Multivariate Statistics

Victor Panaretos

Institut de Mathématiques – EPFL

victor.panaretos@epfl.ch

- Lectures Thu 13.15–15.00
- Exercices Thu 15.15–17.00
- Main reference books (but we go beyond):

  Anderson, T.W. *An Introduction to Multivariate Statistical Analysis*, Wiley
  Muirhead, *Aspects of Multivariate Statistical Theory*, Wiley

- Webpage: moodle
- Bonus (non-compulsory) midterm test on 17 April, 13.15
- Written final exam (cheat sheet allowed)
- Final grade G will be calculated
  - $G = 0.75 \times E + 0.25 \times \max\{E, T\}$
  - $E =$ exam, $T =$ test
  - we round F to obtain G

In short: statistical analysis of random vectors

What does this mean in effect?

- Understanding the probability distribution of a random vector
- Most commonly the vector space is $\mathbb{R}^p$, with $p > 1$.
  (but similar principles can apply to more general vector spaces)
- Random vectors have *internal probabilistic structure* – coordinate dependence

$$X = \left(X_1, \ldots, X_p \right)^\top$$

- Dependence can be *unconditional* or *conditional*
- Need to understand how to encapsulate, model, and infer this dependence

- Our typical setting: sample $X_1, ..., X_n$ of $n > 1$ i.i.d. realisations in $\mathbb{R}^p$

$$\mathrm{X} = \begin{pmatrix} X_{11} & \dots & X_{1p} \\ \vdots & & \vdots \\ X_{n1} & \dots & X_{np} \end{pmatrix}$$

- We will focus on coordinates that are continuous random variables
- Rows are *observations* (a.k.a. individuals) and columns are *variables* (a.k.a. features)
- Central objects (but not *only* ones): covariance Σ and its inverse $\Theta = \Sigma^{-1}$
- Methods/theory depend on whether:
  - $p \ll n$, the so-called *low dimensional case*
  - $p \gg n$, the so-called *high dimensional case*
- Will primarily focus on $p \ll n$ but will also treat selected topics when $p \gg n$

The usual context is when we record p variables (or features) on n individuals:

- The gene expression levels for p genes for n subjects.
- The curvature at p sites on n DNA strands.
- The grades on p courses for n students.
- The portfolio returns on p assets at n times.
- The blood pressure at p times for n patients.

In these cases, we may be interested in:

- Which genes are co-expressed?
- What are the mechanical properties of DNA?
- Are there interesting subgroups based on conditions on variables?
- What is the best portfolio distribution?
- Can we predict the grades in a group of courses from other courses?
- Are there trends? Drivers of variation? Indirect associations?

Often, there are qualitative variables, either recorded or latent:

- Treatment or disease status, subpopulation membership
- DNA Base-pair composition, presence/absence of gold stain
- Gender, race, season, educational background, risk factor ...

- Linear Algebra Recap
- Random Vectors and Matrices
- Gaussian Vectors
- Sampling
- Inference
- Dimension Reduction
- (Gaussian) Graphical Models

Linear Algebra Recap

If Q is an $n \times p$ real matrix, we define the

- *range* (or *column space*) of Q to be the set spanned by its columns:

$$\mathcal{R}(\mathsf{Q}) = \{\mathsf{Q}\beta : \beta \in \mathbb{R}^p\} \subseteq \mathbb{R}^n.$$

- the *null space* (or *kernel*) of Q is the subspace defined as

$$\mathcal{N}(\mathsf{Q}) = \{x \in \mathbb{R}^p : \mathsf{Q}x = 0\};$$

- the *orthogonal complement* of $\mathcal{R}(\mathsf{Q})$, is the subspace defined as

$$\begin{aligned} \mathcal{R}^{\perp}(\mathsf{Q}) &= \{y \in \mathbb{R}^n : y^{\top}\mathsf{Q}x = 0,\ \forall x \in \mathbb{R}^p\} \\ &= \{y \in \mathbb{R}^n : y^{\top}v = 0,\ \forall v \in \mathcal{R}(\mathsf{Q})\}. \end{aligned}$$

The orthogonal complement may be defined for arbitrary subspaces by using the second equality.

Theorem (Singular Value Decomposition)

Any $n \times p$ real matrix Q can be factorised as

$$\underset{n\times p}{\mathsf{Q}} = \underset{n\times n}{\mathsf{U}}\ \underset{n\times p}{\Sigma}\ \underset{p\times p}{\mathsf{V}}^{\top},$$

where U and $\mathsf{V}^{\top}$ are orthogonal with columns called left singular vectors and right singular vectors, respectively, and Σ is diagonal with non-negative real entries called singular values.

Immediate consequence:

1. The left singular vectors corresponding to non-zero singular values form an orthonormal basis for $\mathcal{R}(\mathsf{Q})$.
2. The left singular vectors corresponding to zero singular values form an orthonormal basis for $\mathcal{R}^{\perp}(\mathsf{Q})$.
3. Writing $\{u_i\}_{i=1}^n$ for the left singular vectors and $\{v_j\}_{i=1}^n$ for the right singular vectors, the SVD can also be expressed as

$$\underset{n\times p}{\mathsf{Q}} = \sum_{j=1}^{\mathtt{rank}(\mathsf{Q})} \sigma_j \underbrace{u_j}_{n\times 1} \underbrace{v_j^{\top}}_{1\times p}.$$

Proof.

Since the statement is invariant to transposition, assume wlog that $n \geq p$. We will prove the statement by induction on p. Assume that $p = 1$ so that Q is a column vector. Then the statement holds true trivially, by taking

$$\mathsf{V}^\top = \mathsf{V} = 1, \quad \Sigma = \left(\|\mathsf{Q}\|, \mathbf{0}_{1\times(n-1)}\right)^\top \quad \mathsf{U} = (u_1 \dots u_n), \quad u_1 = \mathsf{Q}/\|\mathsf{Q}\|$$

and $(u_2, ..., u_n)$ an orthonormal basis for $\text{span}^\perp(u_1)$. Thus the statement is true for all $n \geq p$ when $p = 1$. This is the base case for our induction. For the inductive step, assume that the statement is true for some $p > 1$ and all $n \geq p$. Let us prove that it is also true for $p + 1$ and all $n \geq p + 1$.

Let $\mathbb{S}^{p+1} = \{x \in \mathbb{R}^{p+1} : \|x\| = 1\}$ and $q(x) = \|Qx\|$. Since $q(\cdot)$ is continuous and $\mathbb{S}^{p+1}$ is compact, we have that $q(x)$ is bounded over $\mathbb{S}^{p+1}$ and attains its bounds. So there exists $v_1 \in S^{p+1}$ such that

$$q(v_1) = \max_{x\in\mathbb{S}^{p+1}} q(x) = \sigma_1 < \infty.$$

and let $v_1 \in \mathbb{S}^{p+1}$ be maximiser of $q(x)$, i.e. such that $q(v_1) = \max_{x\in\mathbb{S}^{p+1}} q(x)$. Define $u_1 = \sigma_1^{-1} Q v_1$ so $\|u_1\| = 1$. Given any orthonormal bases $\{u_j\}_{j=2}^n$ for $\text{span}^\perp(u_1)$ and $\{v_j\}_{j=2}^p$ for $\text{span}^\perp(v_1)$ define U and V to be orthogonal matrices

$$\mathsf{U} = (u_1 \ u_2 \ \dots \ u_n) = (u_1 \ \mathsf{U}_1) \qquad \& \qquad \mathsf{V} = (v_1 \ v_2 \ \dots \ v_n) = (v_1 \ \mathsf{V}_1).$$

Using block matrix multiplication, we see that

$$\underset{n\times n}{\mathsf{U}^\top}\ \underset{n\times(p+1)}{\mathsf{Q}}\ \underset{(p+1)\times(p+1)}{\mathsf{V}} = \begin{pmatrix} u_1^\top \\ \mathsf{U}_1^\top \end{pmatrix} \mathsf{Q} \begin{pmatrix} v_1 & \mathsf{V}_1 \end{pmatrix} = \begin{pmatrix} u_1^\top \mathsf{Q} v_1 & u_1^\top \mathsf{Q}\mathsf{V}_1 \\ U_1^\top \mathsf{Q} v_1 & \mathsf{U}_1^\top \mathsf{Q}\mathsf{V}_1 \end{pmatrix}$$

$$= \begin{pmatrix} \underset{1\times 1}{\sigma_1} & \underset{1\times p}{\theta^\top} \\ \underset{(n-1)\times 1}{\mathbf{0}} & \underset{(n-1)\times p}{\mathsf{Z}} \end{pmatrix}.$$

Now we claim that $\theta = 0$. To see this, first observe that

$$\sigma_1 = \max_{x\in\mathbb{S}^{p+1}} \|\mathsf{Q}x\| = \max_{x\in\mathbb{S}^{p+1}} \|\mathsf{U}^\top \mathsf{Q}x\| = \max_{x\in\mathbb{S}^{p+1}} \|\mathsf{U}^\top \mathsf{Q}\mathsf{V}x\|.$$

Next, let's consider the norm of $\mathsf{U}^\top \mathsf{Q}\mathsf{V} \begin{pmatrix} \sigma_1 \\ \theta \end{pmatrix}$,

$$\left\| \begin{pmatrix} \sigma_1 & \theta^\top \\ \mathbf{0} & \mathsf{Z} \end{pmatrix} \begin{pmatrix} \sigma_1 \\ \theta \end{pmatrix} \right\| = \left\| \begin{pmatrix} \sigma_1^2 + \theta^\top\theta \\ \mathsf{Z}\theta \end{pmatrix} \right\| = \sqrt{(\sigma_1^2 + \theta^\top\theta)^2 + \|\mathsf{Z}\theta\|^2}$$

$$\geq \sigma_1^2 + \theta^\top\theta = (\sigma_1^2 + \theta^\top\theta)^{1/2} \left\| \begin{pmatrix} \sigma_1 \\ \theta \end{pmatrix} \right\|.$$

Dividing across by $\|(\sigma_1\ \theta)^\top\|$, we see that we must necessarily have

$$(\sigma_1^2 + \theta^\top\theta)^{1/2} \leq \max_{x\in\mathbb{S}^{p+1}} \|\mathsf{U}^\top\mathsf{Q}\mathsf{V}x\| = \sigma_1 = (\sigma_1^2 + 0)^{1/2}.$$

and so it must be that $\theta^\top\theta = 0$. We conclude that

$$\mathsf{U}^\top\mathsf{Q}\mathsf{V} = \begin{pmatrix} \sigma_1 & \mathbf{0}_{1\times p} \\ \mathbf{0}_{(n-1)\times 1} & \mathsf{Z} \end{pmatrix} \overset{thus}{\Longrightarrow} \mathsf{Q} = \mathsf{U}\begin{pmatrix} \sigma_1 & \mathbf{0}_{1\times p} \\ \mathbf{0}_{(n-1)\times 1} & \mathsf{Z} \end{pmatrix}\mathsf{V}^\top.$$

But Z is an $(n-1)\times p$ matrix, and since $n \geq p+1$ it holds that $n-1 \geq p$. So by our inductive hypothesis

$$\mathsf{Z}_{(n-1)\times p} = \mathsf{W}_{(n-1)\times(n-1)}\Omega_{(n-1)\times p}\mathsf{R}^\top_{p\times p}.$$

where W, R are orthogonal and Ω is diagonal. Thus

$$\mathsf{Q}_{n\times p} = \mathsf{U}_{n\times n}\begin{pmatrix} \sigma_1 & \mathbf{0}_{1\times p} \\ \mathbf{0}_{(n-1)\times 1} & \mathsf{W}\Omega\mathsf{R}^\top \end{pmatrix}\mathsf{V}^\top_{p\times p} =$$

$$= \underbrace{\mathsf{U}\begin{pmatrix} 1 & \mathbf{0}_{1\times(n-1)} \\ \mathbf{0}_{(n-1)\times 1} & \mathsf{W}_{(n-1)\times(n-1)} \end{pmatrix}}_{\text{orthogonal}} \underbrace{\begin{pmatrix} \sigma_1 & \mathbf{0}_{1\times p} \\ \mathbf{0}_{(n-1)\times 1} & \Omega_{(n-1)\times p} \end{pmatrix}}_{\text{diagonal}} \underbrace{\begin{pmatrix} 1 & \mathbf{0}_{1\times p} \\ \mathbf{0}_{p\times 1} & \mathsf{R}^\top_{p\times p} \end{pmatrix}\mathsf{V}^\top}_{\text{orthogonal}}$$

