# DSO Screening Exam: 2018 In-Class Exam

Gregory Faletto

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

$$e^{-y}T(y) \xrightarrow{d} \text{Exponential}(\lambda) \iff \lim_{y\to\infty} \Pr(e^{-y}T(y) \le t) = 1 - e^{-\lambda t}$$

$$\iff \lim_{y\to\infty} \Pr\left(\frac{\inf\{x \ge 0 : M(x) \ge y\}}{e^y} > t\right) = e^{-\lambda t} \iff \lim_{y\to\infty} \Pr\left(e^{\log(\inf\{x\ge 0: M(x)\ge y\})-y} > t\right) = e^{-\lambda t}$$

$$\iff \lim_{y\to\infty} \Pr\left(e^{\log(\inf\{x\ge 0: M(x)\ge y\})-y} > t\right) = 1 - \lambda t + \frac{\lambda^2 t^2}{2!} - \frac{\lambda^3 t^3}{3!} + \ldots + \frac{(-\lambda t)^n}{n!} + \ldots$$

$$\vdots$$

$$\lim_{y\to\infty} e^{-y} \inf\{x \ge 0 : M(x) \ge y\} = \lambda e^{-\lambda t}$$

$$\vdots$$

So as $y \to \infty$, $\Pr\left(e^{\log(\inf\{x\ge 0: M(x)\ge y\})-y} > t\right)$ converges to an Exponential($\lambda$) random variable. That is, the probability that the ratio of the first hit time for $y$ ($T(y)$) and $e^y$ exceeds $t$ has a memoryless distribution in $t$ as $y$ grows without bound.

$$\vdots$$

(ii)

(b) **Mohammad:**

Using hint:

$$\int_0^1 t^x \, dt = \frac{1}{x+1}$$

$$\iff \mathbb{E}\int_0^1 t^x \, dt = \mathbb{E}\left(\frac{1}{x+1}\right) \iff \int_0^1 f\mathbb{E}(t^x) \, dt = \mathbb{E}\left(\frac{1}{x+1}\right)$$

$$\vdots$$

$$X \sim \text{Bin}(n, p)$$

Recall that

$$\mathbb{E}(X) = np \iff \sum_{x=0}^{n} x \cdot \binom{n}{x} p^x (1-p)^{n-x} = (1-p)^n \sum_{x=0}^{n} x \binom{n}{x} \left(\frac{p}{1-p}\right)^x$$

$$= (1-p)^n \sum_{x=0}^{n} \frac{n!}{(n-x)!(x-1)!} \left(\frac{p}{1-p}\right)^x = np.$$

Then we have

$$\mathbb{E}\left[\frac{1}{1+X}\right] = \sum_{x=0}^{\infty} \frac{1}{1+x} \cdot \Pr(X = x) = \sum_{x=0}^{n} \frac{1}{1+x} \cdot \binom{n}{x} p^x (1-p)^{n-x} = (1-p)^n \sum_{x=0}^{n} \frac{1}{1+x} \cdot \binom{n}{x} \left(\frac{p}{1-p}\right)^x$$

$$= (1-p)^n \sum_{x=0}^{n} \frac{n!}{(n-x)!(x+1)!} \left(\frac{p}{1-p}\right)^x = (1-p)^n \sum_{x=0}^{n} \frac{1}{(x+1)x} \frac{n!}{(n-x)!(x-1)!} \left(\frac{p}{1-p}\right)^x$$

$$\vdots$$

$$\mathbb{E}\left[\frac{1}{1+X}\right] = \int_0^{\infty} \Pr\left([X+1]^{-1} > x\right) dx = \int_0^1 \Pr\left(\frac{1}{X+1} > x\right) dx = \int_0^1 \Pr\left(X < \frac{1}{x} - 1\right) dx$$

$$= \lim_{a \to 0^+} \int_a^1 \Pr\left(X < \lceil x^{-1} - 1 \rceil\right) dx$$

$$= \lim_{a \to 0^+} \int_a^{(n+1)^{-1}} \Pr\left(X < \lceil x^{-1} - 1 \rceil\right) dx + \int_{(n+1)^{-1}}^1 \Pr\left(X < \lceil x^{-1} - 1 \rceil\right) dx$$

$$= \frac{1}{n+1} \cdot 1 + \int_{(n+1)^{-1}}^1 \Pr\left(X < \lceil x^{-1} - 1 \rceil\right) dx$$

