

DSO Screening Exam: 2018 In-Class Exam

Gregory Faletto

Exercise 1 (Probability). (a) (i) We have

$$T(y) = \inf\{x \geq 0 : M(x) \geq y\}; \quad \lim_{y \rightarrow \infty} \Pr(e^{-y}T(y) \leq a) = 1 - e^{-\lambda a} \quad \forall a \geq 0.$$

Note that

$$\lim_{y \rightarrow \infty} \Pr(e^{-y}T(y) \leq a) = \lim_{y \rightarrow \infty} \Pr(T(y) \leq ae^y)$$

Because $T(y) = \inf\{x \geq 0 : M(x) \geq y\}$ and $M(\cdot)$ is monotonically increasing, we have the inequality $M(z) \geq y$ for all $z \geq x$. Therefore $T(y) \leq ae^y \iff M(ae^y) \geq y$. Let $z = ae^y \implies y = \log(z/a)$; then we have

$$\lim_{y \rightarrow \infty} \Pr(T(y) \leq ae^y) = \lim_{y \rightarrow \infty} \Pr(M(ae^y) \geq y) = \lim_{z \rightarrow \infty} \Pr(M(z) \geq \log(z) - \log(a))$$

Let $b = -\log a$ to get

$$= \lim_{z \rightarrow \infty} \Pr(M(z) - \log(z) \geq b) = 1 - e^{-\lambda a} \implies \lim_{z \rightarrow \infty} \Pr(M(z) - \log(z) \geq b) = 1 - e^{-\lambda e^{-b}}$$

$$\iff \lim_{z \rightarrow \infty} \Pr(M(z) - \log(z) \leq b) = e^{-\lambda e^{-b}}$$

(ii) The distribution function is $F(x) = e^{-\lambda e^{-x}}$, a Gumbel distribution with parameter λ . This is a proper cdf because $\lim_{x \rightarrow \infty} F(x) = 1$ and $\lim_{x \rightarrow -\infty} F(x) = 0$.

(b) Using hint:

$$\begin{aligned} \int_0^1 t^x dt &= \left[\frac{t^{x+1}}{x+1} \right]_0^1 = \frac{1}{x+1} \\ \iff \mathbb{E} \left[\int_0^1 t^X dt \right] &= \mathbb{E} \left(\frac{1}{X+1} \right) \iff \int_0^1 \mathbb{E}(t^X) dt = \mathbb{E} \left(\frac{1}{X+1} \right) \\ \iff \int_0^1 \sum_{i=0}^n t^i \Pr(X=i) dt &= \mathbb{E} \left(\frac{1}{X+1} \right) \iff \int_0^1 \sum_{i=0}^n t^i \binom{n}{i} p^i (1-p)^{n-i} dt = \mathbb{E} \left(\frac{1}{X+1} \right) \\ \iff (1-p)^n \int_0^1 \sum_{i=0}^n \binom{n}{i} \left(\frac{tp}{1-p} \right)^i dt &= \mathbb{E} \left(\frac{1}{X+1} \right) \\ \iff (1-p)^n \int_0^1 \left(1 + \frac{tp}{1-p} \right)^n dt &= \mathbb{E} \left(\frac{1}{X+1} \right) \\ \iff \int_0^1 (1-p+tp)^n dt &= \mathbb{E} \left(\frac{1}{X+1} \right) \end{aligned}$$

Let $u = 1-p+tp \implies du = p dt$. Then we can write

$$\frac{1}{p} \int_{1-p}^1 u^n du = \mathbb{E} \left(\frac{1}{X+1} \right) \iff \frac{1}{p} \left[\frac{u^{n+1}}{n+1} \right]_{1-p}^1 = \mathbb{E} \left(\frac{1}{X+1} \right) \iff \mathbb{E} \left(\frac{1}{X+1} \right) = \frac{1}{p} \left[\frac{1 - (1-p)^{n+1}}{n+1} \right]$$

Now we consider $\mathbb{E} \left[\frac{1}{X+2} \right]$.

$$\begin{aligned}
 \int_0^1 t^{x+1} dt &= \left[\frac{t^{x+2}}{x+2} \right]_0^1 = \frac{1}{x+2} \\
 \iff \mathbb{E} \left[\int_0^1 t^{X+1} dt \right] &= \mathbb{E} \left(\frac{1}{X+2} \right) \iff \int_0^1 \mathbb{E}(t^{X+1}) dt = \mathbb{E} \left(\frac{1}{X+2} \right) \\
 \iff \int_0^1 \sum_{i=0}^n t^{i+1} \Pr(X=i) dt &= \mathbb{E} \left(\frac{1}{X+2} \right) \iff \int_0^1 \sum_{i=0}^n t^{i+1} \binom{n}{i} p^i (1-p)^{n-i} dt = \mathbb{E} \left(\frac{1}{X+2} \right) \\
 \iff (1-p)^n \int_0^1 t \sum_{i=0}^n \binom{n}{i} \left(\frac{tp}{1-p} \right)^i dt &= \mathbb{E} \left(\frac{1}{X+2} \right) \\
 \iff (1-p)^n \int_0^1 t \left(1 + \frac{tp}{1-p} \right)^n dt &= \mathbb{E} \left(\frac{1}{X+2} \right) \\
 \iff \int_0^1 t (1-p+tp)^n dt &= \mathbb{E} \left(\frac{1}{X+2} \right)
 \end{aligned}$$

Let $u = 1-p+tp \implies du = p dt$ and $t = (u+p-1)/p$. Then we can write

$$\begin{aligned}
 \frac{1}{p^2} \int_{1-p}^1 u^n (u+p-1) du &= \mathbb{E} \left(\frac{1}{X+2} \right) \iff \frac{1}{p^2} \int_{1-p}^1 [u^{n+1} + (p-1)u^n] du = \mathbb{E} \left(\frac{1}{X+2} \right) \\
 \iff \frac{1}{p^2} \left[\frac{u^{n+2}}{n+2} + (p-1) \frac{u^{n+1}}{n+1} \right]_{1-p}^1 &= \mathbb{E} \left(\frac{1}{X+2} \right) \\
 \iff \mathbb{E} \left(\frac{1}{X+2} \right) &= \frac{1}{p^2} \left[\frac{1 - (1-p)^{n+2}}{n+2} - (1-p) \frac{1 - (1-p)^{n+1}}{n+1} \right]
 \end{aligned}$$

Exercise 2 (Mathematical Statistics). (a) Let

$$T_n(X_1, \dots, X_n) := \sum_{i=1}^n X_i.$$

(b) Note that if $T_n(X_1, \dots, X_n) = y$, then $\Pr[(X_1, \dots, X_n) = (x_1, \dots, x_n)] = 0$ unless $\sum_{i=1}^n x_i = y$. Therefore we have

$$\Pr[(X_1, \dots, X_n) = (x_1, \dots, x_n) \mid T_n(X_1, \dots, X_n) = y] = \frac{\Pr[(X_1, \dots, X_n) = (x_1, \dots, x_n)]}{\Pr(T_n(X_1, \dots, X_n) = y)}$$

$$\frac{\prod_{i=1}^n p^{x_i} (1-p)^{1-x_i}}{\binom{n}{y} p^y (1-p)^{n-y}} = \frac{p^{\sum_{i=1}^n x_i} (1-p)^{n-\sum_{i=1}^n x_i}}{\binom{n}{y} p^y (1-p)^{n-y}} = \frac{p^y (1-p)^{n-y}}{\binom{n}{y} p^y (1-p)^{n-y}} = \frac{1}{\binom{n}{y}}$$

which does not depend on p . That is, the distribution of X_1, \dots, X_n conditional on $T_n(X_1, \dots, X_n)$ is independent of p . Therefore $T_n(X_1, \dots, X_n)$ is sufficient for p .

