Homework 3

1. (a) The joint probability density function is

$$f_X(x_1, \cdots x_k) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

The likelihood function

$$L(\mu, \Sigma) = \prod_{j=1}^{n} f_X(x_j, \mu, \Sigma)$$

$$= \prod_{j=1}^{n} \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} \exp\left(-\frac{1}{2}(x_j-\mu)^T \Sigma^{-1} (x_j-\mu)\right)$$

$$= (2\pi)^{-\frac{np}{2}} |\Sigma|^{-\frac{n}{2}} \exp\left(-\frac{1}{2}(x_j-\mu)^T \Sigma^{-1} (x_j-\mu)\right)$$

$$\ln(\mu, \Sigma) = \ln(L(\mu, \Sigma))$$

$$= -\frac{np}{2}\ln(2\pi) - \frac{n}{2}\ln\det(\Sigma) - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^T \Sigma^{-1}(x_i-\mu)$$

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

where $S_n = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T$ and C does not depend on μ and Σ.

(b)

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

$$= \operatorname{trace}(A(X(I+X^{-1}\Delta)^{-1})$$

$$\approx \operatorname{trace}(AX^{-1}[I-X^{-1}\Delta])$$

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

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

$$\frac{df(X)}{dX} = \lim_{\Delta \to 0} \frac{f(X+\Delta) - f(X)}{\Delta} = -X^{-1}AX^{-1}$$

(c) $$\log\det(X+\Delta) = \log\det\left(X^{\frac{1}{2}}(I + X^{-\frac{1}{2}}\Delta X^{-\frac{1}{2}})X^{\frac{1}{2}}\right)$$
$$= \log\det X + \log\det(I + X^{-\frac{1}{2}}\Delta X^{-\frac{1}{2}})$$
$$= \log\det X + \sum_{i=1}^{n} \log(1+\lambda_i)$$

where λ_i is the ith eigenvalue of $X^{-\frac{1}{2}}\Delta X^{-\frac{1}{2}}$. Since Δ is small, which implies λ_i is small. $\log(1+\lambda_i) \approx \lambda_i$.

$$\log\det(X+\Delta) \approx \log\det X + \sum_{i=1} \lambda_i$$
$$= \log\det X + \text{trace}(X^{-\frac{1}{2}}\Delta X^{-\frac{1}{2}})$$
$$= \log\det X + \text{trace}(X^{-1}\Delta)$$
$$\frac{dg(X)}{dX} = \lim_{\Delta \to 0} \frac{g(X+\Delta) - g(X)}{\Delta} = X^{-1}$$

(d) $$ln(\mu, \Sigma) = -\frac{n}{2}\text{trace}[\Sigma^{-1}S_n] - \frac{n}{2}\log\det(\Sigma) + C$$
$$\frac{\partial ln}{\partial \mu} = \nabla_\mu\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)\Sigma^{-1}(x_i-\mu)\right)$$
$$= \sum_{i=1}^{n}\Sigma^{-1}(x_i-\mu)$$
$$= \Sigma^{-1}\sum_{i=1}^{n}(x_i-\mu)$$

which is equal to 0 only if

$$\sum_{i=1}^{n} x_i - n\mu = 0 \Rightarrow \mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\frac{\partial l_n}{\partial \Sigma} = \nabla_\Sigma \left[-\frac{n}{2} \operatorname{trace}(\Sigma^{-1} S_n) - \frac{n}{2} \log\det(\Sigma) + C \right)$$

$$= \frac{n}{2} \Sigma^T - \frac{1}{2} \nabla_\Sigma \left[\sum_{i=1}^{n} \operatorname{trace}\left(\Sigma^{-1} S_n\right) \right)$$

$$= \frac{n}{2} \Sigma^T - \frac{1}{2} \nabla_\Sigma \left[\operatorname{trace}\left(\sum_{i=1}^{n} (x_i - u)(x_i - u)^T \right) \Sigma^{-1} \right)$$

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

$$= \frac{n}{2} \Sigma^T - \frac{S_n}{2}$$

Set the equation to 0, we get

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

2. (a) Let $L(u) = \frac{1}{2}\|y - u\|_2^2 + \frac{\lambda}{2}\|u\|_2^2$

$$\frac{\partial L}{\partial u} = (u - y) + \lambda u$$

Set the equation to 0. we get

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

$$MSE(\lambda) = E\left\{ (W_\lambda \hat{\beta} - \beta)^T (W_\lambda \hat{\beta} - \beta) \right\}$$

$$= \sigma^2 \operatorname{trace}\left\{ W_\lambda (X^T X)^{-1} W_\lambda^T \right\} + \beta^T (W_\lambda - I)^T (W_\lambda - I) \beta$$

