Name: Yuxuan Zhang

NetID: yz885

Date: October 1, 2024

HW1

1. (a) *Proof.* We first derive the log-likelihood:

$$L(\mu, \Sigma) = -\frac{1}{2}\left(n\log(|\Sigma|) + \mathrm{tr}\left(\Sigma^{-1}\sum_{i=1}^{n}(Y_i-\mu)(Y_i-\mu)^T\right)\right)$$
$$= -\frac{n}{2}\log|\Sigma| - \frac{n}{2}\mathrm{tr}(\Sigma^{-1}\hat{\Sigma}) - \frac{1}{2}\mathrm{tr}(\Sigma^{-1}(\bar{Y}-\mu)(\bar{Y}-\mu)^T)$$

where $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y})(Y_i-\bar{Y})^T$ is the sample covariance matrix. To estimate the covariance matrix Σ, we plug in the MLE estimate for $\mu = \bar{Y}$. Notice that then the third term

$$\frac{1}{2}\mathrm{tr}(\Sigma^{-1}(\bar{Y}-\mu)(\bar{Y}-\mu)^T) = 0.$$

Then, we have

$$L(\mu, \Sigma) = -\frac{n}{2}\log|\Sigma| - \frac{n}{2}\mathrm{tr}(\Sigma^{-1}\hat{\Sigma}).$$

Since the covariance can be decomposed as $\Sigma = U\Lambda U^T + \sigma^2 I_p$, we can calculate its determinant and inverse as

$$|\Sigma| = |U\Lambda U^T + \sigma^2 I_p|.$$

Since U is orthogonal,

$$|\Sigma| = |\Lambda + \sigma^2 I_p| \cdot (\sigma^2)^{p-r}.$$

Using block matrix inversion, we have

$$\Sigma^{-1} = U(\Lambda + \sigma^2 I_r)^{-1}U^T + \frac{1}{\sigma^2}(I_p - UU^T).$$

Substituting these into the log-likelihood, we have

$$L(\mu, \Sigma) = -\frac{n}{2}(\log|\lambda + \sigma^2 I_r| + (p-r)\log(\sigma^2)) - \frac{n}{2}\mathrm{tr}(\Sigma^{-1}\hat{\Sigma})$$

where $\mathrm{tr}(\Sigma^{-1}\hat{\Sigma}) = \mathrm{tr}(U(\Lambda + \sigma^2 I_r)^{-1}U^T\hat{\Sigma} + \frac{1}{\sigma^2}\mathrm{tr}((I_p - UU^T)\hat{\Sigma})$. To find the MLE, we need to minimize this expression wrt to U. Notice that the first part

$$\frac{\partial}{\partial U}\mathrm{tr}(U(\Lambda + \sigma^2 I_r)^{-1}U^T\hat{\Sigma}) = 2\hat{\Sigma}(\Lambda + \sigma^2 I_r)^{-1},$$

and the second part

$$\frac{1}{\sigma^2}\frac{\partial}{\partial U}\mathrm{tr}((I_p - UU^T)\hat{\Sigma}) = -\frac{2}{\sigma^2}\hat{\Sigma}U.$$

Setting this equal to zero, we have $(\Lambda + \sigma^2 I_r)^{-1} = \frac{1}{\sigma^2}$. From this, we can see that for U to minimize the log-likelihood, it consists of the leading r eigenvectors of $\hat{\Sigma}$. $\square$

(b) We have shown in the last part that $\hat{U} = [x_1, x_2, \dots, x_r]$, the leading r eigen vectors. We have the log-likelihood

$$L(\hat{\mu}, \hat{U}, \hat{\Lambda}, \hat{\epsilon^2}) \propto \sum \ln(a_k) - \sum (a_k \delta_k + C)$$

where a_k is related to the eigenvalues of the sample covariance matrix, δ_k are the eigenvalues with respect to the matrix U_k, and C is constant. To maximize likelihood, we rearrange the terms to have $a_1 \geq a_2 \geq \cdots \geq a_p$, and $\delta_1 \geq \delta_2 \geq \cdots \geq \delta_p$. For $k \leq r$, taking derivative wrt a_k, we have

$$\frac{\partial l}{\partial a_k} = \frac{1}{a_k} - \delta_k.$$

Setting this to zero, we have $\hat{a}_k = \frac{1}{\delta_k}$. Similarly, for $k > r$, we have $\hat{a}_k = \frac{p-r}{\sum_{k>r} \delta_k}$. In our case, with $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})(Y_i - \bar{Y})^T$, we have $\lambda_i = \delta_i$. Then, the formulas for the maximum estimates are

- $\hat{U} = [X_1, \dots, X_r]$, the leading r eigenvectors.
- $\hat{\Lambda} = \text{diag}\{\lambda_i - \hat{\sigma}^2\}_{i=1}^r$, largest r eigenvalues of $\hat{\Sigma}$.
- $\hat{\sigma}^2 = \frac{1}{p-r} \sum_{i=r+1}^p \lambda_i$.