Theorem (Spectral Theorem)

A $p \times p$ matrix A *is symmetric if and only if there exists a $p \times p$ orthogonal matrix* U *and a real diagonal matrix* Λ *such that*

$$\mathsf{A} = \mathsf{U}\Lambda\mathsf{U}^\top.$$

In particular:

1. the orthonormal columns of $\mathsf{U} = (u_1 \ \cdots \ u_p)$ are **eigenvectors** of A, i.e.

$$\mathsf{A}u_j = \lambda_j u_j, \qquad j = 1, \ldots, p$$

   where $\mathrm{diag}(\lambda_1, \ldots, \lambda_p) = \Lambda$ are the corresponding (real) **eigenvalues** of A.
2. the rank of A is the number of non-zero eigenvalues.
3. if the eigenvalues are distinct, the eigenvectors are unique (up to re-ordering and sign flips).
4. The spectral representation can also be expressed as

$$\underset{p\times p}{\mathsf{A}} = \sum_{j=1}^{\mathtt{rank}(\mathsf{A})} \lambda_j \underbrace{u_j}_{p\times 1} \underbrace{u_j^\top}_{1\times p}.$$

Proof.

If $\mathrm{A} = 0$, the statement holds trivially, so let $\mathrm{A} = \mathrm{A}^\top \neq 0$.

First note that the SVD of A guarantees the existence of a singular vector pair (u, v) with non-zero singular value σ, so that

$$\mathrm{A}(v+u) = \mathrm{A}v + \mathrm{A}u = \mathrm{A}v + \mathrm{A}^\top u = \sigma u + \sigma v = \sigma(u+v).$$

hence $w = (u+v)/\|u+v\|$ is a unit eigenvector of A with real eigenvalue σ.
Now the theorem is obviously true for 1×1 matrices (scalars). So use induction.
Assume any non-zero $p \times p$ symmetric matrix satisfies the theorem statement.
Let $\mathrm{A} = \mathrm{A}^\top \neq 0$ be $(p+1) \times (p+1)$. By the displayed equation, A has at least one unit eigenvector $w \in \mathbb{R}^p$ with real eigenvalue $\sigma \neq 0$.
Let $\mathrm{W} = (w\ \mathrm{R})$ where R has p orthonormal columns spanning $\mathrm{span}^\perp(w)$. Then

$$\mathrm{W}^\top \mathrm{A} \mathrm{W} = \begin{pmatrix} w^\top \\ \mathrm{R}^\top \end{pmatrix} \mathrm{A} \begin{pmatrix} w & \mathrm{R} \end{pmatrix} = \begin{pmatrix} w^\top \mathrm{A} w & w^\top \mathrm{A} \mathrm{R} \\ \mathrm{R}^\top \mathrm{A} w & \mathrm{R}^\top \mathrm{A} \mathrm{R} \end{pmatrix}$$

$$= \begin{pmatrix} \sigma & (\mathrm{A}w)^\top \mathrm{R} \\ \mathrm{R}^\top \mathrm{A} w & \mathrm{R}^\top \mathrm{A} \mathrm{R} \end{pmatrix} = \begin{pmatrix} \sigma & \mathbf{0}_{1\times p} \\ \mathbf{0}_{p\times 1} & \mathrm{R}^\top \mathrm{A} \mathrm{R} \end{pmatrix} = \begin{pmatrix} \sigma & \mathbf{0}_{1\times p} \\ \mathbf{0}_{p\times 1} & \mathrm{B} \end{pmatrix}$$

where $\mathrm{B} = \mathrm{R}^\top \mathrm{A} \mathrm{R}$ is a symmetric $p \times p$ matrix.

Since B is symmetric, we have $\mathsf{B} = V \Omega V^\top$ for $\mathsf{V}_{p\times p}$ orthogonal and $\Omega_{p\times p}$ diagonal by our induction hypothesis. In summary

$$\mathsf{A} = \mathsf{W} \begin{pmatrix} \sigma & \mathbf{0}_{1\times p} \\ \mathbf{0}_{p\times 1} & \mathsf{B} \end{pmatrix} W^\top$$

$$= \underbrace{\mathsf{W} \begin{pmatrix} 1 & \mathbf{0}_{1\times p} \\ \mathbf{0}_{p\times 1} & \mathsf{V}_{p\times p} \end{pmatrix}}_{\text{orthogonal}} \underbrace{\begin{pmatrix} \sigma & \mathbf{0}_{1\times p} \\ \mathbf{0}_{p\times 1} & \Omega_{p\times p} \end{pmatrix}}_{\text{diagonal}} \underbrace{\begin{pmatrix} 1 & \mathbf{0}_{1\times p} \\ \mathbf{0}_{p\times 1} & \mathsf{V}_{p\times p}^\top \end{pmatrix} \mathsf{W}_{p\times p}^\top}_{\text{orthogonal}}$$

$$= \mathsf{U}\Lambda\mathsf{U}^\top$$

□

Combining the SVD and the spectral theorem, we notice that:

1. The left singular vectors of Q are eigenvectors of $\mathsf{A} = \mathsf{Q}\mathsf{Q}^\top$.
2. The right singular vectors of Q are eigenvectors of $\mathsf{A} = \mathsf{Q}^\top\mathsf{Q}$.
3. The squared singular values of Q are eigenvalues of both $\mathsf{Q}\mathsf{Q}^\top$ and $\mathsf{Q}^\top\mathsf{Q}$.

A matrix Q is called *idempotent* if $Q^2 = Q$.

An *orthogonal projection* (henceforth projection) onto a subspace $\mathcal{V}$ is a symmetric idempotent matrix H such that $\mathcal{R}(H) = \mathcal{V}$.

Proposition

The only possible eigenvalues of a projection matrix are 0 and 1.

Proposition

Let $\mathcal{V}$ be a subspace and H be a projection onto $\mathcal{V}$. Then $I - H$ is the projection matrix onto $\mathcal{V}^\perp$.

Proof.

$(I - H)^\top = I - H^\top = I - H$ since H is symmetric and,
$(I - H)^2 = I^2 - 2H + H^2 = I - H$. Thus $I - H$ is a projection matrix.

It remains to identify the column space of $I - H$. Let $H = U\Lambda U^\top$ be the spectral decomposition of H. Then $I - H = UU^\top - U\Lambda U^\top = U(I - \Lambda)U^\top$.
Hence the column space of $I - H$ is spanned by the eigenvectors of H corresponding to zero eigenvalues of H, which coincides with $\mathcal{R}^\perp(H) = \mathcal{V}^\perp$. □

Proposition

Let $\mathcal{V}$ be a subspace and H *be a projection onto $\mathcal{V}$. Then* $\mathsf{H}y = y \iff y \in \mathcal{V}$.

Proof.

If $y \in \mathcal{V} \equiv \mathcal{R}(\mathsf{H})$, then $y = \mathsf{H}u$ for some u, so $\mathsf{H}y = \mathsf{H}\mathsf{H}u = \mathsf{H}u = y$. Conversely, if $y = \mathsf{H}y$ then $y \in \mathcal{R}(\mathsf{H}) \equiv \mathcal{V}$ by default (being of the form $\mathsf{H}u$ for $u = y$). □

Proposition

If P *and* Q *are projection matrices onto a subspace $\mathcal{V}$, then* $\mathsf{P} = \mathsf{Q}$.

Proposition

If $x_1, \ldots, x_p$ are linearly independent and are such that span$(x_1, \ldots, x_p) = \mathcal{V}$, then the projection onto $\mathcal{V}$ can be represented as

$$\mathsf{H} = \mathsf{X}(\mathsf{X}^\top \mathsf{X})^{-1} \mathsf{X}^\top$$

where X *is a matrix with columns $x_1, \ldots, x_p$.*

Proposition

Let $\mathcal{V}$ be a subspace of $\mathbb{R}^n$ and H *be a projection onto $\mathcal{V}$. Then*

$$\|x - \mathsf{H}x\| \leq \|x - v\|, \qquad \forall v \in \mathcal{V}.$$

Proof

Let $\mathsf{H} = \mathsf{U}\Lambda\mathsf{U}^\top$ be the spectral decomposition of H, $U = (u_1 \ \cdots \ u_n)$ and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$. Letting $p = \dim(\mathcal{V})$,

1. $\lambda_1 = \cdots = \lambda_p = 1$ and $\lambda_{p+1} = \cdots = \lambda_n = 0$,
2. $u_1, \ldots, u_n$ is an orthonormal basis of $\mathbb{R}^n$,
3. $u_1, \ldots, u_p$ is an an orthonormal basis of $\mathcal{V}$.

$$
\begin{aligned}
\|x - Hx\|^2 &= \sum_{i=1}^{n} (x^\top u_i - (\mathrm{H}x)^\top u_i)^2 && \text{[orthonormal basis]} \\
&= \sum_{i=1}^{n} (x^\top u_i - x^\top \mathrm{H} u_i)^2 && [H \text{ is symmetric}] \\
&= \sum_{i=1}^{n} (x^\top u_i - \lambda_i x^\top u_i)^2 && [u\text{'s are eigenvectors of H}] \\
&= 0 + \sum_{i=p+1}^{n} (x^\top u_i)^2 && \text{[eigenvalues 0 or 1]} \\
&\leq \sum_{i=1}^{p} (x^\top u_i - v^\top u_i)^2 + \sum_{i=p+1}^{n} (x^\top u_i)^2 && \forall v \in \mathcal{V} \\
&= \sum_{i=1}^{n} (x^\top u_i - v^\top u_i)^2 && \forall v \in \mathcal{V} \\
&= \|x - v\|^2 && \forall v \in \mathcal{V}.
\end{aligned}
$$

□

Proposition

Let $\mathcal{V}_1 \subseteq \mathcal{V} \subseteq \mathbb{R}^n$ be two nested linear subspaces. If H_1 is the projection onto $\mathcal{V}_1$ and H is the projection onto $\mathcal{V}$, then

$$\mathsf{H}\mathsf{H}_1 = \mathsf{H}_1 = \mathsf{H}_1\mathsf{H}.$$

Proof.

First we show that $\mathsf{H}\mathsf{H}_1 = \mathsf{H}_1$, and then that $\mathsf{H}_1\mathsf{H} = \mathsf{H}\mathsf{H}_1$. For all $y \in \mathbb{R}^n$ we have $\mathsf{H}_1 y \in \mathcal{V}_1$. But then $\mathsf{H}_1 y \in \mathcal{V}$, since $\mathcal{V}_1 \subseteq \mathcal{V}$.
Therefore $\mathsf{H}\mathsf{H}_1 y = \mathsf{H}_1 y$. We have shown that $(\mathsf{H}\mathsf{H}_1 - H_1)y = 0$ for all $y \in \mathbb{R}^n$, so that $\mathsf{H}\mathsf{H}_1 - \mathsf{H}_1 = 0$, as its kernel is all $\mathbb{R}^n$. Hence $\mathsf{H}\mathsf{H}_1 = \mathsf{H}_1$.

(Or, take n linearly independent vectors $y_1, \ldots, y_n \in \mathbb{R}^n$, and use them as columns of the $n \times n$ matrix Y. Now Y is invertible, and $(\mathsf{H}\mathsf{H}_1 - \mathsf{H}_1)Y = 0$, so $\mathsf{H}\mathsf{H}_1 - \mathsf{H}_1 = 0$, giving $\mathsf{H}\mathsf{H}_1 = \mathsf{H}_1$.)