(c) Likelihood function:

$$\begin{aligned}
 \mathcal{L}(p) &= \prod_{i=1}^n p^{X_i} (1-p)^{1-X_i} \\
 \implies \ell(p) &= \sum_{i=1}^n [X_i \log(p) + (1-X_i) \log(1-p)] = \log(p) \sum_{i=1}^n X_i + \log(1-p) \left(n - \sum_{i=1}^n X_i \right) \\
 \implies \frac{d}{dp} \ell(p) &= \frac{1}{p} \sum_{i=1}^n X_i - \frac{1}{1-p} \left(n - \sum_{i=1}^n X_i \right) = 0 \implies \frac{1}{p} \sum_{i=1}^n X_i = \frac{1}{1-p} \left(n - \sum_{i=1}^n X_i \right) \\
 &\iff \sum_{i=1}^n X_i - p \sum_{i=1}^n X_i = pn - p \sum_{i=1}^n X_i \iff \boxed{\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i.}
 \end{aligned}$$

We can find its asymptotic distribution using the Central Limit Theorem:

Theorem 1. Central Limit Theorem (Grimmett and Stirzaker theorem 5.10.4.) Let X_1, X_2, \dots be a sequence of independent identically distributed random variables with finite mean μ and finite non-zero variance σ^2 , and let $S_n = \sum_{i=1}^n X_i$. Then

$$\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \xrightarrow{d} \mathcal{N}(0, 1)$$

Since $\mathbb{E}(X_i) = p$, $\text{Var}(X_i) = p(1-p)$, we have

$$\begin{aligned}
 \frac{\sum_{i=1}^n X_i - np}{\sqrt{np(1-p)}} &\xrightarrow{d} \mathcal{N}(0, 1) \iff \sum_{i=1}^n X_i - np \xrightarrow{d} \mathcal{N}(0, np(1-p)) \\
 &\iff \frac{1}{n} \sum_{i=1}^n X_i - p \xrightarrow{d} \mathcal{N}\left(0, \frac{p(1-p)}{n}\right) \iff \boxed{\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{d} \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)}
 \end{aligned}$$

Exercise 3 (Mathematical Statistics). (a) We have

$$X \mid \mu \sim \mathcal{N}(\mu, \mathbf{I}_n)$$

Let $X = (X_1, \dots, X_n)^T$ and let $\mu = (\mu_1, \dots, \mu_n)^T$. Notice that

$$\begin{aligned}
 \mathbb{E}(\mathbf{X}^T \mathbf{X}) &= \mathbb{E}[\mathbb{E}(\mathbf{X}^T \mathbf{X} \mid \mu)] = \mathbb{E}[\mathbb{E}(X_1^2 + X_2^2 + \dots + X_n^2 \mid \mu)] = \mathbb{E}\left[\sum_{i=1}^n \mathbb{E}(X_i^2 \mid \mu)\right] \\
 &= \mathbb{E}\left[\sum_{i=1}^n \text{Var}(X_i \mid \mu) + \mathbb{E}(X_i \mid \mu)^2\right] = \mathbb{E}\left[\sum_{i=1}^n 1 + \mu_i^2\right] = n + \mathbb{E}\|\mu\|_2^2 \\
 &\implies \mathbb{E}(\mathbf{X}^T \mathbf{X} - n) = \mathbb{E}\|\mu\|_2^2
 \end{aligned}$$

Therefore $\boxed{\hat{\delta}_{\text{unbiased}} = \mathbf{X}^T \mathbf{X} - n}$ is unbiased for $\mathbb{E}\|\mu\|_2^2$ (and given μ it is unbiased for $\|\mu\|_2^2$).

(b) We will begin by finding the posterior distribution of μ . The prior distribution of μ is

$$f(\mu) = (2\pi)^{-n/2} |k\mathbf{I}_n|^{-1/2} \cdot \exp\left(-\frac{1}{2}\mu^T(k\mathbf{I}_n)^{-1}\mu\right) = \frac{1}{\sqrt{(2\pi k)^n}} \exp\left(-\frac{1}{2k}\mu^T\mu\right).$$

The likelihood is

$$\begin{aligned} f_{\mathbf{X}|\mu}(\mathbf{x} | \mu) &= (2\pi)^{-n/2} |\mathbf{I}_n|^{-1/2} \cdot \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^T(\mathbf{I}_n)^{-1}(\mathbf{x} - \mu)\right) \\ &= (2\pi)^{-n/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^T(\mathbf{x} - \mu)\right). \end{aligned}$$

So the unconditional distribution of \mathbf{X} is

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= \int_{\mathbb{R}^n} f_{\mathbf{X}|\mu}(\mathbf{x} | \mu) f(\mu) d\mu \\ &= \int_{\mathbb{R}^n} (2\pi)^{-n/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^T(\mathbf{x} - \mu)\right) \cdot \frac{1}{\sqrt{(2\pi k)^n}} \exp\left(-\frac{1}{2k}\mu^T\mu\right) d\mu \\ &= \frac{1}{(2\pi\sqrt{k})^n} \int_{\mathbb{R}^n} \exp\left(-\frac{1}{2}\left[\mu^T\mu + \frac{1}{k}\mu^T\mu - 2\mathbf{x}^T\mu + \mathbf{x}^T\mathbf{x}\right]\right) d\mu \\ &= \frac{1}{(2\pi\sqrt{k})^n} \int_{\mathbb{R}^n} \exp\left(-\frac{k+1}{2k}\left[\mu^T\mu - \frac{2k}{k+1}\mathbf{x}^T\mu\right] - \frac{1}{2}\mathbf{x}^T\mathbf{x}\right) d\mu \\ &= \frac{1}{(2\pi\sqrt{k})^n} \int_{\mathbb{R}^n} \exp\left(-\frac{k+1}{2k}\left[\mu^T\mu - \frac{2k}{k+1}\mathbf{x}^T\mu + \left(\frac{k}{k+1}\right)^2\mathbf{x}^T\mathbf{x}\right] - \left(-\frac{k+1}{2k}\right)\left(\frac{k}{k+1}\right)^2\mathbf{x}^T\mathbf{x} - \frac{1}{2}\mathbf{x}^T\mathbf{x}\right) d\mu \\ &= \frac{1}{(2\pi\sqrt{k})^n} \int_{\mathbb{R}^n} \exp\left(-\frac{k+1}{2k}\left[\left(\mu - \frac{k}{k+1}\mathbf{x}\right)^T\left(\mu - \frac{k}{k+1}\mathbf{x}\right)\right] - \frac{1}{2}\left[-\left(\frac{k}{k+1}\right)\mathbf{x}^T\mathbf{x} + \frac{k+1}{k+1}\mathbf{x}^T\mathbf{x}\right]\right) d\mu \\ &= \frac{1}{(2\pi\sqrt{k})^n} \int_{\mathbb{R}^n} \exp\left(-\frac{k+1}{2k}\left[\left(\mu - \frac{k}{k+1}\mathbf{x}\right)^T\left(\mu - \frac{k}{k+1}\mathbf{x}\right)\right]\right) \exp\left(-\frac{1}{2}\frac{1}{k+1}\mathbf{x}^T\mathbf{x}\right) d\mu \\ &= \frac{1}{\sqrt{(2\pi k)^n}} \exp\left(-\frac{1}{2}\frac{1}{k+1}\mathbf{x}^T\mathbf{x}\right) \left(\frac{k}{k+1}\right)^{n/2} \\ &\quad \cdot \int_{\mathbb{R}^n} \frac{1}{\sqrt{(2\pi)^n}} \cdot \left(\frac{k}{k+1}\right)^{-n/2} \exp\left(-\frac{1}{2}\left(\mu - \frac{k}{k+1}\mathbf{x}\right)^T\left(\frac{k}{k+1}\mathbf{I}_n\right)^{-1}\left(\mu - \frac{k}{k+1}\mathbf{x}\right)\right) d\mu \end{aligned}$$

The second row is the integral over \mathbb{R}^n of an n -dimensional multivariate Gaussian distribution with mean $k/(k+1)\mathbf{x}$ and covariance $k/(k+1)\mathbf{I}_n$, so it equals 1. Then we are left with

