Multivariate Statistics – Spring 2025
Solutions

Prof. Victor Panaretos
TA: Leonardo Santoro leonardo.santoro@epfl.ch

Solution 1. 1. Any eigenvector v must satisfy $\lambda v = \mathsf{Q}v$ for some $\lambda \in \mathbb{R}$. Since any projection is idempotent, we apply Q on both sides and obtain:

$$\lambda^2 v = \mathsf{Q}v = \lambda v.$$

which can only be true if λ is 0 or 1.

2. Note that $u = \mathsf{Q}v$ is the projection of v onto the subspace defined by the range of Q, while $w = (\mathsf{I} - \mathsf{Q})v$ is the projection of v onto the orthogonal complement of the range of Q, corresponding to its null space. Suppose there exists another pair of vectors u' and w' such that $v = u' + w'$ and $u' = \mathsf{Q}v$ and $w' = (\mathsf{I} - \mathsf{Q})v$. Note however $\mathsf{Q}(u' - u) = 0$, so that $u' - u \in \ker(\mathsf{Q})$ while $u, u' \in \text{range}(\mathsf{Q})$, so that it need be $u = u'$, and similarly for w, w'.

3. By the above, any vector u may be (uniquely) decomposed as $v + w$ with $v \in V$ and $w \in V^\perp$. Since P, Q are projections onto V, $\mathsf{Q}v = \mathsf{P}v = v$, while $\mathsf{Q}w = \mathsf{P}w = 0$, yielding the claim.

4. Note that Q_v is idempotent. Indeed:

$$\mathsf{Q}_v^2 w = \mathsf{Q}_v \left((v^\top w)v/(v^\top v)\right) = (v^\top w) \cdot (v^\top v)/(v^\top v)v = (v^\top w)v = \mathsf{Q}_v w.$$

Furthermore, it is symmetric:

$$(\mathsf{Q}_v w)^\top u = \frac{1}{v^\top v}(v^\top w)v^\top u = w^\top \mathsf{Q}_v u.$$

This proves that Q_v is indeed a projection. To conclude that it corresponds to the projection onto the span of v we may use the part 3.

Solution 2. 1. Denote the SVD of A as $\mathsf{A} = \mathsf{U}\Sigma\mathsf{V}^\top$, where U is an orthogonal matrix, Σ is a diagonal matrix containing the singular values of A, and V is an orthogonal matrix.

Since A is assumed to be invertible, it means that all the singular values in Σ are non-zero. Therefore, Σ can be inverted by taking the reciprocal of each non-zero singular value:

$$\Sigma^{-1} = \text{diag}\left(\frac{1}{\sigma_1}, \frac{1}{\sigma_2}, \ldots, \frac{1}{\sigma_r}\right)$$

.

Now, the inverse of A can be expressed using the SVD as:

$$\mathsf{A}^{-1} = \mathsf{V}\Sigma^{-1}\mathsf{U}^\top$$

.

2. The determinant of A can be expressed in terms of its SVD as:

$$\det(\mathsf{A}) = \det(\mathsf{U}\Sigma\mathsf{V}^\top) = \det(\mathsf{U})\det(\Sigma)\det(\mathsf{V})^\top)$$

Since U and V are orthogonal matrices, their determinants are either 1 or -1. Hence, we have:

$$|\det(\mathsf{A})| = |\det(\Sigma)|$$

Now, the determinant of a diagonal matrix is the product of its diagonal elements. Therefore, we have:

$$|\det(\Sigma)| = \prod_{i=1}^{n} \sigma_i$$

where σ_i are the singular values of A.

3. Let us consider the SVD decomposition $\mathsf{A} = \mathsf{U}\Sigma\mathsf{V}^\top$, where:

- the singular values are equal to the square roots of the eigenvalues of $\mathsf{A}\mathsf{A}^\top$ (or, equivalently, $\mathsf{A}^\top\mathsf{A}$).
- the right singular vectors (columns of V) are eigenvectors of $\mathsf{A}^\top\mathsf{A}$.
- the left singular vectors (columns of U) are eigenvectors of $\mathsf{A}\mathsf{A}^\top$.

If A is real symmetric then (spectral theorem) it has at least one eigendecomposition $\mathsf{A} = \mathsf{Q}\Lambda\mathsf{Q}^\top$, and its singular values are the absolute values of its eigenvalues. Furthermore, in such case, both the right and left singular vectors (columns of V and U) are eigenvectors of $\mathsf{A}^2 = \mathsf{Q}\Lambda^2\mathsf{Q}^\top$, so they are both eigenvectors of A, and thus equal to vectors in Q up to sign. Therefore, if A is real symmetric and positive definite, Σ is a diagonal matrix containing the eigenvalues, and $\mathsf{U} = \mathsf{V}$.

Note that this argument fails if A is only semi-positive definite! Indeed, in this case,the part of U and V corresponding to the zero eigenvalues can be any orthonormal decomposition of the null space of A, with sign flips allowed independently on U and V.

Solution 3. Let us consider the eigenvalue decomposition of $\mathsf{P} - \mathsf{Q}$:

$$\mathsf{P} - \mathsf{Q} = V\Lambda V^\top$$

where V is the matrix of eigenvectors and Λ is a diagonal matrix of eigenvalues of $\mathsf{P} - \mathsf{Q}$. Since $\mathsf{P} - \mathsf{Q}$ is positive definite, all eigenvalues of $\mathsf{P} - \mathsf{Q}$ are positive. Therefore, all the diagonal elements of Λ are positive. Now let us consider the trace of $\mathsf{P} - \mathsf{Q}$:

$$\operatorname{tr}(\mathsf{P} - \mathsf{Q}) = \operatorname{tr}(V\Lambda V^\top)$$

Using the cyclical property of trace, $\operatorname{tr}(ABC) = \operatorname{tr}(CAB)$, we can rewrite the above expression as:

$$\operatorname{tr}(\mathsf{P} - \mathsf{Q}) = \operatorname{tr}(V^\top V\Lambda) = \operatorname{tr}(\Lambda)$$

Since Λ is a diagonal matrix, the trace of Λ is the sum of its diagonal elements, which are the eigenvalues of $\mathsf{P} - \mathsf{Q}$. Therefore, $\operatorname{tr}(\mathsf{P} - \mathsf{Q}) = \sum \lambda_i$, where λ_i are the eigenvalues of $\mathsf{P} - \mathsf{Q}$. Since all eigenvalues of $\mathsf{P} - \mathsf{Q}$ are positive, $\sum \lambda_i$ is positive. Furthermore, by linearity of the trace:

$$\operatorname{tr}(\mathsf{P}) - \operatorname{tr}(\mathsf{Q}) = \operatorname{tr}(\mathsf{P} - \mathsf{Q}) > 0$$

and hence $\operatorname{tr}(\mathsf{P}) > \operatorname{tr}(\mathsf{Q})$.

Solution 4. 1. $\mathsf{A}^\top = (\mathsf{P}^\top\mathsf{P})^\top = \mathsf{P}^\top\mathsf{P} = \mathsf{A}$, which shows symmetry; furthermore, for any vector v, we have $v^\top\mathsf{A}v = v^t\mathsf{P}^\top\mathsf{P}v = \|\mathsf{P}v\|^2 \geq 0$.

2. If $\lambda v = \mathsf{A}v$, then multiplying both sides by $v^\top$ we obtain: $\lambda = \|\mathsf{P}v\|^2/\|v\|^2 \geq 0$.

3. By the spectral theorem, we may write $\mathsf{A} = \sum_{i=1}^{r} \lambda_i u_i u_i^\top$.We denote $\mathsf{B} = \sum_{i=1}^{r} 1/\lambda_i u_i u_i^T$. We need to show that:

   - $\mathsf{BAB} = \mathsf{B}$
   - $\mathsf{ABA} = \mathsf{A}$

   We prove the first equation – the second is shown almost identically.

$$\mathsf{BAB} = \left(\sum_{i=1}^{r} \frac{1}{\lambda_i} u_i u_i^\top\right)\left(\sum_{j=1}^{r} \lambda_j u_j u_j^\top\right)\left(\sum_{k=1}^{r} \frac{1}{\lambda_k} u_k u_k^\top\right)$$

   Since each $u_i u_i^\top$ is the orthogonal projectors onto u_i and the u_i are linearly independent – because eigenvectors corresponding to different eigenvalues – we have that $u_i u_i^\top u_j = \delta_{ij} u_i$, so that it is easy to see that the above product can be reduced to $\sum_{i=1}^{r} \frac{1}{\lambda_i} u_i u_i^\top = \mathsf{B}$.

