Homework #3 ZHONG, Ziyu Student ID: 20923387

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

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

   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;

$$l_n(\mu,\Sigma) = \log \prod_{i=1}^n f(X_i|\mu,\Sigma) = \log \prod_{i=1}^n \frac{1}{\sqrt{2\pi|\Sigma|}} \exp\left[-\frac{1}{2}(X_i-\mu)^T \Sigma^{-1}(X_i-\mu)\right]$$

$$= -\frac{1}{2}\sum_{i=1}^n (X_i-\mu)^T\Sigma^{-1}(X_i-\mu) - \frac{n}{2}\log|\Sigma| - \frac{n}{2}\log(2\pi)$$

$$= -\frac{1}{2}\sum_{i=1}^n tr\left(\Sigma^{-1}(X_i-\mu)(X_i-\mu)^T\right) - \frac{n}{2}\log|\Sigma| + C$$

$$= -\frac{1}{2} tr\left(\sum_{i=1}^n \Sigma^{-1}(X_i-\mu)(X_i-\mu)^T\right) - \frac{n}{2}\log|\Sigma| + C$$

$$= -\frac{n}{2} tr\left(\Sigma^{-1}\sum_{i=1}^n \frac{(X_i-\mu)(X_i-\mu)^T}{n}\right) - \frac{n}{2}\log|\Sigma| + C$$

$$= -\frac{n}{2} tr(\Sigma^{-1}S_n) - \frac{n}{2}\log|\Sigma| + C$$

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

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

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

$$f(X+\Delta) = tr(A(X+\Delta)^{-1}) = tr\left(A X^{-\frac{1}{2}}(I + X^{-\frac{1}{2}}\Delta X^{-\frac{1}{2}})^{-1} X^{-\frac{1}{2}}\right)$$

$$\approx tr\left(AX^{-\frac{1}{2}}(I - X^{-\frac{1}{2}}\Delta X^{-\frac{1}{2}})X^{-\frac{1}{2}}\right)$$

$$= tr\left(AX - AX^{-1}\Delta X^{-1}\right)$$

$$= tr(AX) - tr(X^{-1}AX^{-1}\Delta)$$

$$= f(X) - tr(X^{-1}AX^{-1}\Delta)$$

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

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

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

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

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

$$= \log|X| + \sum_{i=1}^n \log(1+\lambda_i) \quad , \text{ where } \lambda_i \text{ is eigenvalues of } X^{-\frac{1}{2}}\Delta X^{-\frac{1}{2}}.$$

Δ is small

$$\approx \log|X| + \sum_{i=1}^n \lambda_i$$

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

$$= \log|X| + tr(X^{-1}\Delta) \quad , \text{ which is first-order obviously}$$

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

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

$$\ell_n(\mu,\Sigma) = -\frac{n}{2}\mathrm{tr}(\Sigma^{-1}S_n) - \frac{n}{2}\log|\Sigma| + C$$

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

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

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 $C = I$.

$$\ell(\hat{\mu}) = \frac{1}{2}\|y-\hat{\mu}\|_2^2 + \frac{\lambda}{2}\|\hat{\mu}\|_2^2$$

$$\frac{\partial \ell(\hat{\mu})}{\partial \hat{\mu}_i} = -(y_i - \hat{\mu}_i) + \lambda\hat{\mu}_i = 0$$

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

$$\mathbb{E}[\|\hat{\mu}_i^{ridge} - \mu\|^2] = \sum_{i=1}^{p} \mathbb{E}\left[\left(\frac{1}{1+\lambda}y_i - \mu_i\right)^2\right] = \sum_{i=1}^{p}\mathbb{E}\left[\frac{1}{(1+\lambda)^2}(y_i-\mu_i)^2 + \left(\frac{\lambda}{1+\lambda}\right)^2\mu_i^2\right]$$

$$= \frac{p}{(1+\lambda)^2} + \left(\frac{\lambda}{1+\lambda}\right)^2\sum_{i=1}^{p}\mu_i^2$$

$$= \sum_{i=1}^{p} c_i^2 + \sum_{i=1}^{p}(1-c_i)^2\mu_i^2 \quad , \text{ where } \quad C = \frac{1}{1+\lambda} I_p$$

(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

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

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.

$$0 \in \partial_{\hat{\mu}_i}\ell(\hat{\mu}) = \hat{\mu}_i - y_i + \lambda\,\mathrm{sign}(\hat{\mu}_i)$$

$$\Rightarrow \hat{\mu}_i^{soft} = \mathrm{sign}(y_i)(|y_i| - \lambda)_+$$

Let $y_i = \mu_i + z_i$, $z_i \sim N(0,1)$

$$r_i(\lambda,\mu_i) := \mathbb{E}\left[\hat{\mu}_i(\mu_i+z_i) - \mu_i\right]^2 = \int \left[\hat{\mu}_i(\mu_i+z_i) - \mu_i\right]^2 \phi(z_i)\,dz_i$$