(c) *Proof.* The log-likelihood is

$$l(Y|X) = -\frac{1}{2\sigma^2} \sum_{i,j} (Y_{ij} - X_{ij})^2 + C$$

where C is constant. We assume the singular value decomposition of $X = U \Lambda V^T$. The log-likelihood can be rewritten as

$$l(Y|X) = -\frac{1}{2\sigma^2} ||Y - X||_F^2 + C$$

where $||\cdot||_F$ is the Frobenius norm. Then, this is equivalent to minimizing $||Y - U\Lambda V^T||_F^2$. The best rank-r approximation to Y in terms of minimizing the Frobenius norm, according to the Echart-Young theorem, is given by $\hat{X} = \hat{U}\hat{\Lambda}\hat{V}^T$, where $\hat{U}$ is the matrix of the first r left singular values of Y, $\hat{\Lambda}$ is the diagonal matrix of the first r singular values of Y, and $\hat{V}$ is the matrix of the first r right singular vectors of Y. □

2. *Proof.* Since A is positive definite, we know that $f(A) = A^{-1}$ is convex. We are also given that $\mathbb{E}(A)$ and $\mathbb{E}(A^{-1})$ exist. Then, by Jensen's inequality, we have

$$\begin{aligned} \mathbb{E}(A^{-1}) &\geq \mathbb{E}(A)^{-1} \\ \mathbb{E}(A^{-1}) - \mathbb{E}(A)^{-1} &\geq 0 \\ u^T \left(\mathbb{E}(A^{-1}) - \mathbb{E}(A)^{-1} \right) u &\geq 0. \end{aligned}$$

We have shown that the matrix $\mathbb{E}(A^{-1}) - \mathbb{E}(A)^{-1}$ is an non-negative definite matrix. We will continue to prove that $\mathbb{E}(A^{-1})_{jj} \geq 1/\left(\mathbb{E}(A_{jj})\right)$. From the proof above, Jensen's inequality holds for the entire matrix. Then, we have

$$\begin{aligned} \mathbb{E}(A^{-1})_{jj} &\geq \left(\mathbb{E}(A)^{-1}\right)_{jj} \\ &\geq \frac{1}{\mathbb{E}(A)_{jj}} \end{aligned}$$

for any index j. This completes the proof. □

If A is symmetric but not positive definite, the same conclusion does not hold. This is because the positive definiteness of A is crucial for ensuring that A^{-1} exists and is also positive definite. If A is not positive definite, it could have zero or negative eigenvalues.

3. We will prove that the four conditions are equivalent.

(a) $\to$ (b) *Proof.* We will use the Stirling's approximation to prove this.

$$\begin{aligned}\mathbb{E}(|X|^k) &= \int_0^\infty P(|X|^k > x)dx \\ &= \int_0^\infty P(|X| > x^{1/k})dx.\end{aligned}$$

Using the statement in part (a), we get

$$\begin{aligned}\mathbb{E}(|X|^k) &= \int_0^\infty P(|X| > x^{1/k})dx \\ &\leq \int_0^\infty 2\exp(-(x/K_1)^{1/(k\theta)})dx \\ &= 2K_1 k\theta\Gamma(k\theta) \\ &= 2K_1\Gamma(k\theta+1) \leq 2K_1(k\theta+1)^{k\theta+1}.\end{aligned}$$

Then, take the $k-$th root on both sides of the inequation, we get

$$\begin{aligned}\mathbb{E}(|X|^k)^{1/k} &\leq (2K_1)^{1/k}(k\theta+1)^{\theta+1/k} \\ ||X||_k &\leq (2K_1)^{1/k}(k\theta+1)^{\theta+1/k}.\end{aligned}$$

Let $K_2 = (2K_1)^{1/k}(k\theta+1)^{1/k}$. Then, we have $||X||_k \leq K_2(k\theta+1)^\theta \leq K_2 k^\theta$. □

(b) $\to$ (c) *Proof.* By Taylor expansion, we have

$$\begin{aligned}\mathbb{E}[\lambda|X|^{1/\theta}] &= \mathbb{E}\left[1 + \sum_{i=1}^\infty \frac{(\lambda^{1/\theta}|X|^{1/\theta})^k}{k!}\right] \\ &= 1 + \sum_{i=1}^\infty \frac{\lambda^{k/\theta}}{k!}\mathbb{E}[|X|^{k/\theta}].\end{aligned}$$

