Solution Sheet.

This exercise sheet contains problems from Parts IV and V of the course.
Please submit the exercises highlighted in red before the (extended) deadline.

Problem 1

Estimation of diagonal covariances: Let $(X_i)_{i=1,\ldots,n}$ be an i.i.d. sequence of d-dimensional vectors, drawn from a zero-mean distribution with diagonal covariance matrix $\Sigma = D$. Consider the estimate $\widehat{D} = \text{diag}(\widehat{\Sigma})$, where $\widehat{\Sigma}$ is the usual sample covariance matrix.

When each vector X_i is sub-Gaussian with parameter at most $\sigma = 1$, show that there are universal positive constants c_j $(j = 0, 1, 2)$ such that

$$\mathrm{P}\left(\|\widehat{D} - D\|_2 \geq c_0 \sqrt{\frac{\log(d)}{n}} + \delta \right) \leq c_1 \exp^{-c_2 n \min\left\{\delta, \delta^2\right\}}, \qquad \forall \delta > 0.$$

Solution.

It holds that $X_{ij} \in \text{SubGaussian}(1)$, so X_{ij}^2 is sub-exponential with parameters (2,4). Therefore, $\sum_{i=1}^n X_{ij}^2$ is sub-exponential with parameters $(2\sqrt{n}, 4)$. We know that the sub-exponential tail bound implies that

$$\mathrm{P}\left(|\widehat{D}_{ii} - D_{ii}| \geq t \right) \leq 2 \exp^{-\frac{n}{8} \min\left\{t, t^2\right\}}.$$

Then, by a union bound,

$$\mathrm{P}\left(\|\widehat{D} - D\|_2 \geq t \right) \leq 2d \exp^{-\frac{n}{8} \min\left\{t, t^2\right\}}.$$

Let's consider $\varepsilon \in [0, 1]$. Then,

$$\mathrm{P}\left(\|\widehat{D} - D\|_2 \geq t + \varepsilon \right) \leq 2d \exp^{-\frac{n}{8} \min\left\{t, t^2\right\} - \frac{n}{8} \varepsilon^2}.$$

If we set $\varepsilon = \sqrt{\frac{8 \log(d)}{n}}$ we obtain

$$\mathrm{P}\left(\|\widehat{D} - D\|_2 \geq t + \sqrt{\frac{8 \log(d)}{n}} \right) \leq 2 \exp^{-\frac{n}{8} \min\left\{t, t^2\right\}},$$

which is valid for $n \geq 8 \log(d)$. For $n < 8 \log(d)$, we similarly have

$$\mathrm{P}\left(\|\widehat{D} - D\|_2 \geq t + \frac{8 \log(d)}{n} \right) \leq 2 \exp^{-\frac{n}{8} \min\left\{t, t^2\right\}}.$$

Combining the two, for all $n \geq 1$, we have

$$\mathrm{P}\left(\|\widehat{D} - D\|_2 \geq t + \min\left\{ \frac{8 \log(d)}{n}, \sqrt{\frac{8 \log(d)}{n}} \right\} \right) \leq 2 \exp^{-\frac{n}{8} \min\left\{t, t^2\right\}}.$$

Problem 2

Sparsity is often characterized through matrix norms. The following relationships are helpful to get an idea of how $\|A\| \leq s$ imposes sparsity when A is an adjacency matrix.

For a rectangular matrix A with real entries and a scalar $q \in [1, \infty]$ define the operator norms as

$$\|A\|_q = \sup_{\|x\|_q = 1} \|Ax\|_q,$$

where for a vector x, $\|x\|_r = \sum_{i=1}^d |x_i|^r$. Recall further that $\|A\|_2^2 = \|A^T A\|_2 = \gamma_{\max}\left(A^T A\right)$ with $\gamma_{\max}$ denoting the largest eigenvalue.

(a) Prove that $\|AB\|_q \leq \|A\|_q \|B\|_q$ for any size-compatible matrices A and B.