To prove that $\mathsf{H}_1\mathsf{H} = \mathsf{H}\mathsf{H}_1$, note that symmetry of projection matrices and the first part of the proof give

$$\mathsf{H}_1\mathsf{H} = \mathsf{H}_1^\top \mathsf{H}^\top = (\mathsf{H}\mathsf{H}_1)^\top = (\mathsf{H}_1)^\top = \mathsf{H}_1 = \mathsf{H}\mathsf{H}_1.$$

□

Definition (Pseudoinverse)

Let Q be an $n \times p$ real matrix with SVD

$$\underset{n\times p}{\mathsf{Q}} = \underset{n\times n}{\mathsf{U}}\ \underset{n\times p}{\Sigma}\ \underset{p\times p}{\mathsf{V}}^\top = \mathsf{U}\begin{pmatrix} \Omega_{r\times r} & \mathbf{0}_{p-r} \\ \mathbf{0}_{n-r} & \mathbf{0}_{(n-r)\times(p-r)} \end{pmatrix}\mathsf{V}^\top,$$

where we assume wlog that $n \geq p$ so that $r := \mathtt{rank}(\mathsf{Q}) \leq p$ and Ω is diagonal with non-zero entries. The pseudoinverse of Q is the $p \times n$ matrix $\mathsf{Q}^\dagger$ defined as

$$\mathsf{Q}^\dagger := \left(\mathsf{U}\begin{pmatrix} \Omega^{-1}_{r\times r} & \mathbf{0}_{p-r} \\ \mathbf{0}_{n-r} & \mathbf{0}_{(n-r)\times(p-r)} \end{pmatrix}\mathsf{V}^\top\right)^\top = \mathsf{V}\begin{pmatrix} \Omega^{-1}_{r\times r} & \mathbf{0}_{n-r} \\ \mathbf{0}_{p-r} & \mathbf{0}_{(p-r)\times(n-r)} \end{pmatrix}\mathsf{U}^\top.$$

Intuitively: $\mathsf{Q}^\dagger$ acts as an inverse of Q on $\mathcal{R}(\mathsf{Q}) \subset \mathbb{R}^n$. Its action on $\mathbb{R}^n$ is to first project onto $\mathcal{R}(\mathsf{Q})$ and then acts as the inverse of Q on that range.

When Q is symmetric, then so is $\mathsf{Q}^\dagger$ and the expressions simplify considerably,

$$\mathsf{Q} = \mathsf{U}\begin{pmatrix} \Omega_{r\times r} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathsf{U}^\top \qquad \& \qquad \mathsf{Q}^\dagger = \mathsf{U}\begin{pmatrix} \Omega^{-1}_{r\times r} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathsf{U}^\top$$

The pseudoinverse satisfies the following two properties (exercise):

- $\mathsf{Q}^\dagger\mathsf{Q}$ is the projection onto $\mathcal{R}(\mathsf{Q}^\top)$
- $\mathsf{Q}\mathsf{Q}^\dagger$ is the projection onto $\mathcal{R}(\mathsf{Q})$.

  (so when Q is symmetric, $\mathsf{Q}^\dagger\mathsf{Q} = \mathsf{Q}\mathsf{Q}^\dagger$ by uniqueness of projections)

In fact, the pseudoinverse is the unique matrix satisfying these two properties.

Immediate corrolaries:

- $\mathsf{Q}\mathsf{Q}^\dagger\mathsf{Q} = \mathsf{Q}$
- $\mathsf{Q}^\dagger\mathsf{Q}\mathsf{Q}^\dagger = \mathsf{Q}^\dagger$

Definition (Non-Negative Matrix – Quadratic Form Definition)

A $p \times p$ real symmetric matrix Ω is called non-negative definite (written $\Omega \succeq 0$) if and only if $x^\top \Omega x \geq 0$ for all $x \in \mathbb{R}^p$. If $x^\top \Omega x > 0$ for all $x \in \mathbb{R}^p \setminus \{0\}$, then we call Ω positive definite (written $\Omega \succ 0$).

An equivalent definition is:

Definition (Non-Negative Matrix – Spectral Definition)

A $p \times p$ real symmetric matrix Ω is called non-negative definite (written $\Omega \succeq 0$) if and only the eigenvalues of Ω are non-negative. If the eigenvalues of Ω are strictly positive, then Ω is called positive definite (written $\Omega \succ 0$).

Exercise: prove that the two definitions are equivalent.

Some properties (exercise):

- $\mathsf{Q} \succeq 0$ if and only if $\mathsf{Q} = \sum_j q_j q_j^\top$ for some vectors q_i.
- $\mathsf{Q} \succeq 0$ if and only if there exists $\mathsf{A} \succeq 0$ such that $\mathsf{Q} = \mathsf{A}^2$
  - We call such an A the square root of Q and write it as $\sqrt{\mathsf{Q}}$ or $\mathsf{Q}^{1/2}$
- Any projection P satisfies $\mathsf{P} \succeq 0$.
- When $\mathsf{Q} \succeq 0$, we have

$$v_1, \ldots, v_k \in \mathcal{R}(\mathsf{Q}) \iff \mathcal{R}\left(\sum_{i=1}^k v_k v_k^\top\right) \subseteq \mathcal{R}(\mathsf{Q})$$

- Let $\mathsf{A}, \mathsf{B} \succeq 0$. Then we have:
  - $\mathsf{B} - \mathsf{A} \succeq 0 \implies \mathcal{R}(\mathsf{A}) \subseteq \mathcal{R}(\mathsf{B})$
  - $\mathcal{R}(\mathsf{A}) \subseteq \mathcal{R}(\mathsf{B}) \implies c\mathsf{B} - \mathsf{A} \succeq 0$ for some $c > 0$.

When $\mathsf{B} - \mathsf{A} \succeq 0$, we write $\mathsf{B} \succeq \mathsf{A}$. Non-negative definite matrices are partially ordered with respect to "$\succeq$" (this is called the Loewner order).

- It's clear that any p-vector can be seen as a $p \times 1$ matrix
  (why not $1 \times p$? just a convention)
- Can matrices be viewed as vectors? Yes they can.

The space $\mathbb{R}^{n \times p}$ of $n \times p$ matrices forms a real vector space of dimension np.

- Indeed, this space is isometrically isomorphic with $\mathbb{R}^{np}$
- The isomorphism is given by the vec operation,

$$\text{vec} : \mathbb{R}^{n \times p} \to \mathbb{R}^{np}$$

whose (linear) action is to stack the matrix columns into a tall np-vector,

$$\text{vec}\left\{\begin{pmatrix} v_1 & \dots & v_p \end{pmatrix}\right\} = \begin{pmatrix} v_1 \\ \vdots \\ v_p \end{pmatrix}, \quad v_i \in \mathbb{R}^n$$

- Abstractly, $\mathbb{R}^{n \times p}$ is a real vector space, whose elements are p-vectors with coordinates that are themselves elements of $\mathbb{R}^n$
  (think of partitioned matrix notation).

- The canonical basis of $\mathbb{R}^{n\times p}$ can directly be seen to be the collection

$$\mathsf{E}_{ij} := v_i u_j^\top$$

for $\{v_i\}$ the canonical basis of $\mathbb{R}^n$ and $\{u_j\}$ the canonical basis of $\mathbb{R}^p$

- Some algebra also shows that

$$\langle A, B\rangle_{\mathbb{R}^{n\times p}} := \langle \mathrm{vec}(A), \mathrm{vec}(B)\rangle_{\mathbb{R}^{np}} = \mathrm{vec}(A)^\top \mathrm{vec}(B) = \mathrm{trace}(A^\top B)$$

- A linear transformation on $\mathbb{R}^{n\times p} \to \mathbb{R}^{n\times p}$ will be a matrix transformation $\mathbb{R}^{np} \to \mathbb{R}^{np}$, hence an $np \times np$ matrix.
- In the vectorised perspective, a rank-1 transformation is $\mathrm{vec(U)vec(V)}^\top$, which maps $\mathrm{vec(A)}$ to

$$\mathrm{vec(U)vec(V)}^\top \mathrm{vec(A)} = \mathrm{trace(V}^\top \mathrm{A)vec(U)}$$

which can now easily be re-expressed in matrix form as

$$\mathrm{A} \mapsto \mathrm{trace(V}^\top \mathrm{A)U}.$$

- So by the SVD a linear $f : \mathbb{R}^{n\times p} \to \mathbb{R}^{n\times p}$ is represented as

$$f(\mathrm{A}) = \sum_{i=1}^{\mathrm{rank}(f)} \sigma_i \mathrm{trace}(\mathrm{V}_i^\top \mathrm{A})\mathrm{U}_i, \ \mathrm{vec}\{f(\mathrm{A})\} = \sum_{i=1}^{\mathrm{rank}(f)} \sigma_i \mathrm{vec}(\mathrm{U}_i)\mathrm{vec}(\mathrm{V}_i)^\top \mathrm{vec}(\mathrm{A}),$$

for $\sigma_i > 0$ and $\{\mathrm{U}_i\}$ and $\{\mathrm{V}_j\}$ orthonormal bases of $\mathbb{R}^{n\times p}$.

- Define the Kronecker matrix product as

$$\mathsf{A} \otimes \mathsf{B} = \begin{pmatrix} a_{11}\mathsf{B} & \dots & a_{1p}\mathsf{B} \\ \vdots & \ddots & \vdots \\ a_{n1}\mathsf{B} & \dots & a_{np}\mathsf{B} \end{pmatrix}$$

- Then $\mathrm{vec}(\mathsf{U})\mathrm{vec}(\mathsf{V})^\top \equiv \mathrm{vec}(\mathsf{U}) \otimes \mathrm{vec}(\mathsf{V})^\top$
- Thus, by the SVD, any linear map on $\mathbb{R}^{np}$ can be written (non-uniquely) as

$$\sum_{i=1}^{np} \underset{np\times 1}{\mathsf{A}_i} \otimes \underset{1\times np}{\mathsf{B}_i}$$

- A useful identity (optional exercise) is

$$\mathrm{vec}(\mathsf{AXB}) = (\mathsf{B}^\top \otimes \mathsf{A})\mathrm{vec}(\mathsf{X})$$

- Let's think of vectorization as turning $\mathbb{R}^{n\times p}$ into a vector space whose elements are p-vectors with coordinates that are elements $\mathbb{R}^n$. Think of $np \times np$ matrices as $p \times p$ block matrices with $n \times n$ blocks. Using the previous identity, we can show (exercise) that any linear map acts on $\mathbb{R}^{p\times p}$ as

$$\mathsf{X} \mapsto \sum_{i=1}^{np} \underset{n\times n}{\mathsf{A}_i} \underset{n\times p}{\mathsf{X}} \underset{p\times p}{\mathsf{B}_i}\,.$$

  for (non-unique) A_i and B_i.

Random Vectors and Matrices

A random vector $X = (X_1, \ldots, X_p)^\top$ is a finite collection of jointly distributed real random variables arranged as the coordinates of a vector.

The point is that we may want to make probabilistic statements on the joint behaviour of all these random variables.

- The joint distribution function of a random vector $X = (X_1, \ldots, X_p)^\top$ is

$$F_X(x_1, \ldots, x_p) = \mathbb{P}(X_1 \leq x_1, \ldots, X_p \leq x_p).$$

- Correspondingly, one defines the
  - joint frequency function, if the $\{X_i\}_{i=1}^p$ are all discrete,

$$f_X(x_1, \ldots, x_p) = \mathbb{P}(X_1 = x_1, \ldots, X_p = x_p).$$

  - the joint density function, if there exists $f_X : \mathbb{R}^p \to [0, +\infty)$ such that:

$$F_X(x_1, \ldots, x_p) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_p} f_X(u_1, \ldots, u_p)\, du_1 \ldots du_p$$

    In this case, when f_X is continuous at the point $\mathbf{x}$,

$$f_X(x_1, \ldots, x_p) = \frac{\partial^p}{\partial x_1 \ldots \partial x_p} F_X(x_1, \ldots, x_p)$$

Given the joint distribution of the random vector $X = (X_1, \ldots, X_p)^\top$, we can isolate the distribution of a single coordinate, say X_i.