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n}} \frac{1}{k^{n/2}} \exp\left(-\frac{1}{2}\frac{1}{k+1}\mathbf{x}^T\mathbf{x}\right) \left(\frac{k}{k+1}\right)^{n/2}$$

$$= \frac{1}{\sqrt{(2\pi)^n}} [(k+1)^n]^{-1/2} \exp\left(-\frac{1}{2} \mathbf{x}^T ((k+1)\mathbf{I}_n)^{-1} \mathbf{x}\right) \quad (1)$$

which is the density of an n -dimensional multivariate Gaussian random variable with mean $\mathbf{0}$ and covariance $(k+1)\mathbf{I}_n$. Therefore the posterior distribution of $\boldsymbol{\mu}$ is

$$\begin{aligned} f_{\boldsymbol{\mu}|\mathbf{X}}(\boldsymbol{\mu} | \mathbf{x}) &= \frac{f_{\mathbf{X}|\boldsymbol{\mu}}(\mathbf{x} | \boldsymbol{\mu})f(\boldsymbol{\mu})}{f_{\mathbf{X}}(\mathbf{x})} \\ &= \left[\frac{1}{(2\pi\sqrt{k})^n} \exp\left(-\frac{1}{2} \left[\boldsymbol{\mu}^T \boldsymbol{\mu} + \frac{1}{k} \boldsymbol{\mu}^T \boldsymbol{\mu} - 2\mathbf{x}^T \boldsymbol{\mu} + \mathbf{x}^T \mathbf{x} \right] \right) \right] \bigg/ \left[\frac{1}{\sqrt{(2\pi)^n}} [(k+1)^n]^{-1/2} \exp\left(-\frac{1}{2} \frac{1}{k+1} \mathbf{x}^T \mathbf{x}\right) \right] \\ &= \frac{1}{\sqrt{(2\pi)^n}} \left(\frac{k+1}{k} \right)^{n/2} \exp\left(-\frac{k+1}{2k} \left[\left(\boldsymbol{\mu} - \frac{k}{k+1} \mathbf{x} \right)^T \left(\boldsymbol{\mu} - \frac{k}{k+1} \mathbf{x} \right) \right] - \frac{1}{2} \frac{1}{k+1} \mathbf{x}^T \mathbf{x} + \frac{1}{2} \frac{1}{k+1} \mathbf{x}^T \mathbf{x} \right) \\ &= \frac{1}{\sqrt{(2\pi)^n}} \left(\frac{k+1}{k} \right)^{n/2} \exp\left(-\frac{k+1}{2k} \left(\boldsymbol{\mu} - \frac{k}{k+1} \mathbf{x} \right)^T \left(\boldsymbol{\mu} - \frac{k}{k+1} \mathbf{x} \right) \right) \\ &= \frac{1}{\sqrt{(2\pi)^n}} \cdot \left(\frac{k}{k+1} \right)^{-n/2} \exp\left(-\frac{1}{2} \left(\boldsymbol{\mu} - \frac{k}{k+1} \mathbf{x} \right)^T \left(\frac{k}{k+1} \mathbf{I}_n \right)^{-1} \left(\boldsymbol{\mu} - \frac{k}{k+1} \mathbf{x} \right) \right) \end{aligned}$$

which is an n -dimensional multivariate Gaussian distribution with mean $k/(k+1)\mathbf{x}$ and covariance $k/(k+1)\mathbf{I}_n$. That is, conditional on \mathbf{X} , $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)$, where

$$\begin{aligned} \mu_i | \mathbf{X} &\stackrel{i.i.d.}{\sim} \mathcal{N}\left(\frac{k}{k+1}X_i, \frac{k}{k+1}\right). \\ \iff \left(\mu_i - \frac{k}{k+1}X_i\right) \frac{k+1}{k} | \mathbf{X} &\stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1) \iff \frac{k+1}{k}\mu_i - X_i | \mathbf{X} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1) \\ \implies \mathbb{E}\left(\left[\frac{k+1}{k}\mu_i - X_i\right]^2 | \mathbf{X}\right) &= 1 \iff \mathbb{E}\left(\left[\frac{k+1}{k}\right]^2 \mu_i^2 + X_i^2 - 2\frac{k+1}{k}\mu_i X_i | \mathbf{X}\right) = 1 \quad (2) \end{aligned}$$

So,

$$\begin{aligned} \hat{\delta}_{\text{proper}} &= \mathbb{E}(\|\boldsymbol{\mu}\|_2^2 | \mathbf{X}) = \mathbb{E}\left(\sum_{i=1}^n \mu_i^2 | \mathbf{X}\right) = \left[\frac{k}{k+1}\right]^2 \mathbb{E}\left(\sum_{i=1}^n \left[\frac{k+1}{k}\right]^2 \mu_i^2 | \mathbf{X}\right) \\ &= \left[\frac{k}{k+1}\right]^2 \mathbb{E}\left(\sum_{i=1}^n \left[\frac{k+1}{k}\right]^2 \mu_i^2 + X_i^2 - 2\frac{k+1}{k}\mu_i X_i | \mathbf{X}\right) - \left[\frac{k}{k+1}\right]^2 \mathbb{E}\left(\sum_{i=1}^n X_i^2 - 2\frac{k+1}{k}\mu_i X_i | \mathbf{X}\right) \\ &= \left[\frac{k}{k+1}\right]^2 - \left[\frac{k}{k+1}\right]^2 \sum_{i=1}^n X_i^2 + 2\frac{k+1}{k} \left[\frac{k}{k+1}\right]^2 \sum_{i=1}^n X_i \mathbb{E}(\mu_i | \mathbf{X}) \\ &= \left[\frac{k}{k+1}\right]^2 - \left[\frac{k}{k+1}\right]^2 \sum_{i=1}^n X_i^2 + 2\frac{k}{k+1} \sum_{i=1}^n X_i \cdot \frac{k}{k+1} X_i \\ &= \left[\frac{k}{k+1}\right]^2 + 2\left[\frac{k}{k+1}\right]^2 \sum_{i=1}^n X_i^2 - \left[\frac{k}{k+1}\right]^2 \sum_{i=1}^n X_i^2 = \left[\frac{k}{k+1}\right]^2 (1 + \mathbf{X}^T \mathbf{X}). \end{aligned}$$

- (c) We will again begin by finding the posterior distribution of μ . The (improper) prior distribution of μ is constant; that is, for some $c \in \mathbb{R}$,

$$f(\mu) = c, \quad \forall \mu \in \mathbb{R}^n.$$

The likelihood is

$$\begin{aligned} f_{\mathbf{X}|\mu}(\mathbf{x} | \mu) &= (2\pi)^{-n/2} |\mathbf{I}_n|^{-1/2} \cdot \exp \left(-\frac{1}{2} (\mathbf{x} - \mu)^T (\mathbf{I}_n)^{-1} (\mathbf{x} - \mu) \right) \\ &= (2\pi)^{-n/2} \exp \left(-\frac{1}{2} (\mathbf{x} - \mu)^T (\mathbf{x} - \mu) \right). \end{aligned}$$

So the unconditional distribution of \mathbf{X} is

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= \int_{\mathbb{R}^n} f_{\mathbf{X}|\mu}(\mathbf{x} | \mu) f(\mu) d\mu = c \int_{\mathbb{R}^n} (2\pi)^{-n/2} \exp \left(-\frac{1}{2} (\mathbf{x} - \mu)^T (\mathbf{x} - \mu) \right) d\mu \\ &= c \int_{\mathbb{R}^n} (2\pi)^{-n/2} \exp \left(-\frac{1}{2} (\mu - \mathbf{x})^T (\mu - \mathbf{x}) \right) d\mu \end{aligned}$$

The expression inside the integral is the density of a Gaussian random variable with mean \mathbf{x} and covariance \mathbf{I}_n , so the integral evaluates to 1. Therefore the unconditional distribution of \mathbf{X} is also flat. Therefore the posterior distribution of μ is the same as the likelihood:

$$f_{\mu|\mathbf{X}}(\mu | \mathbf{x}) = \frac{f_{\mathbf{X}|\mu}(\mathbf{x} | \mu) f(\mu)}{f_{\mathbf{X}}(\mathbf{x})} = (2\pi)^{-n/2} \exp \left(-\frac{1}{2} (\mu - \mathbf{x})^T (\mu - \mathbf{x}) \right);$$

that is, conditional on \mathbf{X} , μ is normally distributed with mean \mathbf{X} and covariance \mathbf{I}_n . So conditional on \mathbf{X} , $\mu = (\mu_1, \dots, \mu_n)$, where

$$\mu_i | \mathbf{X} \stackrel{i.i.d.}{\sim} \mathcal{N}(X_i, 1).$$

$$\iff \mu_i - X_i | \mathbf{X} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1) \implies \mathbb{E}([\mu_i - X_i]^2 | \mathbf{X}) = 1 \iff \mathbb{E}(\mu_i^2 + X_i^2 - 2\mu_i X_i | \mathbf{X}) = 1 \quad (3)$$