This is simply the area under $n$ rectangles with heights $1, \Pr(X \leq n-1), \Pr(X \leq n-2), \ldots, \Pr(X = 0)$ and bases $(n+1)^{-1}, n^{-1} - (n+1)^{-1}, (n-1)^{-1} - n^{-1}, \ldots, 1/2$. That is, we can write this as

$$\mathbb{E}\left[\frac{1}{1+X}\right] = \frac{1}{n+1} \cdot 1 + \left(\frac{1}{n} - \frac{1}{n+1}\right) \cdot \Pr(X \leq n-1) + \left(\frac{1}{n-1} - \frac{1}{n}\right) \cdot \Pr(X \leq n-2)$$

$$+ \ldots + \left(\frac{1}{2} - \frac{1}{3}\right) \cdot \Pr(X \leq 1) + \frac{1}{2} \cdot \Pr(X = 0)$$

$$\vdots$$

Recall that

$$\mathbb{E}(X) = np \iff \sum_{x=0}^{n-1} \sum_{j=x+1}^{n} \binom{n}{j} p^j (1-p)^{n-j} = (1-p)^n \sum_{x=0}^{n-1} \sum_{j=x+1}^{n} \binom{n}{j} \left(\frac{p}{1-p}\right)^j = np.$$

**Exercise 2 (Mathematical Statistics).** (a) Let $\boxed{T_n(X_1, \ldots, X_n) := \sum_{i=1}^{n} X_i.}$

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

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

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

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

(c) Likelihood function:

$$\mathcal{L}(p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i}$$

$$\implies \ell(p) = \sum_{i=1}^{n} \left[X_i \log(p) + (1-X_i)\log(1-p)\right] = \log(p)\sum_{i=1}^{n} X_i + \log(1-p)\left(n - \sum_{i=1}^{n} X_i\right)$$

$$\implies \frac{\mathrm{d}}{\mathrm{d}p}\ell(p) = \frac{1}{p}\sum_{i=1}^{n} X_i - \frac{1}{1-p}\left(n - \sum_{i=1}^{n} X_i\right) = 0 \implies \frac{1}{p}\sum_{i=1}^{n} X_i = \frac{1}{1-p}\left(n - \sum_{i=1}^{n} X_i\right)$$

$$\iff \sum_{i=1}^{n} X_i - p\sum_{i=1}^{n} X_i = pn - p\sum_{i=1}^{n} X_i \iff \boxed{\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i.}$$

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

**Theorem 1. Central Limit Theorem (Grimmett and Stirzaker theorem 5.10.4.)** Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables with finite mean $\mu$ and finite non-zero variance $\sigma^2$, 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 $\mathbb{E}(X_i) = p, \mathrm{Var}(X_i) = p(1-p)$, we have

$$\frac{\sum_{i=1}^{n} X_i - np}{\sqrt{np(1-p)}} \xrightarrow{d} \mathcal{N}(0,1) \iff \sum_{i=1}^{n} X_i - np \xrightarrow{d} \mathcal{N}(0, np(1-p))$$

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

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

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

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

$$\mathbb{E}(X^T X \mid \mu) = \mathbb{E}\left(X_1^2 + X_2^2 + \ldots + X_n^2\right) = \sum_{i=1}^{n} \mathbb{E}(X_i^2) = \sum_{i=1}^{n} \left(\mathrm{Var}(X_i) + \mathbb{E}(X_i)^2\right) = \sum_{i=1}^{n} \left(1 + \mu_i^2\right)$$

$$= n + \|\mu\|_2^2 \implies \mathbb{E}(X^T X - n \mid \mu) = \|\mu\|_2^2$$

Therefore given $\mu$, $\boxed{X^T X - n}$ is unbiased for $\|\mu\|_2^2$.

(b) We have

$$\mu \sim \mathcal{N}(0, k\boldsymbol{I}_n)$$

$$\mathbb{E}\left[\left(\|\mu\|_2^2 - t(X)\right)^2 \mid X\right] = \mathbb{E}\left[\|\mu\|_2^4 - 2\|\mu\|_2^2 t(X) + t(X)^2 \mid X\right] = \mathbb{E}\left[\|\mu\|_2^4 \mid X\right] - 2t(X)\mathbb{E}\left[\|\mu\|_2^2 \mid X\right] + t(X)^2$$

