Saheli Bhattacharya
Stuid = 20999867

CSIC 5011 HW (3rd week)

DATE: / / MO TU WE TH FR SA SU

Q1) $x_i \in \mathbb{R}^p \sim N(\mu, \Sigma)$, $i = 1, 2, \cdots n$

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

using the p.d.f of multivariable gaussian:

$$f(x_i \mid \mu, \Sigma) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x_i-\mu)^T \Sigma^{-1}(x_i-\mu)\right)$$

since x_i are i.i.d, taking the products for $i=1, \cdots n$ & then taking the log gives:

$$l_n(\mu, \Sigma) = \underbrace{-\frac{p}{2} \times n \log(2\pi)}_{C = \text{ independent of } \mu \,\&\, \Sigma} - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^T \Sigma^{-1}(x_i-\mu)$$

$$\Rightarrow l_n(\mu, \Sigma) = -\frac{n}{2}\log|\Sigma| - \frac{n}{2} \text{trace}(\Sigma^{-1} S_n) + C$$ {using trace(AB) = trace(BA)}

(b) $f(x) = \text{trace}(A x^{-1})$, $A, x \geq 0$.

To show: $f(x+\Delta) \approx f(x) - \text{trace}(x^{-1} A x^{-1} \Delta)$

let $z = x + \Delta$

$$f(z) = \text{trace}(A(x+\Delta)^{-1}) = \text{trace}(A x^{-1}(I + x^{-1}\Delta)^{-1})$$

$$= \text{trace}(A(x(I + x^{-1}\Delta))^{-1})$$

$$= \text{trace}(A(I + x^{-1}\Delta)^{-1} x^{-1})$$

$$= \text{trace}(A(I - x^{-1}\Delta) x^{-1})$$ [using $(I + x^{-1}\Delta)^{-1} \approx I - x^{-1}A$]

$$= \text{trace}(A x^{-1} - A x^{-1}\Delta x^{-1})$$

$$f(x+\Delta) = \text{trace}(A x^{-1}) - \text{tr}(x^{-1} A x^{-1} A)$$ [using trace(AB) = trace(BA)]

$$\Rightarrow \frac{df(x)}{dx} = -x^{-1} A x^{-1}$$ [$f(z) = f(x) + \nabla f(x)^T (z-x)$]

DATE: / / MO TU WE TH FR SA SU

(c) $g(x) = \log \det(X)$

let $Z = X + \Delta$

$$g(Z) = \log |X + \Delta|$$
$$= \log |X^{1/2}(I + X^{-1/2} \Delta X^{-1/2}) X^{1/2}|$$
$$= \log |X (I + X^{-1/2} \Delta X^{-1/2})| \quad [\text{using } |AB| = |BA|]$$
$$= \log |X| + \log |I + X^{-1/2} \Delta X^{-1/2}|$$
$$= \log |X| + \sum_{i=1}^{n} \log(1 + \lambda_i)$$, where $\lambda_i \to$ i'th eigen value of $X^{-1/2} \Delta X^{-1/2}$

since Δ is small $\Rightarrow \lambda_i$ are small

$$\Rightarrow \log(1 + \lambda_i) \simeq \lambda_i$$

$$\Rightarrow \log|Z| = \log|X| + \sum_{i=1}^{n} \lambda_i = \log|X| + \text{trace}(X^{-1/2} \Delta X^{-1/2})$$
$$= \log|X| + \underbrace{\text{trace}(X^{-1} \Delta)}_{\substack{\downarrow \\ \text{inner product b/w } X^{-1} \text{ \& } \Delta}} \quad [\text{using trace}(AB) = \text{trace}(BA)]$$

$$f(Z) \approx f(X) + \text{trace}(X^{-1}(Z - X))$$
$$\Rightarrow \frac{dg(X)}{dX} = X^{-1}$$

d) $$l_n(\mu, \Sigma) = -\frac{n}{2} \text{trace}(\Sigma^{-1} S_n) - \frac{n}{2} \log|\Sigma| + C$$
$$= -\frac{n}{2} \text{trace}\left(\Sigma^{-1} \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^T (x_i - \mu)\right)$$
$$- \frac{n}{2} \log|\Sigma| + C$$

