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;

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

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

$l(x) = \text{trace}(\log f(x)) = \text{tr}\left(-\frac{np}{2}\log(2\pi)\right) - \frac{n}{2}\log\det(\Sigma)$

$- \frac{1}{2}\sum_{i=1}^{n}\text{trace}\left((x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right)$ cyclicity

$= -\frac{1}{2}\sum_{i=1}^{n}\text{trace}\left(\Sigma^{-1}(x_i-\mu)(x_i-\mu)^T\right)$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

where λ_i is the i-th eigenvalue of $X^{-1/2}\Delta X^{-1/2}$

Since Δ is small λ_i's are small

thus $\log(1+\lambda_i) \approx \lambda_i$

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

$= \log\det(X) + \text{trace}(X^{-1/2}\Delta X^{-1/2})$

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

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

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

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

$l(X) = -\frac{n}{2}\text{trace}(\Sigma^{-1}S_n) - \frac{n}{2}\log(\det(\Sigma)) + C$

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

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

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 $C = I$.

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

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

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

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

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

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

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

$Var(\hat{\mu}_C) = \sigma^2 \, trace(C^TC) = 1 \cdot p = p$

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

(b) Consider the LASSO problem,

$$\min_{\mu} \frac{1}{2}\|y - \mu\|_2^2 + \lambda\|\mu\|_1.$$

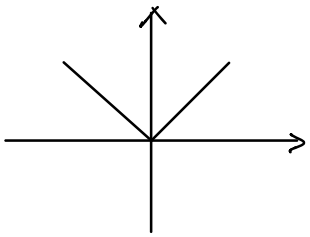
Show that the solution is given by Soft-Thresholding

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

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.



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

for $\|\mu\|_1$ refers to $\|x\|_1$,

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

thus $\partial J(\mu)/\partial \mu|_{\mu^*} = \mu^* - y + \lambda \operatorname{sign}(\mu^*)$

$0 \in \partial J(\mu)$ discuss

If $\mu^* > 0$ $\quad \mu^* - y + \lambda = 0$ $\quad y > \lambda$

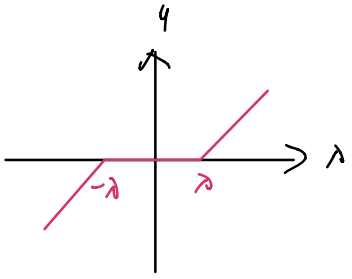
$\mu^* = y - \lambda > 0$

If $\mu^* < 0$ $\quad \mu^* - y - \lambda = 0$ $\quad y + \lambda < 0$

$\mu^* = y + \lambda$

If $\mu^* = 0$ $\quad \mu^* \in y + \lambda[-1, 1]$

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



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

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

Then we found the upperbound of the risk.

suppose $\hat{\mu}_\lambda(y) = y + g_S(y)$ where $g_S(y)$ for soft thresholding

Consider $y = \mu + z \sim N(\mu, 1)$

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

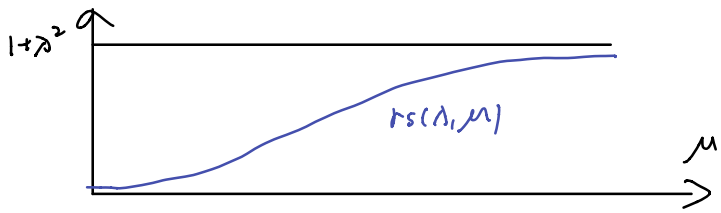
$[\hat{\mu}_\lambda(\mu+z) - \mu]^2 = \begin{cases} (z+\lambda)^2 & \mu < -\lambda - z \\ \mu^2 & \text{otherwise.} \\ (z-\lambda)^2 & \mu > \lambda - z \end{cases}$

$r_S(\lambda, 0) \le 2\lambda^{-1}\phi(\lambda) \le e^{-\lambda^2/2}$

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

$\frac{\partial r_S(\lambda, \mu)}{\partial \mu} = \mu^2 \le (\lambda^2 + 1)$

Thus.