The estimator minimizing this is $t(X) = \mathbb{E}\left[\|\mu\|_2^2 \mid X\right] = \mathbb{E}\left[\mu^T\mu \mid X\right]$, which we need to find. We have that $\mu \sim \mathcal{N}(0, k\boldsymbol{I}_n)$, so

$$f_{\mu^T\mu}(t) = f_{\sum_{i=1}^n \mu_i^2}(t) = f_{\sum_{i=1}^n \left[\frac{\mu_i}{k}\right]^2}\left(\frac{t}{k^2}\right) = f_{\chi_n^2}\left(\frac{t}{k^2}\right)$$

where $f_{\chi_n^2}$ is the density of a $\chi^2$ random variable with $n$ degrees of freedom. Also, the set of vectors $\{\mu' \in \mathbb{R}^n \mid \mu'^T\mu' = \|\mu\|_2^2\}$ is a hypersphere of radius $\|\mu\|_2$ in $\mathbb{R}^n$, so the conditional density of $\mu'$ given $\mu^T\mu$ is uniform over the surface of this hypersphere. Per Muller [1959], this can be generated by drawing $n$ standard Gaussian random variables then dividing each by the $\ell_2$ norm of all of them, then in this case multiplying by the desired $\ell_2$ norm $\|\mu\|_2$. That is,

$$f_{\mu|\mu^T\mu}(m \mid t) = f_{\mu|\mu^T\mu}\big((m_1, \ldots, m_n) \mid t\big) = f_{\mu|\mu^T\mu}\big((m_1, \ldots, m_n) \mid t\big)$$

$$\implies f_{\mu^T\mu|X}(t \mid x) = \frac{f_{\mu^T\mu,X}(t, x)}{f_X(x)} = \frac{f_{\mu^T\mu}(t) f_{\mu|\mu^T\mu}(m \mid t) f_{X|\mu}(x \mid m)}{f_X(x)}$$

$$\implies \mathbb{E}(\mu^T\mu \mid X) = \int_0^\infty t \Pr(\mu^T\mu = t \mid X) dt$$

$$\vdots$$

Given $\mu$, we have

$$X \mid \mu \sim \mathcal{N}(\mu, \boldsymbol{I}_n) \implies (X-\mu)^T(X-\mu) \mid \mu \sim \chi_n^2 \iff X^TX - 2\mu^TX + \mu^T\mu \mid \mu \sim \chi_n^2.$$

$$\vdots$$

Also, we have that $\mu \sim \mathcal{N}(0, k\boldsymbol{I}_n)$. The joint distribution of $X$ and $\mu$ is then

$$f_{X,\mu}(x, m) = f_{X|\mu=m}(x \mid m) f_\mu(m) =$$

(c)

(d)

**Exercise 4 (High-Dimensional Statistics).** (a) **Sparsity in the covariance matrix does not imply sparsity in the precision matrix.** For example, suppose the covariance matrix is the following:

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

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

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

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

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

$$= \begin{pmatrix} \left(\frac{1}{\tau^2} - \frac{1}{\tau^2(\tau^2+s)}\right)\boldsymbol{I}_n & -\frac{1}{\tau^2(\tau^2+s)}\boldsymbol{I}_n & -\frac{1}{\tau^2(\tau^2+s)}\boldsymbol{I}_n & \cdots & -\frac{1}{\tau^2(\tau^2+s)}\boldsymbol{I}_n \\ -\frac{1}{\tau^2(\tau^2+s)}\boldsymbol{I}_n & \left(\frac{1}{\tau^2} - \frac{1}{\tau^2(\tau^2+s)}\right)\boldsymbol{I}_n & -\frac{1}{\tau^2(\tau^2+s)}\boldsymbol{I}_n & \cdots & -\frac{1}{\tau^2(\tau^2+s)}\boldsymbol{I}_n \\ -\frac{1}{\tau^2(\tau^2+s)}\boldsymbol{I}_n & -\frac{1}{\tau^2(\tau^2+s)}\boldsymbol{I}_n & \left(\frac{1}{\tau^2} - \frac{1}{\tau^2(\tau^2+s)}\right)\boldsymbol{I}_n & \cdots & -\frac{1}{\tau^2(\tau^2+s)}\boldsymbol{I}_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{\tau^2(\tau^2+s)}\boldsymbol{I}_n & -\frac{1}{\tau^2(\tau^2+s)}\boldsymbol{I}_n & -\frac{1}{\tau^2(\tau^2+s)}\boldsymbol{I}_n & \cdots & \left(\frac{1}{\tau^2} - \frac{1}{\tau^2(\tau^2+s)}\right)\boldsymbol{I}_n \end{pmatrix}$$

