

$$= \mathbb{E}\left(\sum_{m=1}^{p} X_{im}^2\theta_{im}^2\right) + \mathbb{E}\left(\sum_{m\neq m'} X_{im}X_{im'}\theta_m\theta_{m'}\right) \tag{6.95}$$

$$= \|\theta\|_2^2 \tag{6.96}$$

$$= \theta^T\theta \tag{6.97}$$

and so it suffices to prove

$$\mathbb{P}\left(\max_{l=1,\ldots,N(\frac{1}{3})}\left|\frac{1}{n}\sum_{i=1}^{n}(X\theta)_i^2 - \mathbb{E}\Big((X\theta)_i^2\Big)\right| > \frac{2}{9}\cdot 18(\sqrt{\frac{t+c_0k}{n}} + \frac{t+c_0k}{n})\right) \tag{6.98}$$

We have that (6.98) is bounded by

$$\leq \sum_{l=1}^{N(\frac{1}{3})}\mathbb{P}\left(\left|\sum_{i=1}^{n}(g_i^2-1)\right| > 4(\sqrt{n(t+c_0k)} + (t+c_0k))\right) \tag{6.99}$$

with $g_i = \frac{(X\theta)_i}{\|\theta\|_2} \sim N(0,1)$.

Recall that $\mathbb{P}\big(|X|^2 > u^2\big) \leq e^{-\frac{u}{2}}$ for $X \sim N(0,1)$.

Lemma 6.9. *Let* $X = \sum_{i=1}^{n}(g_i^2 - 1)$ *where* $g_i \sim N(0,1)$. *Then for all* $t > 0, n \in N$,

$$\mathbb{P}(|X| > t) \leq 2e^{-\frac{t^2}{4n+4t}}, \tag{6.100}$$

and for all $z > 0$,

$$\mathbb{P}\big(|X| > 4(\sqrt{nz} + z)\big) \leq 2e^{-z}. \tag{6.101}$$

Proof. Consider, for $g \sim N(0,1)$ and λ such that $|\lambda| < \frac{1}{2}$,

$$\mathbb{E}\Big(e^{\lambda(g^2-1)}\Big) = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{\lambda(x^2-1)}e^{-\frac{x^2}{2}}dx = \frac{e^{-\lambda}}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{-(1-2\lambda)\frac{x^2}{2}}dx = \frac{e^{-\lambda}}{\sqrt{1-2\lambda}} = e^{-\frac{1}{2}[-\log(1-2\lambda)-2\lambda]} \tag{6.102}$$

Using a Taylor expansion of $\log(1-2\lambda)$, we have

$$\log(1-2\lambda) = \log 1 - 2\lambda \cdot 1 - \frac{1}{2}(2\lambda)^2 - \frac{2}{3!}(2\lambda)^3 - \cdots = -\sum_{k=1}^{\infty} \frac{(k+1)!}{(k+2)!}(2\lambda)^{k+2} \tag{6.103}$$

and

$$\frac{1}{2}[-\log(1-2\lambda) - 2\lambda] = \lambda^2[1 + \cdots + \frac{2}{k+2}(2\lambda)^k + \dots] \tag{6.104}$$

$$\leq \frac{\lambda^2}{1-2\lambda} \tag{6.105}$$

for $|\lambda| < \frac{1}{2}$.

So in total, we have

$$\mathbb{E}\left(e^{\lambda(g^2-1)}\right) \leq e^{\frac{\lambda^2}{1-2\lambda}} \tag{6.106}$$

We can also use independence of the g_i to obtain

$$\mathbb{E}\left(e^{\lambda X}\right) = \mathbb{E}\left(e^{\lambda(\Sigma_{i=}^{n}(g_i^2-1))}\right) = \mathbb{E}\left(\prod_{i=1}^{n} e^{\lambda(g_i^2-1)}\right) \tag{6.107}$$

$$= \mathbb{E}\left(e^{\lambda(g_i^2-1)}\right)^n \leq e^{\frac{n\lambda^2}{1-2\lambda}} \tag{6.108}$$

For $\lambda, t > 0$, we have

$$\mathbb{P}(X > t) = \mathbb{P}(\lambda X > \lambda t) = \mathbb{P}\left(e^{\lambda X} > e^{\lambda t}\right) \tag{6.109}$$

$$\leq e^{-\lambda t} E^{e^{\lambda X}} \leq e^{-\lambda t} e^{\frac{n\lambda^2}{1-2\lambda}} \tag{6.110}$$

by Markov's inequality.

We can optimize in λ, and choose

$$\lambda = \frac{t}{2n+2t}, \tag{6.111}$$

in which we obtain the bound

$$e^{-\frac{t^2}{2n+2t}} \exp\left\{\frac{\frac{nt^2}{(2n+2t)^2}}{1-\frac{2t}{2n+2t}}\right\} = e^{-\frac{t^2}{2n+2t}} e^{\frac{t^2}{2(2n+2t)}} = e^{-\frac{t^2}{2(2n+2t)}} \tag{6.112}$$

For the lower deviations, we repeat the above proof with $\lambda \mapsto -\lambda$,

and bound $\mathbb{P}(X > -t) \leq e^{-\frac{t^2}{2(2n+2t)}}$ and so

$$\mathbb{P}(|X| > t) \leq \mathbb{P}(X > t) + \mathbb{P}(X < -t) \leq 2e^{-\frac{t^2}{2(2n+2t)}}. \tag{6.113}$$

