MATH 5473 Homework3 LUO Yuanhui

1. *Maximum Likelihood Method*: consider n random samples from a multivariate normal distribution, $X_i \in \mathbb{R}^p \sim \mathcal{N}(\mu, \Sigma)$ with $i = 1, \ldots, n$.

   (a) Show the log-likelihood function

   $$l_n(\mu, \Sigma) = -\frac{n}{2}\text{trace}(\Sigma^{-1} S_n) - \frac{n}{2}\log\det(\Sigma) + C,$$

   where $S_n = \frac{1}{n}\sum_{i=1}^n (X_i - \mu)(X_i - \mu)^T$, and some constant C does not depend on μ and Σ;

   (b) Show that $f(X) = \text{trace}(AX^{-1})$ with $A, X \succeq 0$ has a first-order approximation,

   $$f(X + \Delta) \approx f(X) - \text{trace}(X^{-1}A'X^{-1}\Delta)$$

   hence formally $df(X)/dX = -X^{-1}AX^{-1}$ (note $(I + X)^{-1} \approx I - X$);

   (c) Show that $g(X) = \log\det(X)$ with $A, X \succeq 0$ has a first-order approximation,

   $$g(X + \Delta) \approx g(X) + \text{trace}(X^{-1}\Delta)$$

   hence $dg(X)/dX = X^{-1}$ (note: consider eigenvalues of $X^{-1/2}\Delta X^{-1/2}$);

   (d) Use these formal derivatives with respect to positive semi-definite matrix variables to show that the maximum likelihood estimator of Σ is

   $$\hat{\Sigma}_n^{MLE} = S_n.$$

A reference for (b) and (c) can be found in Convex Optimization, by Boyd and Vandenbergh, examples in Appendix A.4.1 and A.4.3:

`https://web.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf`

Solution: (a) $f(x_i|\mu,\Sigma) = \frac{1}{\sqrt{2\pi|\Sigma|}} \exp\{-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\}$

The log-likelihood function $L_n(\mu,\Sigma) = \sum_{i=1}^n \log f(x_i|\mu,\Sigma)$

$= \sum_{i=1}^n [-\frac{1}{2}\log(\det(\Sigma)) - \frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu) + C]$

$= -\frac{n}{2}\text{trace}[(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)] - \frac{n}{2}\log(\det(\Sigma)) + C$

$= -\frac{n}{2}\text{trace}(\Sigma^{-1}S_n) - \frac{n}{2}\log\det(\Sigma) + C$

(b) $f(X+\Delta) = \text{trace}[A(X+\Delta)^{-1}] = \text{trace}[A[(I+\Delta X^{-1})X]^{-1}]$

$\approx \text{trace}[AX^{-1}(I-\Delta X^{-1})] = f(X) - \text{trace}(AX^{-1}\Delta X^{-1})$

$= f(X) - \text{trace}(X^{-1}\Delta X^{-1}A') = f(X) - \text{trace}(X^{-1}A'X^{-1}\Delta)$

(c) $g(X+\Delta) = \log|X+\Delta| = \log|X^{\frac{1}{2}}(I + X^{-\frac{1}{2}}\Delta X^{-\frac{1}{2}})X^{\frac{1}{2}}|$

$= \log|X| + \log|I + X^{-\frac{1}{2}}\Delta X^{-\frac{1}{2}}| = g(X) + \sum_{i=1}^{n}\log(1+\lambda_i)$.

where λ_i is the i-th eigervalue of $X^{-\frac{1}{2}}\Delta X^{-\frac{1}{2}}$

Since Δ is small, λ_i's are small, then $\log(1+\lambda_i) \approx \lambda_i$

$\sum_{i=1}^{n}\log(1+\lambda_i) \approx \sum_{i=1}^{n}\lambda_i = tr(X^{-\frac{1}{2}}\Delta X^{-\frac{1}{2}}) = tr(X^{-1}\Delta)$

Then $g(X+\Delta) \approx g(X) + \text{trace}(X^{-1}\Delta)$

(d) $\frac{\partial \ell_n(\mu,\Sigma)}{\partial \Sigma} = \frac{n}{2}\Sigma^{-1}S_n\Sigma^{-1} - \frac{n}{2}\Sigma^{-1} = 0$

$\Rightarrow \hat{\Sigma}_n^{MLE} = S_n$

2. *Shrinkage:* Suppose $y \sim \mathcal{N}(\mu, I_p)$.

(a) Consider the Ridge regression

$$\min_{\mu} \frac{1}{2}\|y-\mu\|_2^2 + \frac{\lambda}{2}\|\mu\|_2^2.$$

Show that the solution is given by

$$\hat{\mu}_i^{ridge} = \frac{1}{1+\lambda}y_i.$$

Compute the risk (mean square error) of this estimator. The risk of MLE is given when $C = I$.