DATE: / / MO TU WE TH FR SA SU

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

$$\frac{\partial \, l_n(\mu, \Sigma)}{\partial \mu} = \sum_{i=1}^{n} \Sigma^{-1} (x_i - \mu) = 0$$

$$\Rightarrow \hat{\mu}_{mL} = \frac{1}{n} \sum_{i=1}^{n} x_i .$$

for $\hat{\Sigma}_{mL}$

$$\frac{\partial}{\partial \Sigma} l_n(\mu, \Sigma) = -\frac{n}{2} \left(-\Sigma^{-1} S_n \Sigma^{-1}\right) - \frac{n}{2} \Sigma^{-1} = 0$$

$$\Rightarrow \Sigma^{-1} S_n \Sigma^{-1} = \Sigma^{-1}$$

$$S_n \Sigma^{-1} = I$$

$$\boxed{\hat{\Sigma} = S_n}.$$

$$\hat{\Sigma}_{mL} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu}_{mL})^T (x_i - \hat{\mu}_{mL})$$

Q2) $y \sim N(\mu, I_p)$

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

Setting the derivative to 0

$$\frac{1}{\cancel{2}} \cancel{2} (y - \mu)(-1) + \frac{\lambda}{\cancel{2}} \times \cancel{2} y = 0 .$$

$$\mu(\lambda + 1) = y \Rightarrow \mu = \frac{y}{\lambda + 1}$$

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

DATE: / / MO TU WE TH FR SA SU

$$\text{var}(\mu^{-1}) = \left[\left(\frac{1}{1+\lambda}y - \frac{1}{1+\lambda}\mu\right)^T \left(\frac{y}{1+\lambda} - \frac{\mu}{1+\lambda}\right)\right]$$

$$= \left(\frac{1}{1+\lambda}\right)^2 \times \left[(y-\mu)^T (y-\mu)\right] = \frac{p}{(1+\lambda)^2}$$

$$\text{Bias}(\hat{\mu}) = \mu - \frac{\mu}{1+\lambda} = \frac{\lambda}{1+\lambda}\mu.$$

$$\text{Risk} = \text{var}(\hat{\mu}) + \text{Bias}(\hat{\mu}))^2 = \frac{p}{(1+\lambda)^2} + \frac{\lambda^2}{(1+\lambda^2)} \times \|\mu\|^2$$

(b) $\min_{\mu} \frac{1}{2}\|y-\mu\|^2 + \lambda\|\mu\|_1 = f(\mu)$

for minimizing $\Rightarrow 0 \in \partial f(\mu)$.

$\partial f(\mu) = (\mu - y) + \lambda \, \text{sign}(\mu)$

if $\mu_i > 0$

$\Rightarrow \mu_i - y_i + \lambda = 0 \Rightarrow \mu_i = y_i - \lambda$ if $y_i > \lambda_i$

if $\mu_i < 0$

$\Rightarrow \mu_i - y_i - \lambda = 0 \Rightarrow \mu_i = y_i + \lambda$ if $y_i + \lambda < 0$

$y_i < -\lambda$.

if $\mu_i = 0$

~~$\Rightarrow \mu_i - y_i - \lambda = 0 \Rightarrow \mu_i = y_i + \lambda$ if $y_i + \lambda < 0$~~

~~$y_i < \lambda$~~

if ~~μ~~ $\Rightarrow y_i \in [-\lambda, \lambda]$

DATE: / / MO TU WE TH FR SA SU

$$\Rightarrow \mu_i = \begin{cases} y_i - \lambda & , \text{if } y_i > \lambda \\ 0 & \text{if } -\lambda \leq y_i \leq \lambda \\ y_i + \lambda & \text{if } y_i < -\lambda \end{cases}$$

$$\Rightarrow \mu_i^{soft} = \operatorname{sgn}(y_i)\left(|y_i| - \lambda\right)_+$$

$$\mathbb{E}\,\|\hat{\mu}^{soft}(y) - \mu\|^2 \qquad \hat{\mu}_{soft} = y + g(y)$$

$$\lambda = \sqrt{2\log p}$$

soft thresholding.

$$g(y) = \begin{cases} -\lambda & \text{if } y_i > \lambda \\ -y & \text{if } -\lambda \leq y_i \leq \lambda \\ \lambda & \text{if } y_i < -\lambda . \end{cases}$$

