① Decision theory { minimaxity, admissibility, James-Stein

② Bayesian Statistics { empirical Bayes, EM, variational Bayes, BVM

③ Model Selection { BIC, AIC, Lasso

④ Hypothesis testing { Neyman-pearson lemma, Chernoff information, Le Cam's method, multiple comparison { global null, FWER, FDR

① E. Ichmann Theory of point Estimation - minimaxity, admissibility

② A van der Vaart Asymptotic Statistics

③ I. Johnstone Gaussian sequence (asymptotic)

④ E. Candes Stat 300c Lecture notes

Decision Theory (A. Wald)

$X_1 \cdots X_n \overset{iid}{\sim} P_\theta$ ← parameter of interest $\quad \theta \in \Theta$ ← parameter space

$\hat{\theta} = \hat{\theta}(X_1 \cdots X_n)$ ← estimator

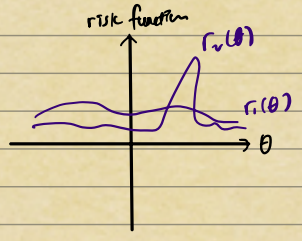
Loss function $L(\hat{\theta}, \theta)$ e.g. $\|\hat{\theta} - \theta\|^2$

risk $R(\hat{\theta}, \theta) = E(L(\hat{\theta}, \theta))$

$= \int L(\hat{\theta}(x), \theta) P_\theta(x) dx$

two estimators: $\hat{\theta}$, $\tilde{\theta}$

$r_1(\theta) = R(\hat{\theta}, \theta)$ $\quad r_2(\theta) = R(\tilde{\theta}, \theta)$



Idea: { average risk $\int R(\hat\theta,\theta)\,\underline{\pi(\theta)}\,d\theta$ ↳ prior dist of θ

maximum risk $\sup_{\theta\in\Theta} R(\hat\theta,\theta)$

def ① $\hat\theta$ is a Bayes estimator if $\hat\theta = \arg\min_{\hat\theta} \int R(\hat\theta,\theta)\pi(\theta)\,d\theta$

↳ $\forall \tilde\theta \quad \int R(\hat\theta,\theta)\pi(\theta)\,d\theta \le \int R(\tilde\theta,\theta)\pi(\theta)\,d\theta$

② $\hat\theta$ is a minimax estimator if $\hat\theta = \arg\min_{\hat\theta} \max_{\theta\in\Theta} R(\hat\theta,\theta)$

↳ $\forall \tilde\theta, \ \sup_{\theta\in\Theta} R(\hat\theta,\theta) \le \sup_{\theta\in\Theta} R(\tilde\theta,\theta)$

Bayes estimator $\hat\theta_\pi = \arg\min_{\hat\theta} \int R(\hat\theta,\theta)\pi(\theta)\,d\theta$

$$\int R(\hat\theta,\theta)\pi(\theta)\,d\theta = \iint L(\hat\theta(x),\theta)\,P_\theta(x)\,dx\,\pi(\theta)\,d\theta$$

$$= \iint L(\hat\theta(x),\theta)\,\underbrace{p(x|\theta)\pi(\theta)}\,dx\,d\theta$$

$$p(x|\theta)\pi(\theta) = \frac{p(x|\theta)\pi(\theta)}{\int p(x|\theta)\pi(\theta)\,d\theta}\int p(x|\theta)\pi(\theta)\,d\theta$$

$$= \pi(\theta|x)\,m(x)$$

($\pi(\theta|x)$: posterior; $m(x)$: marginal density)

$$\int R(\hat\theta,\theta)\pi(\theta)\,d\theta = \iint \underbrace{L(\hat\theta(x),\theta)\,\pi(\theta|x)\,d\theta}_{\text{a function of } x}\,m(x)\,dx$$

Claim: $\hat\theta_\pi(x) = \arg\min_a \int L(a,\theta)\,\pi(\theta|x)\,d\theta$

pf) WTS $\forall \hat\theta$,

$$\int R(\hat\theta_\pi,\theta)\pi(\theta)\,d\theta \le \int R(\hat\theta,\theta)\pi(\theta)\,d\theta$$

$$\int R(\hat\theta_\pi,\theta)\pi(\theta) = \iint L(\hat\theta_\pi(x),\theta)\,\pi(\theta|x)\,m(x)\,d\theta\,dx$$

$$\le \iint L(\hat\theta(x),\theta)\,\pi(\theta|x)\,d\theta\,m(x)\,dx$$

$$= \int R(\hat\theta,\theta)\pi(\theta)\,d\theta$$

an important example $\Theta \subseteq \mathbb{R}$, $L(\hat\theta,\theta) = (\hat\theta-\theta)^2$

$$\hat\theta_\pi(x) = \arg\min_a \int (a-\theta)^2\pi(\theta|x)\,d\theta$$

$= \underset{a}{\operatorname{argmin}}\ \mathbb{E}\left((\theta-\theta)^2 \mid X\right)$

$= E(\theta|X)$ $\quad \left(\text{Since } E\left((\theta-\theta)^2|X\right) = \left(E(\theta|X)-a\right)^2 + \operatorname{Var}(\theta|X)\right)$

e.g. $X_1 \cdots X_n \overset{iid}{\sim} \text{Ber}(p)$ loss $(\hat{p}-p)^2$

prior $\pi = \text{Beta}(\alpha,\beta)$ $\quad \pi(p) \propto p^{\alpha-1}(1-p)^{\beta-1}$

$p|X_1 \cdots X_n \sim \text{Beta}\left(\sum_{i=1}^n X_i+\alpha,\ \sum_{i=1}^n (1-X_i)+\beta\right)$

Bayes estimator $\hat{p} = E(p|X_1 \cdots X_n) = \frac{\sum_{i=1}^n X_i+\alpha}{n+\alpha+\beta}$

$R(\hat{p},p) = E_p(\hat{p}-p)^2 = \operatorname{Var}(\hat{p}) + \left(E(\hat{p})-p\right)^2$

$= \left(\frac{n}{n+\alpha+\beta}\right)^2 \frac{p(1-p)}{n} + \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)^2 \left(\frac{\alpha}{\alpha+\beta}-p\right)$

Risk of MLE $\qquad$ Expectation of prior mean

minimax estimator $\quad \hat{\theta}_{\text{minimax}} = \underset{\hat{\theta}}{\operatorname{argmin}} \sup_{\theta\in\Theta} R(\hat{\theta},\theta)$

Theorem If for some π, $\hat{\theta}$ satisfies

$$\sup_{\theta\in\Theta} R(\hat{\theta},\theta) = \inf_{\tilde{\theta}} \int R(\tilde{\theta},\theta)\pi(\theta)\,d\theta$$

then $\hat{\theta}$ is minimax.

proof) $\forall \tilde{\theta}$

$$\sup_{\theta\in\Theta} R(\tilde{\theta},\theta) \geq \int R(\tilde{\theta},\theta)\pi(\theta)\,d\theta$$

$$\geq \inf_{\tilde{\theta}} \int R(\tilde{\theta},\theta)\pi(\theta)\,d\theta$$

$$= \sup_{\theta\in\Theta} R(\hat{\theta},\theta)$$

Corollary If $\hat{\theta} = \hat{\theta}_\pi$ for some π and $R(\hat{\theta}_\pi,\theta)$ is constant over $\theta\in\Theta$

then $\hat{\theta}$ is minimax.

proof) $\sup_{\theta\in\Theta} R(\hat{\theta},\theta) = \int R(\hat{\theta},\theta)\pi(\theta)\,d\theta$

$$= \inf_{\tilde{\theta}} \int R(\tilde{\theta},\theta)\pi(\theta)\,d\theta$$

$\hat{\theta}$ is minimax.

e.g. $X_1 \cdots X_n \overset{iid}{\sim} \text{Bernoulli}(p)$

loss $(\hat{p}-p)^2$

$$R(\hat{p}, p) = \left[\left(\frac{\alpha+\beta}{n+\alpha+\beta}\right)^2 - \frac{1}{n}\left(\frac{n}{n+\alpha+\beta}\right)^2\right] p^2 + \left[\frac{1}{n}\left(\frac{n}{n+\alpha+\beta}\right)^2 - \left(\frac{\alpha+\beta}{n+\alpha+\beta}\right)^2 \frac{2\alpha}{\alpha+\beta}\right] p$$

$$+ \left(\frac{\alpha+\beta}{n+\alpha+\beta}\right)^2 \left(\frac{\alpha}{\alpha+\beta}\right)^2$$

$$\begin{cases} (\alpha+\beta)^2 = n \\ 2\alpha(\alpha+\beta) = n \end{cases} \Rightarrow \quad \alpha = \beta = \frac{\sqrt{n}}{2}$$

$$\hat{p}_{\text{minimax}} = \frac{\sum X_i + \frac{\sqrt{n}}{2}}{n + \sqrt{n}}$$

$\forall p$ same

Lecture 2

lower bound $\sup_{\theta\in\Theta} R(\hat\theta,\theta) \geq \int R(\hat\theta,\theta)\pi(\theta)d\theta \quad \forall \pi$

Thm. If $\exists \pi$ s.t. $\hat\theta$ satisfies $\sup_{\theta\in\Theta} R(\hat\theta,\theta) = \inf_{\hat\theta} \int R(\hat\theta,\theta)\pi(\theta)d\theta$ then $\hat\theta$ is minimax

Cor. if $\hat\theta$ is Bayes estimator w.r.t. some π_0 and $R(\hat\theta,\theta)$ is constant then $\hat\theta$ is minimax.

e.g. $X_1 \cdots X_n \overset{iid}{\sim} Ber(p)$ loss $(\hat p - p)^2$ $\pi(p) = Beta(\alpha,\beta)$

$\hat p_\pi = E(p|X_1\cdots X_n) = \frac{\sum X_i + \alpha}{n+\alpha+\beta}$

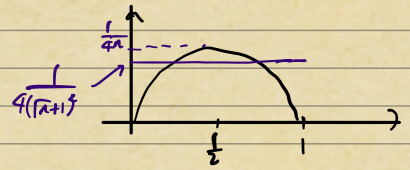
$R(\hat p_\pi, p) = [\quad]p^2 + [\quad]p + [\quad]$ (the p^2 coefficient $\to 0$) $\Rightarrow \begin{cases}\alpha = \frac{\sqrt n}{2}\\ \beta = \frac{\sqrt n}{2}\end{cases}$

$\hat p_{minimax} = \frac{\sum X_i + \frac{\sqrt n}{2}}{n+\sqrt n}$

$R(\hat p_{minimax}, p) = \frac{1}{4(\sqrt n+1)^2}$

$\hat p_{MLE} = \bar X = \frac{1}{n}\sum_{i=1}^n X_i \qquad R(\hat p_{MLE}, p) = E(\hat p_{MLE} - p)^2 = Var(\hat p_{MLE}) = \frac{p(1-p)}{n}$

$\sup_{p\in[0,1]} R(\hat p_{MLE}, p) = \frac{1}{4n} > \frac{1}{4(\sqrt n+1)^2} = R(\hat p_{minimax}, p)$



e.g. $X_1 \cdots X_n \overset{iid}{\sim} Ber(p)$ loss $l(\hat p, p) = \frac{(\hat p - p)^2}{p(1-p)}$ (just extra weight)

Bayes estimator w.r.t. $\pi(p) = 1$

$\hat p(x) = \arg\min_a \int \frac{(a-p)^2}{p(1-p)} \pi(p|x)\,dp$

$= \arg\min_a \int (a-p)^2 \frac{\pi(p|x)}{p(1-p)}\,dp$

$\frac{\pi(p|x)}{p(1-p)} \propto p^{\sum_{i=1}^n X_i - 1}(1-p)^{\sum_{i=1}^n (1-X_i)-1} = Beta\left(\sum X_i, \sum(1-X_i)\right)$

$\hat p(x) = \frac{\sum X_i}{\sum X_i + \sum(1-X_i)} = \frac{\sum X_i}{n} = \bar X \qquad R(\hat p, p) = E\left(\frac{(\hat p - p)^2}{p(1-p)}\right) = \frac{Var(\hat p)}{p(1-p)} = \frac{1}{n}$ (constant)

$\bar{X}$ is minimax

e.g. $X_1 \cdots X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ $\mu \in \mathbb{R}$ loss $(\hat{\mu}-\mu)^2$

Q. Is $\bar{X}$ minimax? $R(\bar{X}, \mu) = E(\bar{X}-\mu)^2 = \frac{\sigma^2}{n}$

Consider $\pi = N(0, \tau^2)$

$\pi(\mu|x) \propto \pi(\mu) \prod_{i=1}^n P(X_i|\mu) \propto e^{-\frac{\mu^2}{2\tau^2}} e^{-\frac{1}{2}\sum_{i=1}^n \frac{(X_i-\mu)^2}{\sigma^2}}$

$f(\mu) = \frac{\mu^2}{\tau^2} + \sum \frac{(X_i-\mu)^2}{\sigma^2}$ $\quad f'(\mu) = \frac{2\mu}{\tau^2} + \frac{1}{\sigma^2} \sum 2(\mu - X_i) = 0$

$\Leftrightarrow \frac{\mu}{\tau^2} + \frac{n\mu}{\sigma^2} = \frac{\sum X_i}{\sigma^2}$

$\Rightarrow \hat{\mu} = \frac{\frac{1}{\sigma^2}\sum X_i}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}} = \frac{\frac{n}{\sigma^2}}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}} \bar{X}$

$R(\hat{\mu}, \mu) = E(\hat{\mu}-\mu)^2 = Var(\hat{\mu}) + (E\hat{\mu}-\mu)^2$

$= \left(\frac{\frac{n}{\sigma^2}}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}}\right)^2 \frac{\sigma^2}{n} + \left(\frac{\frac{1}{\tau^2}}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}}\right)^2 \mu^2$

$\int R(\hat{\mu}, \mu)\pi(\mu)\, d\mu = \left(\frac{\frac{n}{\sigma^2}}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}}\right)^2 \frac{\sigma^2}{n} + \left(\frac{\frac{1}{\tau^2}}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}}\right)^2 \tau^2$

$= \frac{1}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}}$

Recall $R(\bar{X}, \mu) = \frac{\sigma^2}{n}$

pf that $\bar{X}$ is minimax:

$\sup_\mu R(\bar{X}, \mu) = \frac{\sigma^2}{n}$ $\quad \forall \hat{\mu}$

$\sup_\mu R(\hat{\mu}, \mu) \overset{\text{want}}{\geq} \frac{\sigma^2}{n}$

$\sup_\mu R(\hat{\mu}, \mu) \geq \int R(\hat{\mu}, \mu)\, \pi(\mu)\, d\mu$ ($\pi = N(0, \tau^2)$)

$\geq \inf_{\hat{\mu}} \int R(\hat{\mu}, \mu)\, \pi(\mu)\, d\mu$

$= \frac{1}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}} \overset{\tau^2 \to \infty}{\longrightarrow} \frac{\sigma^2}{n}$

Thm $\exists \{\pi_m\}$ s.t. $\sup_{\theta \in \Theta} R(\hat{\theta}, \theta) = \lim_m \inf_{\hat{\theta}} \int R(\hat{\theta}, \theta)\, \pi_m(\theta)\, d\theta$

then $\hat{\theta}$ is minimax

previously... ↑ more general

Thm. If $\exists \pi$ s.t. $\hat{\theta}$ satisfies $\sup_{\theta \in \Theta} R(\hat{\theta}, \theta) = \inf_{\hat{\theta}} \int R(\hat{\theta}, \theta)\pi(\theta)\, d\theta$ then $\hat{\theta}$ is minimax

Admissibility

Def. $\hat{\theta}$ is inadmissible if $\exists\, \tilde{\theta}$ s.t.

$$R(\tilde{\theta}(X), \theta) \le R(\hat{\theta}(X), \theta) \quad \forall \theta \in \Theta$$

$$R(\tilde{\theta}(X), \theta) < R(\hat{\theta}(X), \theta) \quad \text{for some } \theta_0 \in \Theta$$

Thm If $\hat{\theta}$ is Bayes, it is admissible

Suppose $\hat{\theta}$ is inadmissible $\exists\, \tilde{\theta}$

① $R(\tilde{\theta}, \theta) \le R(\hat{\theta}, \theta) \quad \forall \theta \in \Theta$

② $R(\tilde{\theta}, \theta_0) < R(\hat{\theta}, \theta_0)$ for some $\theta_0 \in \Theta$ (continuity assumption)

$\Rightarrow \exists$ open set $\Theta_0 \ni \theta_0$ and $\varepsilon > 0$ such that

$R(\tilde{\theta}, \theta) < R(\hat{\theta}, \theta) \quad \forall \theta \in \Theta_0$.

$$\Rightarrow \int R(\tilde{\theta}, \theta)\, \pi(\theta)\, d\theta = \int_{\Theta_0} R(\tilde{\theta}, \theta)\pi(\theta)\, d\theta + \int_{\Theta_0^c} R(\tilde{\theta}, \theta)\, \pi(\theta)\, d\theta$$

$$< \int_{\Theta_0} R(\hat{\theta}, \theta)\pi(\theta) d\theta + \int_{\Theta_0^c} R(\hat{\theta}, \theta)\pi(\theta) d\theta$$

$$= \int R(\hat{\theta}, \theta)\pi(\theta)\, d\theta \quad \text{contradiction}$$

Complete class theorem (Brown, 1986)

$\hat{\theta}$ is admissible $\Rightarrow \exists \{\pi_m\}$ s.t. $\hat{\theta}_{\pi_m} \to \hat{\theta}$

e.g. $X_1, \dots, X_n \overset{iid}{\sim} N(\theta, \sigma^2) \quad \theta \in \mathbb{R} \quad (\hat{\theta} - \theta)^2$

Q: Is $\bar{X}$ admissible? Wald (1939) wrong proof. Colin Blyth

$X_1, \dots, X_n \overset{iid}{\sim} N(\theta, I_p) \quad \theta \in \mathbb{R}^p \quad \|\hat{\theta} - \theta\|^2 = \sum_{j=1}^{p} (\hat{\theta}_j - \theta_j)^2$

Q. Is $\bar{X}$ admissible? (minimax, limit of Bayes, invariance property)

$p = 1$ admissible Blyth
$p = 2$ admissible Stein
$p \ge 3$ inadmissible Stein

$$\hat{\theta}_{JS} = \left(1 - \frac{p-2}{n\|\bar{x}\|}\right)\bar{X} \quad \text{James-Stein estimator}$$

Larry Brown 1971 Correspondence stochastic process recurrent or not

random walk

Lecture 3 $X_1 \cdots X_n \overset{iid}{\sim} N(\theta, I_p)$ $L(\hat\theta, \theta) = \|\hat\theta - \theta\|^2$ Q. Is $\bar X$ admissible or not?