Solution 5.

$$\begin{aligned}
\underset{\mathsf{X}\,:\,\mathsf{X}\mathsf{X}^\top = \mathsf{I}}{\arg\min} \|\mathsf{XA} - \mathsf{B}\|_F^2 &= \underset{\mathsf{X}\,:\,\mathsf{X}\mathsf{X}^\top = \mathsf{I}}{\arg\min} \langle \mathsf{XA} - \mathsf{B}, \mathsf{XA} - \mathsf{B}\rangle_F \\
&= \underset{\mathsf{X}\,:\,\mathsf{X}\mathsf{X}^\top = \mathsf{I}}{\arg\min} \|\mathsf{XA}\|_F^2 + \|\mathsf{B}\|_F^2 - 2\langle \mathsf{XA}, \mathsf{B}\rangle_F \\
&= \underset{\mathsf{X}\,:\,\mathsf{X}\mathsf{X}^\top = \mathsf{I}}{\arg\min} \|\mathsf{A}\|_F^2 + \|\mathsf{B}\|_F^2 - 2\langle \mathsf{XA}, \mathsf{B}\rangle_F \\
&= \underset{\mathsf{X}\,:\,\mathsf{X}\mathsf{X}^\top = \mathsf{I}}{\arg\max} \langle \mathsf{XA}, \mathsf{B}\rangle_F \\
&= \underset{\mathsf{X}\,:\,\mathsf{X}\mathsf{X}^\top = \mathsf{I}}{\arg\max} \left\langle \mathsf{X}, \mathsf{B}\mathsf{A}^T\right\rangle_F \\
&= \underset{\mathsf{X}\,:\,\mathsf{X}\mathsf{X}^\top = \mathsf{I}}{\arg\max} \left\langle \mathsf{X}, \mathsf{U}\Sigma\mathsf{V}^T\right\rangle_F \\
&= \underset{\mathsf{X}\,:\,\mathsf{X}\mathsf{X}^\top = \mathsf{I}}{\arg\max} \left\langle \mathsf{U}^\top\mathsf{X}\mathsf{V}, \Sigma\right\rangle_F \\
&= \underset{\mathsf{X}\,:\,\mathsf{X}\mathsf{X}^\top = \mathsf{I}}{\arg\max} \langle \mathsf{U}^\top\mathsf{X}\mathsf{V}, \Sigma\rangle_F
\end{aligned}$$

The matrix $\mathsf{U}^\top\mathsf{X}\mathsf{V}$ is an orthogonal matrix, (as it is a product of orthogonal matrices). Therefore, the expression is maximised when this product equals the identity, i.e. when $\mathsf{X} = \mathsf{U}\mathsf{V}^\top$. Indeed, if D is diagonal and W is orthogonal, then:

$$\langle W, D\rangle = \operatorname{tr}(DW) = \sum_i e_i^\top DW e_i = \sum_i e_i D f_i = \sum_i \sum_j a_{ij} e_i^\top D e_j = \sum_i a_{ii} D_{ii}$$

where e_i is some orthogonal basis, and $f_i = We_i = \sum_j a_{ij} e_j$ another orthogonal basis; observing that $|a_{ii}| \leq 1$ and is maximised (equal to 1) only when $f_i = e_i$, i.e. when W is the identity, we conclude.

Solution 6. Solution in Python and R: see this GitHub link.

Solution 7. 1. We first show linearity, wlog in the first argument. The (k, l)-th block of $(A+B)\otimes C$ is

$$\begin{aligned}
(A + B)_{kl} C &= (A_{kl} + B_{kl})\, C \\
&= A_{kl} C + B_{kl} C
\end{aligned}$$

   We then show associativity. Let A be $K \times L$, B be $M \times N$ and C be $O \times P$. Let us first study the structure of $(A \otimes B) \otimes C$. The product $A_{kl} B_{mn}$ is the

$$((k-1)M + m, (l-1)N + n)\text{-th}$$

entry of $A \otimes B$. As a consequence, the product $(A_{kl}B_{mn})\,C_{op}$ is the

$$(((k-1)M+m-1)O+o, ((l-1)N+n-1)P+p)\text{-th}$$

entry of $(A \otimes B) \otimes C$. Let us now study the structure of $A \otimes (B \otimes C)$. The product $B_{mn}C_{op}$ is the

$$((m-1)O+o, (n-1)P+p)\text{-th}$$

entry of $B \otimes C$. Therefore, the product $A_{kl}\,(B_{mn}C_{op})$ is the entry of $A \otimes (B \otimes C)$ that occupies position

$$\begin{aligned}&((k-1)MO+(m-1)O+o, (l-1)NP+(n-1)P+p)\\ =&(((k-1)M+m-1)O+o, ((l-1)N+n-1)P+p)\end{aligned}$$

Thus, the product $A_{kl}B_{mn}C_{op}$ occupies the same position in $(A \otimes B) \otimes C$ and in $A \otimes (B \otimes C)$ for every k, l, m, n, o and p. Therefore,

$$(A \otimes B) \otimes C = A \otimes (B \otimes C)$$

Since $A_{kl}C$ is the (k,l)-th block of $A \otimes C$ and $B_{kl}C$ is the (k,l)-th block of $B \otimes C$, and the above equality holds for every k and l, the claim is true.

To show that it is not commutative it suffices to establish a counterexample. For instance, let us take

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

$$B = \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix}$$

$$A \otimes B = \begin{pmatrix} 5 & 6 & 10 & 12 \\ 7 & 8 & 14 & 16 \\ 15 & 18 & 20 & 24 \\ 21 & 24 & 28 & 32 \end{pmatrix}$$

while

$$B \otimes A = \begin{pmatrix} 5 & 10 & 6 & 12 \\ 15 & 20 & 18 & 24 \\ 7 & 14 & 8 & 16 \\ 21 & 28 & 24 & 32 \end{pmatrix}$$

2. Suppose A is $K \times L$ and C is $L \times M$.

$$\begin{aligned}&(A \otimes B)(C \otimes D)\\ =&\begin{bmatrix} A_{11}B & \dots & A_{1L}B \\ \vdots & \ddots & \vdots \\ A_{K1}B & \dots & A_{KL}B \end{bmatrix}\begin{bmatrix} C_{11}D & \dots & C_{1M}D \\ \vdots & \ddots & \vdots \\ C_{L1}D & \dots & C_{LM}D \end{bmatrix}\\ =&\begin{bmatrix} \left(\sum_{l-1}^{L} A_{1l}C_{l1}\right)BD & \dots & \left(\sum_{l-1}^{L} A_{1l}C_{lM}\right)BD \\ \vdots & \ddots & \vdots \\ \left(\sum_{l-1}^{L} A_{Kl}C_{l1}\right)BD & \dots & \left(\sum_{l=1}^{L} A_{Kl}C_{lM}\right)BD \end{bmatrix}\\ =&\begin{bmatrix} (AC)_{11}BD & \dots & (AC)_{1M}BD \\ \vdots & \ddots & \vdots \\ (AC)_{K1}BD & \dots & (AC)_{KM}BD \end{bmatrix}\\ =&(AC) \otimes (BD)\end{aligned}$$

where we have used the fact that the multiplication of two block matrices can be carried out as if their blocks were scalars, and the definition of matrix multiplication to deduce that

$$(AC)_{km} = \sum_{l=1}^{L} A_{kl} C_{lm}.$$

3. Trivial using 2.

4. *optional*

5. Simply note that:

$$\begin{aligned}
\operatorname{tr}(A \otimes B) &= \operatorname{tr}\left(\begin{bmatrix} A_{11}B & \dots & A_{1K}B \\ \vdots & \ddots & \vdots \\ A_{K1}B & \dots & A_{KK}B \end{bmatrix} \right) \\
&= \sum_{k=1}^{K} \operatorname{tr}\left(A_{kk}B\right) \\
&= \sum_{k=1}^{K} A_{kk} \operatorname{tr}(B) \\
&= \left(\sum_{k=1}^{K} A_{kk} \right) \operatorname{tr}(B) \\
&= \operatorname{tr}(A) \operatorname{tr}(B)
\end{aligned}$$

Solution 8. 1. We first show that if $\mathsf{A} \succeq 0$ then $u \in \ker(\mathsf{A}) \Leftrightarrow u^\top \mathsf{A} u = 0$. The left-to-right implication is trivial; for the other direction, let v_i be a basis of eigenvectors for A with corresponding eigenvalues $\lambda_i \geq 0$. Then we may write $u \sum (u^\top v_i) v_i$ and:

$$u^\top \mathsf{A} u = \sum_i (u^\top v_i)^2 \lambda_i$$

so that if the right-hand-side is null, it need be that $u \perp v_i$ for all i such that $\lambda_i > 0$, proving that $u \in \ker(\mathsf{A})$

Then to show 1. we proceed as follows. Let $0 \neq x \in \ker(\mathsf{Q})$. Then:

$$0 \leq x^\top (\mathsf{Q} - \mathsf{P}) x = -x^\top \mathsf{P} x \leq 0$$

so that it need be:

$$x^\top \mathsf{P} x = 0$$

hence $x \in \ker(\mathsf{P})$ by the remark above. This shows $\ker(\mathsf{Q}) \subset \ker(\mathsf{P})$, and thus $\operatorname{Range}(\mathsf{P}) \subset \operatorname{Range}(\mathsf{Q})$.