From SURE Lemma 2.2 of the notes $\rightarrow$

$$R\left(\hat{\mu}^{soft}, \mu\right) = \mathbb{E}_\mu\left(p + 2\,\nabla^T g(y) + \|g(y)\|^2\right)$$

$$\Delta^T g(y) = \sum_{i=1}^{p} \frac{\partial}{\partial y_i} g(y)$$

$$\frac{\partial g}{\partial y_i}(y) = -I\{|y_i| \leq \lambda\}$$

$$\Rightarrow R\left(\hat{\mu}_{soft}, \mu\right) = \mathbb{E}_\mu\left(p - 2\sum_{i=1}^{p} I\{|y_i| \leq \lambda\} + \sum_{i=1}^{p} \min(\lambda^2, y^2)\right)$$

minimizing the RHS w.r.t λ to get $\hat{\lambda}_{SURE}$

For univariate case, $y = \mu + z \sim N(\mu, 1)$ from [1]

$$\mu(\lambda, \mu) \leq \mu(\lambda, 0) + \min(\mu^2, 1 + \lambda^2)$$

for $\lambda = \sqrt{2\log p}$

$$\mu(\lambda, \mu) \leq \frac{1}{p} + (2\log p + 1)\min(\mu^2, 1)$$

for p-variate distribution, summing over the element

DATE: / / MO TU WE TH FR SA SU

using lemma 2.11 of [1]:

$$R(\hat{\mu}_{soft}, \mu) \leq 1 + \sum_{i=1}^{p} \min(\mu_i^2, 1+\lambda^2)$$
$$\leq 1 + (2\log p + 1) \sum_{i=1}^{p} \{\min(\mu_i^2, 1)$$

If μ is sparse then $R(\hat{\mu}_{soft}, \mu) < R(\hat{\mu}^{MLE}, \mu)$

(c) $\min_{\mu} \|y - \mu\|_2^2 + \lambda^2 \|\mu\|_0 = \min_{\mu} \sum_{i=1}^{n} \left((y_i - \mu_i)^2 + \lambda^2 \mathbb{1}\{\mu_i \neq 0\}\right)$

Solving componentwise:

$$\min_{\mu_i} (y_i - \mu_i)^2 + \lambda^2 \mathbb{1}(\mu_i \neq 0)$$

if $\mu_i = 0 \Rightarrow$ cost $= y_i^2$

if $y_i^2 \leq \lambda^2$ then set $\mu_i = 0$

if $\mu_i \neq 0 \Rightarrow$ min. cost is λ^2 when $\mu_i = y_i$

$$\Rightarrow \mu_i = \begin{cases} y_i & \text{if } y_i^2 \geq \lambda^2 \\ 0 & \text{if } y_i^2 \leq \lambda^2 \end{cases}$$

$$\Rightarrow \mu_i = \begin{cases} y_i, & \text{if } y_i > \lambda \\ 0, & \text{if } -\lambda \leq y_i \leq \lambda \\ y_i, & \text{if } y_i < -\lambda \end{cases}$$

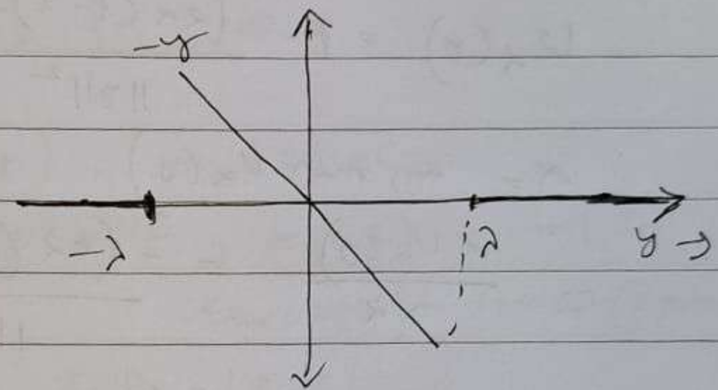
$$\Rightarrow \hat{\mu}_i^{hard} = y_i \, I(|y_i| > \lambda)$$

DATE: / / MO TU WE TH FR SA SU

$\hat{\mu}^{hard} = y + g(y)$.

$$\Rightarrow g(y) = \begin{cases} 0 & \text{if } y > \lambda \\ -y & \text{if } -\lambda \le y \le \lambda \\ 0 & \text{if } y < -\lambda. \end{cases}$$

$g(y) =$



$g(y)$ is not weakly differentiable due to sudden jumps at $-\lambda$ & λ.