(b) For a square matrix A, prove that $\|A\|_2^2 \leq \|A\|_1 \|A\|_\infty$. What happens when A is symmetric?

Note that with A being the adjacency matrix of a covariance matrix Σ, we have $\|A\| \leq s$ whenever Σ has at most s non-zero entries per row.

Solution.

(a) $\|ABx\|_q = \left\| A \frac{Bx}{\|Bx\|_q} \right\|_q \|Bx\|_q \leq \|A\|_q \|Bx\|_q$; the result thus follows by taking the supremum over $\|x\|_q = 1$.

(b) We know that $\|A\|_2^2 = \|A^T A\|_2 = \gamma_{\max}\left(A^T A\right)$. Let v be the eigenvector associated with $\gamma_{\max}\left(A^T A\right)$. By definition $\|A^T A v\|_1 = \gamma_{\max}\left(A^T A\right) \|v\|_1$, and, following the previous exercise, we have that

$$\gamma_{\max}\left(A^T A\right) \|v\|_1 \leq \|A^T A\|_1 \|v\|_1$$

, i.e., $\|A\|_2^2 \leq \|A^T A\|_1$ as $\|v\|_1 \neq 0$. Then,

$$\|A^T A\|_1 \leq \|A^T\|_1 \|A\|_1 = \|A\|_\infty \|A\|_1$$

Problem 3

In class we discussed the so-called hard-thresholding operator

$$T_\lambda(u) = u_{\{|u| > \lambda\}}.$$

There are other methods to apply thresholding. The so-called soft-thresholding operator is defined as follows

$$S_\lambda(u) = \text{sign}(u)(|u| - \lambda)_{\{|u| > \lambda\}},$$

where

$$\text{sign}(u) = \begin{cases} -1, & \text{if } x < 0, \\ 0, & \text{if } x = 0, \\ 1, & \text{if } x > 0. \end{cases}$$

(a) Suppose $\lambda = 1$ and "plot" (simply draw the plot by hand) both the hard- and soft thresholding operator as a function in u.

(b) Write down an estimator for the covariance matrix based on soft-thresholding.

(c) Based on parts (a) and (b), describe the differences between the two shrinkage methods.

(d) Describe in your own words how to choose λ using cross-validation for soft-thrsholding a covariance matrix.

Solution.

(a) see Figure 1 (a).

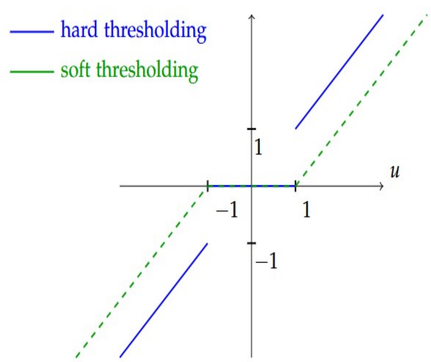

Figure 1: (a)

(b) $\widehat{\Sigma} = \frac{1}{T}\sum_{t=1}^{T}(X_t - \overline{\lambda})(X_t - \overline{\lambda})' = (\sigma_i j)_{i,j=1,\ldots,p}$.
$S_\lambda(\widehat{\Sigma}) = \big(\text{sign}(\widehat{\sigma_{ij}})(|\sigma_{ij}| - \lambda)\mathbb{1}_{|\sigma_{ij}|\geq\lambda}\big)_{i,j=1,\ldots,p}$.

(c) Hard thresholding sets all values whose absolute values are lower than the threshold equal to zero. Soft thresholding is an extension of hard thresholding, first setting to zero the elements whose absolute values are lower than the threshold and then shrinking the non-zero coefficients towards zero.

(d) see lecture notes and replace T_λ with S_λ.

The following two problems are in preparation of the next part of the lecture on "Large inverse covariance matrices".