From the last proof, we know that

$$\mathbb{E}[|X|^{k/\theta}] \leq K_2 k^\theta.$$

Substitute this into the Taylor expansion, we have

$$\mathbb{E}\left[\exp(\lambda|X|^{1/\theta})\right] \leq 1 + \sum_{k=1}^\infty \frac{\lambda^k K_2 k^\theta}{k!}.$$

Notice that $\exists K_3 > 0$ such that

$$\sum_{k=1}^\infty \frac{\lambda^k K_2 k^\theta}{k!} \leq K_3 \sum_{i=1}^\infty \frac{(\lambda K_3)^{k/\theta}}{k!}.$$

Thus, we have

$$\mathbb{E}[\exp(\lambda|X|^{1/\theta})] \leq \exp((\lambda K_3)^{1/\theta}).$$

□

(c) → (d) *Proof.* Let $\lambda = \frac{(log2)^\theta}{K_3}$, and let $K_4 = \frac{1}{\lambda}$. Then, we have

$$\mathbb{E}\left[\exp\left(\frac{|X|}{K_4}\right)^{1/\theta}\right] \leq \left(\frac{K_3}{K_4}\right)^{1/\theta} \leq 2.$$

This completes the proof. □

(d) → (a) *Proof.* From (d), there exists some $K_4 > 0$ such that

$$\mathbb{E}\left[\exp\left(\left(\frac{|X|}{K_4}\right)^{1/\theta}\right)\right] \leq 2.$$

We apply Markov's inequality and get

$$P(|X| \geq x) = P\left(\exp\left(\left(\frac{|X|}{K_4}\right)^{1/\theta}\right) \geq \exp\left(\left(\frac{x}{K_4}\right)^{1/\theta}\right)\right)$$

$$P\left(\exp\left(\left(\frac{|X|}{K_4}\right)^{1/\theta}\right) \geq \exp\left(\left(\frac{x}{K_4}\right)^{1/\theta}\right)\right) \leq \frac{E\left[\exp\left(\left(\frac{|X|}{K_4}\right)^{1/\theta}\right)\right]}{\exp\left(\left(\frac{x}{K_4}\right)^{1/\theta}\right)}.$$

From condition (d), we know that the expectation is bounded by 2, so this becomes

$$P(|X| \geq x) \leq \frac{2}{\exp\left(\left(\frac{x}{K_4}\right)^{1/\theta}\right)} = 2\exp\left(-\left(\frac{x}{K_4}\right)^{1/\theta}\right).$$

This completes the proof. □

4. (a) *Proof.* Using the integral representation of expectation, we have

$$\mathbb{E}[\max_{i\leq n}|X_i|] = \int_0^\infty P\left(\max_{i\leq n}|X_i| \geq t\right) dt.$$

We then apply the union bound for the maximum.

$$P(\max_{i\leq n}|X_i| \geq t) \leq \sum_{i=1}^n P(|X_i| \geq t) \leq 2n\exp\left(-\frac{t^2}{K^2}\right).$$

We can split the integral at a convenient threshold $t_0 = K\sqrt{\log n}$. Then, we have

$$\mathbb{E}[\max_{i\leq n}|X_i|] = \int_o^{t_0} P\left(\max_{i\leq n}|X_i| \geq t\right) dt + \int_{t_0}^\infty P\left(\max_{i\leq n}|X_i| \geq t\right) dt.$$

For the first integral, we can bound the probability by 1. That is

$$\int_0^{t_0} P\left(\max_{i\leq n}|X_i| \geq t\right) dt \leq \int_0^{t_0} 1dt = t_0 = K\sqrt{\log n}.$$

For the second integral, we can use the bound for large t. That is

$$\int_{t_0}^\infty 2n\exp\left(-\frac{t^2}{K^2}\right) dt = 2n\int_{\log n}^\infty \exp(-s)\frac{K}{2\sqrt{s}}ds.$$