- discrete case, the marginal frequency function of X_i is given by

$$f_{X_i}(x_i) = \mathbb{P}(X_i = x_i) = \sum_{x_1} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_d} f_X(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_p)$$

- In the continuous case, the marginal density function of X_i is given by

$$f_{X_i}(x_i) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_X(y_1, \ldots, y_{i-1}, x_i, y_{i+1}, \ldots, y_p)\, dy_1 \ldots dy_{i-1}\, dy_{i+1}\, dy_p.$$

- More generally, we can define the joint frequency/density of a random vector formed by a subset of the coordinates of $X = (X_1, \ldots, X_p)^\top$, say the first k
  - Discrete case:
    $f_{X_1,\ldots,X_k}(x_1, \ldots, x_k) = \sum_{x_{k+1}} \cdots \sum_{x_d} f_X(x_1, \ldots, x_k, x_{k+1}, \ldots, x_p)$.
  - Continuous case
    $f_{X_1,\ldots,X_k}(x_1, \ldots, x_k) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f_X(x_1, \ldots, x_k, x_{k+1}, \ldots, x_p)\, dx_{k+1} \ldots dx_d$.
- I.e. to marginalise we integrate/sum out the remaining random variables from the overall joint density/frequency.
- Marginals do not uniquely determine the joint distribution.

We may wish to make probabilistic statements about the potential outcomes of one random variable, if we already know the outcome of another.

For this we need the notion of a conditional density/frequency function.

If $(X_1, ..., X_p)$ is a continuous/discrete random vector, we define the conditional probability density/frequency function of $(X_1, ..., X_k)$ given $\{X_{k+1} = x_{k+1}, ..., X_p = x_d\}$ as

$$f_{X_1,...,X_k|X_{k+1},...,X_p}(x_1, ..., x_k|x_{k+1}, ..., x_d) = \frac{f_{X_1,...,X_p}(x_1, \ldots, x_k, x_{k+1}, \ldots, x_p)}{f_{X_{k+1},...,X_d}(x_{k+1}, ..., x_d)}$$

provided that $f_{X_{k+1},...,X_d}(x_{k+1}, ..., x_d) > 0$.

The random variables $X_1, \ldots, X_p$ are called independent, denoted if and only if, for all $x_1, \ldots, x_p \in \mathbb{R}$

$$F_{X_1,\ldots,X_p}(x_1, \ldots, x_p) = F_{X_1}(x_1) \times \ldots \times F_{X_p}(x_p).$$

Equivalently, $X_1, \ldots, X_p$ are independent if and only if, for all $x_1, \ldots, x_p \in \mathbb{R}$

$$f_{X_1,\ldots,X_p}(x_1, \ldots, x_p) = f_{X_1}(x_1) \times \ldots \times f_{X_p}(x_p).$$

Note that when random variables are independent, conditional distributions reduce to the corresponding marginal distributions.

Knowing the value of one of the random variables gives us no information about the distribution of the rest.

The random vector X in $\mathbb{R}^p$ is called conditionally independent of the random vector Y given the random vector Z, written

$$X \perp\!\!\!\perp_Z Y \quad \text{or} \quad X \perp\!\!\!\perp Y \mid Z,$$

if and only if, for all $x_1, \ldots, x_p \in \mathbb{R}$

$$F_{X_1,\ldots,X_p|Y,Z}(x_1, \ldots, x_p) = F_{X_1,\ldots,X_p|Z}(x_1, \ldots, x_p).$$

Equivalently, if and only if, for all $x_1, \ldots, x_p \in \mathbb{R}$

$$f_{X_1,\ldots,X_p|Y,Z}(x_1, \ldots, x_p) = f_{X_1,\ldots,X_p|Z}(x_1, \ldots, x_p).$$

Knowing Y in addition to knowing Z gives us no more information about X.

Consequence: if X is conditionally independent of Y given Z, then

$$F_{X,Y|Z} = F_{X|Y,Z} F_{Y|Z} = F_{X|Z} F_{Y|Z}$$

Consequence: $X \perp\!\!\!\perp_Z Y \iff Y \perp\!\!\!\perp_Z X$

Let $X = (X_1, \ldots, X_p)^\top$ be a random vector in $\mathbb{R}^p$ with joint density function $f_X(x_1, \ldots, x_p)$. For any $g : \mathbb{R}^p \to \mathbb{R}$, we define

$$\mathbb{E}\{g(X_1, \ldots, X_p)\} = \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} g(x_1, \ldots, x_p) f_X(x_1, \ldots, x_p)\, dx_1 \ldots dx_p.$$

Similarly, in the discrete case,

$$\mathbb{E}\{g(X_1, \ldots, X_p)\} = \sum_{x_1 \in \mathcal{X}_1} \ldots \sum_{x_p \in \mathcal{X}_p} g(x_1, \ldots, x_p) f_X(x_1, \ldots, x_p).$$

- Consequence $\mathbb{E}[X_1 + \alpha X_2] = \mathbb{E}[X_1] + \alpha \mathbb{E}[X_2]$.

The mean vector or a random vector $X = (X_1, \ldots, X_p)$ is defined as

$$\mathbb{E}[X] = \begin{pmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_p] \end{pmatrix}$$

i.e. it is the vector of means.

A random $n \times p$ matrix

$$\mathsf{X} = \begin{pmatrix} X_{11} & \dots & X_{1p} \\ \vdots & & \vdots \\ X_{n1} & \dots & X_{np} \end{pmatrix}$$

is simply a matrix whose n rows are random vectors in $\mathbb{R}^p$. Equivalently, it is a finite collection of np random variables arranged as the entries of an $n \times p$ matrix.

- Notions of joint densities/frequencies follow immediately.
- Notion of expectation follows suit, as matrix of expectations.

Consequently,

Lemma

Given a random matrix $\mathsf{X}_{n\times p}$ and deterministic matrices $A_{m\times n}$ and $B_{p\times q}$,

- $\mathbb{E}[\mathsf{X}^\top] = \left(\mathbb{E}[\mathsf{X}]\right)^\top$
- *when $n = p$,* $\mathbb{E}[\mathrm{tr}\{\mathsf{X}\}] = \mathrm{tr}\{\mathbb{E}[\mathsf{X}]\}$
- $\mathbb{E}[\mathsf{AXB}] = \mathsf{A}\mathbb{E}[\mathsf{X}]\mathsf{B}$

The covariance of a random variable X_1 with another random variable X_2 expresses the degree of linear dependency between the two.

$$\text{cov}(X_1, X_2) = \mathbb{E}\left[(X_1 - \mathbb{E}(X_1))(X_2 - \mathbb{E}(X_2))\right] \qquad (\text{if } \mathbb{E}[X_i^2] < \infty).$$

The covariance matrix or a random vector $X = (X_1, \ldots, X_p)^\top$, say $\Sigma = \{\Sigma_{ij}\}$, is a $p \times p$ symmetric matrix with entries

$$\Sigma_{ij} = \text{cov}(X_i, X_j) = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])], \quad 1 \leq i \leq j \leq p.$$

That is, the covariance matrix encodes the variances (on the diagonal) and the pairwise covariances (off the diagonal) of the coordinates of X.

Ofteh the following notation is employed:

$$\Sigma_{ii} = \sigma_i^2 \qquad \& \qquad \Sigma_{ij} = \sigma_{ij}, \quad i \neq j.$$

where $\sigma_i = \sqrt{\text{var}(X_i)}$ is the standard deviation of X_i.

It can be easily checked that

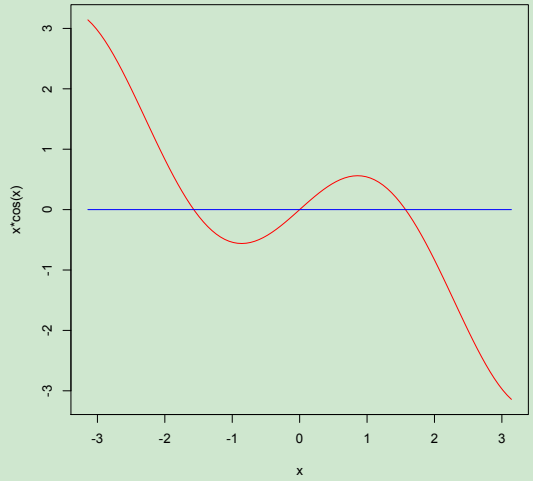
$$\Sigma = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top] = \mathbb{E}[XX^\top] - \mathbb{E}[X]\mathbb{E}[X]^\top.$$

Example ($\mathrm{cov}(X, Y) = 0 \not\Rightarrow$ Independence)

Let $X \sim \mathrm{Unif}[-\pi, \pi]$ and define

$$Y = \cos(X).$$

- Clearly X and Y are not independent.
- To the contrary, they are perfectly dependent.
- Their covariance is, nevertheless, zero!



The function $x\,cos(x)$

Concretely, we calculate

$$\mathbb{P}[Y > 0] = 1/2 \qquad \text{but} \qquad \mathbb{P}[Y > 0 | X \in (-\pi, -2)] = 1.$$

Despite this, we have

$$\mathrm{cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \int_{-\pi}^{+\pi} x \cos(x) \frac{1}{2\pi} dx - 0 = 0.$$

Why: Because some non-linear dependencies cannot be detected by covariance...

Example ($\text{cov}(X, Y) = 0 \not\Rightarrow$ Independence)

Let X and Y have joint density

$$f_{XY}(x, y) = \begin{cases} 1/\pi & \text{si } x^2 + y^2 \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Note that $\mathbb{E}[X] = \mathbb{E}[Y] = 0$ by symmetry. Hence, $\text{cov}\{X, Y\} = \mathbb{E}[XY]$. But

$$\mathbb{E}[XY] = \iint_{x^2+y^2\leq 1} xy \frac{1}{\pi} dxdy = \iint_{x^2+y^2\leq 1, y\geq 0} xy \frac{1}{\pi} dxdy + \iint_{x^2+y^2\leq 1, y<0} xy \frac{1}{\pi} dxdy$$

The two terms are equal, by symmetry. Moreover,

$$\iint_{x^2+y^2\leq 1, y\geq 0} xy \frac{1}{\pi} dxdy = \frac{1}{\pi} \int_{-1}^{1} x \int_{0}^{1-x^2} y dy dx = \frac{1}{\pi} \int_{-1}^{1} x \frac{(1-x^2)^2}{2} dx = 0$$

and so the covariance is zero. But X and Y are clearly dependent, since knowing X restricts the possible values of Y.

Lemma

Let X be a random $p \times 1$ vector such that $\mathbb{E}\|X\|^2 < \infty$ and with covariance Σ. Given a A a $q \times p$ real matrix, the covariance of the $q \times 1$ random vector $\mathrm{A}X$ is $\mathrm{A}\Sigma\mathrm{A}^\top$.

Corollary (Covariance of Projections)

Let Y be a random $d \times 1$ vector such that $\mathbb{E}\|Y\|^2 < \infty$. Let $\beta, \gamma \in \mathbb{R}^d$ be fixed vectors. If Ω denotes the covariance matrix of Y,

- *the variance of $\beta^\top Y$ is $\beta^\top \Omega \beta$;*
- *the covariance of $\beta^\top Y$ with $\gamma^\top Y$ is $\gamma^\top \Omega \beta$.*

Proposition (Non-Negative and Covariance Matrices)

Let Ω be a real symmetric matrix. Then Ω is non-negative definite if and only if Ω is the covariance matrix of some random variable Y.