Here $$\left[\hat{\mu}_i(\mu_i+z_i) - \mu_i\right]^2 = \begin{cases} (z_i+\lambda)^2 & , \ \mu_i+z_i < -\lambda \\ \mu_i^2 & , \ -\lambda \le \mu_i+z_i \le \lambda \\ (z_i-\lambda)^2 & , \ \mu_i+z_i > \lambda \end{cases}$$

$$\frac{\partial r_i(\lambda,\mu_i)}{\partial \mu_i} = 2\mu_i\,\mathbb{P}(|\mu+z| \le \lambda) \le 2\mu_i$$

Thus $r_i(\lambda,\mu_i) - r_i(\lambda,0) \le \mu_i^2$

while

$$r_i(\lambda,0)=2\int_\lambda^\infty (z_i-\lambda)^2\phi(z_i)\,dz_i = 2(\lambda^2+1)\tilde{\Phi}(\lambda) - 2\lambda\phi(\lambda)$$

By $\tilde{\Phi}(\lambda) \le \frac{\phi(\lambda)}{\lambda}$

$$r_i(\lambda,0) \le \frac{2\phi(\lambda)}{\lambda} \le e^{-\frac{\lambda^2}{2}}$$

$\downarrow$ true for $\lambda > 2\phi(0) \approx 0.8$

And $r_i(\lambda,\infty) = 1+\lambda^2$

So $r_i(\lambda,\mu) \le r_i(\lambda,0) + \min\{\mu_i^2, 1+\lambda^2\}$

Let $\lambda=\sqrt{2\log p}$, then $r_i(\lambda,0) \le e^{-\frac{\lambda^2}{2}} = \frac{1}{p}$

thus $\le \frac{1}{p} + \min\{\mu_i^2, 2\log p+1\}$

$\le \frac{1}{p} + (2\log p+1)\min\{\mu_i^2, 1\}$

Finally, $E[\|\hat{\mu}(y)-\mu\|^2] = \sum_{i=1}^p r_i(\lambda,\mu) \le 1 + (2\log p+1)\sum_{i=1}^p \min\{\mu_i^2,1\}$

better than MLE when μ is sparse $(s=\#\{i: \mu_i\neq 0\})$,

we may expect the risk of soft-threshold $\le O(s\log p) \le O(p)$, the risk of MLE

(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

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

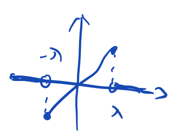
Rewriting $\hat{\mu}^{hard}(y) = (1-g(y))y$, is $g(y)$ weakly differentiable? Why?

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

$\min_{\mu_i=0} \ell(\mu_i) = y_i^2$, $\min_{\mu_i\neq 0} \ell(\mu_i) = \lambda^2$, $\arg\min_{\mu_i\neq 0} \ell(\mu_i) = y_i$

thus $\arg\min_{\mu_i} \ell(\mu_i) = \begin{cases} 0 & y_i^2 > \lambda^2 \\ y_i & y_i^2 \le \lambda^2 \end{cases} = y_i I(|y_i|>\lambda)$

$\hat{\mu}(y) = y + g(y)$, here $g_i(y) = [1 - I(|y_i|>\lambda)]\,y_i$



which is not weakly differentiable.

(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 $\mu \in \mathbb{R}^p$.

$$U_\alpha(y) = p + 2\nabla^T g(Y) + \|g(Y)\|^2 = p + 2\sum_{i=1}^{p}\left[-\alpha\frac{\|y\|^2 - 2y_i^2}{\|y\|^4}\right] + \sum_{i=1}^{p}\frac{\alpha^2 y_i^2}{\|y\|^4}$$

$$= p - 2\alpha\left[\frac{p}{\|y\|^2} - \frac{2}{\|y\|^2}\right] + \frac{\alpha^2}{\|y\|^2}$$

$$= p - \frac{2\alpha(p-2) - \alpha^2}{\|y\|^2}$$

$$\alpha^* = p-2$$

$$\mathbb{E}\, U_{\alpha^*}(y) = p - \mathbb{E}\frac{(p-2)^2}{\|y\|^2} < p = \text{the risk of MLE}$$

(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?

All.

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

(a)

Use the notation $|A| = (A^TA)^{\frac{1}{2}}$,

First prove $\operatorname{tr} A \leq \operatorname{tr}|A|$:

Since $A = UD^{\frac{1}{2}}V^T = \underbrace{UV^T}_{:=\tilde{U}}\underbrace{VD^{\frac{1}{2}}V^T}_{=(A^TA)^{\frac{1}{2}}}$

$= \tilde{U}|A|$, note that $\tilde{U}$ is orthogonal.

Let $u_{ii} := \mathrm{diag}(\tilde{U})_i$, then $|u_{ii}| \le 1$ since $\tilde{U}$ is orthogonal.

$$tr(A) = tr(\tilde{U}|A|) = \sum_{i=1}^{p} u_{ii} \cdot \sigma_i \le \sum_{i=1}^{p} \sigma_i = tr(|A|)$$, where σ_i is eigenvalue of $|A|$.

Then we prove equality holds iff $A = A^T$:

$A = A^T \Rightarrow trA = tr|A|$ is obvious.

For $trA = tr|A| \Rightarrow A = A^T$:

we can infer $u_{ii} = 1$, then $\tilde{U} = I$,

thus $A = |A|$, square on both side, then $AA = A^T A$,

which concludes $A = A^T$.