Solution: $L = \frac{1}{2}\|y-\mu\|_2^2 + \frac{\lambda}{2}\|\mu\|_2^2$, $\frac{\partial L}{\partial \mu} = \mu - y + \lambda\mu = 0$

$\Rightarrow \hat{\mu}_i^{ridge} = \frac{1}{1+\lambda}y_i$

$R(\mu,\hat{\mu}) = E\|\mu-\hat{\mu}\|^2 = Var(\hat{\mu}) + Bias(\hat{\mu})^2 = \sigma^2 tr(C^TC) + \|(I-C)\mu\|_2^2$

$= p + 0 = p$

(b) Consider the LASSO problem,

$$\min_{\mu} \frac{1}{2}\|y-\mu\|_2^2 + \lambda\|\mu\|_1.$$

Show that the solution is given by Soft-Thresholding

$$\hat{\mu}_i^{soft} = \mu_{soft}(y_i;\lambda) := \text{sign}(y_i)(|y_i|-\lambda)_+.$$

For the choice $\lambda = \sqrt{2\log p}$, show that the risk is bounded by

$$\mathbb{E}\|\hat{\mu}^{soft}(y)-\mu\|^2 \le 1 + (2\log p + 1)\sum_{i=1}^{p}\min(\mu_i^2, 1).$$

Under what conditions on μ, such a risk is smaller than that of MLE? Note: see Gaussian Estimation by Iain Johnstone, Lemma 2.9 and the reasoning before it.

Solution: $L = \frac{1}{2}\|y-\mu\|_2^2 + \lambda\|\mu\|_1$, $\frac{\partial L}{\partial \mu} = \mu - y + \lambda \text{sign}(\mu) = 0$

if $\mu_i > 0$, $\mu_i = y_i - \lambda$ and $y_i > \lambda$; if $\mu_i < 0$, $\mu_i = y_i + \lambda$ and $y_i < -\lambda$

if $\mu_i = 0$, $y_i = \lambda \text{sign}(\mu_i) \in [-\lambda, \lambda]$

Therefore, $\hat{\mu}_i^{soft} = \text{sign}(y_i)(|y_i|-\lambda)_+$, Let $y = \mu + \varepsilon$, $\varepsilon \in N(0, I)$

$R(\mu,\lambda) = E\|\hat{\mu}-\mu\|^2 = E\left[\text{sign}(y)(|y|-\lambda)_+ - \mu\right]^2$

$$= \begin{cases} E(\mu+\varepsilon)^2, & \mu < -\lambda-\varepsilon \\ \mu^2, & \text{otherwise} \\ E(\mu-\varepsilon)^2, & \mu > \lambda-\varepsilon \end{cases}$$

Then $R(\lambda,\mu) - R(\lambda,0) \le \mu^2$

Let $\lambda = \sqrt{2\log p}$, $R(\lambda,\mu) \le R(\lambda,0) + \mu^2 \le 1 + (2\log+1)\sum_{i=1}^{p}\min(\mu_i^2, 1)$

Then $R_{LASSO}(\mu,\hat{\mu}) \le R_{MLE}(\mu,\hat{\mu}) \Leftrightarrow 1 + (2\log+1)\sum_{i=1}^{p}\min(\mu_i^2, 1) \le p$

(c) Consider the l_0 regularization

$$\min_{\mu} \|y-\mu\|_2^2 + \lambda^2\|\mu\|_0,$$

where $\|\mu\|_0 := \sum_{i=1}^{p} I(\mu_i \neq 0)$. Show that the solution is given by Hard-Thresholding

$$\hat{\mu}_i^{hard} = \mu_{hard}(y_i;\lambda) := y_i I(|y_i| > \lambda).$$

Rewriting $\hat{\mu}^{hard}(y) = (1-g(y))y$, is $g(y)$ weakly differentiable? Why?

Solution: $L_i = \begin{cases} (y_i-\mu_i)^2+\lambda^2, & \mu_i \neq 0 \\ y_i^2, & \mu_i = 0 \end{cases}$

If $(y_i-\mu_i)^2+\lambda^2 > y_i^2 \iff \Delta = 4y_i^2 - 4\lambda^2 \leq 0 \iff |y_i| \leq \lambda$, the optimal μ_i is 0, otherwise optimal μ_i is y_i

Then $\hat{\mu}_i^{hard} = y_i \mathbb{I}(|y_i|>\lambda) = (1 - \mathbb{I}(|y_i| \leq \lambda))y_i$, $g(y) = \mathbb{I}(|y| \leq \lambda)$ is not weak differentiable since $g(y)$ is not absolutely continuous

(d) Consider the James-Stein Estimator

$$\hat{\mu}^{JS}(y) = \left(1 - \frac{\alpha}{\|y\|^2}\right) y.$$

Show that the risk is

$$\mathbb{E}\|\hat{\mu}^{JS}(y) - \mu\|^2 = \mathbb{E}U_\alpha(y)$$

where $U_\alpha(y) = p - (2\alpha(p-2) - \alpha^2)/\|y\|^2$. Find the optimal $\alpha^* = \arg\min_\alpha U_\alpha(y)$. Show that for $p > 2$, the risk of James-Stein Estimator is smaller than that of MLE for all $\mu \in \mathbb{R}^p$.