Problem 4

Suppose we have a block matrix

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}$$

with $\mathbf{D}$ and $\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C}$ invertible. Show that the matrix $\mathbf{P}$ given below is the inverse of $\mathbf{M}$.

$$\mathbf{P}=\begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{bmatrix}^{-1}=\begin{bmatrix}\left(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right)^{-1} & -\left(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right)^{-1}\mathbf{B}\mathbf{D}^{-1}\\ -\mathbf{D}^{-1}\mathbf{C}\left(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right)^{-1} & \mathbf{D}^{-1}+\mathbf{D}^{-1}\mathbf{C}\left(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right)^{-1}\mathbf{B}\mathbf{D}^{-1}\end{bmatrix}.$$

Solution.

$$\begin{aligned}\mathbf{M}\mathbf{P}&=\begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{bmatrix}\begin{bmatrix}\left(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right)^{-1} & -\left(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right)^{-1}\mathbf{B}\mathbf{D}^{-1}\\ -\mathbf{D}^{-1}\mathbf{C}\left(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right)^{-1} & \mathbf{D}^{-1}+\mathbf{D}^{-1}\mathbf{C}\left(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right)^{-1}\mathbf{B}\mathbf{D}^{-1}\end{bmatrix}\\
&=\begin{bmatrix}\mathbf{A}\left(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right)^{-1}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C}\left(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right)^{-1} & \dots\\ \mathbf{C}\left(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right)^{-1}-\mathbf{C}\left(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right)^{-1} & \dots\end{bmatrix}\\
&=\begin{bmatrix}\left(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right)^{-1}\left(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right) & \dots\\ \mathbf{0} & \dots\end{bmatrix}=\begin{bmatrix}\mathbf{I} & \dots\\ \mathbf{0} & \dots\end{bmatrix}\\
&=\begin{bmatrix}\mathbf{I} & -\mathbf{A}\left(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right)^{-1}\mathbf{B}\mathbf{D}^{-1}+\mathbf{B}\mathbf{D}^{-1}+\mathbf{B}\mathbf{D}^{-1}\mathbf{C}\left(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right)^{-1}\mathbf{B}\mathbf{D}^{-1}\\ \mathbf{0} & -\mathbf{C}\left(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right)^{-1}\mathbf{B}\mathbf{D}^{-1}+\mathbf{I}+\mathbf{C}\left(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right)^{-1}\mathbf{B}\mathbf{D}^{-1}\end{bmatrix}\\
&=\begin{bmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{bmatrix}\end{aligned}$$

Problem 5

Suppose that $X=(X_1,\dots,X_p)'\sim\mathcal{N}(\mu,\Sigma)$. We separate X into two vectors $X(a)\in\mathbb{R}^r$ and $X(b)\in\mathbb{R}^s$ with $r+s=p$ such that

$$X=\begin{bmatrix}X_1\\ \vdots\\ X_p\end{bmatrix}=\begin{bmatrix}X(a)\\ X(b)\end{bmatrix}.$$

Suppose further that Σ has full rank such that the joint density is given by

$$f_{\mathbf{X}}(x_1,\dots,x_k)=\frac{1}{\sqrt{(2\pi)^k|\Sigma|}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right),$$

where $|\Sigma|$ denotes the determinant of Σ. Then,

$$\Sigma=\begin{bmatrix}\Sigma_{aa} & \Sigma_{ab}\\ \Sigma_{ba} & \Sigma_{bb}\end{bmatrix}$$

with

$$\Sigma_{ab}=\mathrm{Cov}(X(a),X(b)).$$

(a) Define

$$\widetilde{X}(b)=X(b)-\Sigma_{ba}\Sigma_{aa}^{-1}X(a)$$

and show that $\mathrm{Cov}(X(a),\widetilde{X}(b))=0$.

(b) Suppose $X(a)$ and $X(b)$ are uncorrelated and show that this implies independence. Independence can be shown by

$$f_X(x_1,\dots,x_p) = f_{X(a)}(x_1,\dots,x_r) * f_{X(b)}(x_{r+1},\dots,x_p).$$

Solution.