So,

$$\begin{aligned} \hat{\delta}_{\text{flat}} &= \mathbb{E}(\|\mu\|_2^2 | \mathbf{X}) = \mathbb{E} \left(\sum_{i=1}^n \mu_i^2 | \mathbf{X} \right) = \mathbb{E} \left(\sum_{i=1}^n \mu_i^2 + X_i^2 - 2\mu_i X_i | \mathbf{X} \right) - \mathbb{E} \left(\sum_{i=1}^n X_i^2 - 2\mu_i X_i | \mathbf{X} \right) \\ &= 1 - \sum_{i=1}^n X_i^2 + 2 \sum_{i=1}^n X_i \mathbb{E}(\mu_i | \mathbf{X}) = 1 - \sum_{i=1}^n X_i^2 + 2 \sum_{i=1}^n X_i^2 = 1 + \mathbf{X}^T \mathbf{X}. \end{aligned}$$

So,

$$\hat{\delta}_{\text{flat}} - \hat{\delta}_{\text{unbiased}} = 1 + \mathbf{X}^T \mathbf{X} - (\mathbf{X}^T \mathbf{X} - n) = 1 + n. \quad (4)$$

(d) If the true distribution of $\boldsymbol{\mu}$ is the prior from part (b), then the marginal distribution of \mathbf{X} is (1):

$$= \frac{1}{\sqrt{(2\pi)^n}} [(k+1)^n]^{-1/2} \exp \left(-\frac{1}{2} \mathbf{x}^T ((k+1)\mathbf{I}_n)^{-1} \mathbf{x} \right);$$

that is, $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, (k+1)\mathbf{I}_n)$. (Note that this means $(k+1)^{-1}X_i$ is standard Gaussian and i.i.d. for all $i \in \{1, \dots, n\}$.) Per the suggestion, let

$$\begin{aligned} \hat{D} = \hat{\delta}_{\text{proper}} - \hat{\delta}_{\text{unbiased}} &= \left[\frac{k}{k+1} \right]^2 \left(1 + \mathbf{X}^T \mathbf{X} \right) - \left(\mathbf{X}^T \mathbf{X} - n \right) = \frac{k^2 - (k^2 + 2k + 1)}{(k+1)^2} \mathbf{X}^T \mathbf{X} + \left[\frac{k}{k+1} \right]^2 + n \\ &= -\frac{2k+1}{(k+1)^2} \mathbf{X}^T \mathbf{X} + \left[\frac{k}{k+1} \right]^2 + n \end{aligned} \quad (5)$$

Since from (4) we have

$$\mathbb{E} \left(\hat{\delta}_{\text{proper}} - \hat{\delta}_{\text{unbiased}} \right)^2 = n^2 + 2n + 1,$$

we seek

$$\mathbb{E} \left(\hat{\delta}_{\text{flat}} - \hat{\delta}_{\text{unbiased}} \right)^2 = \mathbb{E}(\hat{D}^2) = \text{Var}(\hat{D}) + \left[\mathbb{E}(\hat{D}) \right]^2.$$

Note that since $(k+1)^{-1}X_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$ for all $i \in \{1, \dots, n\}$,

$$\sum_{i=1}^n ((k+1)^{-1}X_i)^2 \sim \chi_n^2,$$

so

$$\mathbb{E} \left[\sum_{i=1}^n \left(\frac{1}{k+1} X_i \right)^2 \right] = n \iff \frac{1}{(k+1)^2} \mathbb{E} \left[\sum_{i=1}^n X_i^2 \right] = n \iff \mathbb{E} \left[\sum_{i=1}^n X_i^2 \right] = n(k+1)^2,$$

and

$$\text{Var} \left[\sum_{i=1}^n \left(\frac{1}{k+1} X_i \right)^2 \right] = 2n \iff \frac{1}{(k+1)^4} \text{Var} \left[\sum_{i=1}^n X_i^2 \right] = 2n \iff \text{Var} \left[\sum_{i=1}^n X_i^2 \right] = 2n(k+1)^4.$$

Under the unconditional distribution of \mathbf{X} ,

$$\begin{aligned} \mathbb{E}(\hat{D}) &= \mathbb{E} \left(-\frac{2k+1}{(k+1)^2} \mathbf{X}^T \mathbf{X} + \left[\frac{k}{k+1} \right]^2 + n \right) = -\frac{2k+1}{(k+1)^2} \mathbb{E}(\mathbf{X}^T \mathbf{X}) + \left[\frac{k}{k+1} \right]^2 + n \\ &= -\frac{2k+1}{(k+1)^2} \mathbb{E} \left(\sum_{i=1}^n X_i^2 \right) + \left[\frac{k}{k+1} \right]^2 + n = -\frac{2k+1}{(k+1)^2} n(k+1)^2 + \left[\frac{k}{k+1} \right]^2 + n \\ &= -n \frac{(2k+1)(k^2 + 2k + 1) + k^2}{(k+1)^2} + n = n \left[\frac{k^2 + 2k + 1}{(k+1)^2} - \frac{2k^3 + 4k^2 + 2k + k^2 + 2k + 1 + k^2}{(k+1)^2} \right] \end{aligned}$$

$$= n \left[\frac{-2k^3 - 5k^2 - 2k}{(k+1)^2} \right] = -n \left[\frac{2k^3 + 5k^2 + 2k}{(k+1)^2} \right]$$

Next,

$$\begin{aligned} \text{Var}(\hat{D}) &= \text{Var} \left(-\frac{2k+1}{(k+1)^2} \mathbf{X}^T \mathbf{X} + \left[\frac{k}{k+1} \right]^2 + n \right) = \frac{(2k+1)^2}{(k+1)^4} \text{Var} \left(\sum_{i=1}^n X_i^2 \right) \\ &= \frac{(2k+1)^2}{(k+1)^4} 2n(k+1)^4 = 2n(2k+1)^2 \end{aligned}$$

So

$$\begin{aligned} \mathbb{E} \left(\hat{\delta}_{\text{flat}} - \hat{\delta}_{\text{unbiased}} \right)^2 &= 2n(2k+1)^2 + \left(-n \left[\frac{2k^3 + 5k^2 + 2k}{(k+1)^2} \right] \right)^2 = n^2 \left[\frac{2k^3 + 5k^2 + 2k}{(k+1)^2} \right]^2 + n \cdot 2(2k+1)^2 \\ &\approx 4k^2 n^2 + 8k^2 n, \end{aligned}$$

so in general when $k \geq 1$ and n is large, $\mathbb{E} \left(\hat{\delta}_{\text{flat}} - \hat{\delta}_{\text{unbiased}} \right)^2 > \mathbb{E} \left(\hat{\delta}_{\text{proper}} - \hat{\delta}_{\text{unbiased}} \right)^2$; that is, the flat prior Bayes estimator is further from the unbiased estimator than the proper prior Bayes estimator.

Exercise 4 (High-Dimensional Statistics). (a) **Sparsity in the covariance matrix does not imply sparsity in the precision matrix.** For a simple example, consider a sequence of p Gaussian random variables generated in the following way:

$$X_1 \sim \mathcal{N}(0, 1), \quad X_i = X_{i-1} + Z, i = 2, \dots, p$$

where $Z \sim \mathcal{N}(0, 1)$. Then for all $i \neq j \in \{1, \dots, p\}$, $\text{Cov}(X_i, X_j) \neq 0$, so the covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$ is dense with no zero entries. However, for any $i \neq j \in \{1, \dots, p\}$, $\text{Cov}(X_i, X_j \mid \mathbf{X}_{-i-j})$ (where \mathbf{X}_{-i-j} contains all of the X_k except X_i and X_j ; that is, $\text{Cov}(X_i, X_j \mid \mathbf{X}_{-i-j})$ is the covariance of X_i and X_j conditional on all of the other X_k) is zero for all j except $i-1$ and $i+1$. Since the precision matrix contains in entry σ_{ij} the covariance of X_i and X_j conditional on all the other $p-2$ variables (see the discussion in part (b)), the precision matrix is sparse, with mostly 0 entries except for the diagonal and a band of nonzero entries on each side of the diagonal.

Of course, since the precision matrix and covariance matrix are inverses, had the covariance matrix been sparse except the diagonal and a band of nonzero entries on each side of the diagonal, in general the precision matrix would be dense.