Thm. $p=1$, admissible

pf) (Blyth's method)

Suppose $\hat\theta = \bar X$ is inadmissible

$\exists \tilde\theta$ s.t. ① $R(\tilde\theta, \theta) \le \frac{1}{n}$ $\forall \theta \in \mathbb{R}$

② $R(\tilde\theta, \theta_0) < \frac{1}{n}$ $\exists \theta_0 \in \mathbb{R}$

$\exists a < b$ $\varepsilon > 0$ s.t. $(a,b) \ni \theta_0$

$R(\tilde\theta, \theta) < \frac{1}{n} - \varepsilon$ $\forall \theta \in (a,b)$

Consider $\pi_m = N(0, m)$, $\int R(\hat\theta_{\pi_m}, \theta)\pi_m(\theta)\,d\theta = \frac{1}{n+\frac{1}{m}}$

$$\frac{1}{n} - \int R(\hat\theta_{\pi_m}, \theta)\pi_m(\theta)\,d\theta = \frac{1}{n} - \frac{1}{n+\frac{1}{m}} = \frac{1}{n}\left(1 - \frac{n}{n+\frac{1}{m}}\right)$$

$$= \frac{1}{n}\,\frac{\frac{1}{m}}{n+\frac{1}{m}} \asymp \frac{1}{m} \quad (n \text{ fixed})$$

$$\frac{1}{n} - \int R(\tilde\theta, \theta)\pi_m(\theta)\,d\theta = \frac{1}{n} - \int_{(a,b)} R(\tilde\theta, \theta)\pi_m(\theta)\,d\theta - \int_{(a,b)^c} R(\tilde\theta, \theta)\pi_m(\theta)\,d\theta$$

$$= \int_{(a,b)} \frac{1}{n} - R(\tilde\theta, \theta)\pi_m(\theta)\,d\theta + \int_{(a,b)^c} \frac{1}{n} - R(\tilde\theta, \theta)\pi_m(\theta)\,d\theta$$

$$> \varepsilon \int_{(a,b)} \pi_m(\theta)\,d\theta = \varepsilon\, P(a < N(0,m) < b)$$

$$= \varepsilon\, P\left(\frac{a}{\sqrt m} < N(0,1) < \frac{b}{\sqrt m}\right) \asymp \frac{1}{\sqrt m}$$

$\exists m$ large s.t. $\int R(\hat\theta_{\pi_m}, \theta)\pi_m(\theta)\,d\theta > \int R(\tilde\theta_{\pi_m}, \theta)\pi_m(\theta)\,d\theta$

Contradiction $\Rightarrow \hat\theta = \bar X$ is admissible

$p=2$ $\bar X$ is admissible (Blyth does not work)

proved by Stein

$p=3$ Stein $\bar X$ is inadmissible

James - Stein estimator

$$\hat\theta_{JS} = \underbrace{\left(1 - \frac{p-2}{n\|\bar X\|^2}\right)\bar X}_{\text{Stein Shrinkage}}$$

an empirical Bayes perspective (Efron)

$X_1 \cdots X_n \overset{iid}{\sim} N(\theta, I_p)$ $\quad \theta \sim N(0, \tau^2 I_p)$

$\hat{\theta}_{Bayes} = \mathbb{E}(\theta | X_1 \cdots X_n) = \frac{n}{n+\frac{1}{\tau^2}} \bar{X} = \left(1 - \frac{\frac{1}{\tau^2}}{n+\frac{1}{\tau^2}}\right) \bar{X}$

empirical Bayes framework $\quad P(X|\theta) \quad \pi(\theta|\tau^2)$

$m(x|\tau^2) = \int P(x|\theta)\pi(\theta|\tau^2)\, d\theta$

estimate τ^2 from $m(x|\tau^2)$ ← "marginal likelihood"

$X_1 \cdots X_n | \theta \overset{iid}{\sim} N(\theta, I_p)$

$\theta | \tau^2 \sim N(0, \tau^2 I_p)$

$X_1 \cdots X_n | \tau^2 \sim ?$

$X_i \sim N(\theta, I_p) \iff X_i = \theta + Z_i \quad Z_i \overset{iid}{\sim} N(0, I_p)$ } independent

$\theta \sim N(0, \tau^2 I_p) \iff \theta = \tau W \quad W \sim N(0, I_p)$ }

$\Rightarrow X_i = \tau W + Z_i \quad i = 1 \cdots n$ (dependence)

estimate τ^2 { MLE
Method of moments ⇒ James-Stein

$\bar{X} = \tau W + \bar{Z} \sim N(0, (\tau^2 + \frac{1}{n}) I_p)$

$\Rightarrow \frac{\|\bar{X}\|^2}{\tau^2 + \frac{1}{n}} \sim \chi^2_p \qquad \frac{\tau^2 + \frac{1}{n}}{\|\bar{X}\|^2} \sim \text{inv-}\chi^2_p$

method of moment

$\mathbb{E}\left(\frac{\tau^2 + \frac{1}{n}}{\|\bar{X}\|^2}\right) = \frac{1}{p-2}$

$\Rightarrow \mathbb{E}\left(\frac{p-2}{\|\bar{X}\|^2}\right) = \frac{1}{\tau^2 + \frac{1}{n}}$

We can estimate $\frac{1}{\tau^2 + \frac{1}{n}}$ by $\frac{p-2}{\|\bar{X}\|^2}$

Thm $p \geq 3 \quad \mathbb{E}\|\hat{\theta}_{JS} - \theta\|^2 < p/n = \mathbb{E}\|\bar{X} - \theta\|^2 \quad \forall \theta \in \mathbb{R}^p$

Lemma (Stein's identity) $Z \sim N(0,1)$

$\mathbb{E}\, Z g(Z) = \mathbb{E}\, g'(Z)$ (can prove CLT)

same assumptions ⇐

pf) $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \qquad \phi'(x) = -x\phi(x)$

$\mathbb{E}\, Z g(Z) = \int x g(x) \phi(x)\, dx = -\int g(x) \phi'(x)\, dx$

$= -\left(g(x)\phi(x)\Big|_{-\infty}^{\infty} - \int g'(x)\phi(x)\, dx\right)$

$= \mathbb{E}\, g'(Z)$

$Z \sim N(0, I_p)$ $\quad g: \mathbb{R}^p \to \mathbb{R}^p$ $\quad E(\langle Z, g(Z)\rangle) = E(\langle \nabla, g(Z)\rangle) = \sum_{j=1}^{p} E \frac{\partial}{\partial z_j} g_j(Z)$

pf of thm)

$$E_\theta \|\hat{\theta}_{JS} - \theta\|^2 = E_\theta \left\| \left(1 - \frac{p-2}{n\|\bar{X}\|^2}\right)\bar{X} - \theta \right\|^2$$

$$= E_\theta \left\| \bar{X} - \theta - \frac{p-2}{n\|\bar{X}\|^2}\bar{X} \right\|^2 \qquad \bar{X} \sim N(\theta, \tfrac{1}{n} I_p)$$

$$\Leftrightarrow \bar{X} = \theta + \tfrac{1}{\sqrt{n}} Z \quad Z \sim N(0, I_p)$$

$$= E_\theta \left\| \tfrac{1}{\sqrt{n}} Z - \frac{p-2}{\|\mu+Z\|^2} \tfrac{1}{\sqrt{n}} (\mu + Z) \right\|^2 \qquad = \tfrac{1}{\sqrt{n}}(\sqrt{n}\theta + Z)$$

$$= \tfrac{1}{\sqrt{n}}(\mu + Z)$$

$$= \frac{1}{n} E \left\| Z - \frac{p-2}{\|\mu+Z\|^2}(\mu+Z) \right\|^2 \qquad \mu = \sqrt{n}\theta$$

$$= \frac{1}{n}\left(E\|Z\|^2 + E\frac{(p-2)^2}{\|\mu+Z\|^2} - 2(p-2)\left\langle Z, \frac{\mu+Z}{\|\mu+Z\|^2}\right\rangle \right)$$

$$= \frac{1}{n}\left(p + E\frac{(p-2)^2}{\|\mu+Z\|^2} - 2(p-2)\left\langle Z, \frac{\mu+Z}{\|\mu+Z\|^2}\right\rangle \right)$$

analysis of $E\left(\left\langle Z, \frac{\mu+Z}{\|\mu+Z\|^2}\right\rangle\right)$ $\quad g(Z) = \begin{pmatrix} g_1(Z) \\ \vdots \\ g_p(Z) \end{pmatrix}$ $\quad g_j(Z) = \frac{\mu_j + Z_j}{\|\mu+Z\|^2}$

$$\frac{\partial}{\partial Z_j} g_j(Z) = \frac{\|\mu+Z\|^2 - 2(\mu_j+Z_j)^2}{\|\mu+Z\|^4}$$

$$E\left\langle Z, \frac{\mu+Z}{\|\mu+Z\|^2}\right\rangle = \sum_{j=1}^{p} E \frac{\partial}{\partial Z_j} g_j(Z)$$

$$= E \sum_{j=1}^{p} \frac{\|\mu+Z\|^2 - 2(\mu_j+Z_j)^2}{\|\mu+Z\|^4} = E\frac{p\|\mu+Z\|^2 - 2\|\mu+Z\|^2}{\|\mu+Z\|^4}$$

$$= E\frac{p-2}{\|\mu+Z\|^2}$$

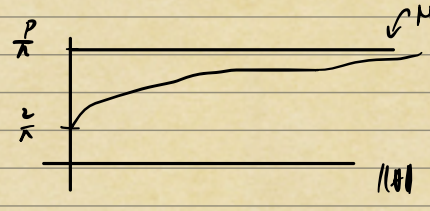
$$E_\theta \|\hat{\theta}_{JS} - \theta\|^2 = \frac{1}{n}\left(p + E\frac{(p-2)^2 - 2(p-2)^2}{\|\mu+Z\|^2}\right)$$

$$= \frac{p}{n} - \frac{1}{n} E\frac{(p-2)^2}{\|\mu+Z\|^2} \qquad \leftarrow \sqrt{n}\theta$$

$$< \frac{p}{n}$$

$$R(\hat{\theta}_{JS}, 0) = \frac{1}{n}\left(p - (p-2)^2 E\frac{1}{\|Z\|^2}\right)$$

$$= \frac{2}{n} \qquad \text{(does not depend on } p\text{)}$$



by symmetry, can show $R(\hat{\theta}_{JS}, \theta)$ is a function of $\|\theta\|$

Also show θ_{ws} is minimax.

You can only break Cramer Rao lower bound upto measure 0 set.

Lecture 4

Shrinkage estimator $\hat{\theta}_c = c\bar{X}$

$$R(\hat{\theta}_c, \theta) = E_\theta \|c\bar{X} - \theta\|^2$$

$$= E_\theta \|c\bar{X} - c\theta\|^2 + \|c\theta - \theta\|^2$$

$$= c^2 \frac{p}{n} + (c-1)^2 \|\theta\|^2 = f(c)$$

$$f'(c) = 2c\frac{p}{n} + 2(c-1)\|\theta\|^2 = 0$$

$$c^* = \frac{\|\theta\|^2}{\frac{p}{n} + \|\theta\|^2}$$

$$\hat{\theta}_{c^*} = \frac{\|\theta\|^2}{\frac{p}{n} + \|\theta\|^2}\bar{X} = \left(1 - \frac{\frac{p}{n}}{\frac{p}{n} + \|\theta\|^2}\right)\bar{X}$$

$$R(\hat{\theta}_{c^*}, \theta) = \left(\frac{b}{a+b}\right)^2 a + \left(\frac{a}{a+b}\right)^2 b \qquad \begin{cases} a = \frac{p}{n} \\ b = \|\theta\|^2 \end{cases}$$

$$= \frac{b^2 a + a^2 b}{(a+b)^2}$$

$$= \frac{ab}{a+b} \leq \min(a, b)$$

$$= \min\left(\frac{p}{n}, \|\theta\|^2\right)$$

$$R(\hat{\theta}_{JS}, \theta) = \frac{1}{n}\left(p - (p-2)^2 E\frac{1}{\|Z+\mu\|^2}\right) \qquad \begin{cases} Z \sim N(0, I_p) \\ \mu = \sqrt{n}\theta \end{cases}$$

$\|Z+\mu\|^2 \sim \chi^2_{p, \|\mu\|^2}$

an important fact $\overset{d}{=} \chi^2_{p+2N}$ $\quad N \sim \text{Poisson}\left(\frac{\|\mu\|^2}{2}\right)$

$$E\frac{1}{\|Z+\mu\|^2} = E\frac{1}{\chi^2_{p+2N}} = E\left(E\left(\frac{1}{\chi^2_{p+2N}} \mid N\right)\right)$$

(χ^2_{p+2N}: random)

$$= E\left(\frac{1}{p+2N-2}\right)$$

Jensen

$$\geq \frac{1}{p+\|\mu\|^2-2}$$

$$R(\hat{\theta}_{JS}, \theta) \leq \frac{1}{n}\left(p - (p-2)^2 \frac{1}{p+\|\mu\|^2-2}\right)$$

$$= \frac{1}{n} \frac{p^2 + p\|\mu\|^2 - 2p - p^2 + 4p - 4}{p+\|\mu\|^2-2}$$

$$= \frac{1}{n} \frac{p\|\mu\|^2 + 2p - 4}{p+\|\mu\|^2-2}$$

$$= \frac{2}{n} + \frac{(p-2)\|\mu\|^2/n}{\|\mu\|^2 + (p-2)} \qquad \mu = \sqrt{n}\theta$$

$$= \frac{2}{n} + \frac{(p-2)\|\theta\|^2/n}{\|\theta\|^2 + (p-2)/n}$$

$$\le \frac{2}{n} + \frac{\frac{p\|\theta\|^2}{n}}{\|\theta\|^2 + \frac{p}{n}} = \frac{2}{n} R(\hat{\theta}_{c^*}, \theta)$$

Theorem (oracle inequality)

$$R(\hat{\theta}_{JS}, \theta) \le \inf_c R(\hat{\theta}_c, \theta^*) + \frac{2}{n}$$ ← dimension free

density estimation

$X_1 \cdots X_n \overset{iid}{\sim} f, \quad f \in S_\alpha(R) \subseteq L^2[0,1]$ ← Sobolev ball

loss: $L(\hat{f}, f) = \|\hat{f} - f\|^2 = \int_0^1 (\hat{f}(x) - f(x))^2 dx = \|\hat{\theta} - \theta\|^2$

Fourier expansion $f(x) = a_0 + \sum_{j=1}^{\infty} \left(a_j \cos(2\pi j x) + b_j \sin(2\pi j x)\right)$

$$= \sum_{j=1}^{\infty} \theta_j \phi_j(x) \qquad \int_0^1 \phi_j(x)\phi_k(x)dx = \begin{cases} 1 & j=k \\ 0 & j \ne k \end{cases}$$

$$f'(x) = \sum_{j=1}^{\infty} \left[-a_j (2\pi j) \sin(2\pi j x) + b_j (2\pi j)\cos(2\pi j x)\right]$$

$$\|f'\|^2 = \int_0^1 (f')^2(x)dx = \sum (2\pi j)^2 (a_j^2 + b_j^2)$$

$$\|f''\|^2 = \sum_{j=1}^{\infty} (2\pi j)^4 (a_j^2 + b_j^2)$$

for $\alpha > 0$, $\|f^{(\alpha)}\|^2 = \sum_{j=1}^{\infty} (2\pi j)^{2\alpha} (a_j^2 + b_j^2)$

$$S_\alpha(R) = \left\{ f = \sum_{j=1}^{\infty} \theta_j \phi_j : \sum_{j=1}^{\infty} j^{2\alpha} \theta_j^2 \le R^2, \int f = 1 \quad f \ge 0 \right\}$$

$\inf_{\hat{f}} \sup_{f \in S_\alpha(R)} E_f \|\hat{f} - f\|^2 = (1+o(1)) C_{\alpha,R} n^{-\frac{2\alpha}{2\alpha+1}}$: "n^{th} parametric rate" cannot reach "parametric rate"

↳ faster constant

(minimax risk)

$E_\theta \|\bar{X} - \theta\|^2 = \frac{p}{n}$ n^{-1} : "parametric rate" goes to 0 faster

If $p \to \infty$, $\frac{p}{n} \to \infty$. In the above, due to regularity condition. It did not blow up

Q. How to achieve minimax rate?

empirical fourier coefficient

$$\theta_j = \int_0^1 f\phi_j = \mathbb{E}(\phi_j(X)) \approx \hat{\theta}_j = \frac{1}{n}\sum_{i=1}^{n} \phi_j(X_i)$$

by CLT : $\frac{\sqrt{n}(\hat{\theta}_j - \theta_j)}{\sqrt{\mathrm{Var}(\phi_j(X))}} \rightsquigarrow N(0,1)$

approximately $\hat{\theta}_j = \theta_j + \frac{1}{\sqrt{n}} Z_j \quad j = 1, \cdots, \infty$

$\rightsquigarrow N(0, \mathrm{Var}(\phi_j(X)))$

↕ difference is minor

infinite dimensional version of Gaussian location model

sufficient stat $\bar{X} \sim N(\theta, \frac{1}{n}\tau_p)$

$\bar{X} = \theta + \frac{1}{\sqrt{n}} Z \quad Z \sim N(0, \tau_p)$

Gaussian Sequence model (Johnstone)

$X_j = \theta_j + \frac{1}{\sqrt{n}} Z_j \quad j = 1 \cdots \infty \iff X_j \sim N(\theta_j, \frac{1}{n})$

$Z_j \sim N(0,1)$

equivalence $X_1 \cdots X_n \sim N(\theta, 1)$

$\Updownarrow$ $\bar{X} \sim N(\bar{X}, 1)$

$\bar{X}$ (sufficient)

(Nussbaum, 1996)

Nonparametric asymptotic equivalence btw two nonparametric models (previously only for parametric models)

density estimation $\overset{\text{asymptotically equivalent}}{\iff}$ Gaussian sequence model

sample $\overset{\text{certain}}{\rightleftarrows}$ sample

Help to connect problems!

non parametric regression $\iff$ white noise model

$\Updownarrow$

density estimation

$\theta \in S_\alpha(R) = \left\{ \theta : \sum_{j=1}^{\infty} j^{2\alpha} \theta_j^2 \le R^2 \right\}$

$L(\hat{\theta}, \theta) = \|\hat{\theta} - \theta\|^2 = \sum_{j=1}^{\infty} (\hat{\theta}_j - \theta)^2$ We can pay the price of bias at the cost of variance ($\frac{1}{n}$ vs. 0)

So set 0 to use the regularity

$\hat{\theta}_j = \begin{cases} X_j & j \le k \\ 0 & j > k \end{cases}$

$E_\theta \|\hat{\theta} - \theta\|^2 = \sum_{j=1}^{\infty} E(\hat{\theta}_j - \theta)^2 = \underbrace{\sum_{j=1}^{k} E(X_j - \theta_j)^2}_{\text{variance}} + \underbrace{\sum_{j=k+1}^{\infty} \theta_j^2}_{\text{bias}}$ balance $\iff$

$= \frac{k}{n} + \sum_{j=k+1}^{\infty} \theta_j^2 \le \frac{k}{n} + R^2 k^{-2\alpha}$

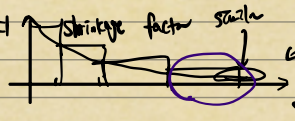
bias^2 $\sum_{j=k+1}^{\infty} \theta_j^2 = \sum_{j>k} j^{-2\alpha} j^{2\alpha} \theta_j^2 \le R^2 k^{-2\alpha}$

Choose k by $\min_k \left(\frac{k}{n} + R^2 k^{-2\alpha}\right)$

$\frac{k}{n} \asymp k^{-2\alpha} \Rightarrow k \asymp n^{\frac{1}{2\alpha+1}}$ (effective dimension of nonparametric problem)

$E_\theta \|\hat{\theta} - \theta\|^2 \le C n^{-\frac{2\alpha}{2\alpha+1}}$ minimax rate optimal

Q. Can we achieve $n^{-\frac{2\alpha}{2\alpha+1}}$ adaptively? using data



$R(\hat{\theta}_{JS}, \theta) = \frac{1}{n}\left(p - (p-2)^2 E \frac{1}{\|Z+\mu\|^2}\right)$ $\begin{cases} Z \sim N(0, I_p) \\ \mu = \sqrt{n}\theta \end{cases}$

large θ: risk increases

Lec 5

Gaussian sequence model

$X_j = \theta_j + \frac{1}{\sqrt{n}} z_j$ $z_j \overset{iid}{\sim} N(0,1)$ $\|\hat{\theta} - \theta\|^2$

$\theta \in \Theta_\alpha(R) = \{\theta : \sum_{j=1}^{\infty} j^{2\alpha} \theta_j^2 \le R^2\}$

$\inf_{\hat{\theta}} \sup_{\theta \in \Theta_\alpha(R)} \mathbb{E}\|\hat{\theta} - \theta\|^2 \asymp n^{-\frac{2\alpha}{2\alpha+1}}$ rate-optimal estimator $\hat{\theta}_j = \begin{cases} X_j & j \le k \\ 0 & j > k \end{cases}$

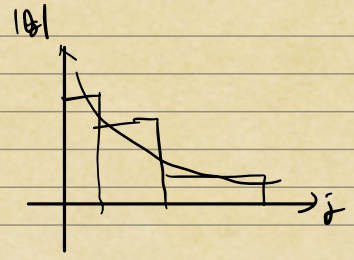
Choose $k \asymp n^{\frac{1}{2\alpha+1}}$ $\sup_{\theta \in \Theta_\alpha(R)} \mathbb{E}_\theta \|\hat{\theta} - \theta\|^2 \lesssim n^{-\frac{2\alpha}{2\alpha+1}}$