Proof.

Exercise. $\square$

Let X and Y be centred random vectors in $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively. The cross-covariance between X and Y is the $n \times m$ matrix

$$\text{cov}\{X, Y\} := \Sigma_{XY} := \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])^\top] = \mathbb{E}[XY^\top] - \mathbb{E}[X]\mathbb{E}[Y]^\top.$$

Note that this is not symmetric (and so, in general, will not be non-negative definite)

$$\text{cov}\{X, Y\} = \text{cov}\{Y, X\}^\top.$$

If we concatenate into an $(n+m)$-dimensional random vector $Z = (X^\top \; Y^\top)^\top$, and use block notation, we see that

$$\Sigma_Z = \begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{XY}^\top & \Sigma_Y \end{pmatrix}.$$

i.e. an $n \times m$ matrix is a cross covariance if and only if it can be represented as the off-diagonal block of some $(n+m) \times (n+m)$covariance matrix.

The support of a random vector X in $\mathbb{R}^p$, is defined as

$$\text{supp}\{X\} := \{x \in \mathbb{R}^p : \mathbb{P}[\|X - x\| < \epsilon] > 0, \forall \epsilon > 0\}$$

Intuitively, the support is the region of $\mathbb{R}^p$ that X can reach. It can be shown (exercise) that $\text{supp}\{X\}$ is a closed set, indeed the smallest closed set F such that $\mathbb{P}[X \in F] = 1$.

The covariance provides some information on the support:

Lemma (Support and Covariance)

Let X be a random vector in $\mathbb{R}^p$ with mean μ_X and covariance Σ_X. Then,

1. $\text{supp}\{X\} \subseteq \mathcal{R}(\Sigma_X) + \mu_X$.
2. $(\Sigma_X \Sigma_X^\dagger)(X - \mu_X) = X - \mu_X$ *almost surely.*
3. $(\Sigma_X \Sigma_X^\dagger)\Sigma_{XY} = \Sigma_{XY}$ *for any random vector Y with finite second moment*

where we recall that $\Sigma_X \Sigma_X^\dagger = \mathsf{H}_X$ is the projection onto $\mathcal{R}(\Sigma_X)$.

Proof.

We first remark that (1) $\iff$ (2). To see this, recall that $\mathsf{H}_X u = u$ if and only if $u \in \mathcal{R}(\mathsf{H}_X)$. Thus (2) is equivalent to stating that $\mathbb{P}\{X - \mu_X \in \mathcal{R}(\mathsf{H}_X)\} = 1$. Since $\mathcal{R}(\mathsf{H}_X) = \mathcal{R}(\Sigma_X)$, and observing that $\mathcal{R}(\mathsf{H}_X) + \mu_X$ is closed, the last statement is equivalent to (1).

To establish (2), write $X - \mu_X = \mathsf{H}_X(X - \mu_X) + (I - \mathsf{H}_X)(X - \mu_X)$ and note that

$$\mathrm{cov}\{(I - \mathsf{H}_X)(X - \mu_X)\} = (I - \mathsf{H}_X)\Sigma_X(I - \mathsf{H}_X) = (\Sigma_X - \Sigma_X\Sigma_X^{\dagger}\Sigma_X)(I - \mathsf{H}_X) = 0.$$

Consequently $(I - \mathsf{H}_X)(X - \mu_X) = \mathbb{E}[(I - \mathsf{H}_X)(X - \mu_X)]$ almost surely. But $\mathbb{E}[(I - \mathsf{H}_X)(X - \mu_X)] = (I - \mathsf{H}_X)\mathbb{E}[X - \mu_X] = 0$. In summary, $X - \mu_X = \mathsf{H}_X(X - \mu_X)$ almost surely, establishing (2).
For (3), it suffices to observe that

$$\mathsf{H}_X\Sigma_{XY} = \mathsf{H}_X\mathbb{E}[(X - \mu_X)(Y - \mu_Y)^\top] = \mathbb{E}[\mathsf{H}_X(X - \mu_X)(Y - \mu_Y)^\top]$$

$$\overset{(2)}{=} \mathbb{E}[(X - \mu_X)(Y - \mu_Y)^\top] = \Sigma_{XY}.$$

□

Lemma

Let Y be a random vector in $\mathbb{R}^p$ with covariance Σ_Y. Then,

- $\mathrm{cov}\{(\Sigma_Y^\dagger)^{1/2} Y\} = \mathsf{H}_Y$, where H_Y is the projection onto $\mathcal{R}(\Sigma_Y)$.
- consequently, when Σ_Y^{-1} exists, we have $\mathrm{cov}\{\Sigma_Y^{-1/2} Y\} = \mathsf{I}_{p\times p}$

Proof.

We calculate

$$\mathrm{cov}\{(\Sigma_Y^\dagger)^{1/2} Y\} = (\Sigma_Y^\dagger)^{1/2}\Sigma_Y(\Sigma_Y^\dagger)^{1/2} = (\Sigma_Y^\dagger)^{1/2}\Sigma_Y^{1/2}\Sigma_Y^{1/2}(\Sigma_Y^\dagger)^{1/2}.$$

By the definition of pseudoinverse and the fact that $\Sigma_Y \succeq 0$,

$$(\Sigma_Y^\dagger)^{1/2}\Sigma_Y^{1/2} = (\Sigma_Y^{1/2})^\dagger\Sigma_Y^{1/2}$$

and the RHS is the projection onto $\mathcal{R}(\Sigma_Y^{1/2})$. Finally, $\mathcal{R}(\Sigma_Y) = \mathcal{R}(\Sigma_Y^{1/2})$, again due to the spectral theorem, and the proof is complete. □

Lemma (Matrix Correlation Inequality)

Let $Z = (X^\top \; Y^\top)^\top$ *be be comprised of two centred random vectors in* $\mathbb{R}^n$ *and* $\mathbb{R}^m$*, respectively, with covariance*

$$\Sigma_Z = \begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{XY}^\top & \Sigma_Y \end{pmatrix}.$$

Then

$$\Sigma_X - \Sigma_{XY}\Sigma_Y^\dagger\Sigma_{XY}^\top \succeq 0.$$

If Σ_Z *is non-singular, then* Σ_Y *is necessarily so too and*

$$\Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^\top \succ 0.$$

The matrix $\Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^\top$ is called the Schur complement of Σ_Y in Σ_Z

When $\Sigma_Z \succ 0$, the last inequality can be re-written as

$$\Sigma_X^{-1}\Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^\top \prec \mathsf{I}$$

which explains the term "correlation inequality"

Proof.

Define the zero-mean random vector $\varepsilon = X - \Sigma_{XY}\Sigma_Y^\dagger Y$, and notice that

$$0 \preceq \Sigma_\varepsilon = \mathbb{E}[\varepsilon\varepsilon^\top] = \mathbb{E}[XX^\top] + \Sigma_{XY}\underbrace{\Sigma_Y^\dagger \mathbb{E}[YY^\top]\Sigma_Y^\dagger}_{=\Sigma_Y^\dagger \Sigma_Y \Sigma_Y^\dagger = \Sigma_Y^\dagger}\Sigma_{XY}^\top$$

$$-\underbrace{\mathbb{E}[XY^\top]}_{=\Sigma_{XY}}\Sigma_Y^\dagger\Sigma_{XY}^\top - \Sigma_{XY}\Sigma_Y^\dagger\underbrace{\mathbb{E}[YX^\top]}_{=\Sigma_{XY}^\top} = \Sigma_X - \Sigma_{XY}\Sigma_Y^\dagger\Sigma_{XY}^\top$$

For the second part, we will argue by contradiction. Assume that $\Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^\top$ is singular. This means that there exists an $x \in \mathbb{R}^n \setminus \{0\}$ such that $x^\top\Sigma_X x - x^\top\Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^\top x = 0$. Now define the $(n+m)\times 1$ vector

$$u = \begin{pmatrix} x \\ -\Sigma_Y^{-1}\Sigma_{XY}^\top x \end{pmatrix}.$$

Since $x \neq 0$, it also holds that $u \neq 0$. Now observe that

$$u^\top\Sigma_Z u = \begin{pmatrix} x^\top & -x^\top\Sigma_{XY}\Sigma_Y^{-1} \end{pmatrix}\begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{XY}^\top & \Sigma_Y \end{pmatrix}\begin{pmatrix} x \\ -\Sigma_Y^{-1}\Sigma_{XY}^\top x \end{pmatrix}$$
$$= x^\top\Sigma_X x - x^\top\Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^\top x = 0.$$

Since $u \neq 0$, this contradicts the assumption that Σ_Z is non-singular. □

Let X be a random vector in $\mathbb{R}^p$ with non-singular covariance Σ_X. The precision matrix of X is defined as

$$\Theta_X := \Sigma_X^{-1}.$$

For now we just need the following (a direct consequence of previous lemma)

Lemma (2×2 Block Precision Matrix)

Let Σ be an $(n+m) \times (n+m)$ non-singular covariance

$$\Sigma = \begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{XY}^\top & \Sigma_Y \end{pmatrix}.$$

Then $\Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^\top$ and $\Sigma_Y - \Sigma_{XY}^\top\Sigma_Y^{-1}\Sigma_{XY}$ are strictly positive-definite and we have the following expression for the precision matrix

$$\Theta = \Sigma^{-1} = \begin{pmatrix} (\Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^\top)^{-1} & -(\Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^\top)^{-1}\Sigma_{XY}\Sigma_Y^{-1} \\ -\Sigma_Y^{-1}\Sigma_{XY}{}^\top(\Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^\top)^{-1} & (\Sigma_Y - \Sigma_{XY}^\top\Sigma_X^{-1}\Sigma_{XY})^{-1} \end{pmatrix}$$

- Proof is immediate once the inverses are well defined (just multiply to verify).
- Notice how there are Schur complements and their inverses appearing everywhere.

The correlation between X_1 and X_2 is defined as

$$\text{corr}(X_1, X_2) = \frac{\text{cov}(X_1, X_2)}{\sqrt{\text{var}(X_1)\text{var}(X_2)}}.$$

Conveys equivalent dependence information to covariance. Advantages: (1) it is invariant to changes of scale, (2) can be be understood in absolute terms (ranges in $[-1, 1]$), as a result of the correlation inequality[1] (Cauchy-Schwarz):

$$|\text{corr}(X_1, X_2)| \leq \sqrt{\text{var}(X_1)\text{var}(X_2)}.$$

The correlation matrix $\mathrm{R} = \{\rho_{ij}\}$ of a random vector $X = (X_1, \ldots, X_p)^\top$, is a $p \times p$ symmetric matrix with entries

$$\rho_{ij} = \text{corr}(X_i, X_j), \quad 1 \leq i \leq j \leq p.$$

Note that the correlation matrix is well-defined whenever $\text{var}(X_i) > 0$ for all $1 \leq i \leq n$, i.e. none of the coordinates are degenerate random variables.

[1] compare now to the *matrix correlation inequality*

Equivalently, the correlation matrix is the covariance matrix of the standardised vector

$$X = (X_1/\sigma_1, \ldots, X_p/\sigma_p)^\top,$$

where $\sigma_i^2 = \mathrm{var}(X_i)$. Thus, recalling that $\mathrm{cov}(\mathsf{A}X) = \mathsf{A}\Sigma\mathsf{A}^\top$, we have

$$\mathsf{R} = \begin{pmatrix} \mathrm{var}(X_1) & & 0 \\ & \ddots & \\ 0 & & \mathrm{var}(X_p) \end{pmatrix}^{-\frac{1}{2}} \Sigma \begin{pmatrix} \mathrm{var}(X_1) & & 0 \\ & \ddots & \\ 0 & & \mathrm{var}(X_p) \end{pmatrix}^{-\frac{1}{2}}$$

$$= \mathrm{diag}\{\sigma_1^{-1}, \ldots, \sigma_p^{-1}\}\, \Sigma\, \mathrm{diag}\{\sigma_1^{-1}, \ldots, \sigma_p^{-1}\}$$

Thus correlation matrices are non-negative definite.