Since $X^T X = I = (X^T X)^{-1}$

$$MSE(\hat{u}_i^{ridge}) = \frac{p}{(1+\lambda)^2} + \frac{\lambda^2}{(1+\lambda)^2} y_i^T y_i$$

(b) $$\min_u \frac{1}{2}\|y-u\|_2^2 + \lambda\|u\|_1$$

Expanding out the first term we get $\frac{1}{2}y^Ty - y^Tu + \frac{1}{2}u^Tu$ and since u^Tu does not contain y, we can consider another equivalent problem

$$\min_u \left(-y^Tu + \frac{1}{2}\|u\|^2\right) + \lambda\|u\|_1.$$

Noting that $\hat{u}^{LS} = y$, the previous problem can be written as

$$\min_u \sum_{i=1}^{p} -\hat{u}_i^{LS}u_i + \frac{1}{2}u_i^2 + \lambda|u_i|.$$

Fix a certain i. Then, we want to minimize

$$L_i = -\hat{u}_i^{LS}u_i + \frac{1}{2}u_i^2 + \lambda|u_i|.$$

If $\hat{u}_i^{LS} > 0$, then we must have $u_i \geq 0$ since otherwise we could flip its sign and get a lower value for the objective function. Likewise if $\hat{u}_i^{LS} < 0$, then we must choose $u_i \leq 0$.

Case 1. $\hat{u}_i^{LS} > 0$. Since $u_i \geq 0$,

$$L_i = -\hat{u}_i^{LS}u_i + \frac{1}{2}u_i^2 + \lambda u_i,$$

and differentiating this with respect to u_i and setting equal to 0, we get $u_i = \hat{u}_i^{LS} - \lambda$ and this is only feasible if the right-hand side is nonnegative, the actual solution is

$$\hat{u}_i^{lasso} = \text{sign}(y_i)(|y_i| - \lambda)^+.$$

Case 2. $\hat{u}_i^{LS} \leq 0$. This implies we must have $u_i \leq 0$ and so

$$L_i = -\hat{u}_i^{LS}u_i + \frac{1}{2}u_i^2 - \lambda u_i.$$

Differentiating with respect to u_i and setting equal to 0, $u_i = \hat{u}_i^{LS} + \lambda = \text{sign}(y_i)(|y_i| - \lambda)$. To ensure this is feasible, we need $u_i \leq 0$, which is achieved by taking

$$\hat{u}_i^{lasso} = \text{sign}(y_i)(|y_i| - \lambda)^+.$$

Thus $\hat{u}_i^{soft} = u_{soft}(y_i;\lambda) = \text{sign}(y_i)(|y_i| - \lambda)_+$.

The risk function

$$r_S(\lambda, u) = \int [\hat{u}_\lambda(u+z) - u]^2 \phi(z)\, dz.$$

Differentiation under the expectation yields

$$0 \le \frac{\partial}{\partial u} r_S(\lambda, u) = 2u\, P(|u+z| \le \lambda) \le 2u$$

the risk at $u=0$ has a simple form

$$r_S(\lambda, 0) = 2\int_\lambda^\infty (z-\lambda)^2 \phi(z)\, dz$$

$$= 2(\lambda^2+1)\bar{\Phi}(\lambda) - 2\lambda\phi(\lambda)$$

and, using the bound for Mills ratio $\bar{\Phi}(\lambda) \le \lambda^{-1}\phi(\lambda)$

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

We have $r_S(\lambda, u) - r_S(\lambda, 0) \le u^2$. Using the bound of ∞, we get

$$r_S(\lambda, u) \le r_S(\lambda, 0) + \min(u^2, 1+\lambda^2).$$

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

$$r_S(\lambda, u) \le 1 + (2\log p + 1)\sum_{i=1}^p \min(u_i^2, 1).$$

(c) The problem can be written as

$$\min_u \|y-u\|_2^2 + \lambda^2\|u\|_0 = \min_u \frac{1}{2}\sum_{i=1}^n (u_i - y_i)^2 + \lambda \mathbb{1}_{\{u_i \neq 0\}}$$

Since it has component wise property, we can solve the scalar problem,

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

In case $u_i = 0$ the cost is $\frac{1}{2}y_i^2$. Hence as long $\frac{1}{2}y_i^2 \le \lambda$ it is worth to set $u_i = 0$. In the other case the minimum value

of the cost function is λ where $u_i = y_i$.