① $R(\hat{\theta}_{JS}, \theta) < \frac{p}{n}$ $\forall \theta \in \mathbb{R}^p$

② oracle inequality $R(\hat{\theta}_{JS}, \theta) \le \inf_c R(cX, \theta) + \frac{2}{n}$ $\forall \theta \in \mathbb{R}^p$

where $\inf_c R(cX, \theta) \le \min(\underbrace{\frac{p}{n}}_{R(\bar{X}, \theta)}, \underbrace{\|\theta\|^2}_{R(0,\theta)})$

better than both



$\{1 \cdots n\} = \bigcup_{j=1}^{m} B_j$ $B_1 = \{1,2,3\}$ $B_2 = \{4,5,\cdots\}$ $B_3 = \{10, \cdots, 27\}$

$B_j = \{3^{j-1}+1, \cdots, 3^j\}$ $|B_j| = 3^j - 3^{j-1} = \frac{2}{3} 3^j$

$\hat{\theta} = \begin{pmatrix} \hat{\theta}_{B_1} \\ \hat{\theta}_{B_2} \\ \vdots \end{pmatrix}$ $B_m = \{\cdots 3^m\}$ $m = \log_3 n$

for $j = 1 \cdots m$

$\hat{\theta}_{B_j} = \left(1 - \frac{|B_j| - 2}{n\|X_{B_j}\|^2}\right) X_{B_j}$

$\hat{\theta}_i = 0$ if $i > n$

$\mathbb{E}_\theta \|\hat{\theta} - \theta\|^2 = \sum_{j=1}^{m} \mathbb{E}\|\hat{\theta}_{B_j} - \theta_{B_j}\|^2 + \sum_{i>n} \theta_i^2$

$\le \sum_{j=1}^{m} \left[\min\left(\frac{|B_j|}{n}, \|\theta_{B_j}\|^2\right) + \frac{2}{n}\right] + \sum_{i>n} \theta_i^2$

$= \sum_{j=1}^{m} \min\left(\frac{|B_j|}{n}, \|\theta_{B_j}\|^2\right) + \frac{2\log_3 n}{n} + \sum_{i>n} \theta_i^2$

③ $\sum_{i>n} \theta_i^2 = \sum_{i>n} i^{-2\alpha} i^{2\alpha} \theta_i^2 \le n^{-2\alpha} \sum i^{2\alpha} \theta_i^2 \le R^2 n^{-2\alpha} = o\left(n^{-\frac{2\alpha}{2\alpha+1}}\right)$

② $= o\left(n^{-\frac{2\alpha}{2\alpha+1}}\right)$

① $\sum_{j=1}^{m} \min\left(\frac{|B_j|}{n}, \|\theta_{B_j}\|^2\right) \le \sum_{j=1}^{m} \min\left(\frac{3^j}{n}, \|\theta_{B_j}\|^2\right)$

$\forall s \in \{1,\dots,m^*\}$ $\leq \sum_{j=1}^{m^*} \frac{3^j}{n} + \sum_{j>m^*} \|\theta_{B_j}\|^2$

Chose m^* to be the smallest int. s.t. $3^m \geq n^{\frac{1}{2\alpha+1}}$

$\Rightarrow 3^{m^*-1} \leq n^{\frac{1}{2\alpha+1}} \leq 3^{m^*}$

$\sum_{j=1}^{m^*} \frac{3^j}{n} = \frac{1}{n} \sum_{j=1}^{m^*} 3^j \lesssim \frac{3^{m^*}}{n} \asymp n^{-\frac{2\alpha}{2\alpha+1}}$

$\sum_{j>m^*} \|\theta_{B_j}\|^2 \leq \sum_{i>3^{m^*-1}} |\theta_i|^2 \leq n^{-\frac{2\alpha}{2\alpha+1}}$

$\asymp n^{\frac{1}{2\alpha+1}}$

$\Rightarrow \exists c>0$ s.t. $E_\theta \|\hat{\theta}-\theta\|^2 \leq C n^{-\frac{2\alpha}{2\alpha+1}}$ $\forall \alpha>0$ Johnstone

Q. How to prove lower bound?

$\inf_{\hat{\theta}} \sup_{\theta \in \Theta_\alpha(R)} E_\theta \|\hat{\theta}-\theta\|^2 \geq C' n^{-\frac{2\alpha}{2\alpha+1}}$ for some $C'>0$

In order to prove lower bound for nonparametric estimation problem, we needs to reduce the problem to the hypothesis testing problem

two-point test (simple vs. simple)

$H_0: X \sim P$ $H_1: X \sim Q$

testing procedure $\phi: X \mapsto \phi(x) \in [0,1]$

Type I error $P\phi = E_{x\sim P} \phi(x)$

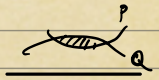
Type II error $Q(1-\phi)$

testing error $P\phi + Q(1-\phi)$

optimal testing error $\inf_{\phi} (P\phi + Q(1-\phi))$

def. total variance distance

$TV(P,Q) = \sup_B |P(B)-Q(B)|$



Theorem: $TV(P,Q) = P(p(x)>q(x)) - Q(p(x)>q(x))$ (deciding fn)

$= \frac{1}{2}\int |p-q| = 1 - \int \min(p,q)$

total variation affinity

(Pf) $A = \{p(x)>q(x)\}$

$TV(P,Q) \geq P(A_0) - Q(A_0)$

$\forall B \quad |P(B) - Q(B)| = \left|\int_B |P-Q|\right|$

$$= \left| \int_{B \cap A_0} (P-Q) + \int_{B \cap A_0^c} (P-Q) \right|$$

$$= \left| \int_{B \cap A_0} (P-Q) - \int_{B \cap A_0^c} (Q-P) \right|$$

$$\leq \max \left(\int_{B \cap A} (P-Q), \int_{B \cap A^c} (Q-P) \right)$$

$$\leq \max \left(\int_{A_0} (P-Q), \int_{A_0^c} (Q-P) \right)$$

$$= \max \left(P(A_0) - Q(A_0), Q(A_0^c) - P(A_0^c) \right)$$

$$= P(A_0) - Q(A_0)$$

Take sup over B.

$$\frac{1}{2}\int |P-Q| = \frac{1}{2}\left(\int_{P>Q} P-Q + \int_{Q \geq P} Q-P \right)$$

$$= \frac{1}{2}(P(A_0) - Q(A_0)) + \frac{1}{2}(Q(A_0^c) - P(A_0^c))$$

$$= P(A_0) - Q(A_0)$$

$$= TV(P,Q)$$

$$\int \min(P,Q) = \int_{P>Q} Q + \int_{P \leq Q} P$$

$$= Q(A_0) + P(A_0^c)$$

$$= Q(A_0) + 1 - P(A_0)$$

$$= 1 - TV(P,Q)$$

Theorem (Neyman-Pearson lemma)

$$\inf_{\phi} (P\phi + Q(1-\phi)) = 1 - TV(P,Q) = \int \min(P,Q)$$

the optimal test ϕ^* is $\phi(x) = \mathbb{1}\{P(x) < Q(x)\}$ ($\leftrightarrow$ MLE)

proof) $\forall \phi$

$$P\phi + Q(1-\phi) = \int P\phi + \int Q(1-\phi)$$

$$= \int P\phi + Q(1-\phi)$$

$$\geq \int \min(P,Q)$$

$$\Rightarrow \inf_{\phi} (P\phi + Q(1-\phi))$$

$$\inf_{\phi} \left(P\phi + Q(1-\phi)\right) \le P(p(x) \le g(x)) + Q(p(x) > g(x))$$
$$= 1 - TV(P,Q) = \int \min(p, q)$$

$H_0 : X_1 \cdots X_n \overset{iid}{\sim} P$

$H_1 : X_1 \cdots X_n \overset{iid}{\sim} Q$

$$\inf_{\phi} \left[P^n \phi + Q^n (1-\phi)\right]$$

$$= 1 - TV(P^n, Q^n)$$

$$= \int \min\left(\prod p(x_i), \prod q(x_i)\right)$$

Q. What is the rate of convergence as $n \to \infty$?

Chernoff information

Def. $C(P,Q) = -\min_{t \in (0,1)} \log \int p^{1-t} q^t$

Theorem Under some mild conditions

$$\inf_{\phi} \left(P^n \phi + Q^n (1-\phi)\right) = \exp\left(-(1+o(1)) n C(P,Q)\right)$$

Coming up: high dimensional linear regression

Lecture 6

$H_0: X_1 \cdots X_n \overset{iid}{\sim} P$

$H_1: X_1 \cdots X_n \overset{iid}{\sim} Q$

$$\inf_{\phi} \left(P^n \phi + Q^n (1-\phi)\right) = 1 - TV(P^n, Q^n)$$

Thm. Under mild assumptions, as $n \to \infty$

$$\inf_{\phi} \left(P^n \phi + Q^n (1-\phi)\right) = \exp\left[-(1+o(1)) n C(P,Q)\right]$$

where $C(P,Q) = -\min_{t \in (0,1)} \log \int p^{1-t} q^t$

proof: Upper bound

$$\inf_{\phi} \left(P^n \phi + Q^n (1-\phi)\right) = P^n\left[\prod_{i=1}^n q(X_i) > \prod_{i=1}^n p(X_i)\right] + Q^n\left(\prod_{i=1}^n q(X_i) \le \prod_{i=1}^n p(X_i)\right)$$

$$P^n\left(\prod_{i=1}^n q(X_i) > \prod_{i=1}^n p(X_i)\right) = P^n\left(\left(\prod_{i=1}^n \frac{q(X_i)}{p(X_i)}\right)^t > 1\right) \quad \forall t > 0$$

$$\le P^n\left(\prod \frac{q(X_i)}{p(X_i)}\right)^t = \int \prod_{i=1}^n p(X_i) \left(\prod_{i=1}^n \frac{q(X_i)}{p(X_i)}\right)^t$$

$$= \left(\int p^{1-t} q^t\right)^n = \exp\left(n \log \int p^{1-t} q^t\right)$$

$$\Rightarrow \inf_{\phi} \left(P^n \phi + Q^n (1-\phi)\right) \le 1 \cdot \exp\left[-n C(P,Q)\right] \quad (o(1) \to 0)$$

$$= \exp\left(-n C(P,Q) + \log 2\right)$$

Lower bound $\quad \tilde{t} = \arg\min_{t \in [0,1]} \log \int p^{1-t} q^t$

$$\inf_{\phi} \left(P^n \phi + Q^n (1-\phi)\right) \ge P^n\left(\prod q(X_i) > \prod p(X_i)\right)$$

$$\ge P^n\left(0 < \sum_{i=1}^n \log \frac{q(X_i)}{p(X_i)} < L\right)$$

$$= \int_{\left\{0 < \sum \log \frac{q(X_i)}{p(X_i)} < L\right\}} \prod p(X_i)$$

$$= \int_{\left\{0 < \sum \log \frac{q(X_i)}{p(X_i)} < L\right\}} \prod_{i=1}^n \frac{p(X_i) \left(\frac{q(X_i)}{p(X_i)}\right)^{\tilde{t}} \int p^{1-\tilde{t}} q^{\tilde{t}}}{\left(\frac{q(X_i)}{p(X_i)}\right)^{\tilde{t}} \int p^{1-\tilde{t}} q^{\tilde{t}}}$$

$$= \left(\int p^{1-\tilde{t}} q^{\tilde{t}}\right)^n \int_{\left\{0 < \sum \log \frac{q(X_i)}{p(X_i)} < L\right\}} \prod_{i=1}^n \frac{1}{\left(\frac{q(X_i)}{p(X_i)}\right)^{\tilde{t}}} \cdot \frac{p(X_i)^{1-\tilde{t}} q(X_i)^{\tilde{t}}}{\int p^{1-\tilde{t}} q^{\tilde{t}}}$$

density $\tilde{p}$

$$\ge e^{-L\tilde{t}}$$

$$\ge \left(\int p^{1-\tilde{t}} q^{\tilde{t}}\right)^n e^{-\tilde{t}L} \int_{\left\{0 < \sum_{i=1}^n \log \frac{q(X_i)}{p(X_i)} < L\right\}} \prod_{i=1}^n \tilde{p}(X_i)$$

$$= \exp\left(-n C(P,Q)\right) e^{-\tilde{t}L} \tilde{P}^n\left(0 < \sum_{i=1}^n \log \frac{q(X_i)}{p(X_i)} < L\right)$$

$\tilde{t} = \arg\min_t \log \int p^{1-t} q^t$ satisfies

$$\frac{d}{dt} \log \int p^{1-t} q^t \Big|_{t=\tilde{t}} = 0$$

(interchange limit and integral)

$$= \frac{\int p \frac{d}{dt}\left(\frac{q}{p}\right)^t}{\int p^{1-t} q^t} = \frac{\int p \left(\frac{q}{p}\right)^t \log \frac{q}{p}}{\int p^{1-t} q^t}$$

$$= \int \tilde{p} \log \frac{q}{p} = 0 \qquad (\text{set } t = \tilde{t})$$

$$\Rightarrow \mathbb{E}_{\tilde{p}^n} \sum_{i=1}^n \log \frac{q(X_i)}{p(X_i)} = 0$$

Set $L = \sqrt{\mathrm{Var}_{\tilde{p}^n}\left(\sum_{i=1}^n \log \frac{q(X_i)}{p(X_i)}\right)} = \sqrt{n \tilde{p} \left(\log \frac{q}{p}\right)^2}$

$$\inf_{\phi} \left(p^n \phi + Q^n (1-\phi)\right) \geq \exp\left(-n C(P,Q) - \tilde{t} \sqrt{n \tilde{p} \left(\log \frac{q}{p}\right)^2}\right) \tilde{p}^n \left(0 < \frac{\sum_{i=1}^n \log \frac{q(X_i)}{p(X_i)}}{\sqrt{n \tilde{p} \left(\log \frac{q(X_i)}{p(X_i)}\right)^2}} < 1\right)$$

assumption ① $\sqrt{\tilde{p}\left(\log \frac{q}{p}\right)^2} = o\left(\sqrt{n}\, C(P,Q)\right)$

② CLT for $\sum_{i=1}^n \log \frac{q(X_i)}{p(X_i)}$ under $\tilde{p}^n$

$$\left(\Rightarrow \tilde{p}^n \left(0 < \frac{\sum \log \frac{q(X_i)}{p(X_i)}}{\sqrt{n \tilde{p} \left(\log \frac{q}{p}\right)^2}} < 1\right) \to \mathbb{P}(0 < N(0,1) < 1)\right.$$

under ① & ②

$$\inf_{\phi} \left(p^n \phi + Q^n (1-\phi)\right) \geq \exp\left(-(1+o(1)) n C(P,Q)\right)$$

Le Cam two point method

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta (\hat{\theta} - \theta)^2 \geq \inf_{\hat{\theta}} \sup_{\theta \in \{\theta_1, \theta_2\}} \mathbb{E} (\hat{\theta} - \theta)^2 \qquad (\theta_1, \theta_2 \in \Theta)$$

$$\geq \frac{(\theta_1 - \theta_2)^2}{4} \int \min(P_{\theta_1}, P_{\theta_2}) \inf_{\phi} \left(P_{\theta_1} \phi + P_{\theta_2} (1-\phi)\right)$$

pf) $\inf_{\hat{\theta}} \sup_{\theta \in \{\theta_1, \theta_2\}} \mathbb{E}_\theta (\hat{\theta} - \theta)^2$

$$\geq \inf_{\hat{\theta}} \frac{1}{2} \mathbb{E}_{\theta_1} (\hat{\theta} - \theta_1)^2 + \frac{1}{2} \mathbb{E}_{\theta_2} (\hat{\theta} - \theta_2)^2$$

$$= \inf_{\hat{\theta}} \frac{1}{2} \int (\hat{\theta} - \theta_1)^2 P_{\theta_1} + (\hat{\theta} - \theta_2)^2 P_{\theta_2}$$

$$\geq \inf_{\hat{\theta}} \frac{1}{2} \int \left[(\hat{\theta} - \theta_1)^2 + (\hat{\theta} - \theta_2)^2\right] \min(P_{\theta_1}, P_{\theta_2}) \qquad (a+b)^2 \leq 2a^2 + 2b^2$$

$$\geq \frac{1}{2}\int \frac{(\theta_1-\theta_2)^2}{2} \min(p_{\theta_1}, p_{\theta_2})$$

$$= \frac{(\theta_1-\theta_2)^2}{4}\int \min(p_{\theta_1}, p_{\theta_2})$$

$X_1 \cdots X_n \overset{iid}{\sim} N(\theta, 1)$

$$\inf_{\hat\theta}\sup_{\theta\in\mathbb{R}} \mathbb{E}_\theta(\hat\theta-\theta)^2 \geq \inf_{\hat\theta}\sup_{\theta\in\{\theta_1,\theta_2\}} \mathbb{E}_\theta(\hat\theta-\theta)^2$$

H_0 $X_1 \cdots X_n \overset{iid}{\sim} N(\theta_1, 1)$

H_1 $x_1 \cdots x_n \overset{iid}{\sim} N(\theta_2, 1)$

$$\inf_{\phi}\left(P_{\theta_1}^n \phi + P_{\theta_2}^n(1-\phi)\right) = \exp\left(-(1+o(1))\, n C(N(\theta_1,1), N(\theta_2,1))\right)$$

$$C(N(\theta_1,1), N(\theta_2,1)) \propto (\theta_1-\theta_2)^2$$

Choose $\theta_1 = 0$ $\theta_2 = \frac{1}{\sqrt{n}}$

$$\inf_{\hat\theta}\sup_{\theta\in\{\theta_1,\theta_2\}} \mathbb{E}_\theta(\hat\theta-\theta)^2 \geq \frac{1}{4n}\inf_{\phi}\left(P_0^n\phi + P_{\frac{1}{\sqrt{n}}}^n(1-\phi)\right)$$

$$\inf_{\phi}\left(P_0^n\phi + P_{\frac{1}{\sqrt{n}}}^n(1-\phi)\right) \geq P_0^n\left(p_0^n(x) < q_0^n(x)\right)$$

$$= P_0^n\left(\prod_{i=1}^n \frac{e^{-\frac{1}{2}(x_i-\frac{1}{\sqrt{n}})^2}}{e^{-\frac{1}{2}x_i^2}} > 1\right)$$

$$= P_0^n\left(\sum_{i=1}^n (x_i-\tfrac{1}{\sqrt{n}})^2 - x_i^2 < 0\right)$$

$$= P_0^n\left(\frac{1}{\sqrt{n}}\sum x_i > \frac{1}{2}\right)$$

$$= P(N(0,1) > \tfrac{1}{2})$$

$$\Rightarrow \inf_{\hat\theta}\sup_{\theta\in\mathbb{R}} \mathbb{E}(\hat\theta-\theta)^2 \geq \frac{P(N(0,1)>\frac{1}{2})}{4}\,\frac{1}{n}$$

Gaussian sequence $X_j = \theta_j + \frac{1}{\sqrt{n}} z_j$ $j = 1, \cdots$

$z_j \sim N(0,1)$

$X_j \sim N(\theta_j, \frac{1}{n})$

$\theta \in \Theta_\alpha(R) = \left\{\theta : \sum_j j^{2\alpha}\theta_j^2 \leq R^2\right\}$

Thm. $\inf_{\hat\theta}\sup_{\theta\in\Theta_\alpha(R)} \mathbb{E}_\theta\|\hat\theta-\theta\|^2 \geq C n^{-\frac{2\alpha}{2\alpha+1}}$

Pf) $\Theta_0 \subset \Theta_\alpha(R)$

$\Theta_0 = \left\{\theta : \theta_j = \begin{cases} 0 \text{ or } \frac{1}{\sqrt{n}} & j \leq k \\ 0 & j > k \end{cases}\right\}$ $|\Theta_0| = 2^k$ hardest possible problem: "lower bound"

$\forall \theta \in \Theta_0$ $\sum_j j^{2\alpha}\theta_j^2 = \sum_{j=1}^k j^{2\alpha}\theta_k^2 \leq \frac{1}{n}\sum_{j=1}^k \theta_k^2 \leq \frac{k^{2\alpha+1}}{n} \leq R^2$