$$= \begin{pmatrix} \frac{\tau^2+s-1}{\tau^2(\tau^2+s)}\boldsymbol{I}_n & -\frac{1}{\tau^2(\tau^2+s)}\boldsymbol{I}_n & -\frac{1}{\tau^2(\tau^2+s)}\boldsymbol{I}_n & \cdots & -\frac{1}{\tau^2(\tau^2+s)}\boldsymbol{I}_n \\ -\frac{1}{\tau^2(\tau^2+s)}\boldsymbol{I}_n & \frac{\tau^2+s-1}{\tau^2(\tau^2+s)}\boldsymbol{I}_n & -\frac{1}{\tau^2(\tau^2+s)}\boldsymbol{I}_n & \cdots & -\frac{1}{\tau^2(\tau^2+s)}\boldsymbol{I}_n \\ -\frac{1}{\tau^2(\tau^2+s)}\boldsymbol{I}_n & -\frac{1}{\tau^2(\tau^2+s)}\boldsymbol{I}_n & \frac{\tau^2+s-1}{\tau^2(\tau^2+s)}\boldsymbol{I}_n & \cdots & -\frac{1}{\tau^2(\tau^2+s)}\boldsymbol{I}_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{\tau^2(\tau^2+s)}\boldsymbol{I}_n & -\frac{1}{\tau^2(\tau^2+s)}\boldsymbol{I}_n & -\frac{1}{\tau^2(\tau^2+s)}\boldsymbol{I}_n & \cdots & \frac{\tau^2+s-1}{\tau^2(\tau^2+s)}\boldsymbol{I}_n \end{pmatrix}.$$

(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 $\boldsymbol{\Sigma}$ is the inverse of $\boldsymbol{\Omega}$.

(c) **double-check that this actually works for $p > n$. Also, 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 $\Sigma$. Let $S$ by the empirical covariance matrix; that is,

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

where $X^{(i)}$ is the $i$th row of $X$ and $\overline{X} = n^{-1}\sum_{i=1}^{n} X^{(i)}$. We will make use of the following partitions of $\hat{\Sigma}$ and $S$:

$$\hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_{11} & \hat{\sigma}_{12} \\ \hat{\sigma}_{12}^T & \hat{\sigma}_{22} \end{pmatrix}, \qquad S = \begin{pmatrix} S_{11} & S_{12} \\ S_{12}^T & S_{22} \end{pmatrix}, \qquad \hat{\Omega} = \begin{pmatrix} \hat{\Omega}_{11} & \hat{\omega}_{12} \\ \hat{\omega}_{12}^T & \hat{\omega}_{22} \end{pmatrix}, \tag{1}$$

as well as the constraint

$$\begin{pmatrix} \hat{\Sigma}_{11} & \hat{\sigma}_{12} \\ \hat{\sigma}_{12}^T & \hat{\sigma}_{22} \end{pmatrix} \begin{pmatrix} \hat{\Omega}_{11} & \hat{\omega}_{12} \\ \hat{\omega}_{12}^T & \hat{\omega}_{22} \end{pmatrix} = \begin{pmatrix} \boldsymbol{I}_{p-1} & 0 \\ 0^T & 1 \end{pmatrix} \tag{2}$$

suggested by $\Sigma\Omega = \boldsymbol{I}_p$. We will optimize the following objective:

$$\hat{\Omega} := \underset{\Omega \in \mathcal{S}^+}{\arg\max} \{\log\det(\Omega) - \mathrm{Tr}(S\Omega) + \lambda\|\Omega\|_1\} \tag{3}$$

where $\mathcal{S}^+$ is the set of nonnegative definite $p \times p$ matrices and $\lambda > 0$ is a penalty parameter. The proposed procedure to optimize (3) is as follows:

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

   $$\hat{\beta} = \underset{\beta \in \mathbb{R}^{p-1}}{\arg\min} \left\{ \frac{1}{2}\|\hat{\Sigma}_{11}^{1/2}\beta - \hat{\Sigma}_{11}^{-1/2}s_{12}\|_2^2 + \lambda\|\beta\|_1 \right\} \tag{4}$$

   Note that this problem takes as input the inner products $\hat{\Sigma}_{11}$ and $s_{12}$.
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 $t \cdot \mathrm{ave}\,|S^{-\mathrm{diag}}|$, where $S^{-\mathrm{diag}}$ 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 (1) using the formulae

   $$\hat{\omega}_{22} = 1/\left(\hat{\sigma}_{22} - \hat{\sigma}_{12}^T\hat{\beta}\right), \qquad \hat{\omega}_{12} = -\hat{\beta}\hat{\omega}_{22}. \tag{5}$$

Formulae (5) are justified as follows: from (2) we have the following identities

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

These yield

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

Then using $\hat{\sigma}_{12} = \hat{\Sigma}_{11}\hat{\beta} \iff \hat{\beta} = \hat{\Sigma}_{11}^{-1}\hat{\sigma}_{12}$, we have (5).