Hence

$$\hat{u}_i^{hard} = u_{hard}(y_i, \lambda) = y_i I(|y_i| > \lambda).$$

$$\hat{u}^{hard}(y) = \begin{cases} 0 & \text{if } |y| \le \lambda \\ y & \text{if } |y| > \lambda \end{cases}$$

$$g(y) = \begin{cases} 1 & \text{if } |y| \le \lambda \\ 0 & \text{if } |y| > \lambda \end{cases}$$

$g(y)$ is weakly differentiable, since $g(y)$ is not differentiable but integrable.

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

$g(y) = -\alpha \|y\|^{-2} y$ is weakly differentiable.

$$Dig_i(y) = -\alpha\left(\frac{1}{\|y\|^2} - \frac{2y_i^2}{\|y\|^4}\right)$$

$\nabla^T g(y) = -\alpha^2 \|y\|^{-2}$ and the unbiased risk estimator

$$U(x) = n - \alpha^2 \|x\|^{-2}$$

Consequently, $r(\hat{u}^{JS}, u) = n - (n-2)^2 E_u \|X\|^{-2}$

The estimator $\hat{u}(y) = (1 - \frac{\beta}{\|x\|^2}) x$ has unbiased risk estimate $U_\beta(x) = n - \frac{2\beta(n-2) - \beta^2}{\|x\|^2}$, the quantity is minimized for each x by the choice $\beta = n-2$.

Thus $E\|\hat{u}^{JS}(y) - u\|^2 = E\, U_\alpha(y)$

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

$\alpha^* = 2$

We begin with the identity

$$\sum_{i=1}^{N}(\hat{u}_i - u_i)^2 = \sum_{i=1}^{N}[(z_i-\hat{u}_i)^2 - (z_i-u_i)^2 + 2(\hat{u}_i-u_i)(z_i-u_i)]$$

Taking expectations on both sides

$$E[\|\hat{u}-u\|^2] = E(\|z-\hat{u}\|^2) - N + 2\sum_{i=1}^{N} Cov(\hat{u}_i, z_i)$$

Using integral by part, we can get

$$Cov(\hat{u}_i, z_i) = E\left(\frac{\partial \hat{u}_i}{\partial z_i}\right)$$

$$E[\|\hat{u}-u\|^2] = E[\|z-\hat{u}\|^2] - N + 2\sum_{i=1}^{N} E\left[\frac{\partial \hat{u}_i}{\partial z_i}\right]$$

Plug in $\hat{u}^{(MLE)} = z$

$$E[\|\hat{u}^{(MLE)} - u\|^2] = E[\|z-z\|^2] - N + 2N = N$$

Plug in $\hat{u}^{(JS)} = z - \frac{n-2}{\|z\|^2} z$

$$\frac{\partial \hat{u}_i^{(JS)}}{\partial z_i} = 1 - \frac{n-2}{\|z\|^2} + \frac{2(n-2)z_i^2}{\|z\|^4}$$

$$\sum_{i=1}^{N} E\left(\frac{\partial \hat{u}_i^{(JS)}}{\partial z_i}\right) = E\left[\sum_{i=1}^{N} \frac{\partial \hat{u}_i^{(JS)}}{\partial z_i}\right]$$

$$= n - E\left[\frac{(n-2)^2}{\|z\|^2}\right].$$

Note that $E[\|z-\hat{u}^{(JS)}\|^2] = E\left(\left\|\frac{n-2}{\|z\|^2} z\right\|^2\right)$

$$E[\|\hat{u}^{(JS)} - u\|^2] = E[\|z-\hat{u}^{(JS)}\|^2] - N + 2\sum_{i=1}^{N} E\left[\frac{\partial \hat{u}_i^{(JS)}}{\partial z_i}\right]$$

$$= n - E\left[\frac{(n-2)^2}{\|z\|^2}\right]$$

Therefore, $E[\|\hat{u}^{(JS)} - u\|^2] < E(\|\hat{u}^{(MLE)} - u\|^2)$ where $n > 2$.

(e) Ridge regression, LASSO regression and the James-Stein Estimator are shrinkage rules.