For a more complicated example, suppose the covariance matrix is the following:

$$\Sigma := \begin{pmatrix} (\tau^2 + 1)\mathbf{I}_n & \mathbf{I}_n & \cdots & \mathbf{I}_n \\ \mathbf{I}_n & (\tau^2 + 1)\mathbf{I}_n & \cdots & \mathbf{I}_n \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{I}_n & \mathbf{I}_n & \cdots & (\tau^2 + 1)\mathbf{I}_n \end{pmatrix},$$

This matrix is relatively sparse. However, its inverse is dense, with every entry nonzero:

$$\Sigma^{-1} = \tau^2 \mathbf{I}_{ns} + \mathbf{1}_s \mathbf{1}_s^T \otimes \mathbf{I}_n = \tau^2 \mathbf{I}_{ns} + (\mathbf{1}_s \otimes \mathbf{I}_n) (\mathbf{1}_s \otimes \mathbf{I}_n)^T$$

Applying the Sherman-Morrison-Woodbury formula with $A = \tau^2 \mathbf{I}_{ns}$, $U = \mathbf{1}_s \otimes \mathbf{I}_n$, $C = \mathbf{I}_n$, and $V = (\mathbf{1}_s \otimes \mathbf{I}_n)^T$ yields

$$\begin{aligned}
\Sigma^{-1} &= \frac{1}{\tau^2} \mathbf{I}_{ns} - \frac{1}{\tau^2} (\mathbf{1}_s \otimes \mathbf{I}_n) \left[\mathbf{I}_n + (\mathbf{1}_s \otimes \mathbf{I}_n)^T \cdot \frac{1}{\tau^2} (\mathbf{1}_s \otimes \mathbf{I}_n) \right]^{-1} (\mathbf{1}_s \otimes \mathbf{I}_n)^T \cdot \frac{1}{\tau^2} \\
&= \frac{1}{\tau^2} \left(\mathbf{I}_{ns} - (\mathbf{1}_s \otimes \mathbf{I}_n) [\tau^2 \mathbf{I}_n + \mathbf{1}_s^T \mathbf{1}_s \otimes \mathbf{I}_n]^{-1} (\mathbf{1}_s \otimes \mathbf{I}_n)^T \right) \\
&= \frac{1}{\tau^2} \left(\mathbf{I}_{ns} - (\mathbf{1}_s \otimes \mathbf{I}_n) [(\tau^2 + s) \mathbf{I}_n]^{-1} (\mathbf{1}_s \otimes \mathbf{I}_n)^T \right) \\
&= \frac{1}{\tau^2} \mathbf{I}_{ns} - \frac{1}{\tau^2(\tau^2 + s)} \mathbf{1}_s \mathbf{1}_s^T \otimes \mathbf{I}_n \\
&= \begin{pmatrix} \left(\frac{1}{\tau^2} - \frac{1}{\tau^2(\tau^2 + s)} \right) \mathbf{I}_n & -\frac{1}{\tau^2(\tau^2 + s)} \mathbf{I}_n & -\frac{1}{\tau^2(\tau^2 + s)} \mathbf{I}_n & \cdots & -\frac{1}{\tau^2(\tau^2 + s)} \mathbf{I}_n \\ -\frac{1}{\tau^2(\tau^2 + s)} \mathbf{I}_n & \left(\frac{1}{\tau^2} - \frac{1}{\tau^2(\tau^2 + s)} \right) \mathbf{I}_n & -\frac{1}{\tau^2(\tau^2 + s)} \mathbf{I}_n & \cdots & -\frac{1}{\tau^2(\tau^2 + s)} \mathbf{I}_n \\ -\frac{1}{\tau^2(\tau^2 + s)} \mathbf{I}_n & -\frac{1}{\tau^2(\tau^2 + s)} \mathbf{I}_n & \left(\frac{1}{\tau^2} - \frac{1}{\tau^2(\tau^2 + s)} \right) \mathbf{I}_n & \cdots & -\frac{1}{\tau^2(\tau^2 + s)} \mathbf{I}_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{\tau^2(\tau^2 + s)} \mathbf{I}_n & -\frac{1}{\tau^2(\tau^2 + s)} \mathbf{I}_n & -\frac{1}{\tau^2(\tau^2 + s)} \mathbf{I}_n & \cdots & \left(\frac{1}{\tau^2} - \frac{1}{\tau^2(\tau^2 + s)} \right) \mathbf{I}_n \end{pmatrix} \\
&= \begin{pmatrix} \frac{\tau^2 + s - 1}{\tau^2(\tau^2 + s)} \mathbf{I}_n & -\frac{1}{\tau^2(\tau^2 + s)} \mathbf{I}_n & -\frac{1}{\tau^2(\tau^2 + s)} \mathbf{I}_n & \cdots & -\frac{1}{\tau^2(\tau^2 + s)} \mathbf{I}_n \\ -\frac{1}{\tau^2(\tau^2 + s)} \mathbf{I}_n & \frac{\tau^2 + s - 1}{\tau^2(\tau^2 + s)} \mathbf{I}_n & -\frac{1}{\tau^2(\tau^2 + s)} \mathbf{I}_n & \cdots & -\frac{1}{\tau^2(\tau^2 + s)} \mathbf{I}_n \\ -\frac{1}{\tau^2(\tau^2 + s)} \mathbf{I}_n & -\frac{1}{\tau^2(\tau^2 + s)} \mathbf{I}_n & \frac{\tau^2 + s - 1}{\tau^2(\tau^2 + s)} \mathbf{I}_n & \cdots & -\frac{1}{\tau^2(\tau^2 + s)} \mathbf{I}_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{\tau^2(\tau^2 + s)} \mathbf{I}_n & -\frac{1}{\tau^2(\tau^2 + s)} \mathbf{I}_n & -\frac{1}{\tau^2(\tau^2 + s)} \mathbf{I}_n & \cdots & \frac{\tau^2 + s - 1}{\tau^2(\tau^2 + s)} \mathbf{I}_n \end{pmatrix}.
\end{aligned}$$

- (b) If $\omega_{jk} = 0$, this means that features X_j and X_k are conditionally independent given all of the other features. (This is in contrast to the meaning of $\sigma_{jk} = 0$, which is that X_j and X_k are unconditionally independent.) This interpretation also gives another answer for part (a) of this question: sparsity in the precision matrix does not necessarily imply sparsity in the covariance matrix (and vice versa) because conditional independence does not necessarily imply unconditional independence (and vice versa).

By the same argument as in part (a), if $\omega_{jk} = 0$ holds it does not necessarily hold that $\sigma_{jk} = 0$, since Σ is the inverse of Ω .

- (c) **consider instead using method from Fan et al. [2008].** I would estimate the precision matrix using the graphical lasso [Friedman et al., 2008]. Let $\hat{\Sigma}$ be an estimate for the covariance matrix Σ . Let S be the empirical covariance matrix; that is,

$$S := \frac{1}{n} \sum_{i=1}^n (X^{(i)} - \bar{X})(X^{(i)} - \bar{X})^T$$

where $X^{(i)}$ is the i th row of X and $\bar{X} = n^{-1} \sum_{i=1}^n X^{(i)}$. In our case, we assume the mean vector is known to be $\mathbf{0}$, so we can instead use

$$S := \frac{1}{n} \sum_{i=1}^n X^{(i)} \left(X^{(i)} \right)^T.$$

To find the function to optimize, we will find the likelihood function. Note that the density for a multivariate p -dimensional Gaussian distribution with known mean $\mathbf{0}$ is

$$f_{\mathbf{X}}(x_1, \dots, x_p) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \cdot \exp \left(-\frac{1}{2} \mathbf{x}^T \Sigma^{-1} \mathbf{x} \right).$$

The likelihood function for n observations from this distribution is given by

$$\begin{aligned} \mathcal{L}(\Sigma^{-1}) &= \prod_{i=1}^n f_{\mathbf{X}_i}(x_{1i}, \dots, x_{pi}) = (2\pi)^{-np/2} |\Sigma|^{-n/2} \cdot \exp \left(-\frac{1}{2} \sum_{i=1}^n \left(\mathbf{X}^{(i)} \right)^T \Sigma^{-1} \mathbf{X}^{(i)} \right) \\ &= (2\pi)^{-np/2} |\Sigma|^{-n/2} \cdot \exp \left(-\frac{1}{2} \sum_{i=1}^n \text{Tr} \left[\left(\mathbf{X}^{(i)} \right)^T \Sigma^{-1} \mathbf{X}^{(i)} \right] \right) \\ &= (2\pi)^{-np/2} |\Sigma|^{-n/2} \cdot \exp \left(-\frac{1}{2} \sum_{i=1}^n \text{Tr} \left[\mathbf{X}^{(i)} \left(\mathbf{X}^{(i)} \right)^T \Sigma^{-1} \right] \right) \\ &= (2\pi)^{-np/2} |\Sigma|^{-n/2} \cdot \exp \left(-\frac{1}{2} \text{Tr} \left[\sum_{i=1}^n \mathbf{X}^{(i)} \left(\mathbf{X}^{(i)} \right)^T \Sigma^{-1} \right] \right) \\ &= (2\pi)^{-np/2} |\Sigma|^{-n/2} \cdot \exp \left(-\frac{1}{2} \text{Tr} [nS\Sigma^{-1}] \right) = (2\pi)^{-np/2} |\Sigma^{-1}|^{n/2} \cdot \exp \left(-\frac{n}{2} \text{Tr} [S\Sigma^{-1}] \right) \end{aligned}$$