(d)

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

$$\underset{\beta\in\mathbb{R}^p}{\text{minimize}} \quad \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \tag{6}$$

as

$$\begin{aligned} \underset{\beta\in\mathbb{R}^p, z\in\mathbb{R}^n}{\text{minimize}} \quad & \frac{1}{2}\|y - z\|_2^2 + \lambda\|\beta\|_1 \\ \text{subject to} \quad & z = X\beta. \end{aligned} \tag{7}$$

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

$$\begin{aligned} \underset{\beta\in\mathbb{R}^p}{\text{minimize}} \quad & \frac{1}{2}\|y - X\beta\|_2^2 \\ \text{subject to} \quad & \|\beta\|_1 \leq t \end{aligned} \tag{8}$$

for some $t > 0$. The Lagrangian of (7) is

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

so the Lagrange dual function is

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

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

This minimization is separable:

$$= \inf_{z}\left\{\frac{1}{2}\left(y^Ty - 2y^Tz + z^Tz\right) + \nu^Tz\right\} + \inf_{\beta}\left\{\lambda\|\beta\|_1 - \nu^TX\beta\right\} \tag{9}$$

We will handle each part of (9) separately. First, the left side:

$$\inf_{z}\left\{\frac{1}{2}\left(y^Ty - 2y^Tz + z^Tz\right) + \nu^Tz\right\} = \inf_{z}\left\{\frac{1}{2}z^Tz + (\nu - y)^Tz + \frac{1}{2}y^Ty\right\}$$

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

$$z + (\nu - y) = 0 \implies z = y - \nu \tag{10}$$

$$\implies \inf_{z}\left\{\frac{1}{2}z^Tz + (\nu - y)^Tz + \frac{1}{2}y^Ty\right\} = \frac{1}{2}(y - \nu)^T(y - \nu) + (\nu - y)^T(y - \nu) + \frac{1}{2}y^Ty$$

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

$$= \frac{1}{2} y^T y - \frac{1}{2}(y^T y - 2\nu^T y + \nu^T \nu) = \frac{1}{2} y^T y - \frac{1}{2}(y-\nu)^T (y-\nu) = \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|y-\nu\|_2^2$$

Next we will minimize the right side of (9):

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

$$= \inf_{\beta}\left\{\sum_{i=1}^{p}\left((-1)^{\mathbb{I}\{\beta_i<0\}}\lambda - \left[\nu^T X\right]_i\right)\beta_i\right\} = \sum_{i=1}^{p}\inf_{\beta_i}\left\{\left((-1)^{\mathbb{I}\{\beta_i<0\}}\lambda - \left[\nu^T X\right]_i\right)\beta_i\right\}$$

where $\mathbb{I}\{\beta_i < 0\}$ is an indicator function. Notice that when $\beta_i$ is negative, if $\left((-1)^{\mathbb{I}\{\beta_i<0\}}\lambda - \left[\nu^T X\right]_i\right) = -\left(\lambda + \left[\nu^T X\right]_i\right)$ is positive there is no lower bound on the quantity we are minimizing; otherwise, when $\beta_i$ is negative the infimum is 0. When $\beta_i$ is positive, if $\left((-1)^{\mathbb{I}\{\beta_i<0\}}\lambda - \left[\nu^T X\right]_i\right) = \left(\lambda - \left[\nu^T X\right]_i\right)$ is negative there is no lower bound on the quantity we are minimizing; otherwise, when $\beta_i$ is negative the infimum is 0. That is, the only dual feasible points satisfy for all $i$

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

which is equivalent to the condition

$$\|\nu^T X\|_\infty \leq \lambda.$$

Therefore the Lagrange dual function is

$$\inf_{\beta,z}\{\mathcal{L}(x,\nu)\} = \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|y-\nu\|_2^2 \tag{11}$$

subject to the constraint

$$\|\nu^T X\|_\infty \leq \lambda.$$

This quantity represents a lower bound on the minimum value of the original optimization problem for all $\nu \in \mathbb{R}^p$. The dual problem is to find the best lower bound by maximizing over $\nu$; that is, the dual problem is