3. (a) We use the notation $|A| = (A^TA)^{\frac{1}{2}}$ and the fact that $trA \leq tr|A|$, with equality only if A is symmetric, $A^T = A$. Let D be defined via the identity $I-D = |I-C|$. D is symmetric and we use the variance-bias decomposition to show that the MSE of $\hat{\theta}_D$ is everywhere better than that of $\hat{\theta}_C$ if C is not symmetric. Since

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

the two estimators have the same bias. Turning to the variance terms, write

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

Comparing with the corresponding variance term for $\hat{\theta}_C$, we see that $tr\, D^TD < tr\, C^TC$ iff

$$tr(I-D) = tr|I-C| > tr(I-C)$$

which occurs iff C fails to be symmetric.

(b) We assume C is symmetric, we can find a decomposition $C = VRV^T$ with V orthogonal and $R = diag(\rho_i)$ containing the eigenvalues of C. Now change variables to $\eta = V^T\theta$ and $x = V^Ty \sim N(\eta, \epsilon^2 I)$. Orthogonality of V implies that $E\|(y-\theta)\|^2 = E\|Rx-\eta\|^2$, so we have

$$r(\hat{\theta}_C, \theta) = r(\hat{\eta}_R, \eta) = \sum_i \epsilon^2\rho_i^2 + (1-\rho_i)^2\eta_i^2$$

$$= \sum_i r_C(\rho_i, \eta_i)$$

if any eigenvalue $\rho_i \notin (0,1)$, a strictly better MSE results by replacing ρ_i by 1 if $\rho_i > 1$ and by 0 if $\rho_i < 0$.

(c) Suppose that $\xi_1 = \cdots = \xi_d = 1 > \xi_i$ for $i > d \geq 3$, and let $x^d = (x_1, \cdots x_d)$. We have note that the James-Stein estimator is everywhere better than $\hat{\eta}_I(x^d) = x^d$.
We define a new estimator $\hat{\eta}$ to use $\hat{\eta}^{JS}$ on x^d and to continue to use $\xi_i x_i$ for $i > d$, then

$$r(\hat{\eta}, \eta) = r(\hat{\eta}^{JS}, \eta^d) + \sum_{i>d} r_I(\xi_i x_i, \eta_i) < r(\hat{\eta}_R, \eta),$$

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

4. For $p=1$ case, the J-S estimator has higher risk than the MLE for any value of n. This is known as the Stein's paradox, where the J-S estimator shrinks the estimates towards the origin, leading to a higher risk than the MLE.
For $p=2$ case, the J-S estimator has a lower risk than the MLE for n in a certain range.

Let $X \sim N(n, I)$. If $p \geq 3$, then

$$E\frac{1}{|X|^2} \leq \frac{1}{p-2}\left(\frac{p}{p+\|n\|^2}\right)$$

$$\frac{R(\hat{n}^{JS}, n)}{R(n, X)} = 1 - \frac{(p-2)^2}{p} E\frac{1}{|X|^2}$$

$$\leq 1 - \frac{(p-2)^2}{p}\left(\frac{1}{p-2+\|n\|^2}\right)$$

$$\therefore R(\hat{n}^{JS}, n) \leq p - \frac{(p-2)^2}{p-2+\|n\|^2} \quad \text{since } R(n, X) \geq p$$

5. (a) $\mu_i \sim N(M, A)$ $(i = 1, \cdots, n)$ $\quad p(\theta) = N(M, A)$

$x_i \sim N(\mu_i, 1)$. $\quad p(x|\theta) = N(\mu_i, 1)$

$$\ln p(x, \theta) = \ln p(x|\theta) + \ln p(\theta)$$
$$= \ln N(x_i|\mu_i, 1) + \ln N(\mu_i|M, A)$$
$$= -\frac{1}{2}\ln 2\pi - \frac{(x_i - \mu_i)^2}{2} - \frac{1}{2}\ln 2\pi - \frac{1}{2}\ln A - \frac{\mu_i^2}{2A}$$

