DSO Screening Exam: 2018 In-Class Exam

Gregory Faletto

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

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

Note that

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

Because $T(y) = \inf\{x \geq 0 : M(x) \geq y\}$ and $M(\cdot)$ is montonically 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

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

Let $b = \log a$ to get

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

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

(ii) The distribution function is $F(x) = e^{-\lambda e^{-x}}$, a Gumbel distribution with parameter λ.

(b) **Mohammad:**

Using hint:

$$\int_0^1 t^x \, dt = \left[\frac{t^{x+1}}{x+1}\right]_0^1 = \frac{1}{x+1}$$

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

$$\iff \int_0^1 \sum_{i=0}^n t^i \Pr(X = i) \, dt = \mathbb{E}\left(\frac{1}{X+1}\right) \iff \int_0^1 \sum_{i=0}^n t^i \binom{n}{i} p^i (1-p)^{n-i} \, dt = \mathbb{E}\left(\frac{1}{X+1}\right)$$

$$\iff (1-p)^n \int_0^1 \sum_{i=0}^n \binom{n}{i} \left(\frac{tp}{1-p}\right)^i \, dt = \mathbb{E}\left(\frac{1}{X+1}\right)$$

$$\iff (1-p)^n \int_0^1 \left(1 + \frac{tp}{1-p}\right)^n \, dt = \mathbb{E}\left(\frac{1}{X+1}\right)$$

$$\iff \int_0^1 (1-p+tp)^n \, dt = \mathbb{E}\left(\frac{1}{X+1}\right)$$

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

$$\int_0^1 t^{x+1}\ dt = \left[\frac{t^{x+2}}{x+2}\right]_0^1 = \frac{1}{x+2}$$

$$\iff \mathbb{E}\left[\int_0^1 t^{X+1}\ dt\right] = \mathbb{E}\left(\frac{1}{X+2}\right) \iff \int_0^1 \mathbb{E}(t^{X+1})\ dt = \mathbb{E}\left(\frac{1}{X+2}\right)$$

$$\iff \int_0^1 \sum_{i=0}^n t^{i+1} \Pr(X=i)\ dt = \mathbb{E}\left(\frac{1}{X+2}\right) \iff \int_0^1 \sum_{i=0}^n t^{i+1}\binom{n}{i}p^i(1-p)^{n-i}\ dt = \mathbb{E}\left(\frac{1}{X+2}\right)$$

$$\iff (1-p)^n \int_0^1 t \sum_{i=0}^n \binom{n}{i}\left(\frac{tp}{1-p}\right)^i\ dt = \mathbb{E}\left(\frac{1}{X+2}\right)$$

$$\iff (1-p)^n \int_0^1 t\left(1+\frac{tp}{1-p}\right)^n\ dt = \mathbb{E}\left(\frac{1}{X+2}\right)$$

$$\iff \int_0^1 t\,(1-p+tp)^n\ dt = \mathbb{E}\left(\frac{1}{X+2}\right)$$

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

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

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

$$\iff \mathbb{E}\left(\frac{1}{X+2}\right) = \frac{1}{p^2}\left[\frac{1-(1-p)^{n+2}}{n+2} - (1-p)\frac{1-(1-p)^{n+1}}{n+1}\right]_{1-p}^1$$

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 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 $\mathbb{E}(X_i) = p, \mathrm{Var}(X_i) = p(1-p)$, 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 μ, $\boxed{X^T X - n}$ 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 χ^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 μ' 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 ℓ_2 norm of all of them, then in this case multiplying by the desired ℓ_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 μ, 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 μ 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 a simple example, consider a sequence of p Gaussian random variables generated in the following way:

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

where $Z \sim \mathcal{N}(0,1)$. Then for all $i \neq j \in \{1, \ldots, p\}$, $\text{Cov}(X_i, X_j) \neq 0$, so the covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$ is dense with no zero entries. However, for any $i \neq j \in \{1, \ldots, p\}$, $\text{Cov}(X_i, X_j \mid \boldsymbol{X}_{-i-j}$ (where $\boldsymbol{X}_{-i-j}$ contains all of the X_k except $X_i and X_j$; that is, $\text{Cov}(X_i, X_j \mid \boldsymbol{X}_{-i-j}$ 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:

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

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

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

$$= \frac{1}{\tau^2}\left(\boldsymbol{I}_{ns} - (\mathbf{1}_s \otimes \boldsymbol{I}_n)\left[\tau^2 \boldsymbol{I}_n + \mathbf{1}_s^T \mathbf{1}_s \otimes \boldsymbol{I}_n\right]^{-1}(\mathbf{1}_s \otimes \boldsymbol{I}_n)^T\right)$$