Proof: Let $g(y) = \frac{\alpha}{\|y\|^2} y$, then $U(Y) \triangleq p + \nabla^T g(y) + \|g(y)\|^2$

we have $\nabla^T g(y) = \sum_{i=1}^{p} \frac{\partial}{\partial y_i} g_i(y) = -\frac{2\alpha(p-2)}{\|y\|^2}$, $\|g(Y)\|^2 = \|-\frac{\alpha}{\|y\|^2} y\|^2 =$

$\alpha^2 \frac{1}{\|y\|^2}$, then $E\|\hat{\mu} - \mu\|^2 = EU_\alpha(y)$

$\frac{\partial U_\alpha(y)}{\partial \alpha} = -\frac{2(p-2) - 2\alpha}{\|y\|^2} = 0 \Rightarrow \alpha^* = p-2$

Then $U_\alpha(y) \geq p - [2(p-2)^2 - (p-2)^2]/\|y\|^2 = p - (p-2)^2/\|y\|^2 < p$

for all $p>2$. Then the risk of J-S Estimator is smaller

than MLE for all $\mu \in \mathbb{R}^p$

(e) In general, an odd monotone unbounded function $\Theta : \mathbb{R} \to \mathbb{R}$ defined by $\Theta_\lambda(t)$ with parameter $\lambda \geq 0$ is called *shrinkage* rule, if it satisfies

[shrinkage] $0 \leq \Theta_\lambda(|t|) \leq |t|$;
[odd] $\Theta_\lambda(-t) = -\Theta_\lambda(t)$;
[monotone] $\Theta_\lambda(t) \leq \Theta_\lambda(t')$ for $t \leq t'$;
[unbounded] $\lim_{t\to\infty} \Theta_\lambda(t) = \infty$.

Which rules above are shrinkage rules?

Solution: By drawing the figure of $\hat{\mu}$, then we can see all above four rules satisfy the four conditions: shrinkage, odd, monotone, and unbounded. Thus, there rules are shrinkage rules.

3. *Necessary Condition for Admissibility of Linear Estimators.* Consider linear estimator for $y \sim \mathcal{N}(\mu, \sigma^2 I_p)$

$$\hat{\mu}_C(y) = Cy.$$

Show that $\hat{\mu}_C$ is admissible only if

(a) C is symmetric;
(b) $0 \leq \rho_i(C) \leq 1$ (where $\rho_i(C)$ are eigenvalues of C);
(c) $\rho_i(C) = 1$ for at most two i.

These conditions are satisfied for MLE estimator when $p = 1$ and $p = 2$.

Reference: Theorem 2.3 in Gaussian Estimation by Iain Johnstone, `http://statweb.stanford.edu/~imj/Book100611.pdf`

Proof: Lemma 2.1, let C be any matrix and P be any orthogonal matrix, then $P^T CP$ is admissible iff Cy is admissible

Let $L(\psi(\theta), \delta(y))$ be the loss for estimating $\psi(\theta)$ by $\delta(y)$, then

risk $\rho(\psi, \delta; \theta) = E_\theta(\psi(\theta), \delta(y))$

Suppose P is orthogonal iff ψ_1, δ_1 given, let $\psi_2(\theta) = \psi_1(P\theta)$, $\delta_2(y) = \delta_1(Py)$, $\rho(\psi_1, \delta_2, \theta) = E_\theta L(\psi_1(P\theta), \delta_1(Py))$, $\rho(\psi_2, \delta_2, \theta) = E_\theta L(\psi_1(P\theta), \delta_1(Py)) = \rho_1(\psi_1, \delta_1; \theta)$.

Thus, ψ_2 is admissible for δ_2 iff ψ_1 is addmissible for δ_1

Let $\psi_1(\theta)=\theta$, $S_1 y = Cy$. then $P'CPy$ is addmissible for θ iff Cy is admissible for θ

Let $C=P^TDP$. $D=diag(d_1,d_2\cdots)$, then Dy is admissible.

Since the replacement $d_i>1 \rightarrow d_i=1$. $d_i<0 \rightarrow -d_i$ and $d_i=1 \Rightarrow (1-(k-2)/\sum_{i=1}^{p} q_i^2)$ will lead to a better estimator Then Cy is also admissible.