Let X and Y be centred random vectors in $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively. The cross-correlation matrix of X and Y is the $n \times m$ matrix R_{XY} with entries

$$\frac{\mathrm{cov}(X_i, Y_j)}{\sqrt{\mathrm{var}(X_i)\mathrm{var}(Y_j)}}, \quad i = 1, ..., n\,;\, j = 1, ..., m.$$

Again, if we concatenate into an $(n+m)$-dimensional random vector $Z = (X^\top\ Y^\top)^\top$, and use block notation, we may write

$$\Sigma_Z = \begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{XY}^\top & \Sigma_Y \end{pmatrix} \quad \& \quad \mathsf{R}_Z = \begin{pmatrix} \mathsf{R}_X & \mathsf{R}_{XY} \\ \mathsf{R}_{XY}^\top & \mathsf{R}_Y \end{pmatrix}$$

We now easily check that:

$$\mathsf{R}_{XY} = (\mathrm{diag}(\Sigma_X))^{-1/2} \Sigma_{XY} (\mathrm{diag}(\Sigma_Y))^{-1/2}.$$

The moment generating function (MGF) $M_X(t) : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ of a scalar random variable X is defined as

$$M_X(t) = \mathbb{E}\Big[e^{tX}\Big], \qquad t \in \mathbb{R}.$$

It need not be finite for $t \neq 0$. But when it is finite zero, magic happens:

Theorem

Let X and Y be scalar random variable, and assume that $M_X(t) < \infty$ and $M_Y(t) < \infty$ for all $t \in I = (-\epsilon, \epsilon)$ for some $\epsilon > 0$. Then, it holds that

1. *M_X is infinitely differentiable on I*
2. *$\mathbb{E}[|X|^k] < \infty$ and $\mathbb{E}[X^k] = \frac{d^k M_X}{dt^k}(0)$, for all $k \geq 1$.*
3. *$F_X = F_Y$ on $\mathbb{R} \iff M_Y = M_X$ on I.*
4. *if $X \perp\!\!\!\perp Y$, then M_{X+Y} is finite and equal to $M_X M_Y$ on I.*

The moment generating function (MGF) of a random vector W in $\mathbb{R}^d$ is defined as

$$M_W(\theta) = \mathbb{E}[e^{\theta^\top W}], \qquad \theta \in \mathbb{R}^d,$$

and need not be finite for $\theta \neq 0$.

When the MGF exists on an open ball at the origin:

- it characterises the distribution of the corresponding random vector, as in the scalar case.
- consequently, it factorizes into two marginal MGFs if and only if the corresponding random vectors are independent:

$$\boxed{X_{n\times 1} \text{ independent of } Y_{m\times 1}}$$

$$\Longleftrightarrow$$

$$\boxed{\mathbb{E}[e^{\beta^\top X+\gamma^\top Y}] = \mathbb{E}[e^{\beta^\top X}] \times \mathbb{E}[e^{\gamma^\top Y}], \qquad \forall\, \beta \in \mathbb{R}^n \,\&\, \gamma \in \mathbb{R}^m}$$

Gaussian Vectors

If for some $\mu \in \mathbb{R}$ and some $\sigma \in (0, +\infty)$ a random variable X has density

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 \right\}, \quad x \in \mathbb{R},$$

then X is called a Gaussian (or Normal) random variable), and we write

$$X \sim N(\mu, \sigma^2).$$

This is indeed a valid probability density, by a simple change-of-variables, establishing existence:

$$\begin{aligned} \left[\int_{\mathbb{R}} f_X(x)\,dx\right]^2 &= \int_{\mathbb{R}}\int_{\mathbb{R}} f(x)f(y)\,dxdx = \frac{1}{2\pi\sigma^2}\int_{\mathbb{R}}\int_{\mathbb{R}} \exp\left\{ -\frac{(x-\mu)^2+(y-\mu)^2}{2\sigma^2} \right\} dxdy \\ &= \frac{1}{2\pi\sigma^2}\int_0^{2\pi} \theta\,d\theta \int_0^{+\infty} r \exp\left\{ -\frac{r^2}{2\sigma^2} \right\} dr = 1. \end{aligned}$$

By convention, a constant $\mu \in \mathbb{R}$ is considered to be a $N(\mu, 0)$ random variable Hence Gaussian random variables need not have density unless $\sigma > 0$.

A random variable $Z \sim N(0,1)$ is called a standard Gaussian random variable. We write ϕ for its PDF and Φ for its CDF.

Lemma

Given $\mu \in \mathbb{R}$ and $\sigma \in (0, +\infty)$, one has $X \sim N(\mu, \sigma^2)$ if and only $X = \sigma Z + \mu$ for some Gaussian random variable $Z \sim N(0,1)$.

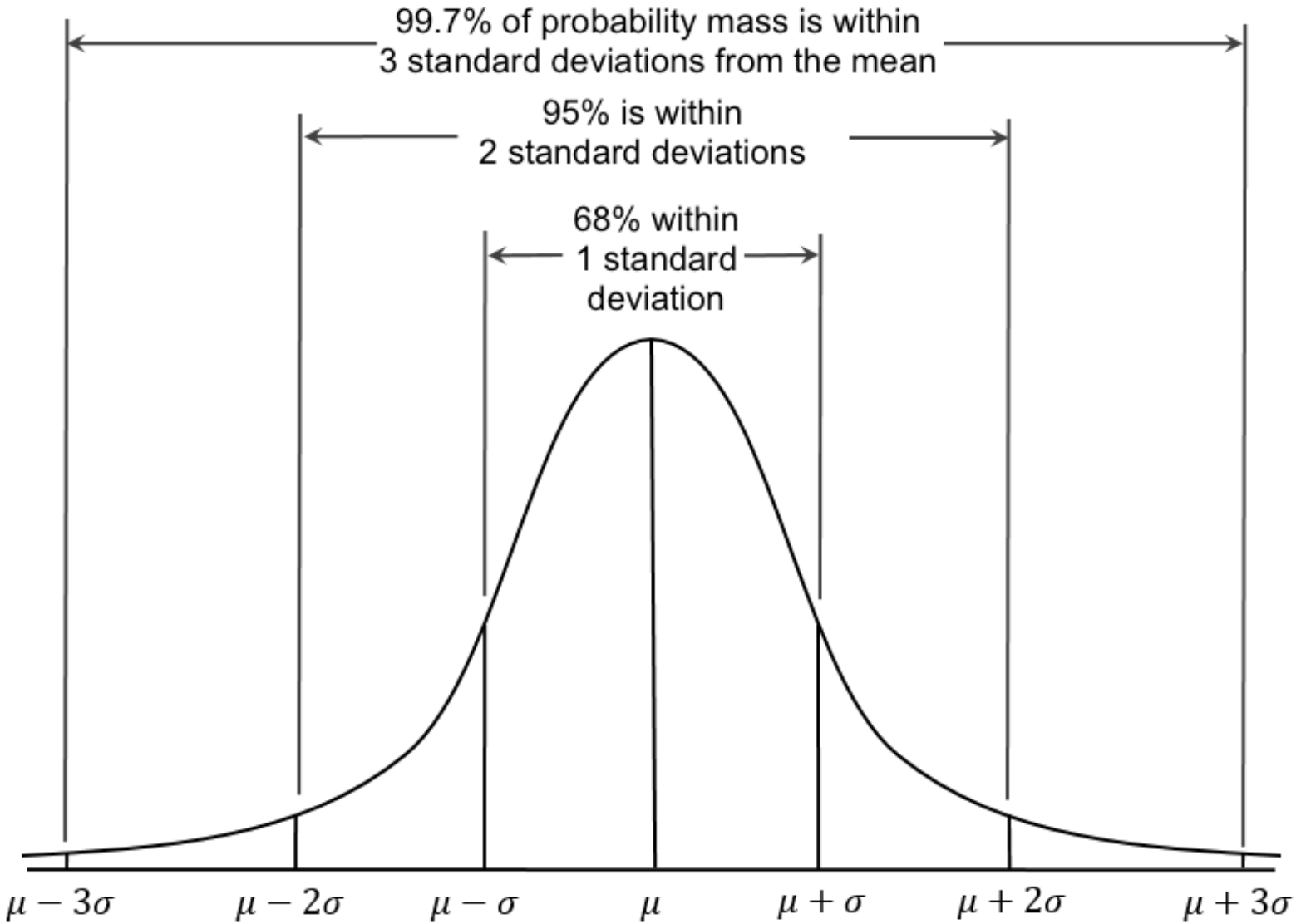
Proof.

Changing variables to $z = (x - \mu)/\sigma$, one has

$$F_X(y) = \int_{-\infty}^{y} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} dy$$

$$= \int_{-\infty}^{\frac{y-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} \exp\{-z^2/2\} dz = \Phi\left(\frac{y-\mu}{\sigma}\right)$$

and so if one defines $Z = \sigma^{-1/2}(X - \mu)$ then the 'if' part is proven. Starting with $Z \sim \mathrm{N}(0,1)$, and following the same steps in reverse gives the 'only if' part, with $X = \sigma Z + \mu$. □


99.7% of probability mass is within
3 standard deviations from the mean
95% is within
2 standard deviations
68% within
1 standard
deviation
$\mu - 3\sigma$
$\mu - 2\sigma$
$\mu - \sigma$
μ
$\mu + \sigma$
$\mu + 2\sigma$
$\mu + 3\sigma$

Lemma (Moment Generating Function)

Given $\mu \in \mathbb{R}$ and $\sigma \in [0, +\infty)$, the moment generating function $M_X(t) = \mathbb{E}[e^{tX}]$ of $X \sim N(\mu, \sigma^2)$ satisfies

$$M_X(t) = \exp\{t\mu + t^2\sigma^2/2\}.$$

This being finite for all $t \in \mathbb{R}$ implies that all moments of X exist, and its central moments are

$$\mathbb{E}\big[(X - \mathbb{E}[X])^k\big] = \mathbb{E}\big[(X - \mu)^k\big] = \mathbb{E}\big[\sigma^k Z^k\big] = \begin{cases} 0 & \text{for } k \text{ odd}, \\ \sigma^k (k-1)!! & \text{for } k \text{ even}. \end{cases}$$

Consequently $\mu = \mathbb{E}[X]$ is the mean and $\sigma^2 = \mathrm{var}[X]$ is the variance.

Proof.

By definition, and the change of variables $y = x - \mu - \sigma^2 t$ we find that

$$\begin{aligned} M_X(t) &= \mathbb{E}[e^{tX}] = \frac{1}{\sigma\sqrt{2\pi}} \int_{\mathbb{R}} \exp\{tx - (x-\mu)^2/(2\sigma^2)\} dx \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_{\mathbb{R}} \exp\left\{-[(x-\mu-\sigma^2 t)^2 - 2\mu\sigma^2 t - \sigma^4 t^2]/2\sigma^2\right\} dx \\ &= \exp\{t\mu + t^2\sigma^2/2\} \underbrace{\frac{1}{\sigma\sqrt{2\pi}} \int_{\mathbb{R}} \exp\left\{-y^2/(2\sigma^2)\right\} dy}_{=1}. \end{aligned}$$

With the MGF in hand, we can now calculate the moments of a $Z \sim N(0,1)$,

$$\mathbb{E}[Z^k] = \left.\frac{dM_X}{du}\right|_{u=0} = \begin{cases} 0 & \text{for } k \text{ odd}, \\ (k-1)!! & \text{for } k \text{ even}. \end{cases}$$

With a change of variables, this yields the central moments of $X \sim \mathrm{N}(\mu, \sigma^2)$ as

$$\mathbb{E}\big[(X - \mathbb{E}[X])^k\big] = \mathbb{E}\big[(X-\mu)^k\big] = \mathbb{E}\big[\sigma^k Z^k\big] = \begin{cases} 0 & \text{for } k \text{ odd}, \\ \sigma^k (k-1)!! & \text{for } k \text{ even}. \end{cases}$$

□

Definition (Multivariate Gaussian Distribution)

A random vector Y in $\mathbb{R}^d$ is Gaussian if and only if $\beta^\top Y$ is a Gaussian random variable for all deterministic vectors $\beta \in \mathbb{R}^d$.