Notice that this integral converges to a constant, which is independent of n. Combining both parts, we get $\mathbb{E}[\max_{i\leq n}|X_i|] \leq K\sqrt{\log n} + C$. Since both K and C are independent of n, we have

$$\mathbb{E}[\max_{i\leq n}|X_i|] \leq C\sqrt{\log n}$$

for some constant C independent of n. □

(b) *Proof.* Consider the lower bound of the probability $P(|X_i| \geq \sigma\sqrt{\log n})$. We know that

$$P(|X_i| \geq \sigma\sqrt{\log n}) = 1 - \text{erf}(\frac{\sqrt{\log n}}{\sqrt{2}})$$

where $\text{erf}(x)$ is the error function. We can use the inequality $\text{erf}(x) \leq \sqrt{1 - \exp(-4/\pi x^2)}$ to approximate the error function. Using this, we can bound the probability

$$P(|X_i| \geq \sigma\sqrt{\log n}) \geq 1 - \sqrt{1 - n\frac{2}{\pi}}.$$

Now we aim to show that this probability is at least $\frac{9}{n}$ by solving the inequality $1 - \sqrt{1 - n^{\frac{2}{\pi}}} \geq 9/n$.

$$\begin{aligned}
-\sqrt{1 - n^{\frac{9}{\pi}}} &\geq \frac{9}{n} - 1 \\
\sqrt{1 - n^{\frac{9}{\pi}}} &\leq 1 - \frac{9}{n} \\
1 - n^{\frac{9}{\pi}} &\leq 1 - \frac{18}{n} + \frac{81}{n^2} \\
n^{2-\frac{2}{\pi}} &\geq 18n - 81.
\end{aligned}$$

Using technology, we see that this inequality holds for $n \geq 2834.88 \approx 2835$. Combining the bounds, we have

$$\mathbb{E}[Y] \geq 0.999(1 - \frac{1}{e^2})\sigma\sqrt{\log n} - 0.001\sigma.$$

Notice that $0.999(1 - \frac{1}{e^2})\sigma\sqrt{\log n} - 0.001\sigma \geq \frac{1}{\sqrt{\pi\log 2}}\sigma\sqrt{\log n}$ holds for any integer $n > 1$. This completes the proof. □

5. (a) *Proof.* We will prove this by contradiction. Let $\hat{\beta}$, $\hat{\beta}'$ be distinct minimizers of the Lasso problem. Suppose, for the sake of contradiction, that $X\hat{\beta} \neq X\hat{\beta}'$. Since they are both minimizers, they satisfy the following

$$\frac{1}{2n}||Y - X\hat{\beta}||_2^2 + \lambda||\hat{\beta}||_1 = \frac{1}{2n}||Y - X\hat{\beta}'||_2^2 + \lambda||\hat{\beta}'||_1.$$

Define a new vector $\tilde{\beta}$ as a convex combination of $\hat{\beta}$ and $\hat{\beta}'$, $\tilde{\beta} = t\hat{\beta} + (1-t)\hat{\beta}'$ for some $t \in (0, 1)$. Since the Lasso objective is convex, $\tilde{\beta}$ would also minimize the objective function. Then, the prediction term for $\tilde{\beta}$ is

$$X\tilde{\beta} = tX\hat{\beta} + (1 - t)\hat{\beta}'.$$

Notice that this is a convex combination of the predictions $X\hat{\beta}$ and $X\hat{\beta}'$, which means that the prediction from $\tilde{\beta}$ lies between the predictions of $\hat{\beta}$ and $\hat{\beta}'$. This leads to a contradiction because, in this case, we could create a new vector that gives a better fit than either $\tilde{\beta}$ or $\tilde{\beta}'$. Therefore, by contradiction, $X\hat{\beta} = X\hat{\beta}'$. □

(b) *Proof.* The Lasso objective function is

$$L(\beta) = \frac{1}{2n}||Y - X\beta||_2^2 + \lambda||\beta||_1.$$

We take the derivative of this objective function with respect to the $j-$th component of β, denoted β_j.

$$\frac{\partial L(\beta)}{\partial \beta_j} = -\frac{1}{n}X_j^T(Y - X\hat{\beta}) + \lambda I_{\beta_j}$$

where X_j is the j-th column of the design matrix X and I_{β_j} is the subdifferential of the L_1-norm. Specifically, $I_{\beta_j} = 1$ if $\beta_j > 0$, $I_{\beta_j} = -1$ if $\beta_j < 0$, and $I_{\beta_j} \in [-1, 1]$ if $\beta_j = 0$.

- When $\hat{\beta}_j > 0$, the gradient becomes

$$\frac{\partial L(\beta)}{\partial \beta_j} = -\frac{1}{n}X_j^T(Y - X\hat{\beta}) + \lambda = 0.$$

  Solving for λ, we have $\lambda = \frac{1}{n}X_j^T(Y - X\hat{\beta})$.
- When $\hat{\beta}_j < 0$, the gradient becomes

$$\frac{\partial L(\beta)}{\partial \beta_j} = -\frac{1}{n}X_j^T(Y - X\hat{\beta}) - \lambda = 0.$$

  Solving for λ, we have $\lambda = -\frac{1}{n}X_j^T(Y - X\hat{\beta})$.
- When $\hat{\beta}_j = 0$, we have $-\frac{1}{n}X_j^T(Y - X\hat{\beta}) \leq \lambda \leq \frac{1}{n}X_j^T(Y - X\hat{\beta})$, or $\lambda \geq \frac{1}{n}|X_j^T(Y - X\hat{\beta})|$.

