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;

Solution. (a) Since $X_i \in \mathbb{R}^p \overset{i.i.d.}{\sim} N(\mu, \Sigma)$, $(i = 1, \cdots, n)$

the density function is $f(X|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^p|\Sigma|}} \exp\left\{-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)\right\}$,

the likelihood function is $P(X_1, \cdots, X_n|\mu, \Sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{(2\pi)^p|\Sigma|}} \exp\left\{-\frac{1}{2}(X_i-\mu)^T\Sigma^{-1}(X_i-\mu)\right\}$,

so the log-likelihood is $l_n(\mu, \Sigma) = \log P(X_1, \cdots, X_n|\mu, \Sigma) = \sum_{i=1}^{n}\left(\log\left((2\pi)^{-\frac{p}{2}}|\Sigma|^{-\frac{1}{2}}\right) - \frac{1}{2}(X_i-\mu)^T\Sigma^{-1}(X_i-\mu)\right)$

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

where $C = -\frac{pn}{2}\log(2\pi)$ is a constant.

Obviously, $l_n(\mu, \Sigma) = \text{trace}(l_n(\mu, \Sigma))$

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

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

   (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 $(I + X)^{-1} \approx I - X$);

(b) Suppose $\Delta = hV$, where V is a matrix. A, X are P.S.D.

So $f(X+\Delta) - f(X) = tr\left(A(X+hV)^{-1}\right) - tr\left(AX^{-1}\right)$

$$= tr\left(A(I+hX^{-1}V)^{-1}X^{-1}\right) - tr\left(AX^{-1}\right)$$

$$= tr\left\{\left[\left(I+hX^{-1}V\right)^{-1} - I\right]X^{-1}A\right\}$$

(h is small) $\approx tr\left\{\left(I - hX^{-1}V - I\right)X^{-1}A\right\}$

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

$\Rightarrow f(X)$ has first-order approximation: $f(X+\Delta) \approx f(X) - tr\left(X^{-1}AX^{-1}\Delta\right)$

Hence $df(X)/dX = -X^{-1}AX^{-1}$

(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 $X^{-1/2}\Delta X^{-1/2}$);

(c) Suppose $\Delta = hV$, h is small. X and A are S.P.D.

$$g(X+\Delta) - g(X) = \log\det(X+hV) - \log\det(X)$$

$$= \log \frac{\det(X)\det(I + hX^{-1}V)}{\det(X)}$$

$$= \log\det(I + hX^{-1}V)$$

Since X^{-1}, V are S.P.D matrix, suppose its eigenvalue decomposition is

$X^{-1}V = U\Lambda U^T$, $UU^T = U^TU = I$, Λ is diagonal matrix, $\Lambda_{ii} = t_i$.

then $\log\det(I + hX^{-1}V) = \log\det(I + h\Lambda)$

$$= \log \prod_{i=1}^{n} (1 + ht_i)$$

$$= \log(1 + h\,\mathrm{tr}(\Lambda) + o(h))$$

$$= h\,\mathrm{tr}(\Lambda) + o(h)$$

$$= h\,\mathrm{tr}(X^{-1}V) + o(h)$$

$$= \mathrm{tr}(X^{-1}\Delta) + o(h)$$

So $g(X)$ has first-order approximation

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

hence $dg(X)/dX = X^{-1}$

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

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

(d) Obviously, S_n, Σ are S.P.D.