Consider the second-order term:

$$-\frac{(x_i - \mu_i)^2}{2} - \frac{\mu_i^2}{2A}$$
$$= -\frac{1}{2}\left[x_i^2 - 2x_i\mu_i + \left(1 + \frac{1}{A}\right)\mu_i^2\right]$$
$$= -\frac{1}{2}\begin{bmatrix} x_i \\ \mu_i \end{bmatrix}^T \begin{pmatrix} 1 & -1 \\ -1 & \left(1 + \frac{1}{A}\right) \end{pmatrix} \begin{bmatrix} x_i \\ \mu_i \end{bmatrix}$$

we get

$$\mathrm{cov}\left[\begin{bmatrix} x_i \\ \mu_i \end{bmatrix}\right] = \begin{pmatrix} 1 & -1 \\ -1 & \left(1 + \frac{1}{A}\right) \end{pmatrix}^{-1} = \begin{pmatrix} A+1 & A \\ A & A \end{pmatrix}$$

we have $E\left[\begin{bmatrix} x_i \\ \mu_i \end{bmatrix}\right] = \begin{pmatrix} M \\ M \end{pmatrix}$

$$p(x) = N(M, A+1)$$

Now we consider the conditional distribution $p(\theta|x)$. pick out all second-order terms of μ_i, where x_i is regarded as a constant, we have

$$-\frac{1}{2}\left[\left(1 + \frac{1}{A}\right)\mu_i^2\right].$$

we can get the variance of $\theta|x$

$$\mathrm{var}(\theta|x) = \frac{A}{1+A}$$

we have $E(\theta(x) = \frac{A}{1+A}x + \frac{M}{1+A}$

The conditional distribution of $\theta|x$ is given by

$$\theta|x \sim N\left(\frac{A}{1+A}x + \frac{M}{1+A}, \frac{A}{1+A}\right)$$

We have $Q = \sum_{i=1}^{N} \frac{x_i}{1+A} \sim \chi^2_N$

Note that $Q \sim \chi^2_N$, $\frac{1}{Q}$ follows Inverse-χ^2 with $df = n$ and $E\left(\frac{1}{Q}\right) = \frac{1}{n-3}$

Therefore, $E\left(\frac{1}{\sum_{i=1}^{N} \frac{x_i^2}{1+A}}\right) = \frac{1}{n-3}$, and we use

$$\hat{M} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \hat{B} = 1 - \frac{n-3}{S} \text{ with } S = \sum_{i=1}^{n} x_i^2$$

This leads to the James-Stein estimator

$$\hat{\mu}_i^{JS} = \bar{x} + \left(1 - \frac{n-3}{S}\right)(x_i - \bar{x}).$$

$$E\|\hat{\mu}^{Bayes} - \mu\|^2 = E\left[\|M + B(x_i - M) - \mu\|^2\right]$$
$$= E\left(\|x - \mu\|^2\right) + nB^2 - nB$$
$$E\left(\|\hat{\mu}^{MLE} - \mu\|^2\right) = E\left(\|x - \mu\|^2\right) - n + 2n = n$$
$$\hat{\mu}^{MLE} = x$$

(6) We assume $\mu \sim g(\cdot)$, where $g(\cdot)$ is an arbitrary PDF.

$$\mu \sim g(\cdot), \quad z|\mu \sim N(\mu, 1).$$

The marginal distribution of z is

$$f(z) = \int_{-\infty}^{\infty} \varphi(z-\mu) g(\mu) d\mu$$

where $\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$

The posterior distribution of μ is

$$g(\mu|z) = \varphi(z-\mu) g(\mu) / f(z).$$

Take derivative on both sizes

$$-\frac{d\psi(\eta)}{d\eta}\exp(-\psi(\eta))\int\exp(\eta x)h_0(x)dx+\exp(-\psi(\eta))\int\exp(\eta x)h_0(x)xdx=0$$

$$-\frac{d\psi(\eta)}{d\eta}+\exp(-\psi(\eta))\int\exp(\eta x)h_0(x)xdx=0$$

and $\exp(-\psi(\eta))\int\exp(\eta x)h_0(x)xdx=\int xh(x)dx=E(x)$

$$\frac{d\psi(\eta)}{d\eta}=E(x)$$

$$g(u|z)=\psi(z-u)g(u)/f(z)$$

$$=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(z-u)^2}{2}\right)g(u)/f(z)$$

$$=[\exp(zu)]\left[\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right)/f(z)\right]\left[\exp\left(-\frac{u^2}{2}\right)g(u)\right]$$

$$=\exp(zu-\psi(z))h_0(u).$$

where $\psi(z)=\log\frac{f(z)}{\frac{1}{\sqrt{2\pi}}\exp(-\frac{z^2}{2})}$ is the cumulant generating function, and $h_0(u)=\exp(-\frac{u^2}{2})g(u)$.

$$E(u|z)=z+\frac{d}{dz}\log f(z).$$