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;

(a) $f(x) = \frac{1}{|2\pi|^{p/2}|\Sigma|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\}$.

$\log(f(x)) = \sum_{i=1}^{n}\left[-p/2\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right]$

i.e. $l_n(\mu,\Sigma) = -\frac{np}{2}\log(2\pi) - n/2\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)$

By cyclicity, $(x_i-\mu)^T\Sigma^{-1}(x_i-\mu) = \text{trace}(\Sigma^{-1}(x_i-\mu)(x_i-\mu)^T) = n\cdot\text{trace}(\Sigma^{-1}S_n)$

Thus, $l_n(\mu,\Sigma) = -n/2\,\text{trace}(\Sigma^{-1}S_n) - n/2\log|\Sigma| + C$.

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

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

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

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

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

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

Thus, $\frac{df(X)}{dX} = -X^{-1}AX^{-1}$.

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

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

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

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

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

$= \log\det(X) + \sum_{i=1}^{n}\log(1+\lambda_i)$, where λ_i are eigenvalues of $X^{-\frac{1}{2}}\Delta X^{-\frac{1}{2}}$.

Since Δ is small, so are its eigenvalues λ_i's,

thus $\log(1+\lambda_i) \sim \lambda_i$, and

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

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

Hence. $\frac{d\, g(X)}{dX} = X^{-1}$

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

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

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`

(d). $\frac{\partial \ln(\mu, \Sigma)}{\partial \Sigma} = \frac{\partial}{\partial \Sigma}\left(-n/2\ \text{trace}(\Sigma^{-1} S_n) - n/2 \cdot \log\det(\Sigma)\right)$

$$= -n/2 \cdot (-\Sigma^{-1} S_n \Sigma^{-1}) - n/2\, \Sigma^{-1} = 0$$

Thus, $\hat{\Sigma} = S_n$

2. Shrinkage : Suppose $y \sim N(\mu, I_p)$

(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 $C = I$.

(a) let $R(\mu) = \frac{1}{2}(y-\mu)^T(y-\mu) + \lambda/2\, \mu^T \mu$

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

$0 = \frac{\partial R}{\partial \mu} = -y + (1+\lambda)\mu$ gives $\mu = \frac{1}{1+\lambda} y$

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

$\hat{\mu} = \frac{1}{1+\lambda} I\, y$ where $C = \frac{1}{1+\lambda} I.$

$MSE = Var(\hat{\mu}) + bias^2(\hat{\mu})$, where $bias^2(\hat{\mu}) = \|(I - (1+\lambda)^{-1} I)\mu\| = 0$

$Var(\hat{\mu}) = \sigma^2 tr(C^T C) = p\sigma^2/(1+\lambda)^2 = p/(1+\lambda)^2$.

(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) Let $R(\mu) = \frac{1}{2}(y-\mu)^T(y-\mu) + \lambda\|\mu\|_1$.

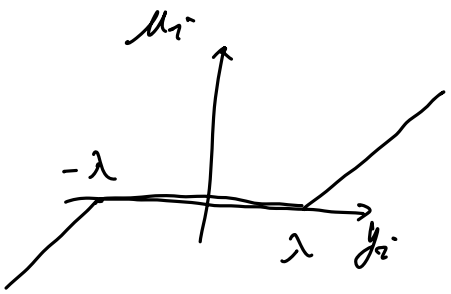
let $0 = \frac{\partial R(\mu)}{\partial \mu_i} = -y_i + \mu_i + \lambda \cdot sgn(\mu_i)$

① $\mu_i > 0$, $\mu_i + \lambda = y_i$

② $\mu_i = 0$, $y_i = \mu_i = 0$

③ $\mu_i < 0$, $\mu_i = y_i + \lambda$

i.e.



So $\hat{\mu}_i^{soft} = sgn(y_i) \cdot (|y_i| - \lambda)_+$

Let $r(\hat{\mu}_i, \mu_i) = E\|\hat{\mu}_i - \mu_i\|^2 = \int \|\hat{\mu}_i - \mu_i\|^2 \phi(z)dz$, where $z \sim N(0,1)$

$$\|\hat{\mu}_i - \mu_i\|^2 = \begin{cases} (y_i - \lambda - \mu_i)^2, & y_i > \lambda \\ \mu_i^2, & y_i = \lambda \\ (y_i + \lambda - \mu_i)^2, & y_i < \lambda \end{cases} \xrightarrow{y_i - \mu_i = z} \begin{cases} (z-\lambda)^2, & z > \lambda - \mu_i \\ \mu^2, & z = \lambda - \mu_i \\ (z+\lambda)^2, & z < \lambda - \mu_i \end{cases}$$

Let $\mu_i \to \infty$. $r(\hat{\mu}_i, \mu_i) = 1 + \lambda^2$.

$\frac{\partial}{\partial \mu_i} r(\hat{\mu}_i, \mu_i) = 2\mu \cdot P(|\mu + z| \le \lambda) \le 2\mu$, thus $r(\hat{\mu}_i, \mu_i) \le 1 + \lambda^2$.

On the other hand,

$r(\hat{\mu}_i, \mu_i) - r(\hat{\mu}_i, 0) \le \int \min(1+\lambda^2, \mu_i^2)\phi(z)dz = \min(1+\lambda^2, \mu_i^2) \le (1+\lambda^2)\min(\mu_i^2, 1)$

and $r(\hat{\mu}_i, 0) \le 2\lambda^{-1}\phi(\lambda) \le e^{-\lambda^2/2} \le 1$

Thus. $r(\hat{\mu}, \mu) = \prod_{i=1}^{p} r(\hat{\mu}_i, \mu_i) \le \prod_{i=1}^{p} (1 + (1+\lambda^2)\min(\mu_i^2, 1))$

$\le 1 + \sum_{i=1}^{p} (1+\lambda^2)\min(\mu_i^2, 1)$

Let $\lambda = \sqrt{2\log p}$. then $r(\hat{\mu}, \mu) \le 1 + (2\log p + 1)\sum_{i=1}^{p} \min(\mu_i^2, 1)$.