(integration; what we want)

Choose $k \asymp n^{\frac{1}{2\beta+1}}$ Choose π to be uniform prior on Θ_0

$$\inf_{\hat\theta} \sup_{\theta \in \Theta_0(k)} \mathbb{E}_\theta \|\hat\theta - \theta\|^2 \geq \inf_{\hat\theta} \sup_{\theta \in \Theta_0} \mathbb{E}_\theta \|\hat\theta - \theta\|^2$$

$$\geq \inf_{\hat\theta} \int \pi(\theta)\, \mathbb{E}_\theta \|\hat\theta - \theta\|^2 \, d\theta$$

$$\geq \inf_{\hat\theta} \int \pi(\theta)\, \mathbb{E}_\theta \sum_{j=1}^{k} (\hat\theta_j - \theta_j)^2 \, d\theta \sim \frac{k}{n} \sim n^{-\frac{2\beta}{2\beta+1}}$$

$$= \inf_{\hat\theta} \sum_{j=1}^{k} \operatorname{ave}_{\theta \in \Theta_0} \mathbb{E}_\theta (\hat\theta_j - \theta_j)^2 \qquad \theta = (\theta_j, \theta_{-j})$$

$$= \inf_{\hat\theta} \sum_{j=1}^{k} \operatorname{ave}_{\theta_{-j}} \operatorname{ave}_{\theta_j} \mathbb{E}_{(\theta_j, \theta_{-j})} (\hat\theta_j - \theta_j)^2$$

$$= \inf_{\hat\theta} \sum_{j=1}^{k} \operatorname{ave}_{\theta_{-j}} \left[\frac{1}{2} \mathbb{E}_{(\theta_j = 0,\, \theta_{-j})} (\hat\theta_j - 0)^2 + \frac{1}{2} \mathbb{E}_{(\theta_j = \frac{1}{n},\, \theta_{-j})} \left(\hat\theta_j - \frac{1}{n}\right)^2 \right]$$

$$\geq \sum_{j=1}^{k} \operatorname{ave}_{\theta_{-j}} \inf_{\hat\theta_j} \left[\frac{1}{2} \mathbb{E}_{(\theta_j = 0,\, \theta_{-j})} (\hat\theta_j - 0)^2 + \frac{1}{2} \mathbb{E}_{(\theta_j = \frac{1}{n},\, \theta_{-j})} \left(\hat\theta_j - \frac{1}{n}\right)^2 \right]$$

previously derived

$$\geq \sum_{j=1}^{k} \operatorname{ave}_{\theta_{-j}} C \frac{1}{n} = C \frac{k}{n} \asymp n^{-\frac{2\beta}{2\beta+1}}$$

Lecture 7

① empirical Bayes

$X|\theta \sim p(x|\theta)$ $\quad \theta|\eta \sim \pi(\theta|\eta)$

$X|\eta \sim p(x|\eta) = \int p(x|\theta)\pi(\theta|\eta)\, d\theta$

"marginal"

MMLE $\hat{\eta} = \arg\max_{\eta} p(x|\eta)$

② model with latent variables

$X|\underline{\theta}, \eta \sim p(x|\theta,\eta)$ $\quad \theta \sim \pi(\theta)$

θ: latent variable (like missing data, has prior); η: model parameter

$X|\eta \sim p(x|\eta) = \int p(x|\theta,\eta)\pi(\theta)\, d\theta$

MMLE $\hat{\eta} = \arg\max_{\eta} p(x|\eta)$

③ In general, $x|\theta, \alpha \sim p(x|\theta,\alpha)$ (Bayesian / frequentist)

$\theta|\beta \sim \pi(\theta|\beta)$

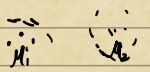
$p(x|\alpha,\beta) = \int p(x|\theta,\alpha)\pi(\theta|\beta)\, d\theta$

example. Gaussian mixture model

$X_i|z_i,\mu,\sigma^2 \sim N(\mu_{z_i}, \sigma^2 I_p)$ $\quad z_i$: latent variable

clustering label $z_i \in \{1 \cdots k\}$

special case $k=2$ $\quad X_i \sim \begin{cases} N(\mu_1, \sigma^2 I_p) & z_i = 1 \\ N(\mu_2, \sigma^2 I_p) & z_i = 2 \end{cases}$



$\mu_1, \mu_2, \cdots, \mu_n \in \mathbb{R}^p$

$z_i \overset{iid}{\sim} \text{Unif}(\{1 \cdots k\})$

$p(x_1 \cdots x_n|\mu,\sigma^2) = \frac{1}{k^n}\sum_{z}\prod_{i=1}^{n} p(x_i|z_i,\mu,\sigma^2)$

MMLE $\max_{\mu,\sigma^2} \frac{1}{k^n}\sum_{z}\prod_{i=1}^{n} p(x_i|z_i,\mu,\sigma^2)$

• EM algorithm

Some basics Kullback-Leibler divergence (relative entropy)

$D(P\|Q) = \int \log\frac{p}{q}\, d\pi$ $\quad (P \ll Q$, if not $\infty)$ $\quad P, Q \ll \pi$

Lemma $D(P\|Q) \geq 0$

Why MLE works?

$X_1 \cdots X_n \overset{iid}{\sim} P_{\theta^*}$ MLE $\hat{\theta} = \arg\max_\theta \frac{1}{n}\sum_{i=1}^n \log P_\theta(X_i)$

$= \arg\max_\theta \frac{1}{n} \sum \log \frac{1}{P_\theta(X_i)}$

$= \arg\max_\theta \left(\frac{1}{n}\sum \log \frac{1}{P_\theta(X_i)} - \frac{1}{n}\sum \log P_{\theta^*}(X_i) \right)$

$= \arg\min_\theta \frac{1}{n} \sum \log \frac{P_{\theta^*}(X_i)}{P_\theta(X_i)}$

LLN: $\frac{1}{n}\sum \log \frac{P_{\theta^*}(X_i)}{P_\theta(X_i)} \longrightarrow \mathbb{E}_{\theta^*} \log \frac{P_{\theta^*}}{P_\theta} = \int P_{\theta^*} \log \frac{P_{\theta^*}}{P_\theta} = D(P_{\theta^*} \| P_\theta)$

$\theta^* = \arg\min_\theta D(P_{\theta^*} \| P_\theta)$

EM algorithm

$\max_{\mu, \sigma^2} \log \sum_z P(X | z, \mu, \sigma^2)$

$\max_\eta \log \int P(x|\theta, \eta)\, \pi(\theta)\, d\theta$

find a lower bound: $\log \int P(X|\theta,\eta)\,\pi(\theta)\,d\theta$

$= \log \int P(X|\theta,\eta) \frac{\pi(\theta)}{g(\theta)} g(\theta)\, d\theta$

Jensen $\geq \int g(\theta) \log \left(P(X|\theta,\eta) \frac{\pi(\theta)}{g(\theta)} \right) d\theta$ (replace log sum with sum log)

$= \int g(\theta) \log P(X|\theta,\eta)\, d\theta - \int g(\theta) \log \frac{g(\theta)}{\pi(\theta)} d\theta$

$= \int g(\theta) \log P(X|\theta,\eta)\, d\theta - D(g \| \pi)$

$= F(\eta, g)$

EM algorithm $\eta^{t+1} = \arg\max_\eta F(\eta, g^t)$ M-step

$g^{t+1} = \arg\max_g F(\eta^{t+1}, g)$ E-step

Theorem (Neal & Hinton): $\max_\eta \log \int P(X|\theta,\eta)\,\pi(\theta) d\theta$

$= \max_{\eta, g} F(\eta, g)$

pf) $F(\eta, q) = \int q(\theta) \log p(x|\theta,\eta)\, d\theta - D(q\|\pi)$

$$= \int q(\theta) \log \frac{p(x|\theta,\eta)\pi(\theta)}{q(\theta)}$$

$$= \int q(\theta) \log \frac{\frac{p(\theta|\pi(\theta,\eta)p(x|\theta,\eta)}{\int p(x|\theta,\eta)\pi(\theta)d\theta}}{q(\theta)} d\theta$$

$$+ \int q(\theta) \log \int p(x|\theta,\eta)\pi(\theta) d\theta$$

$$\log \int p(x|\theta,\eta)\pi(\theta) d\theta$$

$$= -D\left(q \,\middle\|\, \frac{p(x|\theta,\eta)\pi(\theta)}{\int p(x|\theta,\eta)\pi(\theta)}\right) + \log \int p(x|\theta,\eta)\pi(\theta) d\theta \quad \text{Done}$$

$$q^t(\theta) = \frac{p(x|\theta,\eta)\pi(\theta)}{\int p(x|\theta,\eta)\pi(\theta)} : \text{posterior expectation}$$

EM: $\eta^{t+1} = \arg\max_\eta F(\eta, q^t)$

$q^{t+1} = \arg\max_q F(\eta^{t+1}, q)$

$$\begin{cases} \eta^{t+1} = \arg\max_\eta \int q^t(\theta) \log p(x|\theta,\eta)\, d\theta & \text{M-step} \\ q^{t+1} = \frac{p(x|\theta,\eta^{t+1})\pi(\theta)}{\int p(x|\theta,\eta^{t+1})\pi(\theta) d\theta} & \text{E-step} \end{cases}$$

Gaussian mixture

$$\begin{cases} X_i | z_i, \mu \sim N(\mu_{z_i}, I_p) \\ z_i \sim \text{Unif}(\{1,\cdots,k\}) \end{cases}$$

Goal: $\max_\mu \sum_z p(x|z,\mu)$

E-step $q(z_1 \cdots z_n) \propto \prod_{i=1}^n p(x_i|z_i,\mu) \cancel{\frac{1}{k^n}}$

$\Rightarrow q(z_1 \cdots z_n) = \prod q(z_i)$

$q(z_i = j) = q_{ij} \propto p(x_i|z_i=j,\mu)$

$$q_{ij} = \frac{p(x_i|z_i=j,\mu) \sim N(\mu_j, I_p)}{p(x_i|z=1,\cdots\mu) + \cdots + p(x_i|z=k,\mu)}$$

$\sim N(\mu_1, I_p)$ $\quad \sim N(\mu_k, I_p)$

$$= \frac{e^{-\frac{1}{2}\|x_i-\mu_j\|^2}}{e^{-\frac{1}{2}\|x_i-\mu_1\|^2} + \cdots + e^{-\frac{1}{2}\|x_i-\mu_k\|^2}}$$

M-step $\log \prod_{i=1}^{n} p(x_i | z_i, \mu)$

$= \sum_{i=1}^{n} \log p(x_i | z_i, \mu)$

objective $\mathbb{E}_{z \sim q} \left(\sum_{i=1}^{n} \log p(x_i | z_i, \mu) \right)$

$= \sum_{i=1}^{n} \mathbb{E}_{z \sim q} \log p(x_i | z_i, \mu)$

$= \sum_{i=1}^{n} \sum_{j=1}^{k} q(z_i = j) \log p(x_i | z_i = j, \mu)$

$= \sum_{i} \sum_{j} q_{ij} \log \frac{p(x_i | z_i = j, \mu)}{} \propto e^{-\frac{1}{2} \| x_i - \mu_j \|^2}$

$= \sum_{i=1}^{n} \sum_{j=1}^{k} q_{ij} \left(-\frac{1}{2} \| x_i - \mu_j \|^2 \right) + C$

$- \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{k} q_{ij} \| x_i - \mu_j \|^2$

M-step $\Leftrightarrow \min_{\mu_1, \dots, \mu_k} \sum_{j=1}^{k} \sum_{i=1}^{n} q_{ij} \| x_i - \mu_j \|^2$ ← weighted least square problem

$= \sum_{j=1}^{k} \min_{\mu_j} \sum_{i=1}^{n} q_{ij} \| x_i - \mu_j \|^2$

$\Rightarrow \mu_j = \frac{\sum_{i=1}^{n} q_{ij} x_i}{\sum_{i=1}^{n} q_{ij}}$

EM-algorithm for clustering

$$q_{ij}^{t+1} = \frac{e^{-\frac{1}{2} \| x_i - \mu_j^t \|^2}}{\sum_{j=1}^{k} e^{-\frac{1}{2} \| x_i - \mu_j^t \|^2}}$$

$$\mu_j^{t+1} = \frac{\sum_{i=1}^{n} q_{ij}^{t+1} x_i}{\sum_{i=1}^{n} q_{ij}^{t+1}}$$

* if q_i is concentrated on some j-th cluster
  $\mu_j \approx$ sample mean

* non-convex optimization
  So we may run several times

post-processing $\hat{z}_i = \underset{j \in \{1, \dots, k\}}{\operatorname{argmax}} q_{ij}^{T}$.

Lecture 8

EM algorithm (Dempster, Laird, & Rubin)

$X \mid \theta, \eta \sim p(x \mid \theta, \eta) \quad \theta \sim \pi \quad \hat{\eta} = \arg\max_{\eta} \log \int p(x \mid \theta, \eta)\, \pi(\theta)\, d\theta$

$$\max_{\eta} \log \int p(x \mid \theta, \eta)\, \pi(\theta)\, d\theta = \max_{\eta, q} F(\eta, q)$$

$$F(\eta, q) = \int q(\theta) \log p(x \mid \theta, \eta)\, d\theta - D(q \| \pi)$$

Gaussian mixture model

$X_i \mid Z_i, \mu \sim N(\mu_{Z_i}, I_d) \qquad Z_i \sim \mathrm{Unif}(k)$

$\propto e^{-\frac{1}{2}\|x_i - \mu_j\|^2}$

EM

$$\begin{cases} q_{ij}^{t+1} = \dfrac{e^{-\frac{1}{2}\|x_i - \mu_j^t\|^2}}{\sum_{j=1}^{k} e^{-\frac{1}{2}\|x_i - \mu_j^t\|^2}} \\ \mu_j^{t+1} = \dfrac{\sum_{i=1}^{n} q_{ij}^{t+1} x_i}{\sum_{i=1}^{n} q_{ij}^{t+1}} \end{cases}$$

$X \mid \theta, \eta \sim p(x \mid \theta, \eta) \qquad \max_{\theta, \eta} \log p(x \mid \theta, \eta)$

$$\begin{cases} \eta^{t+1} = \arg\max_{\eta} \log p(x \mid \theta^t, \eta) \\ \theta^{t+1} = \arg\max_{\theta} \log p(x \mid \theta, \eta^{t+1}) \end{cases}$$

Gaussian mixture model

$$\sum_{i=1}^{n} \log p(x_i \mid \mu_{z_i}) = -\frac{1}{2} \sum_{i=1}^{n} \|x_i - \mu_{z_i}\|^2 + C$$

$$= -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{k} 1_{\{z_i = j\}} \|x_i - \mu_j\|^2 + C$$

$$= -\frac{1}{2} \sum_{j=1}^{k} \sum_{i=1}^{n} 1_{\{z_i = j\}} \|x_i - \mu_j\|^2 + C$$

$$\max_{z} \sum_{i=1}^{n} \log p(x_i \mid \mu_{z_i}) \Leftrightarrow \min_{z_i} \sum_{j=1}^{k} 1_{\{z_i = j\}} \|x_i - \mu_j\|^2 \quad \text{for } i = 1 \cdots n$$

$$z_i = \arg\min_{j \in \{1 \cdots k\}} \|x_i - \mu_j\|^2$$

$$\max_{\mu} \sum_{i=1}^{n} \log p(x_i \mid \mu_{z_i}) \Leftrightarrow \min_{\mu_j} \sum_{i=1}^{n} 1_{\{z_i = j\}} \|x_i - \mu_j\|^2 \quad \text{for } j = 1 \cdots k$$

$$\mu_j = \frac{\sum_{i=1}^{n} 1_{\{z_i = j\}} x_i}{\sum_{i=1}^{n} 1_{\{z_i = j\}}}$$

$$\Rightarrow \begin{cases} z_i^{t+1} = \arg\min_{j \in \{1 \cdots k\}} \|x_i - \mu_j^t\|^2 \\ \mu_j^{t+1} = \dfrac{\sum_{i=1}^{n} 1_{\{z_i^t = j\}} x_i}{\sum_{i=1}^{n} 1_{\{z_i^t = j\}}} \end{cases}$$

↑ thresholded version of E-step

↑ "M-step"

k-means algorithm

"EM algorithm is a soft version of k-means algorithm"

k-means: purely frequentist

EM: hybrid philosophy

Which is better?

	clustering	parameter estimation
EM	✓	✓
k-means	✓	inconsistency

both compare "likelihood" likelihood test

As long as they don't get stuck at local optimum, they give us asymptotic optimal clusters

you use other cluster's data unless $\mu \to \infty$, $\sigma = \text{constant}$

general case $k=2$, $\mu_1 = -\mu_2 = \mu$

$$X_i \mid z_i, \mu \sim \begin{cases} N(\mu, I_p) & z_i = 1 \\ N(-\mu, I_p) & z_i = 2 \end{cases}$$

$$\ell(\hat{z}, z) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{\hat{z}_i \neq z_i\}} \wedge \frac{1}{n}\sum_{i=1}^{n} 1_{\{\hat{z}_i = z_i\}}$$

all equivalent permutations

$C(N(\mu, I_p), N(-\mu, I_p))$ & $\|\mu\| \to$

$$\inf_{\hat{z}} \sup_{z \in \{1,2\}^n} \mathbb{E}\, \ell(\hat{z}, z) = \exp\left(-(1+o(1))\frac{\|\mu\|^2}{2}\right)$$

$P(X \mid \theta, \eta)$

Neyman-Scott problem $\begin{cases} X_i \sim N(\mu_i, \sigma^2) \\ Y_i \sim N(\mu_i, \sigma^2) \end{cases}$ $i = 1 \cdots n$

goal: estimate σ^2

method ① (like k-means) $\max_{\mu_1 \cdots \mu_n, \sigma^2} P(X, Y \mid \mu_1 \cdots \mu_n, \sigma^2)$

$\hat{\sigma}^2_{MLE} \not\to \sigma^2$ (homework)

method ② REML $X_i - Y_i \sim N(0, 2\sigma^2)$ $i = 1 \cdots n$

$\hat{\sigma}^2 = \frac{1}{2n}\sum_{i=1}^{n}(X_i - Y_i)^2 \xrightarrow{LLN} \sigma^2$

homework: equivalent to putting prior on $\mu_1 \cdots \mu_n$ (EM-algorithm)

Lesson: put prior on local parameter (if you're a frequentist)

EM algorithm

$X \mid \theta, \eta \sim P(X \mid \theta, \eta)$ hybrid

Bayes frequentist

More generally.

$X|\theta,\eta \sim p(X|\theta,\eta)$ $\theta \sim \pi_\tau$ $(\theta|\tau \sim \pi(\cdot|\tau))$

↳ a hyper parameter

$\max_{\eta,\tau} \log \int p(x|\theta,\eta)\pi_\tau(\theta)d\theta$

↳ empirical Bayes principle

$\log \int p(x|\theta,\eta)\pi_\tau(\theta)\,d\theta \geq \int q(\theta)\log p(x|\theta,\eta)d\theta - D(q\|\pi_\tau)$

$= F(\eta,\tau,q)$

$$\begin{cases} \text{E-Step} & q^{t+1} = \arg\max_q F(\eta^t,\tau^t,q) \\ \text{M-Step} & (\eta^{t+1},\tau^{t+1}) = \arg\max_{\eta,\tau} F(\eta,\tau,q^{t+1}) \end{cases}$$

① only η $X|\eta \sim p|X|\eta)$

$\max_\eta \log p(x|\eta)$ MLE

② only θ $x|\theta \sim p(x|\theta)$ $\theta \sim \pi$

$\max_q [\int q(\theta)\log p(x|\theta)\,d\theta - D(q\|\pi)]$

some info abt data — penalty

$\Leftarrow$ $q(\theta) = \dfrac{p(x|\theta)\pi(\theta)}{\int p(x|\theta)\pi(\theta)d\theta}$ Bayes

③ only θ (unknown τ) empirical Bayes

Variational Bayes $X|\theta \sim p(x|\theta)$ $\theta \sim \pi(\theta)$ Computing posterior!