An alternative solution is to use part 2. in this exercise, together with exercise 10. Or, to write $Q = P + (Q - P)$, and use that $\operatorname{Range}(\mathsf{A} + \mathsf{B}) \supset \operatorname{Range}(\mathsf{A})$ when A, B are positive-semidefinite.

2. It suffices to show there exists $c > 0$ such that:

$$\forall\, v \in \operatorname{Range}(\mathsf{P}) \ : c \cdot v^\top \mathsf{Q} v \geq v^\top \mathsf{P} v.$$

If finite, one obvious choice that would work is:

$$\sup_{v \in \operatorname{Range}(\mathsf{P})} \frac{v^\top \mathsf{P} v}{v^\top \mathsf{Q} v}$$

which by the inclusion assumption we can upper bound by:

$$\sup_{v \in \text{Range}(\mathsf{Q})} \frac{v^\top \mathsf{P} v}{v^\top \mathsf{Q} v}$$

which can be upper bounded by

$$0 < \lambda_{\max}(\mathsf{P}) / \lambda_{\min,+}(\mathsf{Q}) < \infty$$

where $\lambda_{\max}(\mathsf{P})$ denotes the maximum eigenvalue of P and $\lambda_{\min,+}(\mathsf{Q})$ denotes the smallest *non zero* eigenvalue of Q.

Solution 9. One direction trivially follows from the spectral theorem. Let us show the other implication, and consider $\mathsf{Q} = \sum_{i=1}^m q_i q_i^\top$ for some vectors $q_1, \ldots, q_m$. Since the sum of PSD matrices is PSD, it suffices to show that the matrix $q_1 q_1^\top$ is PSD. Clearly it is symmetric. Furthermore, for any vector v:

$$v^\top (q_1 q_1^\top) v = (v^\top q_1)^2 \geq 0$$

proving its definiteness.

Solution 10. Let $\mathsf{Q} \in \mathbb{R}^{p \times p}$ be symmetric.

(1) $\Rightarrow$ (2): If $\mathsf{Q}v = \lambda v$ for some $v \neq 0$, then

$$v^\top \mathsf{Q} v = v^\top (\lambda v) = \lambda v^\top v \geq 0.$$

Since $v^\top v > 0$, it follows that $\lambda \geq 0$.

(2) $\Rightarrow$ (1): Since Q is symmetric, it admits an orthonormal eigenbasis $\{v_i\}$ with eigenvalues $\lambda_i \geq 0$. Any $x \in \mathbb{R}^p$ can be written as $x = \sum_i \alpha_i v_i$, so

$$x^\top \mathsf{Q} x = \sum_i \lambda_i \alpha_i^2 \geq 0.$$

Solution 11. The right-to-left implication is easy: indeed, if A, B are PSD, then $\text{Range}(\mathsf{A}) \subset \text{Range}(\mathsf{A}+\mathsf{B})$. So, $\forall\, i = 1, \ldots m \;:\; v_i \in \text{Range}(v_i v_i^\top) \subset \text{Range}(\mathsf{Q})$. We move to the other implication. Let x be in the range of $\left(\sum_{i=1}^m v_i v_i^\top\right)$, i.e.:

$$x = \left(\sum_{i=1}^m v_i v_i^\top\right) y$$

for some y. However, since each $v_i \in \text{Range}(\mathsf{Q})$, we have that $v_i = \mathsf{Q} w_i$, so that we may write:

$$x = \left(\sum_{i=1}^m \mathsf{Q} w_i w_i^\top Q\right) y = Q \left(\sum_{i=1}^m w_i w_i^\top Q y\right) = Qz$$

which shows $x \in \text{Range}(\mathsf{Q})$.

Solution 12. Consider a bivariate Pareto density:

$$f(x, y) = c(x + y - 1)^{-p-2}, \quad \text{for } x, y > 1, \text{ and } p > 2.$$

1. We have that

$$1 = \int_1^\infty \int_1^\infty c(x + y - 1)^{-p-2} dx dy = \int_1^\infty \frac{c}{p+1} y^{-p-1} dy = \frac{c}{p(p+1)},$$

so $c = p(p+1)$.

2. By integration, we obtain
$$f(x) = \int_1^\infty f(x,y)dy = px^{-p-1},$$
for $x > 1$, and the same for $f(y)$.
The calculation of the Ebbected values gives $\mathbb{E}X = \mathbb{E}Y = \int_1^\infty xpx^{-p-1}dx = \frac{p}{p-1}$.

3. We first compute $\mathbb{E}XY$:
$$\begin{aligned}
\mathbb{E}XY &= \int_1^\infty \int_1^\infty cxy(x+y-1)^{-p-2}dxdy \\
&= \int_1^\infty (\frac{1}{-p-1}xy(x+y-1)^{-p-1}\Big|_1^\infty)dy + \int_1^\infty \int_1^\infty c\frac{y}{p+1}(x+y-1)^{-p-1}dxdy \\
&= \int_1^\infty (\frac{c}{p+1}y^{-p}dy + \int_1^\infty c(\frac{y}{-p(p+1)}(x+y-1)^{-p}\Big|_1^\infty)dy \\
&= \frac{p}{p-1} + c\int_1^\infty \frac{y^{-p+1}}{p(p+1)}dy \\
&= \frac{p}{p-1} + \frac{1}{p-2} = \frac{p^2-p-1}{(p-1)(p-2)}.
\end{aligned}$$

We then obtain
$$\mathrm{Cov}(X,Y) = \mathbb{E}XY - \mathbb{E}X\mathbb{E}Y = \frac{p^2-p-1}{(p-1)(p-2)} - \frac{p^2}{(p-1)^2} = \frac{1}{(p-1)^2(p-2)}.$$
We now compute
$$\mathbb{E}X^2 = \int_1^\infty x^2 px^{-p-1}dx = \frac{p}{p-2},$$
and finally
$$\mathrm{Var}X = \mathbb{E}X^2 - \mathbb{E}X^2 = \frac{p}{(p-1)^2(p-2)}.$$
We obtain the variance of Y by following the same steps: $\mathrm{Var}Y = \frac{p}{(p-1)^2(p-2)}$. We find that
$$\Sigma = \frac{1}{(p-1)^2(p-2)}\begin{pmatrix} p & 1 \\ 1 & p \end{pmatrix}.$$

4. The log-likelihood is:
$$L = n\log p + n\log(p+1) + \sum_{i=1}^n (-p-2)\log(x_i+y_i-1).$$
By derivation with respect to p we have
$$\frac{n}{p} + \frac{n}{p+1} - \sum_{i=1}^n \log(x_i+y_i-1) = 0,$$
and by setting $\bar{\alpha} = \frac{1}{n}\sum_{i=1}^n \log(x_i+y_i-1)$ we have the equation
$$p^2\bar{\alpha} + p(\bar{\alpha}-2) - 1 = 0.$$
The product of the roots of this equation $-(\bar{\alpha})^{-1}$ is negative. So we only consider the positive root, $\hat{p} = \frac{1}{\bar{\alpha}} - \frac{1}{2} + \sqrt{\frac{1}{\bar{\alpha}^2} + \frac{1}{4}}$.

Solution 13. 1. Let $\mu = \mathrm{E}(X)$. Then:

$$\mathrm{E}(Y) = \mathrm{E}(AX) = A\mathrm{E}(X) = A\mu$$

and we get:

$$\mathrm{Cov}(Y) = \mathrm{E}[(A(X-\mu))(A(X-\mu))^\top]A\mathrm{Cov}(X)A^\top = A\Sigma A^\top.$$

2. If $\Sigma = \mathbb{E}[XX^\top]$ then $u^\top \Sigma u = \mathbb{E}[u^\top XX^\top u] = \mathbb{E}[(X^\top u)^\top X^\top u] = \mathbb{E}\|X^\top u\|^2 \geq 0$. Fot the other direction, it suffices to take Z to be a random vector with covariance the identity matrix, and pick $X = \sqrt{\Sigma}Z$, where the square root exists because Σ is psd.

Solution 14.

$$E\left[\bar{X}\right] = E\left[\frac{1}{n}\sum_i X_i\right] = \frac{1}{n}\sum_i E\left[X_i\right] = \frac{1}{n}\sum_i \mu = \mu.$$

$$\begin{aligned} S &= \frac{1}{n}\sum_i X_i X_i^\top - \bar{X}\bar{X}^\top \\ &= \frac{1}{n}\sum_i (X_i-\mu)(X_i-\mu)^\top - (\bar{X}-\mu)(\bar{X}-\mu)^\top \\ &= \left(\frac{1}{n} - \frac{1}{n^2}\right)\sum_i (X_i-\mu)(X_i-\mu)^\top - \frac{1}{n^2}\sum_{i\neq j}(X_i-\mu)(X_j-\mu)^\top. \end{aligned}$$

Since $E\left[(X_i-\mu)(X_j-\mu)^\top\right] = 0$ for $i \neq j$, we have

$$E[S] = \frac{n-1}{n}\Sigma.$$

Hence, S is a biased estimate of Σ. If we consider $\tilde{S} = \frac{n}{n-1}S$ then $E[\tilde{S}] = \Sigma$.