The condition of $r(\hat{\mu}^{soft}) < r(\hat{\mu}^{MLE})$ is

$1 + (2\log p + 1)\sum_{i=1}^{p} \min(\mu_i^2, 1) < p.$

(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?

(c) Let $R(\mu) = \|y-\mu\|_2^2 + \lambda^2\|\mu\|_0 = (y-\mu)^T(y-\mu) + \lambda^2 \left(\sum_{i=1}^{p} I\{\mu_i \neq 0\}\right)$

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

Let $0 = \frac{\partial R(\mu_i)}{\partial \mu_i} = \begin{cases} 2(y_i - \mu_i), & \mu_i \neq 0 \\ 0, & \mu_i = 0. \end{cases}$

When $y_i^2 < (y_i - \mu_i)^2 + \lambda^2$ is always true. $\Delta = 4y_i^2 - 4\lambda^2 < 0$, or $|y_i| < \lambda$

this time, set $\mu_i = 0$. Otherwise, let $\mu_i = y_i$.

Thus, $\hat{\mu}_i^{hard} = y_i \cdot I\{|y_i| > \lambda\}$.

$g(y) = 1 - I\{|y| > \lambda\} = I\{|y| \le \lambda\}$. is not weakly 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 $\mu \in \mathbb{R}^p$.

(d) $\hat{\mu}^{JS}(y) = (1 - \alpha/\|y\|^2)y$, $g(y) = -\alpha/\|y\|^2 \cdot y$

Thus $U(y) = p + 2\nabla^T g(y) + \|g(y)\|^2$, and

$$\nabla g^T(y) = \sum_{i=1}^{p} \frac{\partial}{\partial y_i} g(y) = -\alpha(p-2)/\|y\|^2$$

$$\|g(y)\|^2 = \|-\alpha/\|y\|^2\|^2 \cdot \|y\|^2 = \alpha^2/\|y\|^2 .$$

Thus, $U_\alpha(y) = p - (2\alpha(p-2) - \alpha^2)/\|y\|^2$. Risk $R(\hat{\mu}^{JS}, \mu) = \mathbb{E}\, U_\alpha(y)$.

Let $0 = \frac{\partial U_\alpha(y)}{\partial \alpha} = \frac{1}{\|y\|^2}(2(p-2) - 2\alpha)$, $\alpha^* = p-2$

Risk $R(\hat{\mu}^{JS}, \mu) = \mathbb{E}\, U_\alpha(y) = \mathbb{E}\,(p - (p-2)^2/\|y\|^2) < p = R(\hat{\mu}^{MLE}, \mu)$, $p > 2$.

(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). All of the four estimators are shrinkage rules.

We can check these conditions by drawing the picture of $\mu(y) - y$.

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. $\Leftarrow$ If the condition fails, we show a dominating estimator.

If (1) fails:

Lemma: $|A| = (A^TA)^{\frac{1}{2}}$ and $tr(A) \le tr(|A|)$, with equality only if $A = A^T$.

Let $I - D = |I - C|$, D is symmetric.

Since $(I-D)^T(I-D) = |I-C|^2 = (I-C)^T(I-C)$, so $\hat{\theta}_D$ and $\hat{\theta}_C$ has the same bias.

Write $tr\, D^TD = tr\, I - 2tr(I-D) + tr(I-D)^T(I-D)$.

and $tr\, D^TD < tr\, C^TC$ iff $tr(I-D) = tr|I-C| > tr(I-C)$, this occurs only if C is not symmetric. So C is inadmissible.

If (2) fails:

Now C is symmetric, let $C = U\Lambda U^T$ with U orthogonal and $\Lambda = diag(\lambda_i)$

Let $\eta = U^T\mu$, $x = U^Ty \sim N(\eta, \sigma^2 I_p)$. and $E\|Cy - \mu\|^2 = E\|\Lambda - \eta\|^2$.

We have $r(\hat{\mu}_C, \mu) = r(\hat{\eta}_\Lambda, \eta) = \sum_i \sigma^2\lambda_i^2 + (1-\lambda_i)^2\eta_i^2 = \sum_i r(\lambda_i, \eta_i)$

If $\lambda_i \notin [0,1]$, a strictly better MSE results by replacing λ_i by 1 if $\lambda_i > 1$ and by 0 if $\lambda_i < 0$.

If (3) fails:

Suppose that $\lambda_1 = \cdots = \lambda_d = 1 > \lambda_i$ for $i > d \ge 3$. Let $x^d = (x_1, \ldots, x_d)$

Noted that the positive part James-Stein estimator is everybetter than $\hat{\eta}(x^d) = x^d$

So define $\hat{\eta}^{JS}$ on x^d and use $\lambda_i x_i$ for $i > d$, denote by $\hat{\eta}$.

$r(\hat{\eta}, \eta) = r(\hat{\eta}^{JS}, \eta^d) + \sum_{i>d} r(\lambda_i, \eta_i) < r(\Lambda, \eta)$. Thus, $\hat{\eta}$ domain $\hat{\eta}_\Lambda$, hence, $\hat{\eta}_C$.

Thus, condition (i)-(iii) implied $\hat{\mu}_C$ is admissible.

($\Rightarrow$) is similar to ($\Leftarrow$), just from conclusion to condition.

As for MLE, $C = Z_p$. is symmetric, $\lambda_i = 1$.

As for $p = 1, 2$, at most two $\lambda_i = 1$. So $\hat{\mu}_{MLE}$ is admissible.