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;

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

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

$\Rightarrow \mathrm{trace}(\log f(x)) = tr(-\frac{np}{2}\log 2\pi) - \frac{n}{2}\log\det(\Sigma) - \frac{1}{2}\sum_{i=1}^{n} tr[(x_i - \mu)^T \Sigma^{-1} (x_i - \mu)]$

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

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

$S_n = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T$

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

Proof: $f(X+\Delta) = tr[A(X+\Delta)^{-1}] = tr\{A[(I + \Delta X^{-1})X]^{-1}\}$

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

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

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

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

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

$= f(x) - tr(X^{-1}A'X^{-1}\Delta)$

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

proof: $g(X+\Delta) = \log\det(X+\Delta)$

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

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

$= \log\det X + \sum_{i=1}^{n}\log(1+\lambda_i)$

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

$= \log\det X + \mathrm{tr}(X^{-1}\Delta)$

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

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

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

proof: $l(x) = -\frac{n}{2}\mathrm{tr}(\Sigma^{-1}S_n) - \frac{n}{2}\log[\det(\Sigma)] + C$

$\frac{\partial l(x)}{\partial \Sigma} = -\frac{n}{2}(-\Sigma^{-1}S_n\Sigma^{-1}) - \frac{n}{2}\Sigma^{-1} = 0$

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

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 $C = I$.

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

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

Let $\frac{\partial RSS(\mu)}{\partial \mu} = -\frac{1}{2}y + \frac{1+\lambda}{2}\mu = 0$

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

Def of Risk on $\hat{\mu}$ $\quad R(\mu, \hat{\mu}) = L(\mu, \hat{\mu})$

when $L(\mu, \hat{\mu}) = \|\mu - \hat{\mu}\|^2$

MSE $Risk(\mu, \hat{\mu}) = Var(\hat{\mu}) + Bias^2(\hat{\mu})$

since $\hat{\mu}^{ridge} = \hat{\mu}_c = CY$ with $C = X(X^TX + \lambda I)^{-1}X^T$ where $X = I$

so $Bias(\hat{\mu}_c) = \|(I-C)\mu\|^2 = 0$

$Var(\hat{\mu}_c) = \sigma^2 tr(C^TC) = 1 \cdot p = p$

$Risk(\hat{\mu}^{ridge}) = p$

(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.

Proof: $\min_{\mu} J(\mu) = \frac{1}{2}\|y-\mu\|^2 + \lambda\|\mu\|_1$

$$\partial\|x\|_1 = sign(x) = \begin{cases} 1 & x > 0 \\ [-1, 1] & x = 0 \\ -1 & x < 0 \end{cases}$$

$\Rightarrow \partial J(u)|_{u^*} = u^* - y + \lambda \, \text{sign}(u^*)$

if $u^* > 0$ $\quad u^* - y + \lambda = 0$ $\quad y > \lambda$

$u^* = y - \lambda > 0$

if $u^* < 0$ $\quad u^* - y - \lambda = 0$

$u^* = y + \lambda$ $\quad y + \lambda < 0$

if $u^* = 0$ $\quad u^* \in y + \lambda$

$y + * = 0$ $\quad * \in [-\lambda, \lambda]$

$$\Rightarrow u^* = \begin{cases} y - \lambda & y > \lambda \\ 0 & -\lambda \le y \le \lambda \\ y + \lambda & y < -\lambda \end{cases}$$

$u^{soft} = \text{sign}|y| \, (|y| - \lambda)^+$

then we found the upperbound of the risk.

suppose $\hat{u}_\lambda(y) = y + g_s(y)$ $\quad g_s(y)$ is for soft thresholding

Consider $Y = u + z \sim N(u, 1)$

the risk function $r_s(\lambda, u) = \int [\hat{u}_\lambda(u+z) - u]^2 \phi(z) \, dz$

$r_s(\lambda, 0) \le 2\lambda^{-1}\phi(\lambda) \le e^{-\lambda^2/2}$ $\quad r_s(\lambda, \infty) = 1 + \lambda^2$

$$\frac{\partial r_s(\lambda, u)}{\partial u} = 2u P(|u+z| \le \lambda) \le 2u$$

$\therefore r_s(\lambda, u) - r_s(\lambda, 0) \le u^2$

$r_s(\lambda, u) \le r_s(\lambda, 0) + \min(u^2, 1+\lambda^2)$ $\quad$ Let $\lambda = \sqrt{2 \log P}$

$r_s(\lambda, u) \le r_s(\lambda, 0) + (2\log P + 1) \min(u^2, 1)$

when $Y \sim N(u, \varepsilon^2 I)$

$$\text{Risk}(\hat{u}, u) \le \varepsilon^2 + (2\log P + 1) \sum_{i=1}^{P} \min(u_i^2, \varepsilon^2)$$

for $\varepsilon = 1$

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

the condition for $R(\hat{\mu}^{soft}) < R(\hat{\mu}^{MLE})$

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

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

Proof. $R_{ss}(\mu) = (y-\mu)^T(y-\mu) + \lambda^2\left(\sum_{i=1}^{p} \mathbb{I}(\mu_i \neq 0)\right)$

$= y^Ty - y^T\mu - \mu^Ty + \mu^T\mu + \lambda^2\left(\sum_{i=1}^{p} \mathbb{I}(\mu_i \neq 0)\right)$

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

$y_i^2 < (y_i - \mu_i)^2 + \lambda^2$ for all μ_i

$\mu_i^2 - 2y_i\mu_i + \lambda^2 > 0$ for all μ_i

$\Leftrightarrow \Delta = 4y_i - 4\lambda^2 < 0 \Leftrightarrow y_i^2 < \lambda^2$

so when $y_i^2 < \lambda^2$ we adopt $\mu_i = 0$, otherwise $\mu_i = y_i$

$\hat{\mu}^{hard}(y_i) = (1 - g(y_i))\, y_i = (1 - \mathbb{I}(|y_i| \le \lambda))\, y_i = y_i \mathbb{I}(|y_i| > \lambda)$

$g(y) = \mathbb{I}(|y| \le \lambda)$ is not weakly differentiable because g is not absolutely continuous.