(a)

$$\begin{aligned}
\mathrm{Cov}(X(a),\widetilde{X}(b)) =& \mathrm{Cov}\left(X(a), X(b) - \Sigma_{ba}\Sigma_{aa}^{-1}X(a)\right)\\
=& \mathrm{Cov}\left(X(a),X(b)\right) - \mathrm{Cov}\left(X(a), \Sigma_{ba}\Sigma_{aa}^{-1}X(a)\right)\\
=& \mathrm{Cov}\left(X(a),X(b)\right) - \Sigma_{ba}\Sigma_{aa}^{-1}\mathrm{Cov}\left(X(a),X(a)\right)\\
=& \Sigma_{ab} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{aa} = \Sigma_{ab} - \Sigma_{ba} = 0,
\end{aligned}$$

where the last equality holds due to symmetry.

(b) $X(a)$ and $X(b)$ are uncorrelated $\Rightarrow$ $\mathrm{Cov}(X(a),X(b)) = \Sigma_{ab} = 0 \Rightarrow X(a) \sim \mathcal{N}(\mu_a,\Sigma_{aa})$, $X(b) \sim \mathcal{N}(\mu_b,\Sigma_{bb})$ with $\mu = (\mu_1,\dots,\mu_r,\mu_{r+1},\dots,\mu_p)'$ where $\mu_a' = (\mu_1,\dots,\mu_r)'$ and $\mu_b' = (\mu_1 r+1,\dots,\mu_p)'$. We obtain, with $x = (x_1,\dots,x_r,x_{r+1},\dots,x_p)'$ where $x_a' = (x_1,\dots,x_r)'$ and $x_b' = (x_1 r+1,\dots,x_p)'$,

$$\begin{aligned}
(x-\mu)'\Sigma^{-1}(x-\mu) =& (x(a)-\mu_a, x(b)-\mu_b)' \begin{pmatrix}\Sigma_{aa} & 0\\ 0 & \Sigma_{bb}\end{pmatrix} (x(a)-\mu_a, x(b)-\mu_b)\\
=& (x(a)-\mu_a, x(b)-\mu_b)' \begin{pmatrix}\Sigma_{aa}^{-1} & 0\\ 0 & \Sigma_{bb}^{-1}\end{pmatrix} (x(a)-\mu_a, x(b)-\mu_b)\\
=& (x(a)-\mu_a)'\Sigma_{aa}^{-1}(x(a)-\mu_a) - (x(b)-\mu_b)'\Sigma_{bb}^{-1}(x(b)-\mu_b).
\end{aligned}$$

Further, we have

$$|\Sigma| = \det\begin{pmatrix}\Sigma_{aa} & 0\\ 0 & \Sigma_{bb}\end{pmatrix} = \det(\Sigma_{aa})\det(\Sigma_{bb}) = |\Sigma_{aa}||\Sigma_{bb}|.$$

Putting this together, we obtain

$$\begin{aligned}
f_X(x_1,\dots,x_p) =& \frac{1}{\sqrt{(2\pi)^p|\Sigma|}}\exp\Big(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\Big)\\
=& \frac{1}{\sqrt{(2\pi)^{r+p}|\Sigma_{aa}||\Sigma_{bb}|}}\exp\Big(-\frac{1}{2}\big((x(a)-\mu_a)'\Sigma_{aa}^{-1}(x(a)-\mu_a)\\
&+ (x(b)-\mu_b)'\Sigma_{bb}^{-1}(x(b)-\mu_b)\big)\Big)\\
=& \frac{1}{\sqrt{(2\pi)^r|\Sigma_{aa}|}}\exp\left(-\frac{1}{2}(x(a)-\mu_a)'\Sigma_{aa}^{-1}(x(a)-\mu_a)\right)\\
&\times \frac{1}{\sqrt{(2\pi)^s|\Sigma_{bb}|}}\exp\left(-\frac{1}{2}(x(b)-\mu_b)'\Sigma_{bb}^{-1}(x(b)-\mu_b)\right)\\
=& f_{X(a)}(x_1,\dots,x_r)f_{X(b)}(x_{r+1},\dots,x_p).
\end{aligned}$$