Solution 15. The density of $\left(X^\top, Y^\top\right)^\top$ can be written in terms of $\mu = (\mu_X, \mu_Y)$ and the precision matrix $\Sigma^{-1} = \Psi = \begin{pmatrix} \Psi_{11} & \Psi_{12} \\ \Psi_{21} & \Psi_{22} \end{pmatrix}$. We are interested in the density of $X \mid Y = y$ which is given by

$$f_{X|Y=y}(x) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

Note that the denominator carries no information regarding the functional dependence on the variable x. In the following we omit the terms which don't depend on x for the sake of simplicity (the symbol $\overset{x}{\propto}$ means 'proportional to')

$$\begin{aligned} f_{X|Y=y}(x) &= \frac{f_{X,Y}(x,y)}{f_Y(y)} \\ &\overset{x}{\propto} \exp\left(-\frac{1}{2}\left(x-\mu_X\right)^\top \Psi_{11}\left(x-\mu_X\right) - \left(x-\mu_X\right)^\top \Psi_{12}\left(y-\mu_Y\right)\right) \\ &\overset{x}{\propto} \exp\left(-\frac{1}{2}x^\top\Psi_{11}x + x^\top\left(\Psi_{11}\mu_X - \Psi_{12}\left(y-\mu_Y\right)\right)\right), \end{aligned}$$

We recognize the form of a multivariate Gaussian density. We deduce that $X \mid Y = y$ is Gaussian, so it is characterized by two parameters: a mean $\mu_{1|2}$ and its precision matrix $\Psi_{1|2}$ which we need to identify. In general, the density of a multivariate Gaussian with mean $\mu_{1|2}$ and its precision matrix $\Psi_{1|2}$ is proportional to

$$\exp\left(-\frac{1}{2}\left(x-\mu_{1|2}\right)^\top \Psi_{1|2}\left(x-\mu_{1|2}\right)\right) \overset{x}{\propto} \exp\left(-\frac{1}{2}x^\top\Psi_{1|2}x + x^\top\Psi_{1|2}\mu_{1|2}\right)$$

By identification, the precision matrix $\Psi_{1|2}$ is Ψ_{11} and the mean is

$$\mu_X - \Psi_{11}^{-1}\Psi_{12}\left(y - \mu_Y\right)$$

Solution 16. We set $U = Y - \alpha - \beta X$. We will show that X and U are independent using the characteristic function.

$$\begin{aligned}
\Phi_{(X^\top,U^\top)^\top}\begin{pmatrix}t_1\\t_2\end{pmatrix} &= \mathbb{E}\left[\exp\left(\mathrm{i}\begin{pmatrix}t_1 & t_2\end{pmatrix}\begin{pmatrix}X\\U\end{pmatrix}\right)\right] \\
&= \int\int \exp\left[\mathrm{i}t_1x + \mathrm{i}t_2u\right] f_{X,U}(x,u)dxdu \\
&= \int\int \exp\left[\mathrm{i}t_1x + \mathrm{i}t_2u\right] f_{X,Y}(x,u+\beta x+\alpha)dxdu \\
&= \int\int \exp\left[\mathrm{i}t_1x + \mathrm{i}t_2u\right]) f_{Y|X=x}(u+\alpha+\beta x) f_X(x)dxdu \\
&\propto \int\int \exp\left[\mathrm{i}t_1x + \mathrm{i}t_2u\right] \mathbb{E}\left(-\frac{1}{2}u^\top\Sigma^{-1}u\right)\exp\left(-\frac{1}{2}(x-\mu_X)^\top\Sigma_X^{-1}(x-\mu_X)\right)dxdu \\
&\propto \int \exp\left[\mathrm{i}t_1x - \frac{1}{2}(x-\mu_X)^\top\Sigma_X^{-1}(x-\mu_X)\right]dx \int \exp\left[\mathrm{i}t_2u - \frac{1}{2}u^\top\Sigma^{-1}u\right]du
\end{aligned}$$

From line (2) to (3), we have considered the transformation

$$\begin{pmatrix}X\\U\end{pmatrix} = \begin{pmatrix}I_p & 0_{p\times q}\\-\beta & I_q\end{pmatrix}\begin{pmatrix}X\\Y\end{pmatrix} + \begin{pmatrix}0_p\\-\alpha\end{pmatrix},$$

for which the determinant of the Jacobian is 1.

We then obtain

$$\Phi_{(X^\top,U^\top)^\top}\begin{pmatrix}t_1\\t_2\end{pmatrix} = \Phi_X(t_1)\Phi_U(t_2)$$

So U and X are independent and the variance of $(X^\top, U^\top)^\top$ is

$$\mathrm{Var}\begin{pmatrix}X\\U\end{pmatrix} = \begin{pmatrix}\Sigma_X & 0_{p\times q}\\0_{q\times p} & \Sigma\end{pmatrix}.$$

Now, we write

$$\begin{pmatrix}X\\Y\end{pmatrix} = \begin{pmatrix}I_p & 0_{p\times q}\\\beta & I_q\end{pmatrix}\begin{pmatrix}X\\U\end{pmatrix} + \begin{pmatrix}0_p\\\alpha\end{pmatrix}.$$

Hence

$$\mathrm{E}\begin{pmatrix}X\\Y\end{pmatrix} = \begin{pmatrix}\mu_X\\\alpha+\beta\mu_X\end{pmatrix}$$

and

$$\mathrm{Var}\begin{pmatrix}X\\Y\end{pmatrix} = \begin{pmatrix}\Sigma_X & \Sigma_X\beta^\top\\\beta\Sigma_X & \Sigma+\beta\Sigma_X\beta^\top\end{pmatrix}$$

Solution 17. Solution in Python and R: see this GitHub link.

Solution 18. 1. For instance:

$$\begin{aligned}
M_{Z_1}(t) &= \mathbb{E}[e^{t^\top(X-Y)}] \\
&= \mathbb{E}[e^{t^\top X}e^{-t^\top Y}] \\
&= \mathbb{E}[e^{t^\top X}]\mathbb{E}[e^{-t^\top Y}] \quad \text{(since } X \text{ and } Y \text{ are independent)} \\
&= M_X(t)M_X(-t) \quad \text{(using the MGF of } X \text{ and } Y\text{)} \\
&= \exp\left(t^\top\Sigma t\right)
\end{aligned}$$

which is a the MGF of $\mathcal{N}(m, 2\Sigma)$. Proceeding similarly for Z_2 we see that its distribution is $\mathcal{N}(0, 2\Sigma)$.

2. One possibility is to observe that:

$$\begin{aligned} M_{Z_1+Z_2}(t) &= \mathbb{E}[e^{t^\top (2X)}] \\ &= \exp\left(2t^\top m + 2t^\top \Sigma t\right) \\ &= \exp\left(t^\top \Sigma t\right) \cdot \exp\left(2t^\top m + t^\top \Sigma t\right) \\ &= M_{Z_1}(t) M_{Z_2}(t) \end{aligned}$$

yielding their independence.

An alternative solution *not* using the MGF is to note that

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} 1, & -1 \\ 1, & 1 \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}$$

is a jointly gaussian vector, and that $\mathrm{Cov}(Z_1, Z_2) = 0$.

Solution 19. 1. Note that $Z^\top Z = \sum_{i=1}^k Z_i^2$, and the Z_i are i.i.d. standard gaussians, so we may deduce by classical univariate statistics that the sum of squares follows a chi-squared distribution with k degrees of freedom.

2. Using independence of the components Z_i and the MGF of the univariate chi-squared with 1 degree of freedom: $M_{Z^\top Z}(t) = M_{\sum_{i=1}^k Z_i^2}(t) = \prod_{i=1}^k M_{Z_i^2}(t) = \left((1-2t)^{-1/2}\right)^k$

Solution 20. 1. Let $Y = \Sigma^{-1/2} X$. Since Σ is invertible, $\Sigma^{-1/2}$ exists, and Y follows a multivariate normal distribution with mean 0 and covariance matrix I_p. Therefore, $Y_i^2 \sim \chi_1^2$ for $i = 1, \ldots, p$, s and the sum of independent chi-squared variables follows a chi-squared distribution with degrees of freedom equal to the sum of the individual degrees of freedom.

2. Employing the spectral decomposition theorem, we may write $H = UJU^\top$ where J is a diagonal matrix with only zeros and ones, with r non-zero entries. Then:

$$X^\top H X = (U^\top X)^\top J (U^\top X)$$

Note that $Y = U^\top X$ follows the same law as X since U is orthogonal. Therefore:

$$X^\top H X = Y^\top J Y = \sum_{i\,:\,J_{ii} \neq 0} Y_i^2$$

and the conclusion is clear, noting that $|i \ : J_{ii} \neq 0| = \mathrm{rank}(H)$.

3. Denote the eigenvalue decomposition of Σ as $\Sigma = UDU^\top$, where D is a diagonal matrix with the non-zero eigenvalues of Σ on the diagonal.