Observation: From the definition if follows that Y must have some well-defined mean vector μ and some well defined covariance matrix Σ.

To see this note that since $\mathbb{E}\{(\beta^\top Y)^2\} < \infty$ for all β, then we can successively pick β to be equal to each canonical basis vector and conclude that each coordinate has finite variance and thus $\mathbb{E}\| Y \|^2 < \infty$.

So all the means, variances and covariances of its coordinates are well defined.

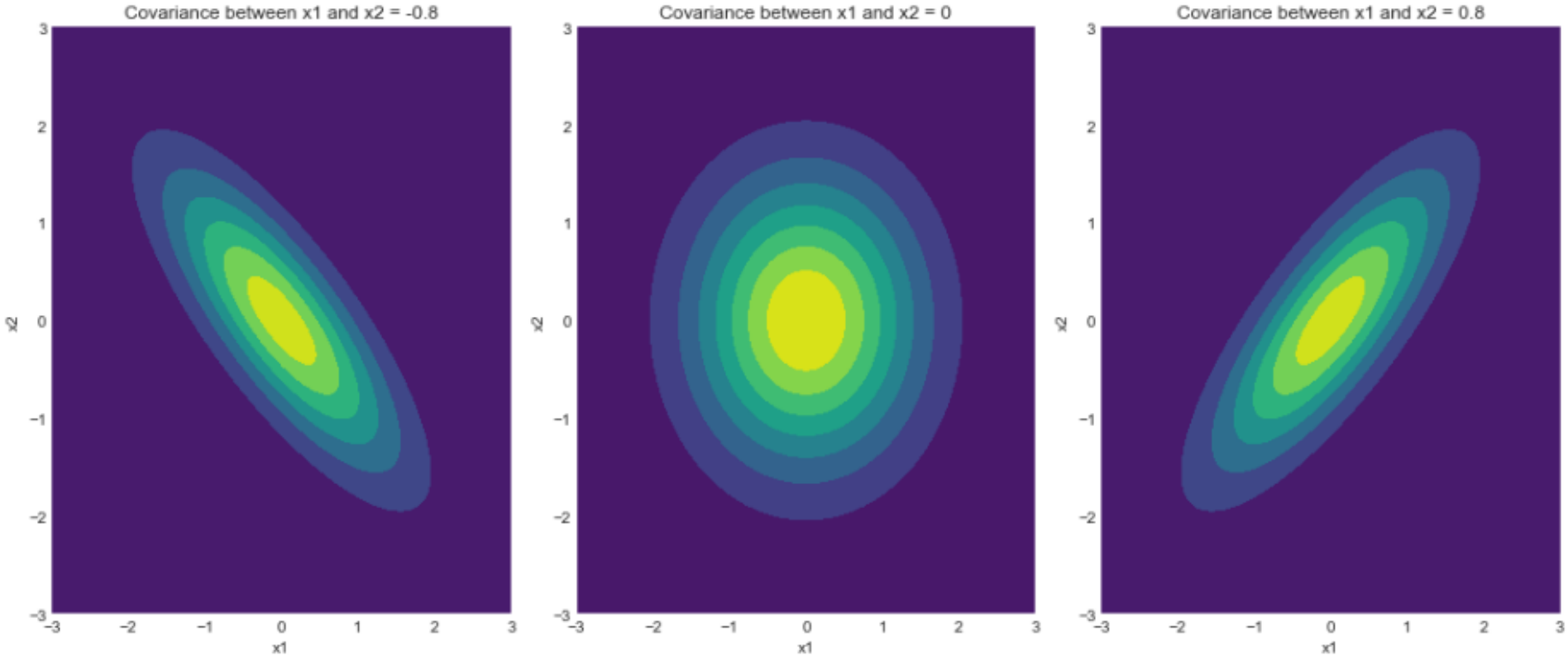
Then, the mean vector μ and covariance matrix Σ can be (uniquely) determined entrywise by equating

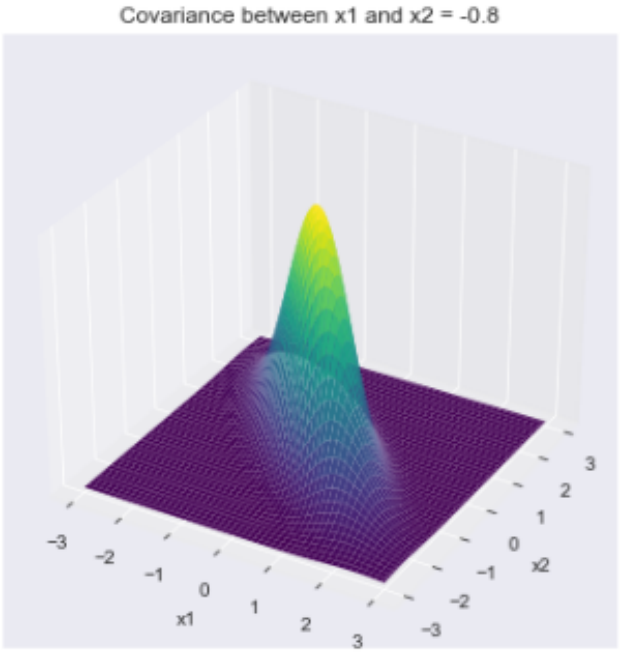
$$\mu_i = \mathbb{E}[\mathrm{e}_i^\top Y] \qquad \& \qquad \Sigma_{ij} = \mathrm{cov}\{\mathrm{e}_i^\top Y, \mathrm{e}_j^\top Y\}.$$

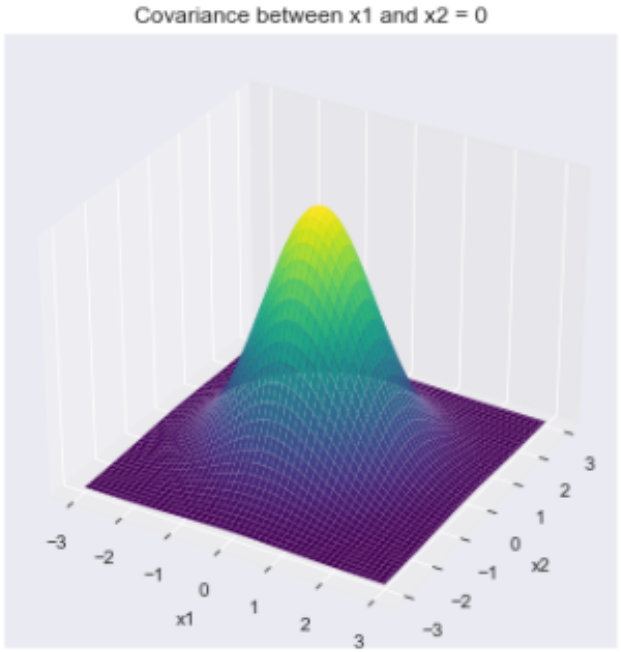
where e_j is the jth canonical basis vector

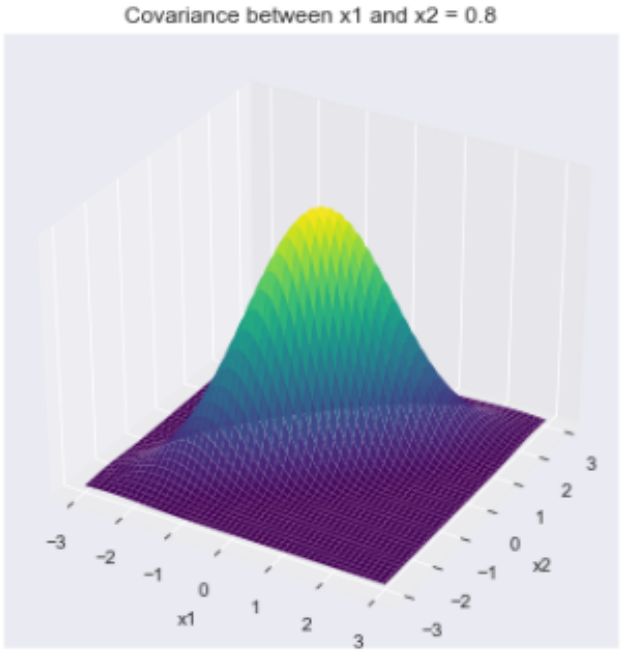
$$\mathrm{e}_j = (0\ ,\ 0\ , \ldots, \underbrace{1}_{j^{th}\ \text{position}}, \ldots,\ 0\ ,\ 0)^\top$$

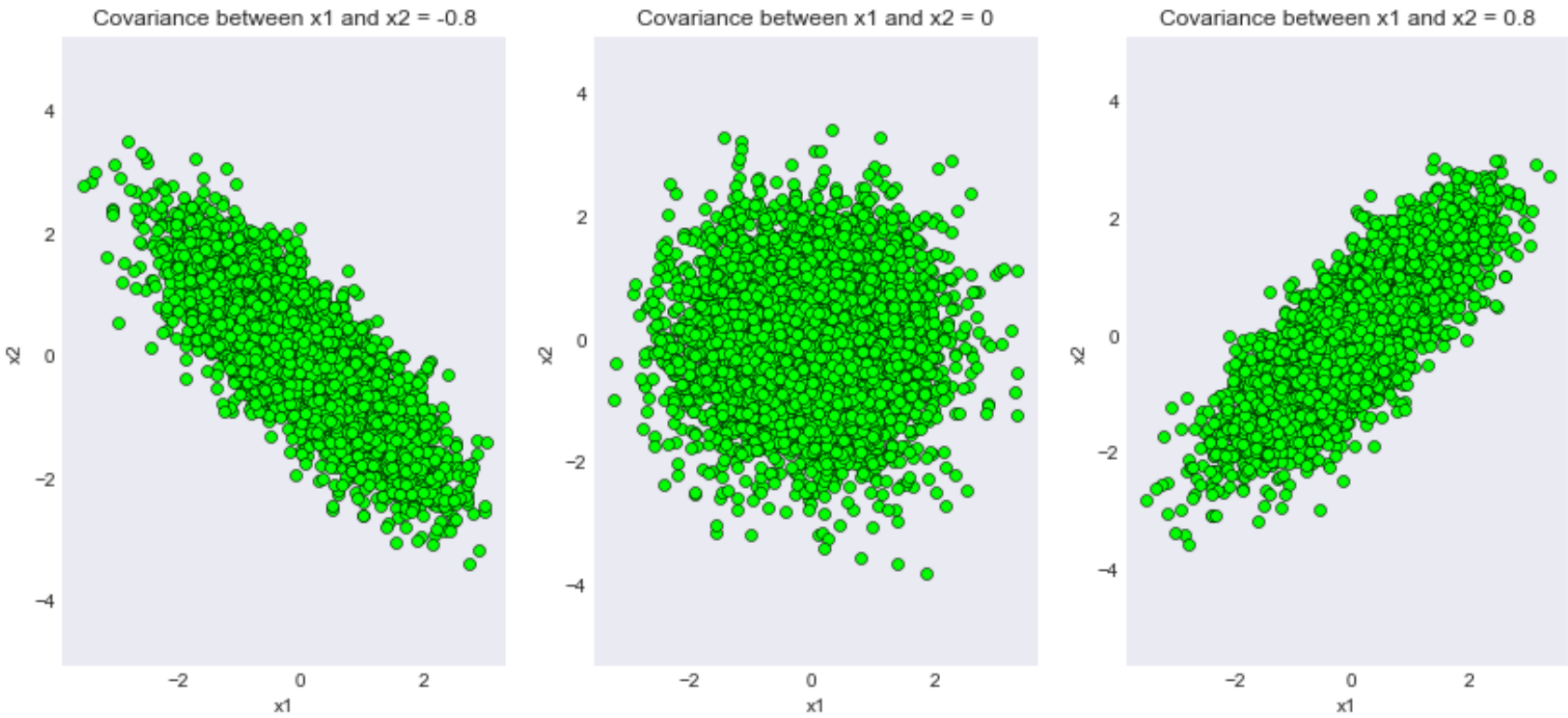
1. MGF of $Y \sim \mathcal{N}(\mu, \Sigma)$: $\boxed{M_Y(u) = \exp\left(u^\top \mu + \frac{1}{2} u^\top \Sigma u\right)}$.
2. $Y \sim \mathcal{N}(\mu_{p\times 1}, \Sigma_{p\times p})$ and given $B_{n\times p}$ and $\theta_{n\times 1}$, then
$$\theta + BY \sim \mathcal{N}(\theta + B\mu, B\Sigma B^\top).$$
3. Marginals are Gaussian (converse NOT true).
4. If $Y \sim \mathcal{N}(\mu_{p\times 1}, \Sigma_{p\times p})$,
$$AY \text{ independent of } BY \iff A\Sigma B^\top = 0.$$
5. Immediate corollary of (4): if $(X^\top Y^\top)^\top$ is a Gaussian vector,
$$Y \perp\!\!\!\perp X \iff \Sigma_{XY} = 0 \iff \operatorname{cov}\left\{\binom{X}{Y}\right\} = \begin{pmatrix} \Sigma_X & 0 \\ 0 & \Sigma_Y \end{pmatrix}$$
6. $\mathcal{N}(\mu, \Sigma)$ PDF, if $\Sigma \succ 0$ $\boxed{f_Y(y) = \frac{1}{(2\pi)^{p/2}\,|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(y-\mu)^\top \Sigma^{-1}(y-\mu)\right\}}$.
7. $Y \sim N(\mu, \Sigma_{p\times p}) \iff Y = \mu + \sum_{j=1}^p \lambda_j^{1/2} Z_j u_j$, for $Z_j \overset{iid}{\sim} N(0,1)$, and and $\{(\lambda_j, u_j)\}_{j=1}^p$ the eigenvalues/vectors of Σ
8. If $Y \sim \mathcal{N}(\mu, \Sigma)$, then $\operatorname{supp}\{Y\} = \mathcal{R}(\Sigma) + \mu$