$\max_{q\in S} [\int q(\theta)\log p(x|\theta)\,d\theta - D(q\|\pi)]$

↖ Variational class

$\Leftarrow$ $\min_{q\in S} D(q\|\pi(\cdot|x))$

mean-field class

$S = \{q(\theta) = \prod_j q_j(\theta_j) : q_j \text{ is any distribution}\}$

e.g. $x|\theta,\eta \sim p(x|\theta,\eta)$ $\theta,\eta \sim \pi$

$S = \{q(\theta,\eta) = f(\theta)g(\eta) : f,g\}$

$\forall\theta$: $\min_{q\in S} D(q\|\pi(\cdot|x))$

$$\Leftrightarrow \max_{f,g} \left[\int f(\theta) g(\eta) \log p(x|\theta,\eta) - D(f \otimes g \| \pi) \right]$$

$$\Leftrightarrow \min_{f,g} \left[\int f(\theta) g(\eta) \log \frac{1}{p(x|\theta,\eta)} d\theta d\eta + D(f \otimes g \| \pi) \right]$$

$$= \min_{f,g} F(f,g)$$

CAVI (coordinate ascent variational inference)

$$\begin{cases} f^{t+1} = \arg\min_f F(f, g^t) \\ g^{t+1} = \arg\min_g F(f^{t+1}, g) \end{cases}$$

$$F(f,g) = \iint f(\theta) g(\eta) \log \frac{1}{p(x|\theta,\eta)} d\theta d\eta + \iint f(\theta) g(\eta) \log \frac{f(\theta) g(\eta)}{\pi(\theta,\eta)} d\theta d\eta$$

$$= \iint f(\theta) g(\eta) \log \frac{1}{p(x|\theta,\eta)} d\theta d\eta + \int f(\theta) \log f(\theta) d\theta + \underbrace{\int g(\eta) \log g(\eta) d\eta}_{\text{constant}} - \int f(\theta) g(\eta) \log \pi(\theta,\eta) d\theta d\eta$$

$$= \iint f(\theta) \left[g(\eta) \log \frac{1}{p(x|\theta,\eta)\pi(\theta,\eta)} d\eta \right] d\theta + C + \int f(\theta) \log f(\theta) d\theta$$

$$= \int f(\theta) \log e^{\int g(\eta) \log \frac{1}{p(x|\theta,\eta)\pi(\theta,\eta)} d\eta} + \int f(\theta) \log f(\theta) d\theta + C$$

$$= \int f(\theta) \log \frac{f(\theta)}{e^{\int g(\eta) \log p(x|\theta,\eta)\pi(\theta,\eta) d\eta}} d\theta + C$$

Solution is $f(\theta) \propto e^{\int g(\eta) \log p(x|\theta,\eta)\pi(\theta,\eta) d\eta}$

Lecture 9

Variational Bayes

$$\min_{q \in S} \int q(\theta) \log \frac{1}{p(x|\theta)} d\theta + D(q \| \pi)$$

$$\Leftrightarrow \min_{q \in S} D(q \| \pi(\cdot|x))$$

$$S = \{ g(\theta,\eta) = f(\theta) g(\eta) : f, g \}$$

(Jensen)

$$F(f,g) = \int f(\theta) g(\eta) \log \frac{1}{p(x|\theta)} d\theta d\eta + D(f \otimes g \| \pi)$$

$$= \int f(\theta) g(\eta) \log \frac{1}{p(x|\theta,\eta)} d\theta d\eta + \iint f(\theta) g(\eta) \log \frac{g(\eta) f(\theta)}{\pi(\theta,\eta)} d\theta d\eta$$

$$= \iint f(\theta) g(\eta) \log \frac{1}{p(x|\theta,\eta)} d\theta d\eta + \int f(\theta) \log f(\theta) d\theta + \underbrace{\int g(\eta) \log g(\eta) d\eta - \iint f(\theta) g(\eta) \log \pi(\theta,\eta) d\theta d\eta}_{\text{constant}}$$

$$= \int f(\theta) \log f(\theta) d\theta + \iint f(\theta) g(\eta) \log \frac{1}{p(x|\theta,\eta) \pi(\theta,\eta)} d\theta d\eta + \text{const}.$$

$$= \int f(\theta) \log \frac{f(\theta)}{\exp \int g(\eta) (\log p(x|\theta,\eta) \pi(\theta,\eta)) d\eta} + \text{const}$$

minimizer is $f(\theta) \propto \exp\left(\int g(\eta) p(x|\theta,\eta) \pi(\theta,\eta) d\eta \right)$

CAVI (Similar to E-step)

$$\begin{cases} f^{t+1}(\theta) \propto \exp\left(\mathbb{E}_{\eta \sim g^t} \log p(x|\theta,\eta) \pi(\theta,\eta) \right) \\ g^{t+1}(\eta) \propto \exp\left(\mathbb{E}_{\theta \sim f^{t+1}} \log p(x|\theta,\eta) \pi(\theta,\eta) \right) \end{cases}$$

$$\Leftrightarrow \quad f^{t+1}(\theta) \propto \exp\left(\mathbb{E}_{\eta \sim g^t} [\log \pi(\theta,\eta|x)] \right)$$

$$g^{t+1}(\eta) \propto \exp\left(\mathbb{E}_{\theta \sim f^t} [\log \pi(\theta,\eta|x)] \right)$$

Gaussian mixture

$$X_i | z_i, M \sim N(M_{z_i}, I_p)$$

$$z_i \sim \text{Unif}(\{1 \cdots k\})$$

$$M_j \sim N(0, \tau^2 I_p) \quad j = 1 \cdots k$$

(known)

$$\pi(M, z|x) \propto \prod_{i=1}^n p(x_i|M_{z_i}) \prod_{j=1}^k p(M_j) \propto \exp\left(-\frac{1}{2} \sum_{i=1}^n \|x_i - M_{z_i}\|^2 - \frac{1}{2\tau^2} \sum_{j=1}^k \|M_j\|^2 \right)$$

find $q(M,z) = f(M) g(z)$

$$\log \pi(\mu, z \mid x) = -\frac{1}{2}\sum_i \sum_j \mathbb{1}_{\{z_i = j\}} \|x_i - \mu_j\|^2 - \frac{1}{2\tau^2}\sum_{j=1}^k \|\mu_j\|^2 + C$$

$$g(z_1 \cdots z_n) \propto \exp\left(-\frac{1}{2}\sum_{i=1}^n \sum_{j=1}^k \mathbb{1}_{\{z_i = j\}} \mathbb{E}_{\mu \sim f} \|x_i - \mu_j\|^2 - \frac{1}{2\tau^2}\sum_{j=1}^k \mathbb{E}_{\mu\sim f}\|\mu_j\|^2\right)$$

$$\propto \prod_{i=1}^n \exp\left(-\frac{1}{2}\sum_{j=1}^k \mathbb{1}_{\{z_i=j\}} \mathbb{E}_{\mu\sim f}\|x_i - \mu_j\|^2\right)$$

$$g(z_1 \cdots z_n) = \prod_{i=1}^n g(z_i)$$

$$g_{ij} = g(z_i = j) \propto \exp\left(-\frac{1}{2}\mathbb{E}_{\mu\sim f}\|x_i - \mu_j\|^2\right) \quad \text{"Like E-step"}$$

$$f(\mu_1 \cdots \mu_k) \propto \exp\left(-\frac{1}{2}\sum_{i=1}^n \sum_{j=1}^k g_{ij}\|x_i - \mu_j\|^2 - \frac{1}{2\tau^2}\sum_{j=1}^k \|\mu_j\|^2\right)$$

$$\propto \prod_{j=1}^k \exp\left(-\frac{1}{2}\sum_{i=1}^n g_{ij}\|x_i - \mu_j\|^2 - \frac{1}{2\tau^2}\|\mu_j\|^2\right)$$

$$f(\mu_1 \cdots \mu_k) = \prod_{j=1}^k f(\mu_j)$$

$$f(\mu_j) \propto \exp\left(-\frac{1}{2}\sum_{i=1}^n g_{ij}\|x_i - \mu_j\|^2 - \frac{1}{2\tau^2}\|\mu_j\|^2\right)$$

$$\sum_{i=1}^n g_{ij}\|x_i - \mu_j\|^2 + \frac{1}{\tau^2}\|\mu_j\|^2$$

$$= \left\|\mu_j - \beta_j\right\|^2 + C$$

$$\beta_j = \frac{\sum_{i=1}^n g_{ij} x_i}{\sum_{i=1}^n g_{ij} + \frac{1}{\tau^2}}$$

$$K_j^{-2} = \sum_{i=1}^n g_{ij} + \frac{1}{\tau^2}$$

$$f(\mu_j) = N(\beta_j, K_j^2 I_p) \quad \begin{cases} \beta_j = \dfrac{\sum_{i=1}^n g_{ij} x_i}{\sum_{i=1}^n g_{ij} + \frac{1}{\tau^2}} \\ K_j^2 = \dfrac{1}{\sum_{i=1}^n g_{ij} + \frac{1}{\tau^2}} \end{cases}$$

$$\mathbb{E}_{\mu\sim f}\|x_i - \mu_j\|^2 = \mathop{\mathbb{E}}_{\mu_j \sim N(\beta_j, K_j^2 I_p)} \|x_i - \mu_j\|^2$$

$$= \|x_i - \beta_j\|^2 + K_j^2 p$$

CAVI for Gaussian mixture

$$g(z_i = j) = g_{ij}^{t+1} \propto e^{-\frac{1}{2}\|x_i - \beta_j^t\|^2 - \frac{1}{2}K_j^{2t}p}$$

$$g_{ij}^{t+1} = \frac{e^{-\frac{1}{2}\|x_i - \beta_j^t\|^2 - \frac{1}{2}K_j^{2t}p}}{\sum_{l=1}^k e^{-\frac{1}{2}\|x_i - \beta_l^t\|^2 - \frac{1}{2}K_l^{2t}p}}$$

$$\begin{cases} \beta_j^t = \dfrac{\sum_{i=1}^n g_{ij}^t x_i}{\sum_{i=1}^n g_{ij}^t + \frac{1}{\tau^2}} \\ K_j^{t2} = \dfrac{1}{\sum_{i=1}^n g_{ij}^t + \frac{1}{\tau^2}} \end{cases}$$

$$\mu_j \sim N(\beta_j^{t+1}, K_j^{2\,t+1} I_p) \quad \begin{cases} \beta_j^{t+1} = \dfrac{\sum_{i=1}^n g_{ij}^{t+1} x_i}{\sum_{i=1}^n g_{ij}^{t+1} + \frac{1}{\tau^2}} \\ \end{cases}$$

$$k_j^{t+1} = \frac{1}{\sum_{i=1}^{n} q_{ij}^{t+1} + \frac{1}{\nu}}$$

As good as EM algorithm. Usually difficult analytically (more than 1 step)

$\pi(\theta_1 \dots \theta_d \mid X) = f(\theta_1 \dots \theta_3)\, g(\theta_4 \dots \theta_d)$ want as small groups as possible (approximation error)

$= \prod_{j=1}^{d} f_j(\theta_j)$ Similar to Gibbs sampling (exact)

Some extensions

$$\max_{q \in S} \max_{\eta} \int q(\theta) \log p(x \mid \theta, \eta)\, d\theta - D(q \| \pi)$$

(p: new, p: old)

"Variational EM algorithm"

$$\max_{q \in T} \int q(\theta) \log p(x \mid \theta)\, d\theta - D(q \| \pi_\eta)$$

Variational empirical Bayes

$$\max_{q \in S} \max_{\eta, \tau} \int q(\theta) \log p(x \mid \theta, \eta)\, d\theta - D(q \| \pi_\tau)$$

- Bayesian asymptotics

review asymptotics of MLE

$X_1 \dots X_n \overset{iid}{\sim} P_\theta$ MLE $\hat{\theta} = \arg\max_\theta L_n(\theta)$

$L_n(\theta) = \sum_{i=1}^{n} \log P_\theta(X_i)$

$\sqrt{n}(\hat{\theta} - \theta^*) \rightsquigarrow N(0, I_{\theta^*}^{-1})$

notation $\ell_\theta(x) = \log P_\theta(x)$ $\dot{\ell}_\theta(x) = \frac{d}{d\theta} \log P_\theta(x)$

$\ddot{\ell}_\theta(x) = \frac{d^2}{d\theta^2} \log P_\theta(x)$

Under regularity conditions $\begin{cases} \mathbb{E}_\theta \dot{\ell}(X) = 0 \\ -\mathbb{E}_\theta \ddot{\ell}_\theta(X) = \mathbb{E}_\theta(\dot{\ell}_\theta(X))^2 = I_\theta \end{cases}$

$L_n(\theta) - L_n(\theta^*) = (\theta - \theta^*) \sum_{i=1}^{n} \dot{\ell}_{\theta^*}(X_i) + \frac{1}{2}(\theta - \theta^*)^2 \sum_{i=1}^{n} \ddot{\ell}_{\theta^*}(X_i) + \text{remainder}$

(TE around θ^*)

$\theta = \theta^* + \frac{h}{\sqrt{n}}$

$$L_n(\theta^* + \tfrac{h}{\sqrt{n}}) - L_n(\theta^*) = h \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{\ell}_{\theta^*}(X_i) - \frac{1}{2} h^2 \frac{1}{n} \sum_{i=1}^{n} (-\ddot{\ell}_{\theta^*}(X_i)) + \text{remainder}$$

CLT: $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{\ell}_{\theta^*}(X_i) \to N(0, I_{\theta^*})$

LLN: $\frac{1}{n} \sum_{i=1}^{n} (-\ddot{\ell}_{\theta^*}(X_i)) \to I_{\theta^*}$

$$\Rightarrow L_n(\theta^* + \tfrac{h}{\sqrt{n}}) - L_n(\theta^*) = h \Delta_n - \frac{1}{2} h^2 I_{\theta^*} + o_p(1)$$

$$\Delta_n \rightsquigarrow N(0, I_{\theta^*})$$

MLE $\hat{\theta} = \operatorname{argmax}_{\theta} L_n(\theta)$

$$\hat{h} = \sqrt{n}(\hat{\theta} - \theta^*) = \operatorname{argmax}_{h} L_n(\theta^* + h/\sqrt{n})$$

$$\overset{\text{"uniform"}}{\approx} \operatorname{argmax}_{h} \left(h \Delta_n - \frac{1}{2} h^2 I_{\theta^*} \right)$$

$$= I_{\theta^*}^{-1} \Delta_n \rightsquigarrow N(0, I_{\theta^*}^{-1})$$

Lecture 11

$X_1 \cdots X_n \overset{iid}{\sim} P_{\theta^*}$

$L_n(\theta) = \sum_{i=1}^n \log P_\theta(X_i) = \sum_{i=1}^n \ell_\theta(X_i)$

$L_n(\theta^* + \frac{h}{\sqrt{n}}) - L_n(\theta^*) = h\Delta_n - \frac{1}{2}h^2 I_{\theta^*} + o_p(1)$

$\Delta_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{\ell}_\theta(X_i) \rightsquigarrow N(0, I_{\theta^*})$

$I_{\theta^*} = -E_{\theta^*} \ddot{\ell}_{\theta^*}(X) = E_{\theta^*}(\dot{\ell}_{\theta^*}(X))^2$

$\hat{\theta} = \operatorname{argmax}_\theta L_n(\theta)$ $\quad \sqrt{n}(\hat{\theta} - \theta^*) = I_{\theta^*}^{-1} \Delta_n + o_p(1)$

$\rightsquigarrow N(0, I_{\theta^*}^{-1})$

$\Rightarrow$ frequentist pov

$X_1 \cdots X_n | \theta \overset{iid}{\sim} P_\theta \quad \theta \sim \pi$

$P(\theta | X_1 \cdots X_n) \propto \pi(\theta) \prod_{i=1}^n P(X_i|\theta)$

Choose $\hat{C}$ such that $\pi(\theta \in \hat{C} | X_1 \cdots X_n) = 0.95$ (credible interval)

Question: Suppose $X_1 \cdots X_n \overset{iid}{\sim} P_{\theta^*}$

Is $P_{\theta^*}(\theta^* \in \hat{C}) \approx 0.95$?

Theorem (Bernstein von-Mises)

Under regularity conditions $\hat{\theta} = \operatorname{argmax}_\theta L_n(\theta)$

$\sqrt{n}(\theta - \hat{\theta}) | X_1 \cdots X_n \overset{P_{\theta^*}}{\rightsquigarrow} N(0, I_{\theta^*}^{-1})$

$\Rightarrow$ frequentist confidence interval $\approx$ Bayesian credible interval

* only for parametric model (High-dim, Non-parametric model: not true)

Comparison $\sqrt{n}(\hat{\theta} - \theta) | \theta = \theta^* \overset{P_{\theta^*}}{\rightsquigarrow} N(0, I_{\theta^*}^{-1})$

BVM $\sqrt{n}(\theta - \hat{\theta}) | X_1 \cdots X_n \overset{P_{\theta^*}}{\rightsquigarrow} N(0, I_{\theta^*}^{-1})$ (frequentist pov)

$\mathbb{P}_{\substack{\theta \sim \pi \\ X \sim P_\theta^*}}(|\sqrt{n}(\theta - \hat{\theta}) - z| > \varepsilon \,|\, X_1 \cdots X_n) \to 0$

Lemma If $\int e^{tx} dP_n(x) \to \int e^{tx} dP(x)$ in probability $\forall t \in \mathbb{R}$

(P_n: random probability (like posterior dist))

then $\sup_{t \in \mathbb{R}} |P_n((-\infty, t]) - P((-\infty, t])| \to 0$ in prob

Kolmogorov Smirnov distance

$\mathbb{P}_{\substack{\theta \sim \pi \\ X \sim P_\theta^*}}\left(\left|\int_{-\infty}^t \sqrt{n}(\theta - \hat{\theta}) d\pi(\theta | X_1 \cdots X_n) - \int_{-\infty}^t N(0, I_{\theta^*}^{-1}) d\theta\right| > \varepsilon\right) \to 0 \ \forall t$

STS $\int e^{t\sqrt{n}(\theta-\hat\theta)}\, d\pi(\theta|X_1,\cdots X_n) \xrightarrow{\text{prob}} e^{\frac{1}{2}I_{\theta^*}^{-1}t^2}$

$$= \frac{\int e^{t\sqrt{n}(\theta-\hat\theta)+L_n(\theta)-L_n(\theta^*)+\log\pi(\theta)}\,d\theta}{\int e^{L_n(\theta)-L_n(\theta^*)+\log\pi(\theta)}\,d\theta}$$

$$= \frac{\int e^{t\sqrt{n}(\theta-\theta^*)+L_n(\theta)-L_n(\theta^*)+\log\pi(\theta)+t\sqrt{n}(\theta^*-\hat\theta)}\,d\theta}{\int e^{L_n(\theta)-L_n(\theta^*)+\log\pi(\theta)}\,d\theta}$$

cov $\sqrt{n}(\theta-\theta^*)=h$ $=$

$$\frac{\int e^{th+L_n(\theta^*+\frac{h}{\sqrt{n}})-L_n(\theta^*)+\log\pi(\theta^*+\frac{h}{\sqrt{n}})+t\sqrt{n}(\theta^*-\hat\theta)}\,dh}{\int e^{L_n(\theta^*+\frac{h}{\sqrt{n}})-L_n(\theta^*)+\log\pi(\theta^*+\frac{h}{\sqrt{n}})}\,dh}$$

$\pi(\theta^*+\frac{h}{\sqrt{n}})\to\pi(\theta^*)$ ≈

$$\frac{\int e^{th+h\Delta_n-\frac{1}{2}h^2I_{\theta^*}+\log\pi(\theta^*+\frac{h}{\sqrt{n}})-tI_{\theta^*}^{-1}\Delta_n}\,dh}{\int e^{h\Delta_n-\frac{1}{2}h^2I_{\theta^*}+\log\pi(\theta^*+\frac{h}{\sqrt{n}})}\,dh}$$

Why prior is "washed out" as we get more data

Since π is cancelled

$$= e^{-tI_{\theta^*}^{-1}\Delta_n}\ \mathbb{E}_{h\sim N(I_{\theta^*}^{-1}\Delta_n,\, I_{\theta^*}^{-1})}\, e^{th}$$

$$= e^{-tI_{\theta^*}^{-1}\Delta_n}\, e^{tI_{\theta^*}^{-1}\Delta_n+\frac{1}{2}t^2I_{\theta^*}^{-1}}$$

$$= e^{\frac{1}{2}t^2I_{\theta^*}^{-1}}$$

↑ MGF of $N(0, I_{\theta^*}^{-1})$

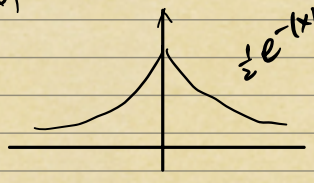
$$\begin{cases}\sqrt{n}(\hat\theta-\theta^*) = I_{\theta^*}^{-1}\Delta_n+o_p(1)\\ (L_n(\theta^*+\frac{h}{\sqrt{n}})-L_n(\theta^*))n = h\Delta_n-\frac{1}{2}h^2I_{\theta^*}+o_p(1)\end{cases}$$

?