Consider the transformation $Y = U^\top X$. Then $Y \sim N(0, U^\top \Sigma U) = N(0, D)$. That is, Y_i are independent centered Gaussians with variance $D_{i,i}$. Then:

$$X^\top \Sigma^\dagger X = Y^\top D^\dagger Y = \sum_{i\,:\,D_{ii} \neq 0} D_{ii}^{-1} Y_i^2$$

and the conclusion is clear,, noting that $|i \ : D_{ii} \neq 0| = \mathrm{rank}(\Sigma)$.

Solution 21. Solution in Python: see this GitHub link

Solution 22. Solution in Python: see this GitHub link

Solution 23 ($1 \Rightarrow 2$) Let us write $W = \xi U$ where $U = W/\|W\|$ is a random unit vector and $\xi = \|W\|$. First note that the distribution of U need be rotationally invariant, and is thus necessarily the uniform distribution over the unit sphere. Furthermore we may also show that $\xi \perp U$.

Heuristica idea. We present here an incomplete argument that however suffices for the porpuses of this course. Let $\epsilon > 0$ and $u_1, u_2, \in \mathbb{S}^{p-1}$ arbitrary. Then there exists an orthogonal matrix O such that $u_2 = O^\top u_1$. Consequently, we may *improperly* but heuristically write:

$$\begin{aligned}
\mathbb{P}(\xi = \epsilon \mid U = u_1) =& \frac{\mathbb{P}(W = \epsilon u_1)}{\mathbb{P}(U = u_1)} \\
\text{(by invariance under rotations)} \quad =& \frac{\mathbb{P}(OW = \epsilon u_1)}{\mathbb{P}(OU = u_1)} \\
=& \frac{\mathbb{P}(W = \epsilon u_2)}{\mathbb{P}(U = u_2)} \\
=& \mathbb{P}(\xi = \epsilon \mid U = u_2)
\end{aligned}$$

Note that this intuitive argument is not rigorous (we are considering events of probability 0). However, it is simple to see that one could consider a generating sigma-algebra of events for which this line of reasoning could be reproduced, at the cost of a more tedious writing (for the angular part consider fixed-length arcs and for the radial part consider intervals).

Rigorous argument.

We want to show that ξ and U are independent, i.e., for any measurable sets $A \subseteq \mathbb{R}_+$ and $B \subseteq \mathbb{S}^{p-1}$,

$$\mathbb{P}(\xi \in \mathcal{E}, U \in \mathcal{U}) = \mathbb{P}(\xi \in \mathcal{E})\mathbb{P}(U \in \mathcal{U}).$$

for suitable measurable sets $\mathcal{E} \subset \mathbb{R}_+$,$\mathcal{U} \subset [0, 2\pi)$.

$$\begin{aligned}
\mathbb{P}(\xi \in \mathcal{E} \mid U \in \mathcal{U}) &= \frac{\mathbb{P}(\|W\| \in \mathcal{E},\, W/\|W\| \in \mathcal{U})}{\mathbb{P}(U \in \mathcal{U})} \\
\text{(by invariance under rotations)} \quad &= \frac{\mathbb{P}(\|OW\| \in \mathcal{E},\, OW/\|OW\| \in \mathcal{U})}{\mathbb{P}(OU \in \mathcal{U})} \\
&= \frac{\mathbb{P}(\|W\| \in \mathcal{E},\, W/\|W\| \in O\mathcal{U})}{\mathbb{P}(U \in O\mathcal{U})} \\
&= \mathbb{P}(\xi \in \mathcal{E} \mid U \in O\mathcal{U}).
\end{aligned}$$

which proves that the law of η is constant under conditioning on U, and hence establishes their independence.

($2 \Rightarrow 3$) Representing v it in polar form $v = \|v\|u$, where u is a unit vector in the direction of v. Then, $v^T W = (\|v\|u)^T(\xi U) = \xi\|v\|u^T U$ and it is easy to see that $u^T U$ has distribution of the marginals U_i (by invariance of the uniform distribution on the sphere wrt the coordinate system). Thus: $v^T W \overset{d}{=} \xi\|v\|U_1 = \|v\|W_1$.

($3 \Rightarrow 1$) First, recall that the distribution of all possible 1-d marginals uniquely characterises the joint distribution of a random vector. That is, the distribution of X is uniquely determined by the family of distributions $v^\top X$ for all vectors v.

Next, for any orthogonal U and vector v, note that $\|(v^\top \mathsf{U})^\top\| = \|v\|$. Then, observe that:

$$v^\top(\mathsf{U}X) = (v^\top \mathsf{U})X \overset{d}{=} \|v\|X_1 \overset{d}{=} v^\top X, \qquad \forall\, v \in \mathbb{R}^p$$

which directly implies that $X \overset{d}{=} \mathsf{U}X$.

Solution 24. Let X be an elliptical random vector. That is, we may write $\mathsf{X} = \mu + \mathsf{AU}$ for some spherical U.

1. $\mathsf{X}_j = e_j^\top \mathsf{X} = \mu_j + (e_j^\top \mathsf{A})\mathsf{U}$, concluding the proof.
2. $\mathsf{BX} + b = (\mathsf{B}\mu + b) + \mathsf{BAU}$, concluding the proof.

Solution 25.

$$(AXB)_{ij} = [A(XB)]_{ij} = \sum_{l=1}^{n} A_{il}(XB)_{lj} = \sum_{l=1}^{n} A_{il} \sum_{k=1}^{p} X_{lk}B_{kj} = \sum_{l=1}^{n}\sum_{k=1}^{p} A_{il}X_{lk}B_{kj}.$$

So, easy to see that each element is univariate normal. But we also require the rows of Y to be independent and to have the same distribution. The heuristic idea is the following. Post-multiplication of X involves adding weighted variables. Hence the rows of $\mathsf{XB}^\top$ are independent. The transformed objects are also independent unless the premultiplication by A introduces some interdependence, so each row can be transformed only by scalar multiplication. To have that the distribution is the same across rows, the scalar factor must be the same.

Concretely, we need each transformed row to have the same mean and covariance, and the cross-covariance between different rows to be zero.

Let Y be the transformed data matrix. We denote its i-th row by Y_i.

1. (*same mean*)

$$\begin{aligned}
\mathbb{E}[\mathsf{Y}_i] &= \mathsf{A}_i \begin{pmatrix} \mu^\top \\ \vdots \\ \mu^\top \end{pmatrix} \mathsf{B} \\
&= \left(\mathsf{A}_i \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_1 \end{pmatrix}, \cdots, \mathsf{A}_i \begin{pmatrix} \mu_p \\ \vdots \\ \mu_p \end{pmatrix} \right) \mathsf{B} \\
&= \big(\mu_1(\mathsf{A}_i 1_n), \cdots, \mu_p(\mathsf{A}_i 1_n)\big)\, \mathsf{B} \\
&= (\mathsf{A}_i 1_n)\mu^\top \mathsf{B}
\end{aligned}$$

   which shows the first part.

2. (*same covariance*) wlog assume X centered.

$$\begin{aligned}
\mathbb{E}[\mathsf{Y}_i^\top \mathsf{Y}_i] &= \mathbb{E}[(\mathsf{A}_i\mathsf{XB})^\top(\mathsf{A}_i\mathsf{XB})] \\
&= \mathsf{B}^\top \mathbb{E}[(\mathsf{X}^\top \mathsf{A}_i^\top \mathsf{A}_i \mathsf{X})]\mathsf{B}
\end{aligned}$$

   Note that:

$$\begin{aligned}
\mathbb{E}[(\mathsf{X}^\top \mathsf{A}_i^\top \mathsf{A}_i \mathsf{X})_{k\ell}] &= \mathbb{E}[(\sum_j \mathsf{X}_{jk}\mathsf{A}_{ij})(\sum_{j'} \mathsf{A}_{ij'}\mathsf{X}_{j'\ell})] \\
&= \mathbb{E}[\sum_{j,j'} \mathsf{X}_{jk}\mathsf{A}_{ij}\mathsf{A}_{ij'}\mathsf{X}_{j'\ell}] \\
&= \Sigma_{\ell k} \sum_j \mathsf{A}_{ij}^2
\end{aligned}$$

   so that:

$$\mathbb{E}[\mathsf{Y}_i^\top \mathsf{Y}_i] = (\mathsf{A}_i\mathsf{A}_i^\top)\mathsf{B}^\top \Sigma \mathsf{B}$$

which does not depend on i iff $\mathsf{A}_i\mathsf{A}_i^\top$ does not depend on i, or $\mathsf{B}^\top\Sigma\mathsf{B} = 0$. Similarly:

$$\mathbb{E}[\mathsf{Y}_i^\top\mathsf{Y}_i] = (\mathsf{A}_i\mathsf{A}_{i'}^\top)\mathsf{B}^\top\Sigma\mathsf{B}$$

which is 0 iff distinct rows of A are orthogonal, or $\mathsf{B}^\top\Sigma\mathsf{B} = 0$.