Take the logarithm of this expression to get the log likelihood function:

$$\log \mathcal{L}(\Sigma^{-1}) = -\frac{np}{2} \log(2\pi) + \frac{n}{2} \log |\Sigma^{-1}| - \frac{n}{2} \text{Tr} (S\Sigma^{-1})$$

Since we are only concerned with the arguments maximizing the log likelihood function, we can disregard the first constant term and the multiplicative constants, leaving a simpler expression maximized by the same matrix

$$\log \mathcal{L}(\Sigma^{-1}) \propto \log |\Sigma^{-1}| - \text{Tr} (S\Sigma^{-1})$$

Lastly, to impose sparsity we will add an ℓ_1 penalty. We will optimize the ℓ_1 -penalized log likelihood function

$$\hat{\Omega} := \arg \max_{\Omega \in \mathcal{S}^+} \{ \log |\Omega| - \text{Tr}(S\Omega) + \lambda \|\Omega\|_1 \} \quad (6)$$

where \mathcal{S}^+ is the set of nonnegative definite $p \times p$ matrices and $\lambda > 0$ is a penalty parameter. Next, we will discuss the algorithm from [Friedman et al. \[2008\]](#) to optimize this function. We will make use of the following partitions of $\hat{\Sigma}$ and S :

$$\hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_{11} & \hat{\Sigma}_{12} \\ \hat{\Sigma}_{12}^T & \hat{\Sigma}_{22} \end{pmatrix}, \quad S = \begin{pmatrix} S_{11} & S_{12} \\ S_{12}^T & S_{22} \end{pmatrix}, \quad \hat{\Omega} = \begin{pmatrix} \hat{\Omega}_{11} & \hat{\omega}_{12} \\ \hat{\omega}_{12}^T & \hat{\omega}_{22} \end{pmatrix}, \quad (7)$$

as well as the constraint

$$\begin{pmatrix} \hat{\Sigma}_{11} & \hat{\sigma}_{12} \\ \hat{\sigma}_{12}^T & \hat{\sigma}_{22} \end{pmatrix} \begin{pmatrix} \hat{\Omega}_{11} & \hat{\omega}_{12} \\ \hat{\omega}_{12}^T & \hat{\omega}_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{I}_{p-1} & 0 \\ 0^T & 1 \end{pmatrix} \quad (8)$$

suggested by the identity $\Sigma\Omega = \mathbf{I}_p$. The proposed procedure to optimize (6) is the following coordinate descent algorithm:

1. Initialize the algorithm with estimate $\hat{\Sigma} = S + \lambda \mathbf{I}_p$. (The diagonal of $\hat{\Sigma}$ remains unchanged for the rest of the algorithm.)
2. For each $j = 1, 2, \dots, p, 1, 2, \dots, p, \dots$, switch the rows and columns of $\hat{\Sigma}$ so that the row and column corresponding to feature j come last, as in partition (7). Then solve the lasso problem

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^{p-1}} \left\{ \frac{1}{2} \left\| \hat{\Sigma}_{11}^{1/2} \beta - \hat{\Sigma}_{11}^{-1/2} s_{12} \right\|_2^2 + \lambda \|\beta\|_1 \right\} \quad (9)$$

3. Fill in the corresponding row and column of $\hat{\Sigma}$ using $\hat{\sigma}_{12} = \hat{\Sigma}_{11} \hat{\beta}$. (Again, the diagonal term $\hat{\sigma}_{22}$ remains as it was after step 1.)
4. Continue until convergence; that is, until the average absolute change in $\hat{\Sigma}$ is less than $t \cdot \text{ave} |S^{-\text{diag}}|$, where $S^{-\text{diag}}$ are the off-diagonal elements of S and t is a fixed threshold ($t = 0.001$ is recommended by Friedman et al. [2008]).
5. Estimate $\hat{\Omega}$ by using $\hat{\Sigma}$ to compute $\hat{\omega}_{22}$ for each feature and filling in the corresponding row of $\hat{\Omega}$ as in (7) using the formulae

$$\hat{\omega}_{22} = 1 / \left(\hat{\sigma}_{22} - \hat{\sigma}_{12}^T \hat{\beta} \right), \quad \hat{\omega}_{12} = -\hat{\beta} \hat{\omega}_{22}. \quad (10)$$

Remark. Formula (9) is justified as follows: Banerjee et al. [2008] show that $\hat{\Sigma}_{12}$ satisfies

$$\hat{\Sigma}_{12} = \arg \min_y \left\{ y^T \hat{\Sigma}^{-1} y : \|y - S_{12}\|_\infty \leq p \right\}. \quad (11)$$

Using strong duality, they then show that the dual problem (9) is equivalent; specifically, if $\hat{\beta}$ solves (9) then $\hat{\sigma}_{12} = \hat{\Sigma}_{11} \hat{\beta}$ solves (11).

Remark. Formulae (10) are justified as follows: from (8) we have the following identities

$$\hat{\Sigma}_{11} \hat{\omega}_{12} + \hat{\sigma}_{12} \hat{\omega}_{22} = 0, \quad \hat{\sigma}_{12}^T \hat{\omega}_{12} + \hat{\sigma}_{22} \hat{\omega}_{22} = 1.$$

These yield

$$\hat{\omega}_{12} = -\hat{\Sigma}_{11}^{-1} \hat{\sigma}_{12} \hat{\omega}_{22}, \quad \hat{\omega}_{22} = 1 / \left(\hat{\sigma}_{22} - \hat{\sigma}_{12}^T \hat{\Sigma}_{11}^{-1} \hat{\sigma}_{12} \right).$$

Then using $\hat{\sigma}_{12} = \hat{\Sigma}_{11} \hat{\beta} \iff \hat{\beta} = \hat{\Sigma}_{11}^{-1} \hat{\sigma}_{12}$, we have (10).

(d)

Exercise 5 (Optimization). (a) We can express the original optimization problem

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \quad \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \quad (12)$$

as

$$\begin{aligned} & \underset{\beta \in \mathbb{R}^p, z \in \mathbb{R}^n}{\text{minimize}} && \frac{1}{2} \|y - z\|_2^2 + \lambda \|\beta\|_1 \\ & \text{subject to} && z = X\beta. \end{aligned} \quad (13)$$

We will also refer to another expression of the lasso optimization problem,

$$\begin{aligned} & \underset{\beta \in \mathbb{R}^p}{\text{minimize}} && \frac{1}{2} \|y - X\beta\|_2^2 \\ & \text{subject to} && \|\beta\|_1 \leq t \end{aligned} \quad (14)$$

for some $t > 0$. The Lagrangian of (13) is

$$\mathcal{L}(\beta, z, u) = \frac{1}{2} \|y - z\|_2^2 + \lambda \|\beta\|_1 + u^T (z - X\beta),$$

so the Lagrange dual function is

$$\begin{aligned} \inf_{\beta, z} \{ \mathcal{L}(\beta, z, u) \} &= \inf_{\beta, z} \left\{ \frac{1}{2} \|y - z\|_2^2 + \lambda \|\beta\|_1 + u^T (z - X\beta) \right\} \\ &= \inf_{\beta, z} \left\{ \frac{1}{2} (y - z)^T (y - z) + u^T z + \lambda \|\beta\|_1 - u^T X\beta \right\} \end{aligned}$$

This minimization is separable:

$$= \inf_z \left\{ \frac{1}{2} (y^T y - 2y^T z + z^T z) + u^T z \right\} + \inf_{\beta} \{ \lambda \|\beta\|_1 - u^T X\beta \} \quad (15)$$

We will handle each part of (15) separately. First, the left side:

$$\inf_z \left\{ \frac{1}{2} (y^T y - 2y^T z + z^T z) + u^T z \right\} = \inf_z \left\{ \frac{1}{2} z^T z + (u - y)^T z + \frac{1}{2} y^T y \right\}$$