$RSS(\lambda,\mu) - RSS(\lambda,0) \le \mu^2$

$RSS(\lambda,\mu) \le RSS(\lambda,0) + \min(\mu^2, 1+\lambda^2)$ choose $\lambda = \sqrt{2\log p}$

$RSS(\lambda,\mu) \le RSS(\lambda,0) + (2\log p+1)\min(\mu^2,1)$

when $Y \sim N(\mu, \sigma^2 I)$

$$\text{Risk}(\hat{\mu},\mu) \le \sigma^2 + (2\log p+1)\sum_{i=1}^{p}\min(\mu_i^2,\sigma^2)$$

for $\sigma=1$ we get the conclusion for this exercise.

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

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

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

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

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

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

for μ_i $\forall i$

$$RSS(\mu_i) = (\eta_i-\mu_i)^2 + \lambda^2 I(\mu_i \neq 0)$$

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

when is $\eta_i^2 < (\eta_i-\mu_i)^2+\lambda^2$ is always true?

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

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

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

$\Leftrightarrow \quad y_i^2 < \lambda^2$

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.

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

$\hat{\mu} \triangleq \hat{\mu}^{TS}(y) = (1 - \frac{\alpha}{\|y\|^2})\, y = y - \frac{\alpha}{\|y\|^2}\, y$

$g(y) = -\frac{\alpha}{\|y\|^2}\, y$ satisfies.

1. g is weakly differentiable
2. $\sum_{i=1}^{p} \int \left| \frac{\partial g_i(y)}{\partial y_i} \right| dy < \infty$

Thus $U(Y) \triangleq p + 2\nabla^T g(y) + \|g(y)\|^2$ we have

Risk $\mathbb{E}\|\hat{\mu}^{TS}(y) - \mu\|^2 = \mathbb{E}U_\alpha(y)$

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

$\|g(y)\|^2 = \|-\frac{\alpha}{\|y\|^2}\, y\|^2 = \alpha^2 \frac{1}{\|y\|^2}$

Thus $U(y) = p - (2\alpha(p-2) - \alpha^2)/\|y\|^2$

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

$\Rightarrow$ the optimal $\alpha^* = \arg\min_\alpha U(y) = p-2$

So $a(y) = p - (2(p-2)^2 - (p-2)^2)/\|y\|^2$

$= p - (p-2)^2/\|y\|^2$

Aim:

$\text{Risk}(\hat{\mu}^{JS}, \mu) = p - E\frac{(p-2)^2}{\|Y\|^2} < p = \text{Risk}(\hat{\mu}^{MLE}, \mu)$

Find the upper bound of the risk of Jame-Stein.

estimator $\|y\|^2 \sim \chi^2(\|\mu\|^2, p)$

$T(a,b)$ $\quad f(x; a, b) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$

$\chi^2(k)$ $\quad f(x; k) = \frac{1}{(\frac{1}{2})^{\frac{k}{2}} \Gamma(\frac{k}{2})} x^{\frac{k}{2}-1} e^{-x}$

$\chi^2(k) \sim T(\frac{k}{2}, \frac{1}{2})$

$\|y\|^2 \sim \chi^2(\|\mu\|^2, p) \sim T(\frac{\|\mu\|^2}{2}, \frac{p}{2}) \sim T(0, \frac{p}{2} + N)$

$N \sim \text{Poisson}(\frac{\|\mu\|^2}{2})$

a.k.a. $Y \sim \chi^2(0, p+2N)$

So, $E(\frac{1}{\|Y\|^2}) = E_N E_Y[\frac{1}{\|Y\|^2} | N] = E(\frac{1}{p+2N-2})$

Jensen's ineq $\geq \frac{1}{p + 2E N - 2}$

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

Thus for $p \geq 3$

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

(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?