For the second inequality, substitute $t = 4(\sqrt{nz} + z)$ into the first inequality to get

$$\mathbb{P}\big(|X| > 4(\sqrt{nz} + z)\big) \leq 2\exp\left\{-\frac{16(\sqrt{nz} + z)^2}{4n + 16\sqrt{nz} + 16z}\right\} \leq^{?} 2e^{-z} \tag{6.114}$$

which follows since

$$16nz + 32\sqrt{n}z^{\frac{3}{2}} + 16z^2 \geq 4nz + 16\sqrt{n}z^{\frac{3}{2}} + 16z^2. \tag{6.115}$$

□

In (6.98) we get (with $z = t + c_0 k$)

$$2N(\frac{1}{3})e^{-t}e^{c_0 k} \leq 2e^{-t}(3A)^k e^{-c_0 k} \leq 1 \tag{6.116}$$

for c_0 large enough. [2] □

□

[2] One shows that for all $\delta > 0$,

$$N(\delta) \leq (\frac{A}{\delta})^k. \tag{6.117}$$

So for $\delta = \frac{1}{3}$, we have

$$N(\frac{1}{3}) \leq (3A)^k \leq e^{c_0 k} \tag{6.118}$$

for c_0 large enough.

Remark 6.10. *We have show that $\theta^T \hat{\Sigma} \theta$ concentrates around its expectation $\theta^T \theta$ uniformly in all $\theta \in B_0(k)$, as long as $\frac{k \log p}{n} \to 0$. This corresponds to the "true" sparse model being of dimension $\leq n$.*

Remark 6.11. *Theorem 6 (in the lecture notes) holds as well if the (X_{ij})'s are **sub**-Gaussian in that they satisfy the tail estimate*

$$\mathbb{P}\big(|X_{ij}| > u\big) \leq Ce^{-\frac{u^2}{2\sigma^2}}, \tag{6.119}$$

for some constants $C, \sigma^2 > 0$. To prove this, one replaces the concentration inequality for squared Gaussians by the Bernstein's inequality for sub-exponential random variables (see Bühlmann and van der Vaart for the inequality).

Remark 6.12. *We see the above proof generalizes to correlated designs such*

that

$$\underset{=}{\mathbb{E}_n(X^TX)}\Sigma \tag{6.120}$$

whee Σ *is any* $\times p$ *matrix with* $\lambda_{min}(\Sigma)$ *bounded away from zero.*

6.2 *Inference with the LASSO*

The construction of a confidence set for θ is an obvious task One construction that can be used is based on **unbiased risk estimation** (and sample splitting).

Split the sample index set I into two subsets I_1, I_2 of approximately equal size, compute $\tilde{\theta}^{(1)}_{LASSO}$ based on I_1, and let $Y^{(2)}, X^{(2)}$ be the observations from I_2. Then compute

$$(Y^{(2)} - X^{(2)}\tilde{\theta}^{(1)}_{LASSO})^T(Y^{(2)} - X^{(2)}\tilde{\theta}^{(1)}_{LASSO}) - 1 + \frac{2Z_\alpha}{\sqrt{n}} \equiv \hat{R}_n \tag{6.121}$$

where Z_α are the $1-\alpha$ quantiles of a $N(0,1)$-distribution.

Our confidence set is then

$$C = \{\theta \in \mathbb{R}^p | \|\theta - \tilde{\theta}^{(1)}_{LASSO}\|_2 \leq \hat{R}_n\} \tag{6.122}$$

which satisfies

$$\lim_{n,p\to\infty} \inf_{\theta \in B_0(h)} \mathbb{P}_\theta(\theta \in C_n) \geq 1 - \alpha \tag{6.123}$$

for all k such that $\frac{k \log p}{n} \to 0$.

One would want

$$\sup_{\theta \in B_0(k)} \mathbb{E}_\theta | \underbrace{C_n}_{l^2 \text{ diameter of } C_n} |_2^2 \lesssim \frac{k}{n} \log p \tag{6.124}$$

so that the confidence interval reflects the accuracy of estimation of $\tilde{\theta}_{LASSO}$.

One can show

$$\mathbb{E}_\theta |C_n|^2 \lesssim \frac{k}{n} \log p + \frac{1}{\sqrt{n}}, \tag{6.125}$$

which gives the result that

$$\frac{k}{n}\log p \gg \frac{1}{\sqrt{n}} \iff k\log p \gg \sqrt{n}. \tag{6.126}$$

Uniformly in $B_0(k)$, no further improvement is possible.[3]

[3] An information theoretic bound.

7
Conclusion

7.1 Outlook on Nonparameterics

What if we relax our parametric assumption on p?

(i) We have $X_1, \ldots, X_n \sim^{iid} p$, where we can parameterize p by cumulative distribution functions or probability densities.

(ii) We have $Y_i = f(X_i) + \epsilon_i$, with $\epsilon_i \sim N(0, I)$, with nothing known about f.

7.2 Relevant previous Tripos questions

(i) 2013 - 1, 2

(ii) 2012 - 1, 2

(iii) 2011 - 1, 3, 4

(iv) 2010 - 1, 2

(v) 2009 - 1, 2

8

Bibliography