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

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

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

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

(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) **consider instead using method from Fan et al. [2008].** I would estimate the precision matrix using the graphical lasso [Friedman et al., 2008]. Let $\hat{\boldsymbol{\Sigma}}$ be an estimate for the covariance matrix Σ. 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 ith row of X and $\overline{X} = n^{-1}\sum_{i=1}^n X^{(i)}$. 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

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

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

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

$$= (2\pi)^{-np/2}|\mathbf{\Sigma}|^{-n/2} \cdot \exp\left(-\frac{1}{2}\operatorname{Tr}\left[nS\mathbf{\Sigma}^{-1}\right]\right) = (2\pi)^{-np/2}\left|\mathbf{\Sigma}^{-1}\right|^{n/2} \cdot \exp\left(-\frac{n}{2}\operatorname{Tr}\left[S\mathbf{\Sigma}^{-1}\right]\right)$$

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

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

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

$$\log \mathcal{L}(\mathbf{\Sigma}^{-1}) \propto \log\left|\mathbf{\Sigma}^{-1}\right| - \operatorname{Tr}\left(S\mathbf{\Sigma}^{-1}\right)$$

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

$$\hat{\Omega} := \underset{\Omega \in \mathcal{S}^+}{\arg\max}\left\{\log|\Omega| - \operatorname{Tr}(S\Omega) + \lambda\|\Omega\|_1\right\} \tag{1}$$

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 to optimize this function. We will make use of the following partitions of $\hat{\mathbf{\Sigma}}$ and S:

$$\hat{\mathbf{\Sigma}} = \begin{pmatrix} \hat{\Sigma}_{11} & \hat{\Sigma}_{12} \\ \hat{\Sigma}_{12}^T & \hat{\sigma}_{22} \end{pmatrix}, \qquad S = \begin{pmatrix} S_{11} & S_{12} \\ S_{12}^T & S_{22} \end{pmatrix}, \qquad \hat{\Omega} = \begin{pmatrix} \hat{\Omega}_{11} & \hat{\omega}_{12} \\ \hat{\omega}_{12}^T & \hat{\omega}_{22} \end{pmatrix}, \tag{2}$$

as well as the constraint

$$\begin{pmatrix} \hat{\Sigma}_{11} & \hat{\sigma}_{12} \\ \hat{\sigma}_{12}^T & \hat{\sigma}_{22} \end{pmatrix}\begin{pmatrix} \hat{\Omega}_{11} & \hat{\omega}_{12} \\ \hat{\omega}_{12}^T & \hat{\omega}_{22} \end{pmatrix} = \begin{pmatrix} \boldsymbol{I}_{p-1} & 0 \\ 0^T & 1 \end{pmatrix} \tag{3}$$

suggested by the identity $\Sigma\Omega = \boldsymbol{I}_p$. The proposed procedure to optimize (1) is the following coordinate descent algorithm:

1. Initialize the algorithm with estimate $\hat{\mathbf{\Sigma}} = S + \lambda\boldsymbol{I}_p$. (The diagonal of $\hat{\mathbf{\Sigma}}$ 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 $\hat{\mathbf{\Sigma}}$ so that the row and column corresponding to feature j come last, as in partition (2). Then solve the lasso problem

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

3. Fill in the corresponding row and column of $\hat{\mathbf{\Sigma}}$ 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 $\hat{\mathbf{\Sigma}}$ 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 $\hat{\mathbf{\Sigma}}$ to compute $\hat{\omega}_{22}$ for each feature and filling in the corresponding row of $\hat{\Omega}$ as in (2) using the formulae

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

Remark. Formula (4) is justified as follows: Banerjee et al. [2008] show that $\hat{\Sigma}_{12}$ satisfies

$$\hat{\Sigma}_{12} = \arg\min_{y} \left\{ y^T \hat{\mathbf{\Sigma}}^{-1} y : \|y - S_{12}\|_\infty \leq p \right\}. \tag{6}$$

Using strong duality, they then show that the dual problem (4) is equivalent; specifically, if $\hat{\beta}$ solves (4) then $\hat{\sigma}_{12} = \hat{\Sigma}_{11}\hat{\beta}$ solves (6).