Solution 26. Given that X is an $n \times p$ data matrix from a multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$, and we define $\mathsf{Y} = \mathsf{AXB}$ and $\mathsf{Z} = \mathsf{CXD}$, where A, B, C, and D are matrices of appropriate dimensions.

Using vectorization, $\mathrm{vec}(Y)$ and $\mathrm{vec}(Z)$ can be written as $(\mathsf{B}^\top \otimes \mathsf{A})\mathrm{vec}(\mathsf{X})$ and $(\mathsf{D}^\top \otimes \mathsf{C})\mathrm{vec}(\mathsf{X})$, respectively. Hence, the covariance between $\mathrm{vec}(\mathsf{Z})$ and $\mathrm{vec}(\mathsf{Y})$ is given by:

$$\mathrm{Cov}(\mathrm{vec}(\mathsf{Y}), \mathrm{vec}(\mathsf{Z})) = (\mathsf{B}^\top \otimes \mathsf{A})\mathrm{Cov}(\mathrm{vec}(\mathsf{X}))(\mathsf{D}^\top \otimes \mathsf{C})^\top.$$

Given that $\mathrm{Cov}(\mathrm{vec}(\mathsf{X})) = \Sigma \otimes I$ (where I is the identity matrix), the above expression becomes:

$$\begin{aligned}\mathrm{Cov}(\mathrm{vec}(\mathsf{Z}), \mathrm{vec}(\mathsf{Y})) &= (\mathsf{B}^\top \otimes \mathsf{A})(\Sigma \otimes I)(\mathsf{D}^\top \otimes \mathsf{C})^\top \\ &= ((\mathsf{B}^\top\Sigma) \otimes \mathsf{A})(\mathsf{D} \otimes \mathsf{C}^\top) \\ &= \mathsf{B}^\top\Sigma\mathsf{D} \otimes \mathsf{AC}^\top.\end{aligned}$$

Thus, the elements of $\mathrm{vec}(\mathsf{Y})$ and $\mathrm{vec}(\mathsf{Z})$ are uncorrelated if and only if the above matrix is the zero matrix, i.e., if and only if we have $\mathsf{B}^\top\Sigma\mathsf{D} = 0$ or $\mathsf{AC}^\top = 0$.

Solution 27. 1.

$$\mathsf{X}^\top\mathsf{X} = \begin{bmatrix} X_1^\top & \dots & X_n^\top \end{bmatrix} \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} = \sum_{i=1}^n X_i^\top X_i$$

Defining $\mathsf{W}_i = \mathsf{X}_i^\top\mathsf{X}_i \sim W_p(\Sigma, 1)$ and since the rows are independent, the matrices W_i are independent this completes the claim.

2. $\mathbb{E}[\mathsf{W}] = \mathbb{E}[X^\top X]. = \sum_{i=1}^n \mathbb{E}[\mathsf{X}_i^\top\mathsf{X}_i]. = n\Sigma.$

3. When $n < p$, the Wishart matrix $W = \sum_{i=1}^n X_i X_i^\top$ is the sum of at most n rank-one matrices. Since the rank of each $X_i X_i^\top$ is at most 1, the total rank of W is at most n. Therefore, W is almost surely singular because its rank never reaches the full dimension p. If it is almost surely singular, then it is constrained to a lower-dimensional subset of the space of symmetric $p \times p$ matrices. This means it cannot have a density, since it is supported on a Lebesgue-zero measure set.

4. Write $\mathsf{W} = X^\top X$ where X is an $n \times p$ matrix with i.i.d. rows $\mathsf{X}_i \sim \mathcal{N}(0, \Sigma)$, and for any $\theta \notin \ker(\Sigma)$, define $Y = X\theta.$, which satisfies $Y_i \sim \mathcal{N}(0, \theta^\top\Sigma\theta).$. Then,

$$\theta^\top\mathsf{W}\theta = \theta^\top X^\top X\theta = Y^\top Y = \sum_{i=1}^n Y_i^2.$$

Since $Y_i \sim \mathcal{N}(0, \theta^\top\Sigma\theta)$, normalizing by $\theta^\top\Sigma\theta$ gives:

$$\frac{\theta^\top\mathsf{W}\theta}{\theta^\top\Sigma\theta} = \sum_{i=1}^n \left(\frac{Y_i}{\sqrt{\theta^\top\Sigma\theta}}\right)^2.$$

Since each term follows $\mathcal{N}(0, 1)$, we conclude:

$$\frac{\theta^\top\mathsf{W}\theta}{\theta^\top\Sigma\theta} \sim \chi_n^2.$$

Solution 28. 1. Note that if $\mathsf{W} \sim W_p(\Sigma, n)$ then

$$\Sigma^{-1/2\top}\mathsf{W}\Sigma^{-1/2} \sim W_p(I, n) \quad \text{and} \quad \Sigma^{-1/2}(X-\mu) \sim N(I, 0).$$

By definition of the Hotelling distribution we have that:

$$n(\Sigma^{-1/2}(X-\mu))^\top((\Sigma^{-1/2}W\Sigma^{-1/2})^{-1}(\Sigma^{-1/2}(X-\mu)) \sim T^2(p,n)$$

but the above expression is simply

$$n(X-\mu)^\top W^{-1}(X-\mu).$$

2. Similar, using the Gaussian Sampling Theorem on slide 111.

Solution 29. Consider the random Gaussian vectors X, Y with same marginals, but different covariance structure. Then the sequence $Z_{2n} := X$ and $Z_{2n+1} = Y$ establishes a counterexample.

Solution 30. 1. Since Σ is diagonal, $\sigma_{ik} = 0$ when $i \neq k$, thus:

$$\text{cov}\{w_{ij}, w_{kl}\} = \sigma_{ii}\sigma_{jj}(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk})$$

hence: $\text{cov}\{w_{ij}, w_{ij}\} = \text{cov}\{w_{ij}, w_{ji}\} = \sigma_{ii}\sigma_{jj}$ and otherwise 0 (recall W is symmetric!).

2. Let us first consider the diagonal case. Then, the covariance of the vectorisation of the lower-diagonal part of W has diagonal covariance, with strictly positive elements on the diagonal, which is clearly positive definite. For the general case, it suffices to observe that the reduction can be obtained by the svd $\Sigma = U^T\Lambda U$ and observing that $U\mathsf{W}U^\top$ is Wishart with diagonal scale Λ.

Solution 31. Solution in Python: see this GitHub link

Solution 32. Let $\overline{X}$ be the MLE of μ. Note it is clearly unbiased for μ. Let T be any other unbiased estimator of μ. We want to show that:

$$v^\top \text{Cov}(\overline{X})v \leq v^\top \text{Cov}(T)v \quad \forall\, v \in \mathbb{R}^p.$$

The key is to observe that $v^\top\overline{X}$ is the UMVUE for $v^\top\mu$, by known 1-d results. Therefore, observing that $v^\top T$ is unbiased for $v^\top\mu$ by linearity, we have that:

$$\text{Cov}(v^\top\overline{X}) \leq \text{Cov}(v^\top T)$$

or equivalently

$$v^\top\mathbb{E}[(\overline{X}-\mu)(\overline{X}-\mu)^\top]v \leq v^\top\mathbb{E}[(T-\mu)(T-\mu)^\top]v \quad \forall\, v \in \mathbb{R}^p.$$

which is precisely the claim.

Solution 33. Consider statistical model parameterized by θ, and let $\hat{\theta}$ be the MLE of θ:

$$\hat{\theta} = \arg\max_\theta L(\theta|\mathbf{x})$$

Define for $g(\theta)$ the induced likelihood function L^*, given by

$$L^*(\eta \mid \mathbf{x}) = \sup_{\{\theta: g(\theta)=\eta\}} L(\theta \mid \mathbf{x}).$$

Let $\hat{\eta}$ denote the value that maximizes $L^*(\eta \mid \mathbf{x})$, which is the MLE of $\eta = g(\theta)$. We must show that $L^*(\hat{\eta} \mid \mathbf{x}) = L^*[g(\hat{\theta}) \mid \mathbf{x}]$. Note that:

$$L^*(\hat{\eta} \mid \mathbf{x}) = \sup_\eta \sup_{\{\theta: g(\theta)=\eta\}} L(\theta \mid \mathbf{x}) = \sup_\theta L(\theta \mid \mathbf{x}) = L(\hat{\theta} \mid \mathbf{x}),$$

where the second equality follows because the iterated maximization is equal to the unconditional maximization over θ, which is attained at $\hat{\theta}$. Furthermore

$$L(\hat{\theta} \mid \mathbf{x}) = \sup_{\{\theta : g(\theta) = g(\hat{\theta})\}} L(\theta \mid \mathbf{x}) = L^*[g(\hat{\theta}) \mid \mathbf{x}].$$

Hence, the string of equalities shows that $L^*(\hat{\eta} \mid \mathbf{x}) = L^*(g(\hat{\theta}) \mid \mathbf{x})$ and that $g(\hat{\theta})$ is the MLE of $g(\theta)$.