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

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

$$= \mathbb{E}\left(\|y - \mu\|^2 + 2(y-\mu)^T \frac{\alpha y}{\|y\|^2} + \frac{\alpha^2 \|y\|^2}{\|y\|^4}\right)$$

$$= p + 2\mathbb{E}_\mu\left[\frac{\alpha (y-\mu)^T y}{\|y\|^2}\right] + E\left[\frac{\alpha^2}{\|y\|^2}\right]$$

$$= p + \mathbb{E}_\mu\left[\frac{2\alpha\|y\|^2 - 2\alpha\mu^T y + \alpha^2}{\|y\|^2}\right]$$

$$= \mathbb{E}_y 2\alpha\left(\|y\|^2 - \mu^T y\right)$$

$$= \mathbb{E}_y 2\alpha\left(y^T y - \mu^T y\right) = 2\alpha\left((y-\mu)^T y\right)$$

DATE: / / MO TU WE TH FR SA SU

$$= \mathbb{E}_y\, 2\alpha\left(\|y-\mu\|^2 - \|\mu\|^2 - \mu^T y\right)]$$

$$= 2\alpha(p-2)$$

$$\Rightarrow \mathbb{E}\|\hat{\mu}^{JS}(y)-\mu\|^2 = \mathbb{E}\left[p - \frac{(2\alpha(p-2)-\alpha^2}{\|y\|^2}\right]$$

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

$$\alpha = \arg\min U_\alpha(y)$$

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

$$\Rightarrow (p-2)-\alpha = 0 \Rightarrow \alpha^* = p-2$$

$$R_{MLE}^{(\hat{\mu}_{MLE})} = \mathbb{E}\|\hat{\mu}_{MLE}-\mu\|^2$$

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

$$R_{JS} = p - \mathbb{E}\frac{(2\alpha(p-2)-\alpha^2)}{\|y\|^2}$$

for $p>2 \Rightarrow \mathbb{E}\left(\frac{2\alpha(p-2)-\alpha^2}{\|y\|^2}\right) > 0$

$$\Rightarrow R_{JS} < p = R_{MLE}$$

$$\Rightarrow R_{JS} < R_{MLE}.$$

DATE: / / MO TU WE TH FR SA SU

(c) ℓ-1 norm minimization, i.e. soft thresholding & James Stein are shrinkage rule which satisfy

(a) $\theta_\lambda(|t|) \leq |t|$

(b) $\theta_\lambda(-t) = \theta_\lambda(t)$

(c) $\theta_\lambda(t) \leq \theta_\lambda(t')$ for $t \leq t'$

(d) $\lim_{t \to \infty} \theta_\lambda(t) = \infty$

Q) 3) $y \sim N(\mu, \sigma^2 I_p)$, $\hat{\mu}_C(y) = Cy$

(i) let $|A| = (A^T A)^{1/2}$, $\operatorname{tr}(A) \leq \operatorname{tr}|A|$

equality iff $A = A^T$ (symmetric)

let D be such that $I - D = |I - C|$

$\Rightarrow D$ is symmetric

MSE, $\Rightarrow E[\|\hat{\mu} - \mu\|^2] = E\|\hat{\mu} - E\hat{\mu}\|^2 + \|E\hat{\mu} - \mu\|^2 = \operatorname{var}(\hat{\mu}) + \operatorname{bias}^2(\hat{\mu})$.

for linear estimators, $\operatorname{var}(\hat{\mu}) = \operatorname{tr}(\operatorname{cov}(\hat{\mu})) = \operatorname{tr}(\sigma^2 CC^T)$

$\Rightarrow \operatorname{var}(\hat{\mu}) = \sigma^2 \operatorname{tr}(C^T C)$

Bias $= E\hat{\mu} - \mu = |C - I|\mu$

$\Rightarrow \text{MSE} = \sigma^2 \operatorname{tr}(C^T C) + \|(I - C)\mu\|^2$ — (A)

Claim: MSE of $\hat{\mu}_D$ is everywhere better than μ_C if C is not symmetric.