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

**still need to finish; actually trivial based on (10):** Lastly, suppose $\hat{\beta}$ and $\hat{\nu}$ satisfy

$$\hat{\beta} = \underset{\beta\in\mathbb{R}^p}{\arg\min} \quad \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \quad ,$$

$$\hat{\nu} = \begin{array}{ll} \underset{\nu \in \mathbb{R}^p}{\arg\max} & \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|y-\nu\|_2^2 \\ \text{subject to} & \|\nu^T X\|_\infty \leq \lambda \end{array} = \begin{array}{ll} \underset{\nu \in \mathbb{R}^p}{\arg\min} & -\frac{1}{2}\|y\|_2^2 + \frac{1}{2}\|y-\nu\|_2^2 \\ \text{subject to} & \|\nu^T X\|_\infty \leq \lambda \end{array}$$

$$\vdots$$

(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 (6). **jacob 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 $\ell_1$ could be smaller by setting one equal to 0.**

(ii) **Necessarily unique. flesh out more** By a result from part (a), $\hat{u} = y - X\hat{\beta}$. By part (iii), even though $\hat{\beta}$ is not unique, $X\hat{\beta}$ is (see also Lemma 1 in Tibshirani [2013]). Therefore $\hat{u}$ is unique.

**Jacob's solution: strictly convex optimization problem, so argument maximizing is unique. dual is always convex;**

(iii) **Necessarily unique** (except in the trivial case $\lambda = 0$). Per part 5(b)(iv), $\|\hat{\beta}\|_1$ is unique. (6) 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: if $\hat{\nu}$ is unique then its $\ell_2$ norm is unique.**

(iv) **Necessarily unique change this argument: if you solve 1 and take the norm of that and choose that value for $t$ then you will get the same solution. because if there existed a better solution then it would have made the objective function 1 better.** (except in the trivial case $\lambda = 0$). Whenever $\lambda > 0$, the lasso objective function (6) is the dual of (8) with $t$ less than the $\ell_1$ norm of the OLS solution (if it exists). (8) is convex and Slater's condition holds because every point $\{\beta \in \mathbb{R}^p \mid \|\beta\|_1 < t\}$ is feasible. Therefore for this dual function, strong duality holds; that is, the optimal values of (6) and (8) are equal. Further, they are optimized over the same variable and they are both convex, so the solution set of (6) is identical to the solution set of (8).

Since the objective function of (8) is continuous and the feasible region $\|\beta\|_1 \leq t$ is compact, a minimum is guaranteed to exist. Since (8) is convex and the global minimum lies outside the region **for** $t < t_0$, the minimum will lie on the boundary; that is, $\|\hat{\beta}\|_1 = t$ for (8) and therefore for (6). See Osborne et al. [2000] for details.

(c) (i) Since $\beta^*$ is clearly feasible for (6) 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) From part (a), since the optimal value of the dual (11) is a lower bound for the optimal value of the primal (6), we have

$$\frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|y - \hat{\nu}\|_2^2 \leq \frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1. \tag{12}$$

**Jacob's solution: plug in $\varepsilon$ for $\nu$ in the lower bound equation, then you get $X\beta^*$**

so we are done if $\|X\beta^*\|_2^2 \geq \|y - \hat{\nu}\|_2^2$. Also by part (a), $\hat{\nu} = y - X\hat{\beta}$, so $\|y - \hat{\nu}\|_2^2 = \|X\hat{\beta}\|_2^2$, so we are done if $\|X\beta^*\|_2^2 \geq \|X\hat{\beta}\|_2^2$. From part (c)(i), we have

$$\frac{1}{2}\left[(y - X\hat{\beta})^T(y - X\hat{\beta}) - (y - X\beta^*)^T(y - X\beta^*)\right] \leq \lambda(\|\beta^*\|_1 - \|\hat{\beta}\|_1)$$

$$\iff 2y^T X(\beta^* - \hat{\beta}) + (\hat{\beta})^T X^T X\hat{\beta} - (\beta^*)^T X^T X\beta^* \leq 2\lambda(\|\beta^*\|_1 - \|\hat{\beta}\|_1)$$

$$\iff \|X\beta^*\|_2^2 - \|X\hat{\beta}\|_2^2 \geq 2y^T X(\beta^* - \hat{\beta}) - 2\lambda(\|\beta^*\|_1 - \|\hat{\beta}\|_1) \tag{13}$$