Second derivative may not exist

e.g. $P_\theta(x) = \frac{1}{2}e^{-|x-\theta|}$

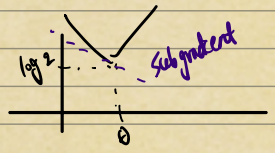
Laplace density



MLE $\hat\theta = \arg\min_\theta \sum_{i=1}^n |X_i-\theta|$

Sample median

$-l_\theta(x) = -\log P_\theta(x) = |x-\theta| + \log 2$



$$-\dot{l}_\theta(x) = \begin{cases} -1 & \theta < x \\ [-1,1] & \theta = x \\ 1 & \theta > x \end{cases}$$

$$-\ddot{l}_\theta(x) = \begin{cases} 0 & \theta < x \\ \infty ? & \theta = x \\ 0 & \theta > x \end{cases}$$

$$I_{\theta^*} = -E_\theta \ddot{l}_\theta(x) = E_{\theta^*}(\dot{l}_{\theta^*}(x))^2$$

$-E_\theta \ddot{l}_\theta(x)$: ? — not well-defined

$E_{\theta^*}(\dot{l}_{\theta^*}(x))^2$: 1

Le Cam & Hajek

LAN (local asymptotic normality)

$$n\left(L_n(\theta + \frac{h}{\sqrt{n}}) - L_n(\theta)\right) = h\Delta_{n,\theta} - \frac{h^2}{2} I_\theta + o_p(1) \quad \text{(previously seen)}$$

for some $\Delta_{n,\theta}$ and I_θ such that $\Delta_{n,\theta} \rightsquigarrow N(0, I_\theta)$

DQM (differentiability under quadratic mean)

$$\exists\, \dot{l}_\theta \text{ s.t. } \int \left(\sqrt{P_{\theta+h}} - \sqrt{P_\theta} - \frac{1}{2} h \dot{l}_\theta \sqrt{P_\theta}\right)^2 = o(h^2) \text{ as } h \to 0$$

e.g. can take $\dot{l}_\theta = \frac{\dot{P}_\theta}{P_\theta} = \frac{d}{d\theta}\log P_\theta(x)$ (if it exists)

$\frac{1}{2}\dot{l}_\theta \sqrt{P_\theta} = \frac{d}{d\theta}\sqrt{P_\theta}$ (Slightly stronger than the first derivative)

for $P_\theta = \frac{1}{2} x^{|x-\theta|}$ this $\dot{l}_\theta(x) = \begin{cases} -1 & \theta < x \\ 0 & \theta = x \\ 1 & \theta > x \end{cases}$

Can check DQM is satisfied

Theorem If DQM, then

① $E_\theta \dot{l}_\theta(x) = 0$ → call this as score

② $I_\theta = E_\theta(\dot{l}_\theta(x))^2 < \infty$ (call this as Fisher information)

③ LAN holds with $\Delta_{n,\theta} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{l}_\theta(X_i)$

$I_\theta = \mathbb{E}_\theta(\dot{\ell}_\theta(X))^2$

enough to cover Laplace

What this condition Le Cam proved first

We can create an alternative estimator that only needs LAN for asymptotic normality

moreover, under extra regularity conditions (weak condition)

$$\sqrt{n}(\hat{\theta}_{MLE} - \theta^*) = I_{\theta^*}^{-1} \Delta_{n,\theta^*} + o_p(1)$$

$$\rightsquigarrow N(0, I_{\theta^*}^{-1})$$

Theorem (Le Cam) If

① LAN

② $\Pi(|\sqrt{n}(\theta - \theta^*)| > M_n \mid X_1 \cdots X_n) = o_p(1)$ for some $M_n \to \infty$ "$\sqrt{n}$-consistent"

Chebyshev

$\Leftarrow |\theta - \theta^*| = O_p(\frac{1}{\sqrt{n}})$ under P_θ^*

$X = O_p(1)$

$\Leftrightarrow P(X > M_n) \to 0 \quad \forall M_n \to \infty$

then BVM holds

$$TV\left(P_{\sqrt{n}(\theta-\theta^*)|X_1\cdots X_n}, N\left(I_{\theta^*}^{-1}\Delta_{n,\theta^*}, I_{\theta^*}^{-1}\right)\right) \xrightarrow{P_\theta^*} 0$$

($\sup_{t \in \mathbb{R}} |P_n((-\infty,t]) - P((-\infty,t])| \to 0$ vs. $\sup_A |P_n(A) - P(A)|$, doesn't assume MLE)

Stronger

1; 2; >0

Corollary: If additionally $\sqrt{n}(\hat{\theta}_{MLE} - \theta^*) = I_{\theta^*}^{-1}\Delta_{n,\theta^*} + o_p(1)$ then

$$TV\left(P_{\sqrt{n}(\theta - \hat{\theta}_{MLE})|X_1\cdots X_n}, N(0, I_{\theta^*}^{-1})\right) \xrightarrow{P_\theta^*} 0$$

$\hat{\theta} = \theta^* + \frac{1}{\sqrt{n}} I_{\theta^*}^{-1}\Delta_n + o_p(\frac{1}{\sqrt{n}})$

↖ can be posterior mean/mode MLE, etc

$X_1 \cdots X_n \overset{iid}{\sim} \text{Unif}([0,\theta]) \in LAE$

$|\hat{\theta}_{MLE} - \theta^*| = O_p(\frac{1}{n})$

(local asymptotic exponentiality)

LAMN: financial mathematics

$n(\hat{\theta}_{MLE} - \theta^*) \rightsquigarrow \text{Exp} \cdots$

Lecture 12

$(P_\theta : \theta \in \Theta)$

LAN : $L_n(\theta) = \sum_{i=1}^n \log p_\theta(X_i)$

$L_n(\theta + \frac{h}{\sqrt{n}}) - L_n(\theta) = h\Delta_{n,\theta} - \frac{1}{2}h^2 I_\theta + o_p(1)$

for some $\Delta_{n,\theta}$ and I_θ s.t. $\Delta_{n,\theta} \rightsquigarrow N(0, I_\theta)$

Thm (BvM, Le Cam) If ① LAN

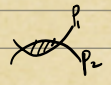
② $\pi(|\sqrt{n}(\theta-\theta^*)| > M_n \mid X_1 \cdots X_n) = o_p(1)$ for some $M_n \to \infty$

③ π is smooth at θ^*

Then $TV\left(P_{\sqrt{n}(\theta-\theta^*)|X_1\cdots X_n},\ N(I_{\theta^*}^{-1}\Delta_{n,\theta}, I_{\theta^*}^{-1})\right) \xrightarrow{P_{\theta^*}} 0$

proof) Let $h = \sqrt{n}(\theta-\theta^*)$, $C_n = \{\sqrt{n}|\theta-\theta^*| \le M_n\} = \{|h| \le M_n\}$

$$\Pi^{C_n}(B|X) = \frac{\pi(B \cap C_n | X)}{\pi(C_n|X)}$$



$$TV(\Pi^{C_n}(\cdot|X), \pi(\cdot|X)) = \sup_B |\Pi^{C_n}(B|X) - \pi(B|X)|$$

$$= \sup_B \left| \frac{\pi(B\cap C_n|X) - \pi(B|X)\pi(C_n|X)}{\pi(C_n|X)} \right|$$

$$\le \sup_B \left| \frac{\pi(B\cap C_n|X) - \pi(B|X)}{\pi(C_n|X)} \right| + \sup_B \left| \frac{\pi(B|X) - \pi(B|X)\pi(C_n|X)}{\pi(C_n|X)} \right|$$

$$\le \left| \frac{2\pi(C_n^c|X)}{\pi(C_n|X)} \right| \xrightarrow{P_{\theta^*}} 0 \quad \text{by ②}$$

($2\pi(C_n^c|X) = o_p(1)$)

also have $TV\left(N(I_{\theta^*}^{-1}\Delta_{n,\theta^*}, I_{\theta^*}^{-1}),\ N^{C_n}(I_{\theta^*}^{-1}\Delta_{n,\theta^*}, I_{\theta^*}^{-1})\right) \xrightarrow{P_{\theta^*}} 0$

STS to bound

$$TV\left(P^{C_n}_{h|X},\ N^{C_n}(I_{\theta^*}^{-1}\Delta_{n,\theta^*}, I_{\theta^*}^{-1})\right)$$

density $\frac{1_{C_n}(h)\prod_{i=1}^n p_{\theta^*+\frac{h}{\sqrt{n}}}(X_i)\,\pi(\theta^*+\frac{h}{\sqrt{n}})}{\int_{C_n}\prod_{i=1}^n p_{\theta^*+\frac{\vartheta}{\sqrt{n}}}(X_i)\,\pi(\theta^*+\vartheta/\sqrt{n})\,d\vartheta}$

density $dN^{C_n}(I_{\theta^*}^{-1}\Delta_{n,\theta^*}, I_{\theta^*})(h)$

$$TV(P,Q) = \frac{1}{2}\int|p-q| = \frac{1}{2}\int_{p>q}(p-q) + \frac{1}{2}\int_{q>p}(q-p)$$

$$= \int (p-q)_+ \qquad X_+ = \max(X, 0)$$

$$= \int \left(1 - \frac{q}{p}\right)_+ dP$$

$$TV(\quad) = \int \left(1 - \frac{dN^{C_n}(I_{\theta^*}^{-1}\Delta_{n,\theta^*}, I_{\theta^*}^{-1})(h) \int_{C_n} \prod_{i=1}^n p_{\theta^*+\vartheta/\sqrt{n}}(X_i)\,\pi(\theta^*+\vartheta/\sqrt{n})\,d\vartheta}{1_{C_n}(h) \prod_{i=1}^n p_{\theta^*+\frac{h}{\sqrt{n}}}(X_i)\,\pi(\theta^*+\frac{h}{\sqrt{n}})}\right) dP^C_{H|X}$$

$$= \int \left(1 - \int_{C_n} \frac{dN^{C_n}(I_{\theta^*}^{-1}\Delta_{n,\theta^*}, I_{\theta^*}^{-1})(h) \prod_{i=1}^n p_{\theta^*+\vartheta/\sqrt{n}}(X_i)\,\pi(\theta^*+\vartheta/\sqrt{n})}{\cancel{1_{C_n}(h)} \prod_{i=1}^n p_{\theta^*+\frac{h}{\sqrt{n}}}(X_i)\,\pi(\theta^*+\frac{h}{\sqrt{n}})} d\vartheta\right)_+ dP^C_{H|X}$$

outside of support

$$= \int \left(1 - \int_{C_n} \frac{\prod_{i=1}^n p_{\theta^*+\vartheta/\sqrt{n}}(X_i)\; dN^{C_n}(I_{\theta^*}^{-1}\Delta_{n,\theta^*}, I_{\theta^*}^{-1})(h)\,\pi(\theta^*+\vartheta/\sqrt{n})}{C_n \prod_{i=1}^n p_{\theta^*+\frac{h}{\sqrt{n}}}(X_i)\; dN^{C_n}(I_{\theta^*}^{-1}\Delta_{n,\theta^*}, I_{\theta^*}^{-1})(\vartheta)\,\pi(\theta^*+\frac{h}{\sqrt{n}})} dN^{C_n}(I_{\theta^*}^{-1}\Delta_{n,\theta^*}, I_{\theta^*}^{-1})(\vartheta)\right)_+ dP^{C_n}_{H|X}$$

$$\overset{\text{Jensen}}{\leq} \int \left(1 - \frac{\prod_{i=1}^n p_{\theta^*+\vartheta/\sqrt{n}}(X_i)\; dN^{C_n}(I_{\theta^*}^{-1}\Delta_{n,\theta^*}, I_{\theta^*}^{-1})(h)\,\pi(\theta^*+\vartheta/\sqrt{n})}{C_n \prod_{i=1}^n p_{\theta^*+\frac{h}{\sqrt{n}}}(X_i)\; dN^{C_n}(I_{\theta^*}^{-1}\Delta_{n,\theta^*}, I_{\theta^*}^{-1})(\vartheta)\,\pi(\theta^*+\frac{h}{\sqrt{n}})}\right)_+ dN^{C_n}(I_{\theta^*}^{-1}\Delta_{n,\theta^*}, I_{\theta^*}^{-1})(\vartheta)\, dP^{C_n}_{H|X}$$

$\frac{1}{C_n}$ cancelled

$|\vartheta| \le M_n$, $M_n \to \infty$ slower than $\sqrt{n}$ ($\to \pi(\theta^*)$ by Slutsky)

$$\frac{\pi(\theta^*+\vartheta/\sqrt{n})}{\pi(\theta^*+h/\sqrt{n})} \to 1 \quad \text{(prior effect washed out)}$$

$$= \frac{e^{L_n(\theta^*+\vartheta/\sqrt{n}) - L_n(\theta^*)}\; dN(I_{\theta^*}^{-1}\Delta_{n,\theta^*}, I_{\theta^*}^{-1})(h)}{e^{L_n(\theta^*+h/\sqrt{n}) - L_n(\theta^*)}\; dN(I_{\theta^*}^{-1}\Delta_{n,\theta^*} I_{\theta^*}^{-1})(\vartheta)} \to 1 \text{ by } \textcircled{D}$$

$\to 0$ by DCT

(there is subtlety, the measures depend on n, need to assign)

Other Bayesian asymptotic results

① Least-informative prior

$D(\pi(\theta|X) \| \pi(\theta))$ $\quad P(X|\theta) \Rightarrow$ predictive. (why)

$m(x) = \int P(x|\theta)\pi(\theta)\,d\theta$

$$\max_{\pi} \int D(\pi(\theta|x) \| \pi(\theta))\, m(x)\, dx$$

$\pi(\theta|x)$ "

$$\frac{\pi(\theta|x)}{\pi(\theta)} = \frac{p(x|\theta)}{m(x)}$$

$$\int D(\pi(\theta|x)\,\|\,\pi(\theta))\, m(x)\,dx$$

$$= \int\int \pi(\theta|x)\log\frac{\pi(\theta|x)}{\pi(\theta)}\,d\theta\, m(x)\,dx$$

$$= \iint \pi(\theta|x)\, m(x)\left(\log\frac{\pi(\theta|x)}{\pi(\theta)}\right)d\theta\,dx$$

$$= \iint \pi(\theta)\, p(x|\theta)\log\frac{p(x|\theta)}{m(x)}\,dx\,d\theta$$

$$= \iint D(p(x|\theta)\,\|\,m(x))\,\pi(\theta)\,dx\,d\theta$$

$X_1 \cdots X_n \mid \theta \overset{iid}{\sim} p(x|\theta)$

$$\frac{m(x)}{p(x|\theta)} = \frac{\int \pi(\tilde\theta)\prod p(x_i|\tilde\theta)\,d\tilde\theta}{\prod p(x_i|\theta)} \qquad (\theta, \tilde\theta \text{ are in neighborhood of } \theta^* \text{ (not rigorous)})$$

$$= \int \pi(\tilde\theta)\, e^{L_n(\tilde\theta)-L_n(\theta)}\,d\tilde\theta$$

$$\approx \pi(\theta)\int e^{L_n(\tilde\theta)-L_n(\theta)}\,d\tilde\theta$$

$$\approx \pi(\theta)\int e^{\sqrt{n}(\tilde\theta-\theta)\Delta_{n,\theta} - \frac{1}{2}n(\tilde\theta-\theta)^2 I_\theta}\,d\tilde\theta$$

$h = \sqrt{n}(\tilde\theta-\theta)$

$$h\Delta_{n,\theta} - \frac{1}{2}h^2 I_\theta = -\frac{1}{2}I_\theta\left(h^2 - 2I_\theta^{-1}h\Delta_{n,\theta}\right)$$

$$= -\frac{1}{2}I_\theta\left(h - I_\theta^{-1}\Delta_{n,\theta}\right)^2 + \frac{\Delta_{n,\theta}^2}{2I_\theta}$$

$$= \pi(\theta)\, e^{\frac{1}{2}\frac{\Delta_{n,\theta}^2}{I_\theta}}\int e^{-\frac{1}{2}I_\theta\left(\sqrt{n}(\tilde\theta-\theta)-I_\theta^{-1}\Delta_{n,\theta}\right)^2}\,d\tilde\theta$$

$$= \pi(\theta)\, e^{\frac{1}{2}\frac{\Delta_{n,\theta}^2}{I_\theta}}\sqrt{\frac{2\pi}{nI_\theta}}$$

In summary, $\dfrac{m(x)}{p(x|\theta)} \approx \pi(\theta)\, e^{\frac{1}{2}\frac{\Delta_{n,\theta}^2}{I_\theta}}\sqrt{\frac{2\pi}{nI_\theta}}$

$$\int D(\pi(\theta|x)\,\|\,\pi(\theta))\, m(x)\,dx$$

$$= \iint \pi(\theta)\, p(x|\theta)\log\frac{p(x|\theta)}{m(x)}\,dx\,d\theta$$

$$\approx \mathop{\mathbb{E}}_{x,\theta}\left(-\log\pi(\theta)\, e^{\frac{1}{2}\frac{\Delta_{n,\theta}^2}{I_\theta}}\sqrt{\frac{2\pi}{nI_\theta}}\right)$$

$$= \mathop{\mathbb{E}}_{x,\theta}\log\left(\frac{1}{\pi(\theta)}\, e^{-\frac{1}{2}\frac{\Delta_{n,\theta}^2}{I_\theta}}\sqrt{\frac{nI_\theta}{2\pi}}\right)$$

$$= \mathop{\mathbb{E}}_{x,\theta}\left(-\frac{1}{2}\frac{\Delta_{n,\theta}^2}{I_\theta} + \frac{1}{2}\log\frac{n}{2\pi} + \log\frac{\sqrt{I_\theta}}{\pi(\theta)}\right)$$

$$= \underset{x,\theta}{\mathbb{E}} \frac{1}{2} \log \frac{n}{2\pi e} + \int \pi(\theta) \log \frac{\sqrt{I_\theta}}{\pi(\theta)} d\theta \qquad \theta \in \mathbb{R}$$

multivariate case $\theta \in \mathbb{R}^p$

Clarke - Barron expansion

$$\int D(\pi(\theta|X) \| \pi(\theta)) \, m(x) \, dx$$

$$= \frac{p}{2} \log \frac{n}{2\pi e} + \int \pi(\theta) \log \frac{\sqrt{\det(I_\theta)}}{\pi(\theta)} d\theta + o(1)$$

the maximizer of C-B expansion is $\pi(\theta) \propto \sqrt{\det(I_\theta)}$

called Jeffrey's prior e.g. $N(\theta, 1)$

"Jeffrey's prior" is Lebesgue measure ↩ not required

② Bayesian model selection

$X | \theta_m, m \quad m \in \mathcal{M} \quad \theta_m \in \mathbb{R}^{d_m}$

$\theta_m \sim \pi_m$

$$\int p(x|\theta_m, m) \pi_m(\theta_m) d\theta_m = p(x|m)$$

$$\hat{m} = \underset{m \in \mathcal{M}}{\operatorname{argmax}} \, p(x|m)$$

BIC

Lec 13-14

$X|\theta_m, m \quad m \in \mathcal{M} \quad \theta_m \sim \pi_m \quad \pi_m$ a prior on $\mathbb{R}^{d_m}$

$$\int p(x|\theta_m, m)\pi_m(\theta_m)\, d\theta_m = p(x|m)$$

$$\hat{m} = \underset{m\in\mathcal{M}}{\operatorname{argmax}}\ p(x|m)$$

Laplace approximation

e.g. $e^{nx_1} + e^{nx_2} + \cdots + e^{nx_k} = (1+o(1))e^{nx_{max}}$

$x_{max} = \max_{1\le j\le k} x_j$ WLOG assume $x_{max} = x_1$

$$\sum_{l=1}^{k} e^{nx_l} = e^{nx_1}\left(1 + e^{n(x_2-x_1)} + \cdots + e^{n(x_k-x_1)}\right) = e^{nx_1}(1+o(1))$$

all negative

$$\int e^{nf(\theta)}\, d\theta \qquad \hat{\theta} = \arg\max_\theta f(\theta) \Rightarrow f'(\hat{\theta}) = 0$$

$$= e^{nf(\hat{\theta})} \int e^{n(f(\theta)-f(\hat{\theta}))}\, d\theta$$

$$\approx e^{nf(\hat{\theta})} \int e^{n(\theta-\hat{\theta})\underbrace{f'(\hat{\theta})}_{=0} + \frac{1}{2}n(\theta-\hat{\theta})^2 f''(\hat{\theta})}\, d\theta$$