Remark. Formulae (5) are justified as follows: from (3) 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

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \quad \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \tag{7}$$

as

$$\begin{aligned} &\underset{\beta \in \mathbb{R}^p, z \in \mathbb{R}^n}{\text{minimize}} \quad \frac{1}{2}\|y - z\|_2^2 + \lambda\|\beta\|_1 \\ &\text{subject to} \quad z = X\beta. \end{aligned} \tag{8}$$

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

$$\begin{aligned} &\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \quad \frac{1}{2}\|y - X\beta\|_2^2 \\ &\text{subject to} \quad \|\beta\|_1 \leq t \end{aligned} \tag{9}$$

for some $t > 0$. The Lagrangian of (8) 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

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

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

This minimization is separable:

$$= \inf_{z} \left\{ \frac{1}{2} \left(y^T y - 2 y^T z + z^T z \right) + u^T z \right\} + \inf_{\beta} \left\{ \lambda \|\beta\|_1 - u^T X \beta \right\} \tag{10}$$

We will handle each part of (10) separately. First, the left side:

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

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

$$z + (u - y) = 0 \implies z = y - u \tag{11}$$

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

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

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

Next we will minimize the right side of (10):

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

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

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

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

which is equivalent to the condition

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

Therefore the Lagrange dual function is

$$\inf_{\beta, z} \left\{ \mathcal{L}(x, u) \right\} = \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|y-u\|_2^2 \tag{12}$$

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

$$\begin{aligned} \underset{u \in \mathbb{R}^p}{\text{maximize}} \quad & \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|y-u\|_2^2 \\ \text{subject to} \quad & \|u^T X\|_\infty \leq \lambda. \end{aligned} \tag{13}$$

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

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

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

Then by (11) 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 (7). **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 (13) 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. (7) 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 (11) $\hat{u} = y - X\hat{\beta}$, 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 (7) is the Lagrangian of (9). We will prove a useful lemma about the relationship between these functions.

Lemma 2. For a given $\lambda > 0$, let $\hat{\beta}$ minimize (7). Then there is exactly one $t = \|\hat{\beta}\|_1$ such that any $\hat{\beta}$ minimizing (7) also minimizes (9).

Proof. This must be true by contradiction. First of all, since the objective function of (9) is continuous and the feasible region $\|\beta\|_1 \leq t$ is compact, a minimum of (9) is guaranteed to exist. Now suppose $\hat{\beta}$ minimizes (7) for a fixed λ, with $\|\hat{\beta}\|_1 = t$, but there is a different solution $\hat{\beta}^*$ that is feasible for (9) 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 (9) 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 (7).

□

Now the result follows in a simple way:

Proposition 3. Let $\mathcal{B}$ be the set of all $\hat{\beta}$ that minimize (7) 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 (7), and (without loss of generality) $\|\hat{\beta}_1\|_1 < \|\hat{\beta}_2\|_1$. By Lemma 2, these values both minimize (9) with $t = \|\hat{\beta}_2\|_1$ (we cannot choose $t = \|\hat{\beta}_1\|_1$ because $\hat{\beta}_1$ is not feasible for that problem). But $\hat{\beta}_1$ does not lie on the boundary of the feasible region of (9), so $\hat{\beta}_1$ minimizing (9) contradicts the convexity of (9) (and the fact that the global minimum lies outside the feasible region). Therefore $\|\hat{\beta}_1\|_1 = \|\hat{\beta}_2\|_1$. □

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

(c) (i) Since β^* is clearly feasible for (7) 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 (13) is a lower bound for the solution of the primal problem (7) for any u feasible for (13) (that is, any u satisfying $\|u^T X\|_\infty \leq \lambda$). Therefore we have

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

Since by assumption $\lambda \geq \|X^T\varepsilon\|_\infty$, ε is feasible for (13). Therefore we have

$$\frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\|\beta\|_1 \geq \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|y - \varepsilon\|_2^2 = \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|X\beta^*\|_2^2 \tag{14}$$

as desired.

(iii) We can rewrite the right side of (14) as

$$\frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|X\beta^*\|_2^2 = \frac{1}{2}\|X\beta^*\|_2^2 + \frac{1}{2}\|\varepsilon\|_2^2 + \varepsilon^T X\beta^* - \frac{1}{2}\|X\beta^*\|_2^2 = \frac{1}{2}\|\varepsilon\|_2^2 + \varepsilon^T X\beta^*. \tag{15}$$

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 (14) and using the identity in (15) 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 $\sigma^2 = \mathbb{E}\left[n^{-1}\|\varepsilon\|_2^2\right]$.

(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 ℓ_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

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. 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=377946{&}acc=ACTIVESERVICE{&}key=B63ACEF81C6334F5.C52804B674E616B8.4D4702B0C3E38B35.4D4702B0C3E38B35{&}{_}{_}acm{_}{_}=1559508442{_}3ad6c8548bc08a0694a945e1d7f5e5e5.

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.