Since this is a convex quadratic form, differentiate with respect to z and set equal to zero:

$$z + (u - y) = 0 \implies z = y - u \quad (16)$$

$$\begin{aligned} \implies \inf_z \left\{ \frac{1}{2} z^T z + (u - y)^T z + \frac{1}{2} y^T y \right\} &= \frac{1}{2} (y - u)^T (y - u) + (u - y)^T (y - u) + \frac{1}{2} y^T y \\ &= \frac{1}{2} (y^T y - 2u^T y + u^T u) + 2u^T y - y^T y - u^T u + \frac{1}{2} y^T y = -\frac{1}{2} u^T u + u^T y = \frac{1}{2} y^T y - \frac{1}{2} y^T y + u^T y - \frac{1}{2} u^T u \\ &= \frac{1}{2} y^T y - \frac{1}{2} (y^T y - 2u^T y + u^T u) = \frac{1}{2} y^T y - \frac{1}{2} (y - u)^T (y - u) = \frac{1}{2} \|y\|_2^2 - \frac{1}{2} \|y - u\|_2^2 \end{aligned}$$

Next we will minimize the right side of (15):

$$\inf_{\beta} \{ \lambda \|\beta\|_1 - u^T X\beta \} = \inf_{\beta} \left\{ \lambda \sum_{i=1}^p |\beta_i| - \sum_{i=1}^p [u^T X]_i \beta_i \right\} = \inf_{\beta} \left\{ \sum_{i=1}^p \left(\lambda |\beta_i| - [u^T X]_i \beta_i \right) \right\}$$

$$= \inf_{\beta} \left\{ \sum_{i=1}^p \left(\text{sgn}(\beta_i) \lambda - [u^T X]_i \right) \beta_i \right\} = \sum_{i=1}^p \inf_{\beta_i} \left\{ \left(\text{sgn}(\beta_i) \lambda - [u^T X]_i \right) \beta_i \right\}.$$

Notice that when β_i is negative, if $\left(\text{sgn}(\beta_i) \lambda - [u^T X]_i \right) = -\left(\lambda + [u^T X]_i \right)$ is positive there is no lower bound on the quantity we are minimizing; otherwise, when β_i is negative the infimum is 0. When β_i is positive, if $\left(\text{sgn}(\beta_i) \lambda - [u^T X]_i \right) = \left(\lambda - [u^T X]_i \right)$ is negative there is no lower bound on the quantity we are minimizing; otherwise, when β_i is positive the infimum is 0. That is, the only dual feasible points satisfy for all i

$$-\left(\lambda + [u^T X]_i \right) \leq 0, \quad \lambda - [u^T X]_i \geq 0 \iff [u^T X]_i \geq -\lambda, \quad [u^T X]_i \leq \lambda$$

which is equivalent to the condition

$$\|u^T X\|_{\infty} \leq \lambda.$$

Therefore the Lagrange dual function is

$$\inf_{\beta, z} \{ \mathcal{L}(x, u) \} = \frac{1}{2} \|y\|_2^2 - \frac{1}{2} \|y - u\|_2^2 \quad (17)$$

subject to the constraint $\|u^T X\|_{\infty} \leq \lambda$. This quantity represents a lower bound on the minimum value of the original optimization problem for all $u \in \mathbb{R}^p$. The dual problem is to find the best lower bound by maximizing over u ; that is, the dual problem is

$$\begin{aligned} & \underset{u \in \mathbb{R}^p}{\text{maximize}} && \frac{1}{2} \|y\|_2^2 - \frac{1}{2} \|y - u\|_2^2 \\ & \text{subject to} && \|u^T X\|_{\infty} \leq \lambda. \end{aligned} \quad (18)$$

Lastly, suppose $\hat{\beta}$ and \hat{u} satisfy

$$\begin{aligned} \hat{\beta} &= \arg \min_{\beta \in \mathbb{R}^p} \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1, \\ \hat{u} &= \underset{u \in \mathbb{R}^p}{\arg \max} \frac{1}{2} \|y\|_2^2 - \frac{1}{2} \|y - u\|_2^2 = \underset{u \in \mathbb{R}^p}{\arg \min} -\frac{1}{2} \|y\|_2^2 + \frac{1}{2} \|y - u\|_2^2 \\ & \text{subject to} \quad \|u^T X\|_{\infty} \leq \lambda \quad \text{subject to} \quad \|u^T X\|_{\infty} \leq \lambda \end{aligned}$$

Then by (16) and strong duality we have $\hat{u} = y - X\hat{\beta}$.

- (b) (i) **Not necessarily unique.** Per Tibshirani [2013], if $\text{rank}(X) < p$, the lasso solution is not necessarily unique. Intuitively, this is because the columns of X are linearly dependent, so there may exist more than one linear combination of the columns that minimizes (12). **Jacob's suggestion: counterexample. X is two columns that are equal; then convex combinations of two solutions are equal as long as same sign (can't be opposite sign because then ℓ_1 could be smaller by setting one equal to 0).**
- (ii) **Necessarily unique.** The dual problem (18) is strictly concave, so the value \hat{u} that maximizes it is unique.
- (iii) **Necessarily unique** (except in the trivial case $\lambda = 0$). Per part 5(b)(iv), $\|\hat{\beta}\|_1$ is unique. (12) is convex, so the minimum $\frac{1}{2} \|y - X\hat{\beta}\|_2^2 + \lambda \|\hat{\beta}\|_1$ is unique. Therefore $\|y - X\hat{\beta}\|_2^2$ must be unique. **Jacob's solution:** Since \hat{u} is unique and by (16) $\hat{u} = y - X\hat{\beta}$, we must have that $\|\hat{u}\| = \|y - X\hat{\beta}\|$ is unique.

- (iv) **Necessarily unique** (except in the trivial case $\lambda = 0$). Whenever $\lambda > 0$, the lasso objective function (12) is the Lagrangian of (14). We will prove a useful lemma about the relationship between these functions.

Lemma 2. For a given $\lambda > 0$, let $\hat{\beta}$ minimize (12). Then there is exactly one $t = \|\hat{\beta}\|_1$ such that any $\hat{\beta}$ minimizing (12) also minimizes (14).

Proof. This must be true by contradiction. First of all, since the objective function of (14) is continuous and the feasible region $\|\beta\|_1 \leq t$ is compact, a minimum of (14) is guaranteed to exist. Now suppose $\hat{\beta}$ minimizes (12) for a fixed λ , with $\|\hat{\beta}\|_1 = t$, but there is a different solution $\hat{\beta}^*$ that is feasible for (14) and achieves a lower value. That is,

$$\frac{1}{2}\|y - X\hat{\beta}^*\|_2^2 < \frac{1}{2}\|y - X\hat{\beta}\|_2^2$$

and $\|\hat{\beta}^*\|_1 \leq \|\hat{\beta}\|_1 = t$. Since $\lambda > 0$, $\|\hat{\beta}\|_1 < \|\hat{\beta}_{global}\|_1$, where $\hat{\beta}_{global}$ is a global minimum for $\frac{1}{2}\|y - X\hat{\beta}\|_2^2$. Since (14) is convex and all global minima lie outside the feasible region, $\hat{\beta}^*$ lies on the boundary; that is, $\|\hat{\beta}^*\|_1 = \|\hat{\beta}\|_1 = t$. But then

$$\frac{1}{2}\|y - X\hat{\beta}^*\|_2^2 < \frac{1}{2}\|y - X\hat{\beta}\|_2^2 \iff \frac{1}{2}\|y - X\hat{\beta}^*\|_2^2 + \lambda\|\hat{\beta}^*\|_1 < \frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1$$

which contradicts the fact that $\hat{\beta}$ minimizes (12). □

Now the result follows in a simple way:

Proposition 3. Let \mathcal{B} be the set of all $\hat{\beta}$ that minimize (12) for some fixed $\lambda > 0$. Then for any two $\hat{\beta}_1, \hat{\beta}_2 \in \mathcal{B}$, $\|\hat{\beta}_1\|_1 = \|\hat{\beta}_2\|_1$. That is, $\|\hat{\beta}\|_1$ is unique.