(a) plot the graph of $\mu(y)$

we see clearly that all estimators above Ridge. LASSO, l_0-Regularizer. James-Stein. satisfy Shrinkage, odd, monotone, unbounded

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 $p = 2$.

Lemma 2.1 Let C be any matrix and let P be any orthogonal matrix. Then:

$P^T C P$ admissible iff Cy is admissible

Let $L(\psi(\theta), \delta(y))$ be the loss for estimating $\psi(\theta)$ by $\delta(y)$

then Risk $\rho(\psi, \delta; \theta) = E_\theta(L(\psi(\theta), \delta(y)))$

suppose P is orthogonal if ψ_1, δ_1 given

define $\psi_2(\theta) = \psi_1(P\theta)$ $\delta_2(y) = \delta_1(Py)$

$$\begin{aligned}\rho(\psi_2, \delta_2, \theta) &= E_\theta L(\psi_1(P\theta), \delta_1(Py)) \\ &= \rho(\psi_1, \delta_1; P(\theta)) \\ &= \rho(\psi_1, \delta_1; \theta)\end{aligned}$$

Thus. ψ_2 is admissible for δ_2

iff ψ_1 is admissible for δ_1

Set $\psi_1(\theta) = \theta \quad \delta_1(y) = Cy$

So Cpy is admissible for $P\theta$ iff

Cy is admissible for θ

$L(p\psi_1, P\delta) = L(\psi, \delta)$

$\Rightarrow P'CPy$ is admissible for θ iff

CPy is admissible for $P\theta$

$\Rightarrow P'CPy$ is admissible for θ iff

Cy is admissible for θ

$C = P^T D P \quad D = diag(d_1\ d_2 \cdots)$

Dy is the unique Bayes solution w.r.t. prior distribution

$$dg(\theta) = (2\pi)^{-p/2} \prod_{i=1}^{p} \lambda_i^{\frac{1}{2}} e^{-\lambda_i \theta_i^2/2} \prod_{i=1}^{p} d\theta_i$$

where $\lambda_i = \frac{(1-d_i)}{d_i}$

Following replacement leads to better estimator

$d_i > 1 \rightarrow d_i = 1$

$d_i < 0 \rightarrow -d_i$

three $d_i = 1 \rightarrow \left(1 - (k-2) / \sum_{i=1}^{p} y_i^2\right)$

thus, it only remains to prove Dy is admissible

if one or two $d_i = 1$ while others $0 \le d_i < 1$

Let. $d_1 = d_2 = 1 \quad 0 \le d_i < 1 \quad i = 3, 4 \cdots$

Suppose such a Dy is inadmissible

Then, there exists an estimate $H'(y) = (h_1(y), \cdots h_p(y))$

s.t. $H(y)$ is better than δ_y

$$R(\theta, H(y)) \le R(\theta, \delta_y)$$

$$\sum_{i=1}^{p} \int_{\mathbb{R}^p} (h_i(y) - \theta_i)^2 d\Phi(y; \theta, I) \le 2 + \sum_{i=3}^{p} (d_i^2 + \theta_i^2 (d_i - 1)^2)$$

Multiply Both sides by

$$\varphi(\theta) = \prod_{i=3}^{p} (\lambda_i / 2\pi)^{\frac{1}{2}} e^{-\lambda_i \theta_i^2 / 2}$$

$$\Rightarrow \int_{\mathbb{R}^p} [(h_1(y) - \theta_1)^2 + (h_2(y) - \theta_2)^2] d\Phi(y; \mu, \Sigma) \le 2$$

where $\mu' = (\theta_1, \theta_2, 0 \cdots 0)$ $\Sigma = \text{diag}(1, 1, \frac{\lambda_3 + 1}{\lambda_3}, \cdots)$

Thus, δ_y is admissible $\Rightarrow$ c_y is admissible ∎