$(I - D)^T (I - D) = |I - C|^2 = (I - C)^T (I - C)$

DATE: / / MO TU WE TH FR SA SU

$\Rightarrow$ the bias squared is same for both estimators (from A)

Now, for the variance term is

$tr(D^TD) = \cancel{tr(}\ tr\, I - 2\,tr(I-D) + tr\,(I-D)^T(I-D)$

Now, $tr(D^TD) < tr(C^TC)$ iff

$tr\,(I-D) = \cancel{tr}\ tr\,|I-C| > tr(I-C)$

$\Rightarrow$ It occurs only if C is not symmetric

ii) EVD of symmetric C (proved in (i))

$\Rightarrow$ $C = U \Lambda U^{-1}$ All the eigenvalues are real

let $\eta = U^T\mu$ & $x = U^T y \sim N(\eta, \sigma^2 I_p)$ since $U^TU = I$

Now, $E\|Cy-\mu\|^2 = E\|U\Lambda U^Ty - \mu\|^2$

$= E\|\Lambda x - \eta\|^2$

$\Rightarrow r(\hat{\mu}, \mu) = r(\hat{\eta}_\Lambda, \eta) = \sum_{i=1}^{p} \sigma^2\lambda_i + (1-\lambda_i)^2\eta_i$

If $\lambda_i \notin [0,1]$, a strictly better MSE can be obtained by replacing λ_i by 1 if $\lambda_i > 1$ & by 0 if $\lambda_i < 0$.

DATE: / / MO TU WE TH FR SA SU

iii) let $\lambda_1 = \lambda_2 = \cdots = \lambda_d = 1 > \lambda_i$, $i > d \geq 3$ &

let $x^d = (x_1, \ldots x_d)$

positive part of JS estimator is everywhere better than

$\hat{\eta}_I(x^d) = x^d$

If a new estimator $\hat{\eta}$ is defined to use $\hat{\eta}$ JS on x^d

& to continue to use $\lambda_i x_i$ for $i > d$ then

$$r(\hat{\eta}, \eta) = r(\hat{\eta}^{JS}, \eta^d) + \sum_{i>d} r(\lambda_i, \eta_i) \quad \mathcal{R}(\lambda, \eta)$$

$\Rightarrow$ so, $\hat{\eta}$ dominates $\hat{\eta}_\lambda$ & hence $\hat{\mu}_c$.

Q4) For $p = 1$ $Y \sim N(\mu, \sigma^2)$

$$R(\hat{\mu}^{JS}, \mu) = P - E\mu \frac{(p-2)^2}{\|Y\|^2} \quad \text{—— (1)}$$

$\|Y\|^2 \Rightarrow$ follows non central Chi squared distribution with non-centrality param $\|\mu\|^2$

Non-central distn can be realized as a mixture of central Chi squared distn χ^2_{p+2N}, where N is a Poisson variable with mean $\|\mu\|^2/2$

$$\& \quad \mathbb{E}\left[1/\chi^2_p\right) = \frac{1}{p-2} \quad \text{—— (A)}$$

for $P = 1$ & 2

$$\text{for } P=1 \Rightarrow R(\hat{\mu}_{JS}, \mu) = 1 - \frac{1}{(-1)} = 2 > 1 = \text{MLE}$$

$$\text{for } P=2, \; R(\hat{\mu}_{JS}, \mu) = 2 = R(\hat{\mu}_{MLE}, \mu)$$

DATE: / / MO TU WE TH FR SA SU

Now, conditioning on N & using (A)

$$E\left(\frac{1}{\|\gamma\|^2}\right) = E\left(\frac{1}{\chi^2_{p+2N}}\right) = E\left[\frac{1}{p+2N-2}\right]$$

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

Substituting in (1)

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

$$= 2+(p-2) - \frac{(p-2)^2}{(p-2)+\|\mu\|^2}$$

$$= 2 + \frac{(p-2)\|\mu\|^2}{(p-2)+\|\mu\|^2}.$$