Laplace approximation

$$= e^{nf(\hat{\theta})} \int e^{\frac{1}{2}n(\theta-\hat{\theta})^2 f''(\hat{\theta})}\, d\theta \qquad f''(\hat{\theta}) < 0 \text{ (assumption)}$$

$$= e^{nf(\hat{\theta})} \sqrt{\frac{2\pi}{-nf''(\hat{\theta})}}$$

multivariate case $\theta \in \mathbb{R}^d$

$$\int e^{nf(\theta)}\, d\theta \approx e^{nf(\hat{\theta})} \left(\frac{2\pi}{n}\right)^{d/2} \sqrt{\frac{1}{\det(-\nabla^2 f(\hat{\theta}))}}$$

$$\int \pi(\theta) \prod_{i=1}^{n} p(x_i|\theta)\, d\theta = \int e^{nf(\theta)}\, d\theta$$

$$f(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log p(x_i|\theta) + \frac{1}{n}\log \pi(\theta)$$

$$\hat{\theta} = \arg\max_\theta f(\theta) = \hat{\theta}_{MAP}$$

Maximum a posteriori estimation

$$\nabla f(\theta) = \frac{1}{n}\sum_{i=1}^{n} \nabla^2 \log p(x_i|\theta) + \frac{1}{n}\nabla^2 \log \pi(\theta)$$

$$= -I_\theta + o_p(1)$$

$\Rightarrow \quad \nabla^2 f(\hat\theta) = -I_{\hat\theta} + o_p(1)$ not rigorous; need to verify this goes to zero

apply Laplace approximation

$$\Rightarrow \int \pi(\theta) \prod_{i=1}^n p(X_i|\theta)\, d\theta \approx \pi(\hat\theta) \prod_{i=1}^n p(X_i|\hat\theta) \left(\frac{2\pi}{n}\right)^{\frac{d}{2}} \sqrt{\frac{1}{\det(I_{\hat\theta})}}$$

$$\left(\text{using} \quad \int e^{nf(\theta)}\, d\theta \approx e^{f(\hat\theta)} \left(\frac{2\pi}{n}\right)^{d/2} \sqrt{\frac{1}{\det(-\nabla^2 f(\hat\theta))}} \right)$$

Choose π to be Jeffrey's prior

① $\pi(\hat\theta)\sqrt{\frac{1}{\det(I_{\hat\theta})}} = \text{const}$

② $\hat\theta = \hat\theta_{MAP} \approx \hat\theta_{MLE}$ (the two estimators are close in second order)

$$\int \pi(\theta) \prod_{i=1}^n P(X_i|\theta)\, d\theta = C \prod_{i=1}^n P(X_i|\hat\theta_{MLE}) \left(\frac{2\pi}{n}\right)^{\frac{d}{2}}$$

$\Rightarrow \max_{m\in\mathcal{M}} p(X|m)$ is approximately

// $\lg n - \lg 2\pi \doteq (1+o(1)) \log n$

$$\max_{m\in\mathcal{M}} \sum_{i=1}^n \log P(X_i|\hat\theta_m) - \frac{d_m}{2}\log\left(\frac{n}{2\pi}\right)$$

$\hat\theta_m$: MLE for model m

BIC

$$\max_{m\in\mathcal{M}} \left[\sum_{i=1}^n \log P(X_i|\hat\theta_m) - \frac{d_m}{2}\log n \right]$$

regression $y \sim N(X\beta, \sigma^2 I_n)$ $\quad X \in \mathbb{R}^{n\times p}$ $\quad \beta \in \mathbb{R}^p$

$P(y|\beta) \propto e^{-\frac{1}{2\sigma^2}\|y-X\beta\|^2}$

⇒ more conservative : for large n, get sparser model

BIC: $\min_\beta \|y-X\beta\|^2 + \sigma^2 p \log n$

↳ dimension of our model

AIC: $\min_\beta \|y-X\beta\|^2 + 2\sigma^2 p$

SURE (Stein's unbiased risk estimate)

$y \sim N(\mu, \sigma^2 I_n)$, $\mu \in \mathbb{R}^n$

$\hat\mu = \hat\mu(y)$ $\qquad \sigma Z = y - \mu \sim N(0, \sigma^2 I_n)$, $Z \sim N(0, I_n)$

$$\mathbb{E}\|\hat{\mu}-\mu\|^2 = \mathbb{E}\|\hat{\mu}-y+y-\mu\|^2$$

$$= \mathbb{E}\left(\|\hat{\mu}-y\|^2 + \|y-\mu\|^2 + 2\langle\hat{\mu}-y, y-\mu\rangle\right)$$

$$= \mathbb{E}\|\hat{\mu}-y\|^2 + n\sigma^2 + 2\,\mathbb{E}\langle\hat{\mu}, y-\mu\rangle - 2\,\mathbb{E}\langle\mu+\sigma Z, \sigma Z\rangle$$

$$= \mathbb{E}\|\hat{\mu}-y\|^2 - n\sigma^2 + 2\,\mathbb{E}\langle\hat{\mu}, \sigma Z\rangle$$

$$= \mathbb{E}\|\hat{\mu}-y\|^2 - n\sigma^2 + 2\sigma\,\mathbb{E}\sum_{i=1}^n \frac{\partial\hat{\mu}_i}{\partial z_i}$$

$$= \mathbb{E}\|\hat{\mu}-y\|^2 - n\sigma^2 + 2\sigma\,\mathbb{E}\sum_{i=1}^n \frac{\partial\hat{\mu}_i}{\partial y_i}\cdot\frac{\partial y_i}{\partial z_i}$$

$$= \mathbb{E}\|\hat{\mu}-y\|^2 - n\sigma^2 + 2\sigma^2\,\mathbb{E}\sum_{i=1}^n \frac{\partial\hat{\mu}_i}{\partial y_i}$$

$$\mathrm{SURE}(\hat{\mu}) = \underbrace{\|\hat{\mu}-y\|^2}_{\text{fitness}} - n\sigma^2 + 2\sigma^2\underbrace{\sum_{i=1}^n \frac{\partial\hat{\mu}_i}{\partial y_i}}_{\text{Complexity}}$$

"df of $\hat{\mu}$"

Theorem $\mathbb{E}\,\mathrm{SURE}(\hat{\mu}) = \mathbb{E}\|\hat{\mu}-\mu\|^2$

μ: true data generating process

Linear model $y \sim N(X\beta, \sigma^2 I_n)$ (working model)

$\hat{\mu} = X\hat{\beta} = Hy$ $\quad H = X(X^TX)^{-1}X^T$ if X^TX is invertible

$\hat{\beta}$: LSE $\quad = P_X$ (in general)

$$\mathrm{df}(\hat{\mu}) = \sum_{i=1}^n \frac{\partial\hat{\mu}_i}{\partial y_i} = \sum_{i=1}^n H_{ii} = \mathrm{Tr}(H) = p \text{ if } X^TX \text{ invertible}$$

$$= \mathrm{rank}(X) \quad \text{(in general)}$$

$$\mathrm{SURE}(X\hat{\beta}) = \|X\hat{\beta}-y\|^2 - n\sigma^2 + 2\sigma^2 p$$

$\Updownarrow$

AIC $\quad \min_{\beta} \|y-X\beta\|^2 + 2\sigma^2 p$

all-subset selection $\quad y \sim N(X\beta, \sigma^2 I_n)$

↳ full model

assume β is sparse: $\beta_j = 0$ for many j's

can we find the best $X_S\beta_S$ for $S \subseteq [p] = \{1, \dots, p\}$

ℓ_0 penalty $\quad \|\beta\|_0 = \sum_{j=1}^p \mathbb{1}\{\beta_j \neq 0\}$

$$\min\left(\|y-X\beta\|^2 + \underbrace{\pi\|\beta\|_0}_{\text{penalty}}\right)$$

$$= \min_{1 \le s \le p} \min_{\{S \subseteq [p] \,:\, |S| = s\}} \min_{\beta} \|y - X_S \beta_S\|^2 + \pi s$$

AIC: $\pi = 2\sigma^2$

BIC: $\pi = \sigma^2 \log n$

$$\inf_{\hat{\beta}} \sup_{\beta \in \mathbb{R}^p} \mathbb{E} \|X(\hat{\beta} - \beta)\|^2 = \sigma^2 p$$

$$\inf_{\hat{\beta}} \sup_{\|\beta\|_0 \le S} \mathbb{E} \|X(\hat{\beta} - \beta)\|^2 \asymp \sigma^2 S \log\left(\frac{ep}{S}\right)$$

grows slowly

(Birge & Massart)

$S \log\left(\frac{ep}{S}\right) \asymp \log\binom{p}{S}$

cost to pay for not knowing the location of the nonzero β_j's.

rule of thumb
$\Rightarrow$ minimax rate

achieved by

$$\min_{\beta} \left[\|y - X\beta\|^2 + C\sigma^2 \|\beta\|_0 \log\left(\frac{ep}{\|\beta\|_0}\right) \right]$$

Some Constant

(hard to solve)

BIC seems to be a bit better. Usually $n \asymp p$ (in an extremely high dimensional model)

Convex relaxation $\ell_0 \to \ell_1$ — convex

Lasso $\min_{\beta} \left[\|y - X\beta\|^2 + \pi \|\beta\|_1 \right]$ $\quad \|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$

Statistical perf. equivalent. Optimization scale, scale

Tibshirani ($\min \|y - X\beta\|^2$ s.t. $\|\beta\|_1 \le \rho$)

earlier works { Donoho & Chen
Frank & Friedman

How to solve Lasso ?

special case $X = I_p$ $n = p$

$$\min_{\beta} \left[\|y - \beta\|^2 + \pi \|\beta\|_1 \right]$$

$$= \min_{\beta} \sum_{j=1}^{p} \left[(y_j - \beta_j)^2 + \pi |\beta_j| \right]$$

$$= \sum_{j=1}^{p} \min_{\beta_j} \left[(y_j - \beta_j)^2 + \pi |\beta_j| \right]$$

one-dimensional problem

$$\min_b \left[(y-b)^2 + \pi|b| \right] = \min_b f(b)$$

$$f'(b) = \begin{cases} 2(b-y)+\pi & b>0 \\ 2(b-y)-\pi & b<0 \end{cases}$$

$2(b-y)+\pi=0 \iff b = y-\frac{\pi}{2}$

$2(b-y)+\pi=0 \iff b = y+\frac{\pi}{2}$

Case 1 $y > \frac{\pi}{2}$ $\quad f'(y-\frac{\pi}{2})=0 \Rightarrow \hat{b} = y-\frac{\pi}{2}$

case 2 $y < \frac{-\pi}{2}$ $\quad f'(y+\frac{\pi}{2})=0 \Rightarrow \hat{b} = y+\frac{\pi}{2}$

case 3 $y \in [\frac{-\pi}{2}, \frac{\pi}{2}]$

When $b>0 \Rightarrow b-y \geq -\frac{\pi}{2} \Rightarrow 2(b-y) \geq -\pi$

$\Rightarrow f'(b) \geq 0$

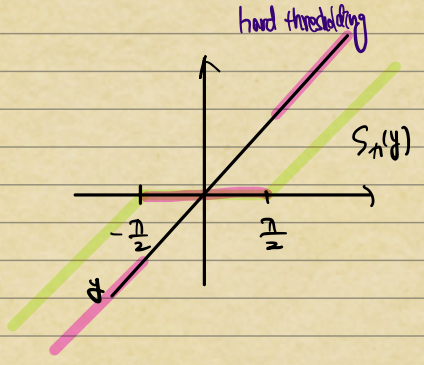
Similarly When $b<0 \Rightarrow f'(b) \leq 0$

$\Rightarrow \hat{b} = 0$

soft thresholding function

$$\operatorname{argmin}_b \left[(y-b)^2 + \pi|b| \right] = S_\pi(y)$$

$$= \begin{cases} y-\frac{\pi}{2} & y > \frac{\pi}{2} \\ 0 & y \in [-\frac{\pi}{2}, \frac{\pi}{2}] \\ y+\frac{\pi}{2} & y < -\frac{\pi}{2} \end{cases}$$



general case: convex optimization

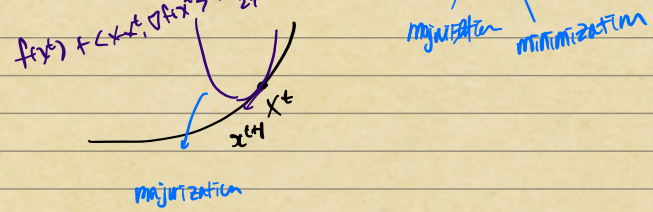
$\min_x f(x)$ $\quad f$ is convex

gradient descent

$$x^{t+1} = x^t - \eta \nabla f(x^t)$$ (step size; differentiate)

$$= \operatorname{argmin}_x \left[f(x^t) + \langle x - x^t, \nabla f(x^t) \rangle + \frac{1}{2\eta}\|x-x^t\|^2 \right]$$

$f(x^t) + \frac{1}{2\eta}\|x-x^t\|^2$ — MM algorithm



$$\min_x (f(x)+g(x))$$

at x_t, an upper bound of $f(x)+g(x)$ is $f(x^t)+\langle x-x^t, \nabla f(x^t)\rangle + \frac{1}{2\eta}\|x-x^t\|^2 + g(x)$

$$x^{t+1} = \arg\min_x \left[f(x^t)+\langle x-x^t, \nabla f(x^t)\rangle + \frac{1}{2\eta}\|x-x^t\|^2 + g(x)\right]$$

Lasso: $\min_\beta [\|y-X\beta\|^2 + \pi\|\beta\|_1]$

$$\begin{cases} f(\beta) = \|y-X\beta\|^2 & \nabla f(\beta) = 2X^T(X\beta - y) \\ g(\beta) = \pi\|\beta\|_1 \end{cases}$$

$$\beta^{t+1} = \arg\min_\beta \left[\|\beta - [\beta^t - 2\eta X^t(X\beta^t - y)]\|^2 + 2\eta\pi\|\beta\|_1 \right]$$

$$= S_{2\eta\pi}(\beta^t - 2\eta X^t(X\beta^t - y)) \Rightarrow \text{explain why Lasso solution is sparse}$$

Iterative Soft thresholding algorithm (Beck & Teboulle)

risk analysis of Lasso

$y \sim N(X\beta, \sigma^2 I)$

$$\hat\beta = \arg\min_\beta [\|y-X\beta\|^2 + \pi\|\beta\|_1]$$

basic inequality

$$\|y-X\hat\beta\|^2 + \pi\|\hat\beta\|_1 \le \|y-X\beta\|^2 + \pi\|\beta\|_1$$

(β: true β)

$$\|y-X\hat\beta\|^2 = \|y-X\beta + X\beta - X\hat\beta\|^2$$

$$= \|y-X\beta\|^2 + \|X(\hat\beta-\beta)\|^2 + 2\langle y-X\beta, X(\beta-\hat\beta)\rangle$$

$$\Rightarrow \|X(\hat\beta-\beta)\|^2 + 2\langle y-X\beta, X\beta - X\hat\beta\rangle + \pi\|\hat\beta\|_1 \le \pi\|\beta\|_1$$

$$\Rightarrow \|X(\hat\beta-\beta)\|^2 \le \underbrace{2\langle y-X\beta, X(\hat\beta-\beta)\rangle}_{2(y-X\beta)^T X(\hat\beta-\beta)} + \pi\|\beta\|_1 - \pi\|\hat\beta\|_1$$

$$\leq 2\|X^T(y-X\beta)\|_\infty \|\hat\beta-\beta\|_1 \quad \text{(Hölder)}$$

If $\pi \geq 2\|X^T(y-X\beta)\|_\infty$

then $\|X(\hat\beta-\beta)\|^2 \leq \pi\|\hat\beta-\beta\|_1 + \pi\|\beta\|_1 - \pi\|\hat\beta\|_1$

$$\leq 2\pi\|\beta\|_1$$

Lemma If $\pi \geq 2\|X^T(y-X\beta)\|_\infty$ then $\|X(\hat\beta-\beta)\|^2 \leq 2\pi\|\beta\|_1$

$y \sim N(X\beta, \sigma^2 I_n)$ $\quad y = X\beta + \sigma z$ $\quad z \sim N(0, I_n)$

$2\|X^T(y-X\beta)\|_\infty = 2\sigma\|X^T z\|_\infty$

$X = \underset{n\times p}{(X_1 \cdots X_p)}$

$$\|X^T z\|_\infty = \max_{1\leq j\leq p} |X_j^T z|$$

$$= \max_{1\leq j\leq p} \|X_j\| \left|\left(\frac{X_j}{\|X_j\|}\right)^T z\right|$$

$$\leq \max_{1\leq j\leq p} \|X_j\| \cdot \max_{1\leq j\leq p} \left|\left(\frac{X_j}{\|X_j\|}\right)^T z\right|$$

$\sim N(0,1)$

assume $\max_{1\leq j\leq p} \|X_j\| \leq L\sqrt{n}$ (constant)

$$\leq L\sqrt{n} \max_{1\leq j\leq p} \left|\left(\frac{X_j}{\|X_j\|}\right)^T z\right|$$

Claim: for $w_1, w_2, \cdots, w_p \sim N(0,1)$ (not necessarily independent)

then $\max_{1\leq j\leq p} |w_j| \lesssim \sqrt{\log p}$ w.h.p.

$$2\|X^T(y-X\beta)\|_\infty = 2\sigma\|X^T z\|_\infty$$

$$\leq 2\sigma L\sqrt{n} \max_{1\leq j\leq p} \left|\frac{|X_j^T z|}{\|X_j\|}\right|$$

$$\lesssim \sigma\sqrt{n\log p} \quad \text{(w.h.p.)}$$

Choose $\pi = C\sigma\sqrt{n\log p}$ then $2\|X^T(y-X\beta)\|_\infty \leq \pi$ w.h.p.