So (13) is a sufficient condition for the result. Examining the right side of the right side of (13), we have

$$2y^T X(\beta^* - \hat{\beta}) - 2\lambda(\|\beta^*\|_1 - \|\hat{\beta}\|_1) = 2y^T(y - \varepsilon - X\hat{\beta}) - 2\lambda(\|\beta^*\|_1 - \|\hat{\beta}\|_1) \tag{14}$$

Borrowing notation from Osborne et al. [2000], note that the subdifferential of (6) is given by

$$\partial_\beta \mathcal{L}(\beta, \lambda) = \partial_\beta \left(\frac{1}{2}(y - X\beta)^T(y - X\beta) + \lambda\|\beta\|_1\right) = \partial_\beta \left(\frac{1}{2}(y^T y - 2y^T X\beta + \beta^T X^T X\beta) + \lambda\|\beta\|_1\right)$$

$$= -y^T X + X^T X\beta + \lambda v = -X^T(y - X\beta) + \lambda v$$

where $v = (v_1, \ldots, v_p)^T$ with $v_i = \text{sgn}(\beta_i)$ if $\beta_i \neq 0$ and $v_i \in [-1, 1]$ if $\beta_i = 0$. By the convexity of (6), for $\hat{\beta}$ minimizing (6) it must hold that

$$0 = -X^T(y - X\beta) + \lambda v \iff \lambda v^T\hat{\beta} = (y - X\hat{\beta})^T X\hat{\beta} \iff \lambda = \frac{(y - X\hat{\beta})^T X\hat{\beta}}{\|\hat{\beta}\|_1}$$

$$\iff \|\hat{\beta}\|_1 = \frac{(y - X\hat{\beta})^T X\hat{\beta}}{\lambda} \tag{15}$$

where we used $v^T\hat{\beta} = \|\hat{\beta}\|_1$. Substituting (15) into (14), we have

$$2y^T(y - \varepsilon - X\hat{\beta}) - 2\lambda\left(\|\beta^*\|_1 - \frac{(y - X\hat{\beta})^T X\hat{\beta}}{\lambda}\right)$$

$$= 2y^T y - 2y^T\varepsilon - 2y^T X\hat{\beta} - 2\lambda\|\beta^*\|_1 + 2y^T X\hat{\beta} - 2\hat{\beta}^T X^T X\hat{\beta} = 2y^T(y - \varepsilon) - 2\lambda\|\beta^*\|_1 - 2\hat{\beta}^T X^T X\hat{\beta}$$

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

$$= 2\|X\beta^*\|_2^2 + 2\varepsilon^T X\beta^* - 2\lambda\|\beta^*\|_1 - 2\|X\hat{\beta}\|_2^2 = 2\left(\|X\beta^*\|_2^2 - \|X\hat{\beta}\|_2^2\right) - 2\left(\lambda\|\beta^*\|_1 - \varepsilon^T X\beta^*\right)$$

$$= 2\left(\|X\beta^*\|_2^2 - \|X\hat{\beta}\|_2^2\right) - 2\sum_{i=1}^{p}\left((-1)^{\mathbb{I}\{\beta_i < 0\}}\lambda - \left[\varepsilon^T X\right]_i\right)\beta_i^*$$

Because $\lambda \geq \|X^T\varepsilon\|_\infty$ (that is, for all $i \in \{1, \ldots, p\}$, $\lambda - \left|\left[X^T\varepsilon\right]_i\right| \geq 0$ ) we have that for all $i \in \{1, \ldots, p\}$,

$$\left((-1)^{\mathbb{I}\{\beta_i < 0\}}\lambda - \left[\varepsilon^T X\right]_i\right)\beta_i^* \geq 0$$

by the following argument. If $\beta_i^* = 0$, the result is trivial. If $\beta_i^* > 0$, we have

$$(-1)^{\mathbb{I}\{\beta_i<0\}}\lambda - \left[\varepsilon^T X\right]_i = \lambda - \left[\varepsilon^T X\right]_i \geq \lambda - \left|\left[X^T\varepsilon\right]_i\right| \geq 0$$

$$\implies \left((-1)^{\mathbb{I}\{\beta_i<0\}}\lambda - \left[\varepsilon^T X\right]_i\right)\beta_i^* \geq 0.$$