Proof. Suppose $\hat{\beta}_1$ and $\hat{\beta}_2$ both minimize (12), and (without loss of generality) $\|\hat{\beta}_1\|_1 < \|\hat{\beta}_2\|_1$. By Lemma 2, these values both minimize (14) with $t = \|\hat{\beta}_2\|_1$ (we cannot choose $t = \|\hat{\beta}_1\|_1$ because $\hat{\beta}_1$ is not feasible for that problem). Because the global minimum of (14) lies outside the feasible region and (14) is convex, all solutions to (14) lie on the boundary of the feasible region. But $\|\hat{\beta}_1\|_1 < \|\hat{\beta}_2\|_1$, so $\hat{\beta}_1$ is not on the boundary of the feasible region, contradiction. Therefore $\|\hat{\beta}_1\|_1 = \|\hat{\beta}_2\|_1$ for all solutions $\hat{\beta}_1, \hat{\beta}_2$ to (12); that is, $\|\hat{\beta}\|_1$ is unique. □

(See Osborne et al. [2000] for more details.)

- (c) (i) Since β^* is clearly feasible for (12) and $\hat{\beta}$ achieves the minimum, we have

$$\frac{1}{2}\|y - X\beta^*\|_2^2 + \lambda\|\beta^*\|_1 \geq \frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 \iff \frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{2}\|\varepsilon\|_2^2 + \lambda\|\beta^*\|_1$$

- (ii) We know that the expression in the dual problem (18) is a lower bound for the solution of the primal problem (12) for any u feasible for (18) (that is, any u satisfying $\|u^T X\|_\infty \leq \lambda$). Therefore we have

$$\frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 \geq \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|y - u\|_2^2.$$

Since by assumption $\lambda \geq \|X^T \varepsilon\|_\infty$, ε is feasible for (18). Therefore we have

$$\frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 \geq \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|y - \varepsilon\|_2^2 = \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|X\beta^*\|_2^2 \quad (19)$$

as desired.

(iii) We can rewrite the right side of (19) as

$$\frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|X\beta^*\|_2^2 = \frac{1}{2}\|X\beta^*\|_2^2 + \frac{1}{2}\|\varepsilon\|_2^2 + \varepsilon^T X\beta^* - \frac{1}{2}\|X\beta^*\|_2^2 = \frac{1}{2}\|\varepsilon\|_2^2 + \varepsilon^T X\beta^*. \quad (20)$$

By assumption, we have

$$\lambda \geq \|X^T \varepsilon\|_\infty \iff \lambda \mathbf{1} - X^T \varepsilon \succeq 0 \implies \lambda \mathbf{1} \beta^* - X^T \varepsilon \beta^* \succeq 0$$

$$\iff -\lambda \|\beta^*\|_1 \leq \varepsilon^T X\beta^* \leq \lambda \|\beta^*\|_1.$$

By Hölder's Inequality, we have for any two vectors $u, v \in \mathbb{R}^n$, $|u^T v| \leq \|u\|_\infty \|v\|_1$. Therefore

$$|\varepsilon^T X\beta^*| = |(X^T \varepsilon)^T \beta^*| \leq \|X^T \varepsilon\|_\infty \|\beta^*\|_1 \leq \lambda \|\beta^*\|_1$$

where the last step used the assumption $\|X^T \varepsilon\|_\infty \leq \lambda$. So we have

$$\frac{1}{2}\|\varepsilon\|_2^2 + \lambda \|\beta^*\|_1 \leq \frac{1}{2}\|\varepsilon\|_2^2 + \varepsilon^T X\beta^*.$$

Substituting in to (19) and using the identity in (20) yields

$$\frac{1}{2}\|\varepsilon\|_2^2 + \lambda \|\beta^*\|_1 \leq \frac{1}{2}\|\varepsilon\|_2^2 + \varepsilon^T X\beta^* = \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|X\beta^*\|_2^2 \leq \frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda \|\beta\|_1$$

as desired.

(iv) We see from parts (i) and (iii) that

$$\frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda \|\beta\|_1 - \lambda \|\beta^*\|_1 \leq \frac{1}{2}\|\varepsilon\|_2^2 \leq \frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda \|\beta\|_1 + \lambda \|\beta^*\|_1$$

$$\iff \frac{1}{n}\|y - X\hat{\beta}\|_2^2 + \frac{2}{n}\lambda \|\beta\|_1 - \frac{2}{n}\lambda \|\beta^*\|_1 \leq \frac{1}{n}\|\varepsilon\|_2^2 \leq \frac{1}{n}\|y - X\hat{\beta}\|_2^2 + \frac{2}{n}\lambda \|\beta\|_1 + \frac{2}{n}\lambda \|\beta^*\|_1$$

that is, we can lower bound and upper bound $\frac{1}{n}\|\varepsilon\|_2^2$ by taking the quantity $\frac{1}{n}\|y - X\hat{\beta}\|_2^2 + \frac{2}{n}\lambda \|\beta\|_1$ and adding or subtracting $\frac{2}{n}\lambda \|\beta^*\|_1$. Therefore it seems that the quantity in the middle of this interval, $\frac{1}{n}\|y - X\hat{\beta}\|_2^2 + \frac{2}{n}\lambda \|\beta\|_1$, is a reasonable estimator for $\sigma^2 = \mathbb{E}[\|\varepsilon\|_2^2]$.

(d) *Proof.*

Definition 1 (Convex function in \mathbb{R}^n). Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$. We say that f is **convex** if, for any $x, y \in \mathbb{R}^n$ and for any $t \in [0, 1]$, we have

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y). \quad (21)$$

Note that

$$\|tx + (1-t)y\|_1 \leq \|tx\|_1 + \|(1-t)y\|_1 = t\|x\|_1 + (1-t)\|y\|_1 \quad (22)$$

where the first step follows by the Triangle Inequality (which all norms satisfy, including the ℓ_1 norm) and the second step follows by the homogeneity property of norms. Therefore $\|\theta\|_1$ is convex. Next, by (22) and the monotonicity of $g(\theta) = \theta^2$ when $\theta \geq 0$,

$$\begin{aligned}
f(tx + (1-t)y) &= (\|tx + (1-t)y\|_1)^2 \leq (t\|x\|_1 + (1-t)\|y\|_1)^2 \\
&= t^2\|x\|_1^2 + (1-t)^2\|y\|_1^2 + 2t(1-t)\|x\|_1\|y\|_1
\end{aligned}$$

and

$$tf(x) + (1-t)f(y) = t\|x\|_1^2 + (1-t)\|y\|_1^2$$

Taking the difference of these yields

$$\begin{aligned}
tf(x) + (1-t)f(y) - f(tx + (1-t)y) &\geq t\|x\|_1^2 + (1-t)\|y\|_1^2 - (t^2\|x\|_1^2 + (1-t)^2\|y\|_1^2 + 2t(1-t)\|x\|_1\|y\|_1) \\
&= (t-t^2)\|x\|_1^2 + [(1-t) - (1-t)^2]\|y\|_1^2 - 2t(1-t)\|x\|_1\|y\|_1 \\
&= (t-t^2)(\|x\|_1^2 + \|y\|_1^2 - 2\|x\|_1\|y\|_1) = t(1-t)(\|x\|_1 - \|y\|_1)^2 \geq 0 \\
&\iff tf(x) + (1-t)f(y) \geq f(tx + (1-t)y)
\end{aligned}$$

which proves convexity. □

References

- O. Banerjee, L. E. Ghaoui, and A. Edu. Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data Alexandre d'Aspremont. *Journal of Machine Learning Research*, 9:485–516, 2008. URL http://delivery.acm.org.libproxy2.usc.edu/10.1145/1400000/1390696/p485-banerjee.pdf?ip=154.59.124.74&id=1390696&acc=OPEN&key=B63ACEF81C6334F5.C52804B674E616B8.4D4702B0C3E38B35.6D218144511F3437&{}_{}_acm{}_{}_=1559857568{}_ade261c54a2f13c3a609b160683e627e.
- J. Fan, Y. Fan, and J. Lv. High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147:186–197, 2008. doi: 10.1016/j.jeconom.2008.09.017. URL www.elsevier.com/locate/jeconom.
- J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008. doi: 10.1093/biostatistics/kxm045. URL <https://academic.oup.com/biostatistics/article-abstract/9/3/432/224260>.
- M. R. Osborne, B. Presnell, and B. A. Turlach. On the LASSO and its Dual. *Journal of Computational and Graphical Statistics*, 9(2):319–337, 2000. ISSN 1537-2715. doi: 10.1080/10618600.2000.10474883. URL <https://www.tandfonline.com/action/journalInformation?journalCode=uugs20>.
- R. J. Tibshirani. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7(1):1456–1490, 2013. ISSN 19357524. doi: 10.1214/13-EJS815.