$\Rightarrow$ by the lemma $\|X(\hat{\beta}-\beta)\|^2 \lesssim \sigma\|\beta\|_1\sqrt{n\log p}$

$\Leftrightarrow \frac{1}{n}\|X(\hat{\beta}-\beta)\|^2 \lesssim \sigma\|\beta\|_1\sqrt{\frac{\log p}{n}}$

Lec 15

$y = X\beta + \sigma Z \quad X \in \mathbb{R}^{n \times p} \quad Z \sim N(0,1) \quad \beta \in \mathbb{R}^p$

Lasso $\hat{\beta} = \arg\min_{\beta} \left[\|y - X\beta\|^2 + \pi \|\beta\|_1 \right]$

$\frac{1}{n} \|X(\hat{\beta} - \beta)\|^2 \lesssim \sigma \|\beta\|_1 \sqrt{\frac{\log p}{n}} \qquad \pi = C\sigma\sqrt{\log p}$

w.h.p. $\Rightarrow 2\|X^T(y - X\beta)\|_\infty = 2\sigma\|X^T Z\|_\infty \le \pi$

Claim: for $W_1 \cdots W_p \sim N(0,1)$

$\max_{1 \le j \le p} |W_j| \le \sqrt{\log p}$ w.h.p.

pf) $\mathbb{P}(|W_1| > t) = \mathbb{P}(W_1 > t) + \mathbb{P}(W_1 < -t)$

$= 2\mathbb{P}(W_1 > t)$

$= 2\mathbb{P}(e^{\rho W_1} > e^{\rho t}) \quad \rho > 0$

$\le 2e^{-\rho t} \mathbb{E} e^{\rho W_1}$

$= 2e^{-\rho t + \frac{1}{2}\rho^2}$

$(\rho = t) \quad = 2e^{-\frac{1}{2}t^2}$

$\mathbb{P}(\max_{1 \le j \le p} |W_j| > t) = \mathbb{P}(\bigcup_{j=1}^p E_j) \qquad E_j = \{|W_j| > t\}$

$\le \sum_{j=1}^p \mathbb{P}(E_j)$

$\le 2p e^{-\frac{1}{2}t^2}$

$= 2^{-\frac{1}{2}t^2 + \log(2p)}$

Choose $t^2 = 4\log(2p)$

$\mathbb{P}(\max_{1 \le j \le p} |W_j| > 2\sqrt{\log(2p)}) \le \frac{1}{2p}$

$\Leftarrow \quad \max_{1 \le j \le p} |W_j| \le 2\sqrt{\log(2p)}$ w.p. at least $1 - \frac{1}{2p}$

Birge & Massart $\inf_{\hat{\beta}} \sup_{\|\beta\|_0 \le s} \mathbb{E} \|X\hat{\beta} - X\beta\|^2 \asymp \frac{\sigma^2 s \log(\frac{ep}{s})}{n}$

$\inf_{\hat{\beta}} \sup_{\|\beta\|_1 \le R} \mathbb{E} \frac{1}{n} \|X(\hat{\beta} - \beta)\|^2 \asymp \sigma R \sqrt{\frac{\log p}{n}}$ so for L^1, optimal

Candes & Tao "Dantzig Selector"

$$\min \|\beta\|_1$$

$$\text{s.t.} \quad \|X^T(y - X\beta)\|_\infty \le \pi \qquad \left(\text{previously } 2\|X^T(y - X\beta)\|_\infty = 2\sigma\|X^T z\|_\infty \le \pi\right)$$

$$\frac{1}{n}\|X(\hat\beta - \beta)\|^2 \lesssim \frac{\sigma^2 s \log p}{n k^2} \quad \text{with high probability}$$

(nonparametric rate; can go to 0 for bad design matrix)

Bickel, Ritov & Tsybakov

$$\{\beta \in \mathbb{R}^p : \|\beta\|_0 \le s\}$$

$$\|y - X\hat\beta\|^2 + \pi\|\hat\beta\|_1 \le \|y - X\beta\|^2 + \pi\|\beta\|_1 \quad (\beta \text{ true})$$

$$\Rightarrow \|X(\hat\beta - \beta)\|^2 \le 2\sigma\|X^T z\|_\infty \|\hat\beta - \beta\|_1 + \pi\|\beta\|_1 - \pi\|\hat\beta\|_1$$

Last time, we require

Choose $\pi \ge 4\sigma\|X^T z\|_\infty$ $\left(\pi = C\sigma\sqrt{n \log p}\right)$

$$\le \frac{\pi}{2}\|\hat\beta - \beta\|_1 + \pi\|\beta\|_1 - \pi\|\hat\beta\|_1$$

$$\le \frac{\pi}{2}\|\hat\beta\|_1 + \frac{\pi}{2}\|\beta\|_1 + \pi\|\beta\|_1 - \pi\|\hat\beta\|_1$$

$$\le \frac{3}{2}\pi\|\beta\|_1$$

If $\pi \ge 4\sigma\|X^T z\|_\infty$, then $\|X(\hat\beta - \beta)\|^2 \le \frac{3\pi}{2}\|\beta\|_1$

⇓

deduce the same ℓ_1-norm bound last time

notation $\Delta = \hat\beta - \beta$, $S = \text{supp}(\beta) = \{i \in [p] : \beta_i \ne 0\}$

$\forall\, v \in \mathbb{R}^p \quad \|v_S\|_1 = \sum_{i \in S}|v_i| \quad \|v_{S^c}\| = \sum_{j \in S^c}|v_i|$

$\|v\|_1 = \|v_S\| + \|v_{S^c}\|_1$

$$\|X(\hat\beta - \beta)\|^2 \le \frac{\pi}{2}\|\Delta\|_1 + \pi\|\beta\|_1 - \pi\|\beta + \Delta\|_1$$

$$= \frac{\pi}{2}\|\Delta_S\|_1 + \frac{\pi}{2}\|\Delta_{S^c}\|_1 + \pi\|\beta_S\|_1 - \pi\|\beta_S + \Delta_S\| - \pi\|\Delta_{S^c}\|$$

$$\le -\frac{\pi}{2}\|\Delta_{S^c}\|_1 + \frac{\pi}{2}\|\Delta_S\|_1 + \pi\|\Delta_S\|$$

$$= -\frac{\pi}{2}\|\Delta_{S^c}\| + \frac{3\pi}{2}\|\Delta_S\|_1$$

To summarize, $\|X(\hat\beta - \beta)\|^2 \le -\frac{\pi}{2}\|\Delta_{S^c}\| + \frac{3\pi}{2}\|\Delta_S\|$

Lemma If $n \geq 4c \|X^T z\|_\infty$, then

$$\begin{cases} \|X(\hat\beta - \beta)\|^2 \leq \frac{3n}{2} \|\Delta_S\|_1 \\ \|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1 \quad \text{(cone condition)} \end{cases}$$

$$\|X(\hat\beta-\beta)\|^2 \leq \frac{3n}{2}\|\Delta_S\|_1 \leq \frac{3n}{2}\sqrt{s}\|\Delta\|_2$$

↑ Cauchy Schwartz

$$\frac{1}{n}\|X\Delta\|^2 = \Delta^T\underbrace{\left(\frac{1}{n}X^TX\right)}_{\text{sample covariance}}\Delta \geq \pi_{\min}\left(\frac{1}{n}X^TX\right)\|\Delta\|^2$$

$$\Rightarrow \|\Delta\| \leq \frac{\frac{1}{\sqrt{n}}\|X\Delta\|}{\sqrt{\pi_{\min}\left(\frac{1}{n}X^TX\right)}}$$

$$\|X\Delta\|^2 \leq \frac{3n}{2}\sqrt{s}\|\Delta\| \leq \frac{3n}{2}\sqrt{s}\,\frac{\frac{1}{\sqrt{n}}\|X\Delta\|}{\sqrt{\pi_{\min}\left(\frac{1}{n}X^TX\right)}}$$

$$\|X\Delta\| \leq \frac{3n}{2}\frac{\sqrt{s}}{\sqrt{\pi_{\min}\left(\frac{1}{n}X^TX\right)}} \qquad \text{Choose } n = C\sigma^2 s\log p$$

$$\to \frac{1}{n}\|X\Delta\|^2 \lesssim \frac{\sigma^2 s\log p}{n\,\pi_{\min}\left(\frac{1}{n}X^TX\right)}$$

$\frac{\sigma^2 s\log\left(\frac{ep}{s}\right)}{n}$ $\qquad \pi_{\min} \neq 0$

need $n \geq p$ so that $\pi_{\min}\left(\frac{1}{n}X^TX\right) > 0$

$$\lambda_{\min}\left(\frac{1}{n}X^TX\right) = \min_{\Delta\neq 0} \frac{\Delta^T\left(\frac{1}{n}X^TX\right)\Delta}{\|\Delta\|^2}$$

restricted eigenvalue

$$\text{define } K^2 = \min_{\substack{\Delta\neq 0 \\ \|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1}} \frac{\Delta^T\left(\frac{1}{n}X^TX\right)\Delta}{\|\Delta\|^2}$$

With the same analysis

$$\frac{1}{n}\|X(\hat\beta-\beta)\|^2 \lesssim \frac{\sigma^2 s\log p}{nK^2}$$

Theorem: $X_{ij} \overset{iid}{\sim} N(0,1)$, if $\frac{s\log p}{n}$ is small $\overset{< 0.01}{}$ then

"each column is independent" ↗

$K^2 \geq 1$ w.h.p.

Thm (Lasso error bound)

Choose $\lambda = C\sigma\sqrt{n \log p}$ then w.h.p.

hold for arbitrary design matrix

we have

$$\frac{1}{n}\|X(\hat{\beta}-\beta)\|^2 \lesssim \min\left(\frac{\sigma^2 \|\beta\|_0 \log p}{n K^2}, \|\beta\|_1 \sqrt{\frac{\sigma^2 \log p}{n}} \right)$$

$$\|\hat{\beta}-\beta\|^2 \lesssim \min\left(\frac{\sigma^2 \|\beta\|_0 \log p}{n K^4}, \|\beta\|_1 \frac{1}{K^2}\sqrt{\frac{\sigma^2 \log p}{n}} \right)$$

Zhang, Wainwright & Jordan

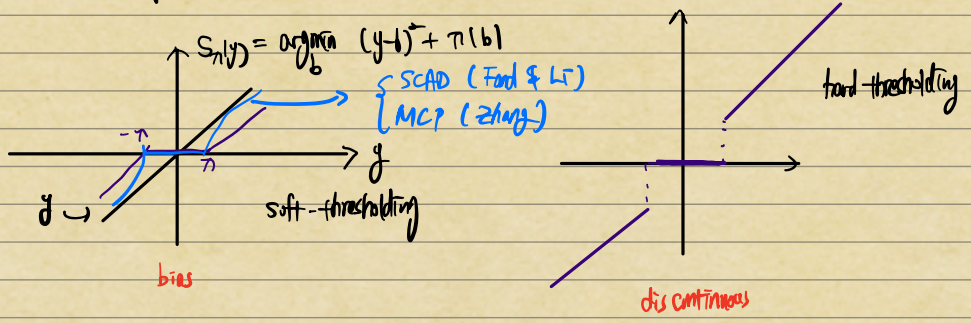
K^2 cannot be completely removed by any polynomial time algorithm

(due to $\frac{1}{n}\|X\Delta\|^2 \geq \pi_{\min}(\frac{1}{n}X^TX)\|\Delta\|^2$ before)

extensions & improvements

① l_0 penalty

② non-convex penalty



for SCAD or MCP (small coeff)

$$\frac{1}{n}\|X(\hat{\beta}-\beta)\|^2 \lesssim \frac{\sigma^2(S_1 + S_2 \log p)}{n K^2}$$

$$S_1 = \left|\left\{ j \in S : |\beta_j| > C\sqrt{\frac{\log p}{n}} \right\}\right| \qquad S_2 = \left|\left\{ j \in S, |\beta_j| \leq C\sqrt{\frac{\log p}{n}} \right\}\right|$$

Lec 16

$y \sim N(X\beta, \sigma^2 I_n)$ $\quad \hat{\beta} = \arg\min_\beta \left(\|y - X\beta\|^2 + \pi \|\beta\|_1 \right)$

Thm. If $\max_{1 \le j \le p} \|X_j\| \lesssim \sqrt{n}$ choose $\pi = C\sigma\sqrt{n \log p}$ (don't know σ)

for some large $C > 0$, we have

$$\frac{1}{n}\|X(\hat{\beta} - \beta)\|^2 \lesssim \min\left(\frac{\sigma^2 \|\beta\|_0 \log p}{n\kappa^2}, \|\beta\|_1 \sqrt{\frac{\sigma^2 \log p}{n}} \right)$$

where $\kappa = \min_{\Delta \neq 0, \ \|\Delta_{S^c}\|_1 \le 3\|\Delta_S\|_1} \dfrac{\frac{1}{\sqrt{n}}\|X\Delta\|}{\|\Delta\|}$

extension & improvements

① ℓ_0 penalty

② non convex penalty

③ SLOPE (Bogdan, Van den Berg, Sabatti, Su, & Candès)

achieves $\frac{1}{n}\|X(\hat{\beta} - \beta)\|^2 \lesssim \frac{\sigma^2 s \log(\frac{ep}{s})}{n\kappa^2}$

how to achieve? $\min\left(\|y - X\beta\|^2 + \sum_{j=1}^p \pi_j |\beta|_{(j)} \right)$

$|\beta|_{(1)} \ge |\beta|_{(2)} \ge \cdots \ge |\beta|_{(p)}$

④ Scaled Lasso / Square-root Lasso

(Sun & Zhang) (Belloni, Chernozhukov & Wang)

$$F(\beta, \sigma) = \frac{\|y - X\beta\|^2}{\sigma} + n\sigma + \pi\|\beta\|_1$$

jointly convex

Choose $\pi = C\sqrt{n \log p}$

$$\begin{cases} \beta^{t+1} = \arg\min_\beta F(\beta, \sigma^t) = \arg\min_\beta \left(\|y - X\beta\|^2 + \pi\sigma^t \|\beta\|_1 \right) \\ \sigma^{t+1} = \arg\min_\sigma F(\beta^{t+1}, \sigma) = \sqrt{\frac{1}{n}\|y - X\beta^{t+1}\|^2} \end{cases}$$

1-d problem

$\min_\sigma F(\beta, \sigma) = G(\beta)$ ← squared root lasso

$\min_\beta G(\beta) \Leftrightarrow \min_\beta \left(\|y - X\beta\| + \rho\|\beta\|_1 \right)$

⑤ Bayesian methods (Bayesian sparse linear regression)

$$\hat{\beta}_{Lasso} = \text{argmax}\left[e^{-\frac{1}{2\sigma^2}\|y - X\beta\|^2 - \pi\|\beta\|_1} \right]$$

likelihood + Laplace prior

25:00

a useless posterior distribution

Ghosh, Van der Vaart & Zhou

$y|\beta \sim N(X\beta, \sigma^2 I_n)$

$\beta \sim \Pi$ is specified by the following sampling process

① $S \sim \pi(S)$ supported on $\{1 \cdots p\}$

$$\pi(S) \propto e^{-c\, s \log(ep/s)} \frac{\Gamma(s)}{\Gamma(s/2)}$$

② given S, sample $S \subseteq \{1 \cdots p\}$ uniformly over all subsets s.t. $|S| = S$

③ given (S, S), $\beta = \begin{pmatrix} \beta_S \\ \beta_S^c \end{pmatrix}$, $\beta_{S^c} = 0$

$\beta_S \sim e^{-\pi\|X_S\beta_S\|}$ (elliptical Laplace distribution)

$n \times s$

given $X_1 \cdots X_n \sim N(\mu, 1)$

$\mu \sim N(0, \tau^2)$ $\quad \hat{\mu} = E(\mu | X_1 \cdots X_n)$

$\Rightarrow \sup_{\mu} E(\hat{\mu} - \mu) = \infty$

truth

Theorem $\Pi(\|X(\beta - \beta^*)\|^2 > C\sigma^2 s^* \log(\frac{ep}{s^*}) \,|\, y) \xrightarrow{p} 0$

under posterior $\quad \|\beta^*\|_0$

density of elliptical Laplace is

$$\frac{1}{2}\sqrt{\det(X_S^T X_S)}\left(\frac{\pi}{\sigma}\right)^{s/2}\frac{\Gamma(s/2)}{\Gamma(s)} e^{-\pi\|X_S\beta_S\|}$$

any prior that is computable?

spike-and-slab prior (George & McCulloch)

① $\eta \sim$ Beta(a, b) ← conjugate

② $\gamma_1 \cdots \gamma_p | \eta \overset{iid}{\sim}$ Bernoulli(η)

③ $\beta_j | \gamma_j \sim N(0, \sigma^2 v_0)$, $\gamma_j = 0$ spike (very small → tending to 0, best computational reason)

$N(0, \sigma^2 v_1)$, $\gamma_j = 1$ slab

(very large (heavy tail))

can do MCMC

but can be computed by EM (Rockova & George)

Kim & Stephen's method → empirical Bayes

$y|\beta,\sigma^2 \sim N(X\beta, \sigma^2 I_n)$

$\beta_i | \sigma^2, g \overset{iid}{\sim} g_\sigma \qquad g_\sigma(t) = \frac{1}{\sigma} g(\frac{t}{\sigma})$

$\downarrow$

$g = \sum_{k=1}^{K} \pi_k N(0, \sigma_k^2)$

π_k: $\sum_{k=1}^{K} \pi_k = 1$

fix $\sigma_1^2, \sigma_2^2, \cdots \sigma_K^2$ (grid)

Hypothesis test (multiple test & high dimension)

$X \sim N(\theta, I_p) \qquad P_\theta = N(\theta, I_p)$

$H_0: \theta = 0 \qquad H_1: \|\theta\| \geq \varepsilon$

minimax risk $\inf_{\phi} \left[P_0\phi + \sup_{\|\theta\| \geq \varepsilon} P_\theta(1-\phi) \right] = R(\varepsilon)$

minimax separation rate

def. ε^* is minimax separation rate if

① $\forall \delta \in (0,1), \exists C > 0$ s.t. $R(C\varepsilon^*) < \delta$ (upper bound)

② $\exists c_1, c_2 > 0$ such that $R(c\varepsilon^*) > c_2$ (lower bound)

Theorem The minimax separation rate is $\varepsilon^* = p^{\frac{1}{4}}$ $\quad 1:1+\infty$

estimation: $\frac{p}{n}$ should be harder than $(p^{\frac{1}{4}})^2 = p^{\frac{1}{2}}$

the optimal testing procedure is

$\phi = 1\{\|X\|^2 > p + C\sqrt{p}\}$ (expectation; order of sd)

(also an exact optimal) $\quad "\theta^2 + p"$

Lec 19

$X \sim N(\theta, I_p)$ $\quad H_0: \theta = 0 \quad H_1: \|\theta\| > \varepsilon$

minimax separation $\varepsilon^* \sim p^{\frac{1}{4}}$

$H_0: \theta = 0 \quad H_1: \theta \neq 0 \quad \theta$ is sparse

Ingster formulation

$H_0: X_1 \cdots X_p \overset{iid}{\sim} N(0,1)$

$H_1: X_1 \cdots X_p \overset{iid}{\sim} (1-\varepsilon) N(0,1) + \varepsilon N(\mu, 1)$ $\quad (\varepsilon \to 0$ as $p \to \infty)$

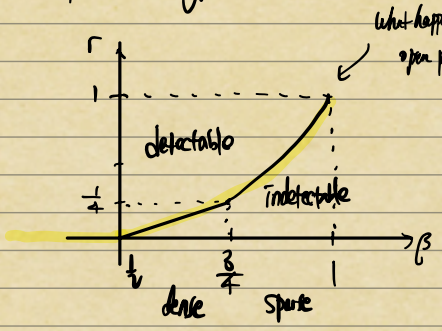
ε or μ small $\to$ hard

Theorem (Ingster) Set $\varepsilon = p^{-\beta} \quad \mu = \sqrt{2r \log p}$

If $r > r^*(\beta)$ can have a consistent test

If $r < r^*(\beta)$ does not exist consistent test

phase diagram



what happen on the boundary? open problem

$$r^*(\beta) = \begin{cases} \beta - \frac{1}{2} & \frac{1}{2} < \beta \leq \frac{3}{4} \\ (1 - \sqrt{1-\beta})^2 & \frac{3}{4} < \beta < 1 \end{cases}$$

when $\beta \leq \frac{1}{2}$ (pretty dense case): at least $\sqrt{p}$ non-zero coefficients

under H_1, $\|\theta\|^2 \sim p^{1-\beta} \mu^2$ ("$p^{\frac{1}{2}}$")

$\sim p^{1-\beta} \log p$

$\geq \sqrt{p} \log p \gg \sqrt{p}$

as long as $r > 0$, detectable using Chi square test

$\beta < 1$ necessary

$W_1 \cdots W_p \overset{iid}{\sim} N(0,1)$

$\max_{1 \leq j \leq p} W_j = (1 + o_p(1)) \sqrt{2 \log p}$