Lastly, if $\beta_i^* < 0$, we have

$$(-1)^{\mathbb{I}\{\beta_i<0\}}\lambda - \left[\varepsilon^T X\right]_i = -\left(\lambda + \left[\varepsilon^T X\right]_i\right) \leq -\left(\lambda - \left|\left[X^T\varepsilon\right]_i\right|\right) \leq 0$$

$$\implies \left((-1)^{\mathbb{I}\{\beta_i<0\}}\lambda - \left[\varepsilon^T X\right]_i\right)\beta_i^* \geq 0.$$

Therefore we have

$$2\left(\|X\beta^*\|_2^2 - \|X\hat{\beta}\|_2^2\right) - 2\sum_{i=1}^{p}\left((-1)^{\mathbb{I}\{\beta_i<0\}}\lambda - \left[\varepsilon^T X\right]_i\right)\beta_i^* \leq 2\left(\|X\beta^*\|_2^2 - \|X\hat{\beta}\|_2^2\right)$$

$$\vdots$$

$$\|X\beta^*\|_2^2 - \|X\hat{\beta}\|_2^2 \geq 2y^T X(\beta^* - \hat{\beta}) - 2\lambda(\|\beta^*\|_1 - \|\hat{\beta}\|_1)$$

$$\iff \|X\beta^*\|_2^2 - \|X\hat{\beta}\|_2^2 \geq 2(X\beta^* + \varepsilon)^T X\beta^* - 2y^T X\hat{\beta} - 2\lambda(\|\beta^*\|_1 - \|\hat{\beta}\|_1)$$

$$\iff \|X\beta^*\|_2^2 - \|X\hat{\beta}\|_2^2 \geq 2\|X\beta^*\|_2^2 + 2\varepsilon^T X\beta^* - 2y^T X\hat{\beta} - 2\lambda(\|\beta^*\|_1 - \|\hat{\beta}\|_1)$$

(iii) **haven't finished, but this is just a bonus question** We already have from part (c)(i)

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

Note that

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

(iv) **haven't finished, but this is just a bonus question**

(d) *Proof.*

> **Definition 1 (Convex function in $\mathbb{R}^n$).** Let $f : \mathbb{R}^n \to \mathbb{R}$. We say that $f$ is **convex** if, for any $x, y \in \mathbb{R}^n$ and for any $t \in [0, 1]$, we have
>
> $$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y). \tag{16}$$

Note that

$$\|tx + (1-t)y\|_1 \leq \|tx\|_1 + \|(1-t)y\|_1 = t\|x\|_1 + (1-t)\|y\|_1 \tag{17}$$

where the first step follows by the Triangle Inequality (which all norms satisfy, including the $\ell_1$ norm) and the second step follows by the homogeneity property of norms. Therefore $\|\theta\|_1$ is convex. Next, by (17) and the monotonicity of $g(\theta) = \theta^2$ when $\theta \geq 0$,

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

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

and

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

Taking the difference of these yields

$$tf(x)+(1-t)f(y)-f(tx+(1-t)y) \geq t\|x\|_1^2+(1-t)\|y\|_1^2-\left(t^2\|x\|_1^2+(1-t)^2\|y\|_1^2+2t(1-t)\|x\|_1\|y\|_1\right)$$

$$= (t-t^2)\|x\|_1^2 + [(1-t)-(1-t)^2]\|y\|_1^2 - 2t(1-t)\|x\|_1\|y\|_1$$

$$= (t-t^2)\left(\|x\|_1^2 + \|y\|_1^2 - 2\|x\|_1\|y\|_1\right) = t(1-t)(\|x\|_1 - \|y\|_1)^2 \geq 0$$

$$\iff tf(x) + (1-t)f(y) \geq f(tx + (1-t)y)$$

which proves convexity.

□

# References

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. E. Muller. A Note on a Uniformly Method for Generating Points on N-Dimensional Spheres. *Communications of the ACM*, 2:19–20, 1959. URL http://delivery.acm.org.libproxy1.usc.edu/10.1145/380000/377946/p19-muller.pdf?ip=132.174.255.3&id=377
SERVICE&key=B63ACEF81C6334F5.C52804B674E616B8.4D4702B0C3E38B35.4D4702B0C3E38B35&__acm__=1559508442_3ad6

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=ucgs20.

R. J. Tibshirani. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7(1):1456–1490, 2013. ISSN 19357524. doi: 10.1214/13-EJS815.