Concerning uniqueness, note that if $\hat{\eta}, \hat{\eta}'$ both maximise the transformed likelihood, then necessarily $g^{-1}(\hat{\eta}) = g^{-1}(\hat{\eta}') = \hat{\theta}$ so that, by injectivity of the inverse $\hat{\eta} = \hat{\eta}'$.

Solution 34. Solution in Python: see this GitHub link

Solution 35. Solution in Python: see this GitHub link

Solution 36. 1. Let us first consider the case $\Sigma = I$. Recall that the determinant corresponds to product of the eigenvalues. Then, note that the matrix $I + vv^\top$ has v as eigenvector, with corresponding eigenvalue $1 + v^\top v$. Being a positive matrix, we may complete to find a basis of orthogonal eigenvectors, all of which will be orthogonal to v, and thus with eigenvalue 1. Therefore:

$$|I + vv^\top| = 1 + v^\top v$$

The general case then follows by using the multiplicativity of the determinant, and taking $u = \Sigma^{1/2} v$:

$$|\Sigma + uu^\top| = |\Sigma^{1/2}(I + vv^\top)\Sigma^{1/2}| = |\Sigma||I + vv^\top| = |\Sigma||1 + v^\top v| = |\Sigma||1 + u^\top \Sigma^{-1} u|$$

2. First note that when $\Sigma = I$ we may expand by power series:

$$(I + vv^\top)^{-1} = I + \sum_{j \geq 1} (-vv^\top)^j = I - v \sum_{j \geq 0} (-v^\top v)^j v^\top = I - \frac{1}{1 + v^\top v} vv^\top$$

so that the general case is follows by taking $u = \Sigma^{1/2} v$:

$$\begin{aligned}(\Sigma + uu^\top)^{-1} =& \Sigma^{-1/2}(I + vv^\top)^{-1}\Sigma^{-1/2} \\ =& \Sigma^{-1/2}(I - \frac{1}{1 + v^\top v} vv^\top)\Sigma^{-1/2} \\ =& \Sigma^{-1} - \frac{1}{1 + u^\top \Sigma^{-1} u} \Sigma^{-1} uu^\top \Sigma^{-1}.\end{aligned}$$

Solution 37. Let P_n denote the projection onto $\text{span}\,(1_n, (Z_1, \ldots, Z_n))$, and write $H_n = I - P_n$. Consider its SVD decomposition $H_n = U\Omega U^\top$. Note that $\hat{\epsilon}_X = H_n(X_1, \ldots, X_n)$ and similarly for Y. Note that $(W, V) = \Omega U^\top X, \Omega U^\top Y)$ satisfy $W, V \sim \mathcal{N}(0, I_{n-2,n-2})$ and are independent under H_0 and thus the claim follows as in the proof in slide 160-161.

Solution 38. Let $X_1, \ldots, X_n$ be a random sample from $N(\mu, \Sigma_{p \times p})$ with $\Sigma \succ 0$. The restricted and unrestricted MLEs are, respectively, $(\bar{X}, \hat{\lambda} I)$ and $(\bar{X}, \hat{\Sigma})$, where:

$$\hat{\lambda} = \text{tr}(\hat{\Sigma})/p.$$

This is easy to see, using equivariance of the MLE. Thus:

$$\ell_0^* = \ell\left(\bar{X}, \hat{\lambda} I\right) = -\frac{n}{2} \log\left|2\pi\hat{\lambda} I\right| - \frac{n}{2} \text{tr}\left(\hat{\lambda}^{-1}\hat{\Sigma}\right)$$

and:

$$\ell^* = \ell(\bar{X}, \hat{\Sigma}) = -\frac{n}{2} \log|2\pi\hat{\Sigma}| - \frac{np}{2}$$

which yields:

$$\begin{aligned} 2\log\hat{\Lambda} =& n\left(\log\left|2\pi\hat{\lambda}I\right| + \mathrm{tr}(\hat{\lambda}^{-1}\hat{\Sigma}) - \log|2\pi\hat{\Sigma}| - p\right) \\ =& n\left(-\log|\hat{\lambda}^{-1}\hat{\Sigma}| + \mathrm{tr}(\hat{\lambda}^{-1}\hat{\Sigma}) - p\right) \\ =& \log\left(\frac{\frac{(\mathrm{tr}(\hat{\Sigma}))}{p}}{|\hat{\Sigma}|^{1/p}}\right)^{pn} \end{aligned}$$

Solution 39. Solution in Python: see this GitHub link

Solution 40. We decompose X using SVD, i.e.

$$\mathsf{X} = \mathsf{U}\Gamma\mathsf{V}^T$$

and find that we can write the covariance matrix as

$$\mathsf{C} = \frac{1}{n}\mathsf{X}\mathsf{X}^T = \frac{1}{n}\mathsf{U}\Gamma^2\mathsf{U}^T.$$

In this case U is a $n \times m$ matrix. Assuming that the singular values are ordered descending order we know that, if $n < m$, the first n columns in U correspond to the sorted eigenvalues of C and if $m \geq n$, the first m corresponds to the sorted non-zero eigenvalues of C. The transformed data can thus be written as

$$\mathsf{Y} = \tilde{\mathsf{U}}^T\mathsf{X} = \tilde{\mathsf{U}}^T\mathsf{U}\Gamma\mathsf{V}^T,$$

where $\tilde{\mathsf{U}}^T\mathsf{U}$ is a simple $n \times m$ matrix which is one on the diagonal and zero everywhere else. To conclude, we can write the transformed data in terms of the SVD decomposition of X.

Solution 41. Solution in Python: see this GitHub link

Solution 42.

$$\begin{aligned} \|\hat{\Sigma} - \Sigma\|^2 &= \|\Sigma\|^2 + \|\hat{\Sigma}\|^2 - 2\mathrm{tr}(\hat{\Sigma}\Sigma) \\ \textit{(by the trace inquality)} \quad &\geq \sum\lambda_i^2 + \sum\hat{\lambda}_i^2 - 2\sum\lambda_i\hat{\lambda}_i \\ &= \sum(\lambda_i - \hat{\lambda}_i)^2 \end{aligned}$$

and since:

$$\sum(\lambda_i - \lambda_i)^2 = (\lambda_j - \hat{\lambda}_j)^2\left(1 + \sum_{i\neq j}\frac{(\lambda_i - \hat{\lambda}_i)^2}{(\lambda_j - \hat{\lambda}_j)^2}\right) \geq (\lambda_j - \hat{\lambda}_j)^2, \qquad \forall\, j = 1, \ldots, p$$

we find that:

$$|\lambda_j - \hat{\lambda}_j| \leq \|\hat{\Sigma} - \Sigma\|, \qquad \forall\, j = 1, \ldots, p.$$

Solution 43. An equivalent formulation of the statement in the exercise is that:

$$\sum_{j=1}^{p}\|X_j - \mathrm{H}_k X_j\|^2 \leq \sum_{j=1}^{p}\|X_j - \mathrm{Q}X_j\|^2$$

for any $n\times n$ projection operator Q or rank at most k, where $\mathrm{H}_k = \sum_{i=1}^{k}\hat{u}_i\hat{u}_i^\top$. However, this is a direct consequence of the Optimal Linear Dimension Reduction Theorem (slide 174), taking expectations with respect to the empirical (discrete) measure $P_n = \sum_{i=1}^{p}\delta_{X_i}$.

Solution 44. Note that $n\hat{\Sigma} \sim \text{Wishart}(\Sigma, n-1)$, and so we have the following identity in law:

$$n\hat{\Sigma} \overset{d}{=} \sum_{i=1}^{n-1} W_i, \quad \text{for} \quad W_i \overset{\text{IID}}{\sim} \text{Wishart}(\Sigma, 1)$$

By the content in slide 129, we thus know that Therefore:

$$\text{Cov}(\sqrt{n}(\hat{\Sigma} - \Sigma))_{(ij)(kl)} = \frac{n-1}{n}(\Sigma_{ik}\Sigma_{jl} + \Sigma_{il}\Sigma_{jk}).$$

In particular, this gives us the covariance – taking the limit – of Z:

$$\text{Cov}(Z)_{(ij)(kl)} = (\Sigma_{ik}\Sigma_{jl} + \Sigma_{il}\Sigma_{jk}).$$

which may be succintly written as:

$$\text{Cov}(\text{vec}(Z)) = (\Sigma \otimes \Sigma)(I + K)$$

where I is the identity and K the commutator, i.e. $K\,\text{vec}(A) = \text{vec}(A^\top)$.