Let $\frac{d\ln(\mu,\Sigma)}{d\Sigma} = -\frac{n}{2}\frac{d\,\mathrm{tr}(S_n\Sigma^{-1})}{d\Sigma} - \frac{n}{2}\frac{dg(\Sigma)}{d\Sigma} = 0$

then $\frac{d\,\mathrm{tr}(S_n\Sigma^{-1})}{d\Sigma} + \frac{dg(\Sigma)}{d\Sigma} = -\Sigma^{-1}S_n\Sigma^{-1} + \Sigma^{-1} = 0.$

$$\Rightarrow -\Sigma^{-1}S_n + I = 0$$

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

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

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 (a) $I(\mu) = \frac{1}{2}\sum_{i=1}^{p}(y_i-\mu_i)^2 + \frac{\lambda}{2}\mu_i^2$, Ridge regression is to $\min_{\mu} I(\mu)$

$$\frac{\partial I(\mu)}{\partial \mu_i} = (\mu_i - y_i) + \lambda\mu_i = 0$$

$$\Rightarrow \hat{\mu}_i^{ridge} = \frac{1}{1+\lambda} y_i \quad \text{so } \hat{\mu}_n = \frac{1}{1+\lambda} y, \quad \mathbb{E}\hat{\mu}_n = \frac{1}{1+\lambda}\mathbb{E}y = \frac{1}{1+\lambda}\mu$$

The risk is $\mathbb{E}\|\hat{\mu}_n - \mu\|^2 = \mathbb{E}\|\hat{\mu}_n - \mathbb{E}\hat{\mu}_n\|^2 + \|\mathbb{E}\hat{\mu}_n - \mu\|^2$

$$= \mathbb{E}\left\|\frac{1}{1+\lambda}y - \frac{1}{1+\lambda}\mu\right\|^2 + \left\|\frac{1}{1+\lambda}\mu - \mu\right\|^2$$

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

$$= \frac{1}{(1+\lambda)^2}\,tr(\Sigma) + \left(\frac{\lambda}{1+\lambda}\right)^2 \mu^T\mu$$

$$= \frac{p}{(1+\lambda)^2} + \frac{\lambda^2}{(1+\lambda)^2}\mu^T\mu$$

(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.

(b) To solve $\min_{\mu} J(\mu) = \frac{1}{2}\|y-\mu\|_2^2 + \lambda\|\mu\|_1$,

we compute the subgradient of $J(\mu)$:

$\partial_\mu J(\mu) = \mu - y + \lambda\,\text{sign}(\mu)$, where $\text{sign}(\mu)\begin{cases} = 1, & \mu > 0 \\ \in [-1,1], & \mu = 0 \\ = -1, & \mu < 0 \end{cases}$, and $0 \in \partial_\mu J(\mu)$

if $\hat{\mu}_i > 0$, since $0 = \hat{\mu}_i - y_i + \lambda$, then $\hat{\mu}_i = y_i - \lambda > 0$, $y_i > \lambda > 0$

so $\hat{\mu}_i = \text{sign}(y_i)(|y_i|-\lambda)_+$

if $\hat{\mu}_i < 0$, since $0 = \hat{\mu}_i - y_i - \lambda$, then $\hat{\mu}_i = y_i + \lambda < 0$, $y_i < -\lambda < 0$

so $\hat{\mu}_i = -(-y_i - \lambda) = \text{sign}(y_i)(|y_i|-\lambda)_+$

if $\hat{\mu}_i = 0$, then $\hat{\mu}_i = 0 \in y_i + \lambda[-1,1]$, so $y_i \in [-\lambda, \lambda]$, $(|y_i|-\lambda)_+ = 0$

In conclusion, $\hat{\mu}_i^{soft} = \mu_{soft}(y_i;\lambda) = sign(y_i)\,(|y_i|-\lambda)_+$

If we choose $\lambda = \sqrt{2\log p}$,

By Stein's unbiased risk estimate, $\hat{\mu}^{soft} = y + g(y)$,

so $g_i(y) = \begin{cases} -\lambda, & y_i > \lambda \\ -y_i, & |y_i| \le \lambda \\ \lambda, & y_i < -\lambda \end{cases}$, $\partial_i g_i(y) = -I(|y_i| \le \lambda)$,

the risk $\mathbb{E}\,\|\hat{\mu}^{soft}(y) - \mu\|^2 = \mathbb{E}\,(p - 2\sum_{i=1}^{p} I(|y_i| \le \lambda) + \sum_{i=1}^{p} y_i^2 \wedge \lambda^2$

$(\frac{1}{2}a\wedge b \le \frac{ab}{a+b} \le a\wedge b)$ $\le 1 + (2\log p + 1)\sum_{i=1}^{p} \mu_i^2 \wedge 1$

the risk of MLE is p.