□

(c) *Proof.* From the last proof, we have shown that the first derivative of the Lasso objective function is

$$\begin{aligned} -\frac{1}{n}X^T(Y - X\hat{\beta}_\lambda) + \lambda I_\beta &= 0 \\ \frac{1}{n}X^T Y &= \frac{1}{n}X^T X\hat{\beta}_\lambda + \lambda I_\beta. \end{aligned}$$

When $\lambda_\beta = 0$, this becomes

$$\frac{1}{n}X^T Y = \lambda I_\beta.$$

If $\lambda > ||\frac{1}{n}X^T Y||_\infty$, then the right-hand side satisfies the condition for all $I_\beta \in [-1, 1]$. Therefore, the only feasible solution is $\hat{\beta}_\lambda = 0$. □

6. (a) *Proof.* We have the partition of the covariance matrix and precision matrix as follow: $\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}$, $\Theta = \Sigma^{-1} = \begin{pmatrix} \Theta_{aa} & \Theta_{ab} \\ \Theta_{ba} & \Theta_{bb} \end{pmatrix}$. Notice that we can express the inverse of Σ with terms involving Σ_{bb}^{-1} and the inverse of Shur's complement. Specifically, we have

$$\Sigma^{-1} = \Theta = \begin{bmatrix} (\Sigma/\Sigma_{bb})^{-1} & -(\Sigma/\Sigma_{bb})^{-1}\Sigma_{ab}\Sigma_{bb}^{-1} \\ -\Sigma_{bb}^{-1}\Sigma_{ba}(\Sigma/\Sigma_{bb})^{-1} & \Sigma_{bb}^{-1} + \Sigma_{bb}^{-1}\Sigma ba(\Sigma/\Sigma_{bb})^{-1}\Sigma_{ab}\Sigma_{bb}^{-1} \end{bmatrix}.$$

Notice that $\Theta_{aa} = (\Sigma/\Sigma_{bb})^{-1} = (\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1}$. Notice that the terms inside the parenthesis $\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba} = \Sigma_{a.b}$. Therefore, $\Sigma_{a.b} = \Theta_{aa}^{-1}$. This completes the proof. □

(b) *Proof.* We know that diag(Θ) corresponds to the variances. Therefore, multiplying diag$(\Theta)^{-1/2}$ normalizes the diagonal elements of Θ to 1. It follows that

$$\begin{aligned} R_{jk} &= \left(\text{diag}(\Theta)^{-1/2}\Theta\text{diag}(\Theta)^{-1/2}\right)_{jk} \\ &= \frac{\Theta_{jk}}{\sqrt{\Theta_{jj}\Theta_{kk}}}. \end{aligned}$$

We also know that the off-diagonal elements of Θ represent the negative partial correlations conditioned on the remaining variables. That is

$$\rho_{jk}|\text{rest} = -\frac{\Theta_{jk}}{\sqrt{\Theta_{jj}\Theta_{kk}}}.$$

Therefore, $R_{jk} = -\rho_{jk}|\text{rest}$. This completes the proof. □

7. *Proof.* Since $r(X) = r(X_{11})$, we know that $r(X) = r(X_{11}X_{12})$. It follows that the second block of the matrix X, $(X_{21}\ X_{22})$ can be written as a linear transformation of the first block $(X_{11}\ X_{12})$. That is, $\exists Z$ such that

$$(X_{21}\ X_{22}) = Z(X_{11}\ X_{12}).$$

It then follows that $X_{21} = ZX_{11}$. Similarly, since $r(X_{11}) = r(X)$, X_{12} can be written as a linear combination of X_{11}. That is, $\exists U$ such that $X_{12} = X_{11}U$. By the similar reasoning, we can write $\begin{bmatrix} X_{12} \\ X_{22} \end{bmatrix} = \begin{bmatrix} X_{11} \\ X_{21} \end{bmatrix} U$. Then, we know that $X_{22} = X_{21}U = ZX_{11}U$. Then, by the definition of Moore-Penrose pseudoinverse, we know that $X_{11} = X_{11}\dagger X_{11}X_{11}$. Then, we have

$$X_{22} = ZX_{11}U = ZX_{11}\dagger X_{11}X_{11}U = X_{21}\dagger X_{11}X_{12}.$$

This completes the proof. □