Now note that:

$$\begin{aligned}
\mathbb{E}\left[\langle Zu_i, u_i\rangle \langle Zu_j, u_j\rangle\right] &= \mathbb{E}\left[\left\langle Z, u_i u_i^\top\right\rangle \left\langle Z, u_j u_j^\top\right\rangle\right] \\
&= \langle \text{Cov}(Z) u_i u_i^\top, u_j u_j^\top \rangle \\
&= \langle \text{Cov}(\text{vec}(Z))\text{vec}(u_i u_i^\top), \text{vec}(u_j u_j^\top)\rangle \\
&= \langle (\Sigma \otimes \Sigma)(I + K)\text{vec}(u_i u_i^\top), \text{vec}(u_j u_j^\top)\rangle \\
&= 2\langle (\Sigma \otimes \Sigma)\text{vec}(u_i u_i^\top), \text{vec}(u_j u_j^\top)\rangle \\
(A \otimes B)\,\text{vec}\,(V) = \text{vec}(BVA^\top) \qquad &= 2\langle \text{vec}(\Sigma u_i u_i^\top \Sigma), \text{vec}(u_j u_j^\top)\rangle \\
&= 2\lambda_i \langle \text{vec}(u_i u_i^\top \Sigma), \text{vec}(u_j u_j^\top)\rangle \\
&= 2\lambda_i \text{tr}(u_i u_i^\top \Sigma u_j u_j^\top) \\
&= 2\lambda_i \lambda_j \text{tr}(u_i u_i^\top u_j u_j^\top) \\
&= 2\lambda_i \lambda_j \delta_{ij} \text{tr}(u_i u_j^\top) \\
&= 2\lambda_i \lambda_j \delta_{ij}
\end{aligned}$$

Be mindful that the inner product in the first two equations is not the same (we go from vectors to matrices!)

Solution 45. Solution in Python: see this GitHub link

Solution 46. Recall that $\text{Cov}(\mathsf{Q}_n)_{(ij)(k\ell)} = \text{Cov}(\sqrt{n}(\mathsf{U}^\top \hat{\Sigma} \mathsf{U} - \Lambda))_{(ij)(k\ell)} = \frac{n}{n-1}\lambda_i\lambda_j\delta_{(ij)(k\ell)}$ thus:

$$\text{Cov}(\mathsf{Q})_{(ij)(k\ell)} = \lambda_i\lambda_j(\delta_{(ij)(k\ell)} + \delta_{(ij)(\ell k)}).$$

Furthermore, note that we established in the proof that

$$\mathsf{W}_{ij}(\lambda_j - \lambda_i) \overset{d}{=} \mathsf{Q}_{ij} \quad \text{when } i \neq j, \text{ and 0 otherwise.}$$

Hence:

$$\text{Cov}(W_i W_i)_{k\ell} = \mathbb{E}[\mathsf{W}_{ki}\mathsf{W}_{\ell i}] = \mathbb{E}[\mathsf{Q}_{ki}\mathsf{Q}_{\ell i}]\frac{1}{(\lambda_i - \lambda_k)(\lambda_i - \lambda_\ell)} = \frac{\text{Cov}(\mathsf{Q})_{(ki)(\ell i)}}{(\lambda_i - \lambda_k)(\lambda_i - \lambda_\ell)}$$

or equivalently:

$$\text{Cov}(W_i W_i) = \sum_{k \neq i} \frac{\lambda_k \lambda_i}{(\lambda_i - \lambda_k)^2} e_k e_k^\top$$

Proceeding similarly:

$$\begin{aligned}\mathrm{Cov}(W_iW_j)_{k\ell} = \mathbb{E}[\mathsf{W}_{ki}\mathsf{W}_{\ell j}] &= \mathbb{E}[\mathsf{Q}_{ki}\mathsf{Q}_{\ell j}]\frac{1}{(\lambda_i-\lambda_k)(\lambda_j-\lambda_\ell)}\\ &= \frac{\mathrm{Cov}(\mathsf{Q})_{(ki)(\ell j)}}{(\lambda_i-\lambda_k)(\lambda_\ell-\lambda_j)}\\ &= -\frac{\lambda_i\lambda_j\delta_{i\ell}\delta_{kj}}{(\lambda_i-\lambda_j)^2}\end{aligned}$$

or equivalently:

$$\mathrm{Cov}(W_iW_j) = -\frac{\lambda_i\lambda_j\delta_{i\ell}\delta_{kj}}{(\lambda_i-\lambda_j)^2}e_je_i^\top$$

Solution 47. Let $X_1,\dots,X_n$ be IID copies of the random vector $X\in\mathbb{R}^p$, with $n\geq p$. Suppose that $\hat\Sigma_n$ is *not* full rank. Then, there exists $v\in\mathbb{R}^p$ such that :

$$0 = v^\top\hat\Sigma_n v = \frac{1}{n-1}\sum_{j=1}^n v^\top X_iX_i^\top v = \sum_{j=1}^n\langle v,X_i\rangle^2$$

which yields that v is orthogonal to all X_i. But this occurs with probability zero, since $\mathbb{P}(\mathrm{span}(X_1,\dots,X_p)=\mathbb{R}^p)=1$. Indeed, the probability of the complementary is upper bounded by that of drawing a vector in a $p-1$ dimensional subspace, which is 0 for all absolutely continuous measures. To conclude, observe that finite intersection of almost sure events is almost sure.

Solution 48. Suppose that we have a statistical model M_k of some data, with k the number of estimated parameters in the model. Let $L(\hat M_k)$ be the maximized value of the likelihood function for the model. Then the AIC value of the model is

$$\mathrm{AIC}_{(M_k)} \;=\; 2k - 2\ln(L(\hat M_k)$$

Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value. Thus, AIC rewards goodness of fit (as assessed by the likelihood function), but it also includes a penalty that is an increasing function of the number of estimated parameter, which discourages overfitting.

In our setting, the model M_k assumes that the covariance function corresponding to the data generating measure has rank k. Assumning gaussianity, up to constants idependent of k we have that:

$$2\ln(L(\hat M_k)\simeq -n\sum_{j=k+1}^p\log\hat\lambda_j + nk$$

where $\hat\lambda_1,\dots,\hat\lambda_p$ are the eigenvalues of the empirical covariance $\hat\sigma$, and n the number of observed samples. Inclusion of an additional principal component, i.e. increasing the number of parameters of the model, is justified if the criterion value decreases. That is the model M_{k+1} is preferred to the model M_k if $\mathrm{AIC}_{(M_{k+1})}<\mathrm{AIC}_{(M_k)}$. That is, if:

$$2(k+1)+n\sum_{j=k+2}^p\log\hat\lambda_j - n(k+1) < 2k+n\sum_{j=k+1}^p\log\hat\lambda_j - nk.$$

which, more explicitly, corresponds to:

$$\lambda_{k+1} > e^{-\frac{n-2}{n}}.$$

Solution 49.

Solution 50. To show that the matrix functions $\Sigma \mapsto \Sigma^2$, $\Sigma \mapsto \Sigma^{1/2}$, and $\Sigma \mapsto \Sigma^{-1}$ are continuously differentiable (C^1) at $\Sigma \succ 0$ we will use the fact that these functions are smooth functions of the eigenvalues of the matrix.

The argument is very similar for all three cases. Wlog, consider the function $\Sigma \mapsto \Sigma^{-1}$. For $\Sigma = U\Lambda U^T$, the inverse is given by:

$$\Sigma^{-1} = U\Lambda^{-1}U^T,$$

where Λ^{-1} is the diagonal matrix with entries λ_i^{-1}. The function $x \mapsto x^{-1}$ is smooth and continuously differentiable on $\mathbb{R}_{>0}$. This, with the smoothness of the "svd mapping" $\Sigma \mapsto U\Lambda U^T$ with respect to Σ ensures that $\Sigma \mapsto \Sigma^{-1}$ is C^1.

Solution 51. Let $\rho := \mathrm{corr}(\xi_k, \xi_{k+1})$, which is independent of k by stationarity. Assume $\rho \notin \{-1, 1\}$, otherwise the claim is trivial. Then, by the regression representation, for all k there exists a centered random variable ϵ_k, independent of ξ_k with variance $\sigma^2 > 0$ given by:

$$\epsilon_k := \xi_{k+1} - \rho\xi_k.$$

Note that ϵ_k need be Gaussian, since Gaussians are closed under linear modifications, and that the sequence of ϵ_k need be iid by stationarity. By assumption, all ξ_k are centered and Gaussian, say $\mathcal{N}(0, \nu^2)$. By stationarity:

$$\xi_1 \overset{d}{=} \xi_2 \overset{d}{=} \rho\xi_1 + \epsilon_k$$

so that we must have $\nu^2 = \rho^2\nu^2 + \sigma^2$; solving for ν gives:

$$\nu^2 = \sigma^2/(1-\rho^2).$$

Solution 52.

$$\begin{aligned} l(\rho, \sigma) &= \log f_{\xi_i} + \sum_{j=1}^{p-1} \log f_{\xi_{j+1}|\xi_j}(\xi_{j+1}; \rho, \sigma) \\ &= -\frac{1}{2}\log 2\pi\frac{\sigma^2}{1-\rho^2} - \frac{\xi_1}{2\frac{\sigma^2}{1-\rho^2}} - (p-1)\frac{1}{2}\log 2\pi\sigma^2 - \sum_{j=1}^{p-1}\frac{\xi_{j+1}-\xi_j}{2\sigma^2}. \end{aligned}$$

Solution 53. Solution in Python: see this GitHub link.