If $\mathbb{E}\,(\sum_{i=1}^{p} y_i^2 \wedge \lambda^2) \le 2\mathbb{E}\,(\sum_{i=1}^{p} I(|y_i| \le \lambda))$,

then the risk of LASSO is smaller than that 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?

(c) To solve $\min_{\mu} J_0(\mu) = \frac{1}{2}\|y-\mu\|_2^2 + \lambda\|\mu\|_0 = \begin{cases} \frac{1}{2}\|y-\mu\|^2 + \lambda, & \mu > 0 \\ \frac{1}{2}\|y-\mu\|^2 - \lambda, & \mu < 0 \end{cases}$

we compute the gradient of $J(\mu)$:

1) $\mu^* > 0$, if $y > \lambda$, then $\mu^* = y$

2) $\mu^* < 0$, if $y < -\lambda$, then $\mu^* = y$

otherwise, $\mu^* = 0$

$\hat{\mu}_i^{hard} = \mu_i^* = \begin{cases} y_i, & |y| > \lambda \\ 0, & |y| < \lambda \end{cases} = y_i\, I(|y_i| > \lambda)$.

$\hat{\mu}^{hard} = (1 - g(y))\,y$, then $g(y) = 1 - I(|y| > \lambda)$

$= I(|y| < \lambda)$

$g(y)$ is not weakly differentiable,

because g 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 $\mu \in \mathbb{R}^p$.

(d) $\hat{\mu}^{JS}(y) = (1 - \frac{\alpha}{\|y\|^2})y = y + g(y) \Rightarrow g(y) = -\frac{\alpha}{\|y\|^2} y$, g is weakly differentiable

$$\nabla^T g(y) = -\sum_i \frac{\partial}{\partial y_i}\left(\frac{\alpha}{\|y\|^2} y\right) = -\alpha \sum_i \left(\frac{1}{\|y\|^2} - \frac{2y_i^2}{\|y\|^4}\right)$$

$$= -\alpha\left(\frac{p}{\|y\|^2} - \frac{2\|y\|^2}{\|y\|^4}\right) = \frac{-2\alpha(p-2)}{\|y\|^2}.$$

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

so $U_\alpha(y) = p - \frac{2\alpha(p-2)}{\|y\|^2} + \frac{\alpha^2}{\|y\|^2} = p + 2\nabla^T g(y) + \|g(y)\|^2$

By Stein's unbiased risk estimate,

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

Let $\frac{\partial U_\alpha(y)}{\partial \alpha} = \frac{-2(p-2) + 2\alpha}{\|y\|^2} = 0$, then $\alpha^* = \arg\min_\alpha (U_\alpha(y)) = p-2$.

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

For $p > 2$, $\mathbb{E}\|\hat{\mu}^{JS}(y) - \mu\|^2 = p - \mathbb{E}\frac{(p-2)^2}{\|y\|^2} < p = R(\hat{\mu}^{MLE}, \mu)$

(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?

(e) For (a), $\theta_\lambda(t) = \frac{1}{1+\lambda} t$, $0 \leq \theta_\lambda(|t|) \leq |t|$; $\theta_\lambda(-t) = -\theta_\lambda(t)$;

$\theta_\lambda(t) \leq \theta_\lambda(t')$ for $t \leq t'$; $\lim_{t\to+\infty} \theta_\lambda(t) = \infty$.