If C is a linear estimator, then $r(\hat{\mu}_C(y), \mu) = \sigma^2 tr(C^T C) + \|(I-C)\mu\|^2$.

We define $D = I - |I-C|$, D is symmetric obviously.

Then $\|(I-D)\mu\|^2 = \mu^T (I-D)^T (I-D)\mu = \mu^T |I-C|^2 \mu = \mu^T (I-C)^T (I-C)\mu = \|(I-C)\mu\|^2$,

which indicates the biases are equal.

$$tr(D^T D) = tr I - 2 tr(I-D) + tr (I-D)^T (I-D)$$

then $$tr(D^T D) < tr(C^T C) \iff tr(I-D) = tr|I-C| > tr(I-C)$$
$$\iff (I-C)^T \ne I-C$$
$$\iff C \ne C^T$$

Thus C has to be symmetric if admissible.

(b) Since C is symmetric, $C = P\Lambda P^T$, P is orthogonal, $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \cdots \lambda_p)$.

$$r(\hat{\mu}_C(y), \mu) = \sigma^2 tr(C^T C) + \|(I-C)\mu\|^2$$
$$= \sum_{i=1}^{p} \sigma^2 \lambda_i^2 + (1-\lambda_i)^2 \cdot \eta_i^2$$, where $P^T \mu = \begin{pmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_p \end{pmatrix}$

When $\lambda_i > 1$, let $\lambda_i = 1$ or when $\lambda_i < 0$, let $\lambda_i = 0$, the estimation will be everywhere better.

(c)
$$\mathbb{E}\|Cy - \mu\|^2 = \mathbb{E}\|\Lambda x - \eta\|^2$$, where $x = P^T y \sim N(\eta, \sigma^2 I_p)$
$$= \sum_{i=1}^{p} \sigma^2 \lambda_i^2 + (1-\lambda_i)^2 \cdot \eta_i^2$$

If $\lambda_1 = \lambda_2 = \cdots = \lambda_d = 1$, $d \geqslant 3$,

we know J-S estimator is everywhere better than MLE, i.e. $C = I$,

$$\mathbb{E}\|\Lambda x - \eta\|^2 = \mathbb{E}\|I_d x_d - \eta_d\|^2 + \mathbb{E}\|\Lambda_{-d} x - \eta_{-d}\|^2$$
$$< r(\hat{\mu}_{J\text{-}S}(x_d), \eta_d) + \mathbb{E}\|\Lambda_{-d} x - \eta_{-d}\|^2$$

Thus $d \leqslant 2$.

4. *James Stein Estimator for $p = 1, 2$ and upper bound:*
If we use SURE to calculate the risk of James Stein Estimator,

$$R(\hat{\mu}^{\mathrm{JS}}, \mu) = \mathbb{E}U(Y) = p - \mathbb{E}_\mu \frac{(p-2)^2}{\|Y\|^2} < p = R(\hat{\mu}^{\mathrm{MLE}}, \mu)$$

it seems that for $p = 1$ James Stein Estimator should still have lower risk than MLE for any μ. Can you find what will happen for $p = 1$ and $p = 2$ cases?
Moreover, can you derive the upper bound for the risk of James-Stein Estimator?

$$R(\hat{\mu}^{\mathrm{JS}}, \mu) \leq p - \frac{(p-2)^2}{p-2+\|\mu\|^2} = 2 + \frac{(p-2)\|\mu\|^2}{p-2+\|\mu\|^2}.$$

$$\hat{\mu}^{JS}(Y) = \left(1 - \frac{p-2}{\|Y\|^2}\right) Y$$

When $p = 1$,

$\hat{\mu}^{JS}(Y) = \left(1 + \frac{1}{Y^2}\right) Y = Y + \frac{1}{Y}$, $g(Y) = \frac{1}{Y}$, there is a singularity at $Y = 0$.

$g(Y)$ is not weakly differentiable.

When $p = 2$, $\hat{\mu}^{JS}(Y) = Y =$ MLE.

$$\|Y\| \sim \chi^2_{p+2N} \,\Big|\, N \sim P\left(\frac{\|\mu\|}{2}\right)$$

By $\mathbb{E}\left[\frac{1}{\chi^2_p}\right] = \frac{1}{p-2}$, condition on N:

$$\mathbb{E}\left[\frac{1}{\chi^2_{p+2N}}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{1}{\chi^2_{p+2N}} \,\middle|\, N\right]\right] \overset{\text{independent}}{=\!=} \mathbb{E}\left[\mathbb{E}\left[\frac{1}{\chi^2_{p+2\tilde{n}}}\right]\Big|_{\tilde{n}=N}\right] = \mathbb{E}\left[\frac{1}{p-2+2N}\right]$$
$$\overset{\text{Jensen's}}{\geqslant} \frac{1}{p-2+2\mathbb{E}N}$$
$$= \frac{1}{p-2+\|\mu\|^2}$$

Thus $r(\hat{\mu}^{JS}, \mu) \leqslant p - \frac{(p-2)^2}{p-2+\|\mu\|^2}$