For (b), $\theta_\lambda(t) = \text{sign}(t)(|t| - \lambda)_+$, $0 \leq \theta_\lambda(|t|) \leq |t| - \lambda \leq |t|$; $\theta_\lambda(-t) = -\theta_\lambda(t)$;

$\theta_\lambda(t) \leq \theta_\lambda(t')$ for $t \leq t'$; $\lim_{t\to+\infty} \theta_\lambda(t) = \infty$

For (c), $\theta_\lambda(t) = t\, I(|t| > \lambda)$, $0 \leq \theta_\lambda(|t|) \leq |t|$; $\theta_\lambda(-t) = -\theta_\lambda(t)$;

$\theta_\lambda(t) \leq \theta_\lambda(t')$ for $t \leq t'$; $\lim_{t\to+\infty} \theta_\lambda(t) = \infty$

For (d), $\theta_\lambda(t) = (1 - \frac{\alpha}{t^2})t$. $\theta_\lambda'(t) = 1 + \frac{\alpha}{t^2} > 0$. $0 \leq \theta_\lambda(|t|) \leq |t|$; $\theta_\lambda(-t) = -\theta_\lambda(t)$;

$\theta_\lambda(t) \leq \theta_\lambda(t')$ for $t \leq t'$; $\lim_{t\to+\infty} \theta_\lambda(t) = \infty$

So (a) (b) (c) (d) are all 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) $0 \le \rho_i(C) \le 1$ (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 $p = 2$.

Proof. In order to show the necessasity of (a)(b)(c),
we show that if the conditions fail, then there exists a dominating estimator.

(a). Let $\widetilde{A} = (A^TA)^{\frac{1}{2}}$, then $tr(A) \le tr\widetilde{A}$, with equity only if $A^T = A$.
Let $D = I - \widetilde{(I-C)}$, then $D^T = D$.

Since $(I-D)^T(I-D) = (\widetilde{I-C})^2 = (I-C)^T(I-C)$,
the two estimators have the same squared bias.
For the variance terms, $tr\, D^TD = tr\, I - 2tr(I-D) + tr(I-D)^T(I-D)$

$$\le tr\, I - 2tr(I-C) + tr\,(I-C)^T(I-C)$$
$$= tr\,(C^TC).$$

If C is not symmetric, $tr\, D^TD < tr(C^TC)$.
From the variance-bias decomposition, we know $\hat{\mu}_C$ is not admissible,
it can be dominated by $\hat{\mu}_D$. So (a) is necessary.

(b) Now assume C is symmetric, consider the eigenvalue decomposition $C = U\Lambda U^T$,
where $\Lambda = diag(\lambda_i)$, $UU^T = U^TU = I_p$. $\lambda_i = \rho_i(C)$.

Let $\eta = U^T\mu$, $x = U^Ty \sim N(\eta, \sigma^2 I_p)$

Since $\mathbb{E}\|Cy - \mu\|^2 = \mathbb{E}\|\Lambda - \eta\|^2$,
we have $R(\hat{\mu}_C, \mu) = R(\hat{\eta}_\Lambda, \eta) = \sum_{i=1}^{p} \sigma^2\lambda_i^2 + (1-\lambda_i)^2\eta_i^2 = \sum_{i=1}^{p} R(\lambda_i, \eta_i)$

If $\exists\, \lambda_i \notin [0,1]$, replace λ_i by 1 if $\lambda_i > 1$ and by 0 if $\lambda_i < 0$,
then the new estimator dominates $\hat{\mu}_C$.

So (b) is also necessary

(c) Suppose $\lambda_1 = \cdots = \lambda_d = 1 > \lambda_i$, $i > d \ge 3$. Let $x^d = (x_1, \cdots, x_d)$
Note that the positive part of J-S estimator is everywhere better than $\hat{\eta}_I(x^d) = x^d$.
Define a new estimator $\hat{\eta}$ to use $\hat{\eta}^{JS}$ on x^d, use $\lambda_i x_i$ for $i > d$

then $R(\hat{\eta}, \eta) = R(\hat{\eta}^{JS}, \eta^d) + \sum_{i>d} r(\lambda_i, \eta_i) < r(\Lambda, \eta)$

so $\hat{\eta}$ dominates $\hat{\eta}_\Lambda$ and hence $\hat{\theta}_C$.