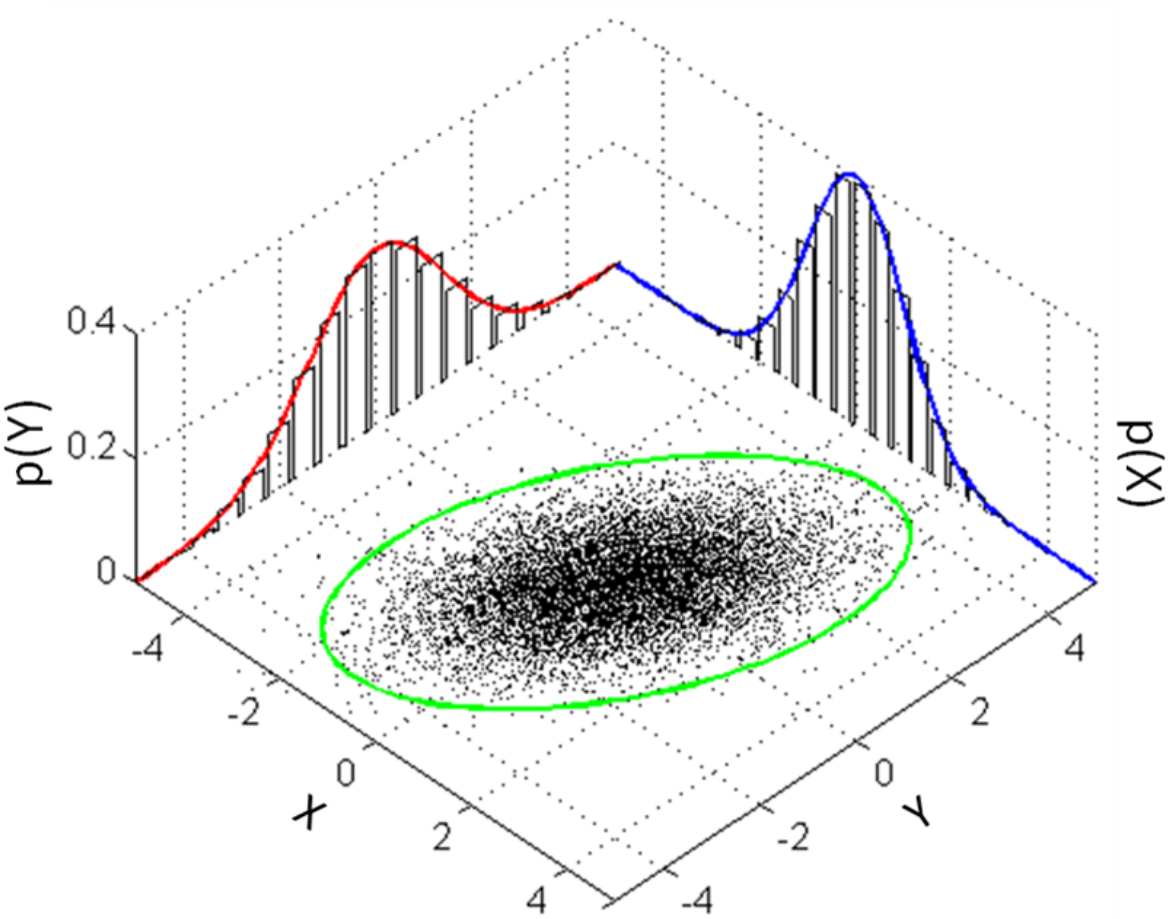

Covariance between x1 and x2 = -0.8
Covariance between x1 and x2 = 0
Covariance between x1 and x2 = 0.8
x1
x2
3
2
1
0
−1
−2
−3

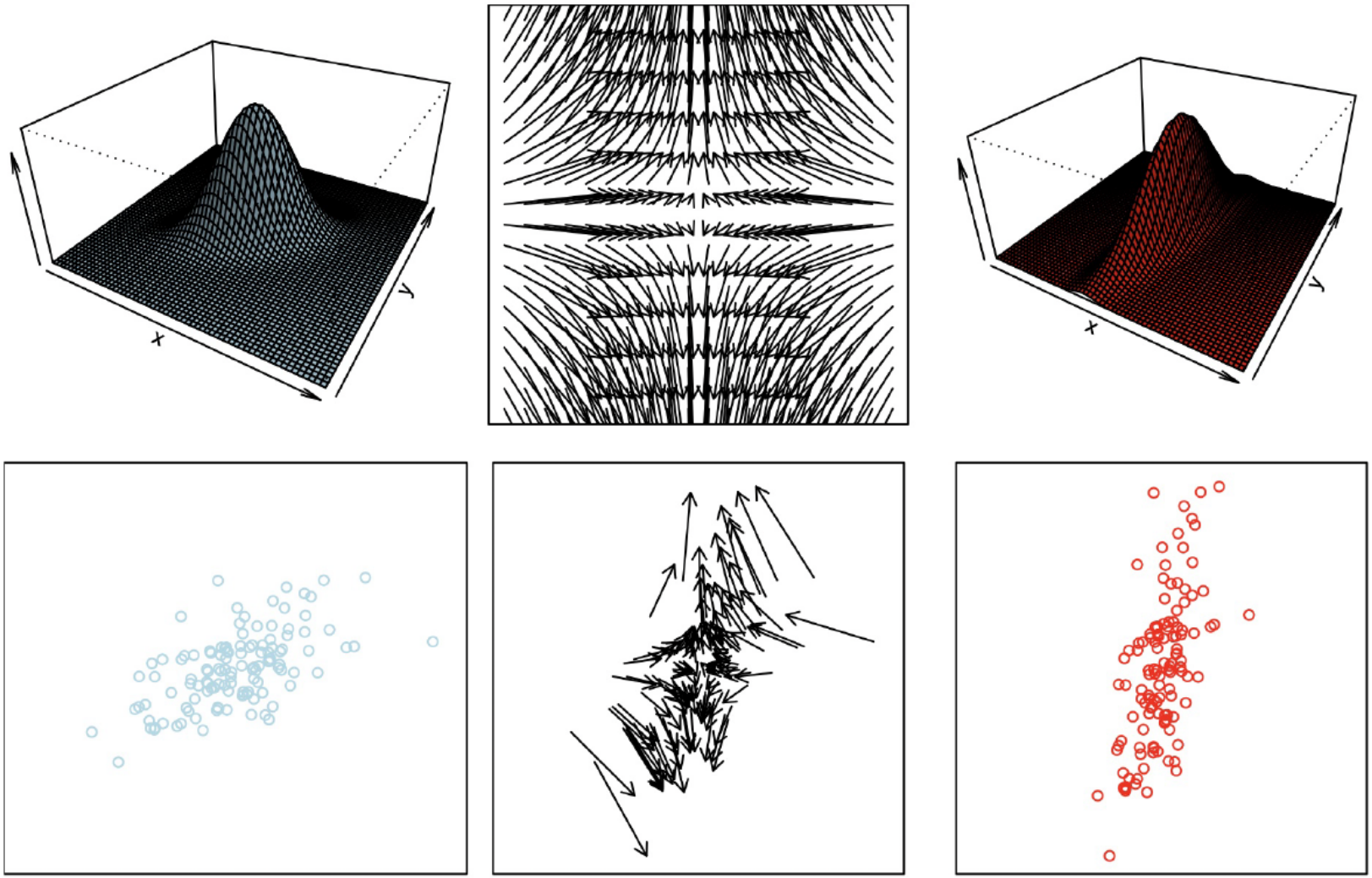

Covariance between x1 and x2 = -0.8
x1
x2


Covariance between x1 and x2 = 0
x1
x2


Covariance between x1 and x2 = 0.8
x1
x2


Covariance between x1 and x2 = -0.8
Covariance between x1 and x2 = 0
Covariance between x1 and x2 = 0.8
x1
x2
4
2
0
−2
−4


0.4
0.2
0
p(Y)
-4
-2
0
2
4
X
4
2
0
-2
-4
Y
p(X)



$$Y = \mathrm{A}X$$

Proposition (Moment Generating Function)

The moment generating function of $Y \sim \mathcal{N}(\mu, \Sigma)$ *is*

$$M_Y(u) = \exp\left(u^\top \mu + \tfrac{1}{2} u^\top \Sigma u\right)$$

Proof.

Let $v \in \mathbb{R}^d$ be arbitrary. Then $v^\top Y$ is scalar Gaussian with mean $v^\top \mu$ and variance $v^\top \Sigma v$. Hence it has moment generating function:

$$M_{v^\top Y}(t) = \mathbb{E}\left(e^{t v^\top Y}\right) = \exp\left\{t(v^\top \mu) + \frac{t^2}{2}(v^\top \Sigma v)\right\}.$$

Now take $t = 1$ and observe that

$$M_{v^\top Y}(1) = \mathbb{E}\left(e^{v^\top Y}\right) = M_Y(v).$$

Combining the two, we conclude that

$$M_Y(v) = \exp\left(v^\top \mu + \frac{1}{2} v^\top \Sigma v\right), \quad v \in \mathbb{R}^d.$$

□

Proposition (Closure Under Affine Transformation)

For $Y \sim \mathcal{N}(\mu_{p\times 1}, \Sigma_{p\times p})$ and given $B_{n\times p}$ and $\theta_{n\times 1}$, we have

$$\theta + BY \sim \mathcal{N}(\theta + B\mu, B\Sigma B^\top)$$

Proof.

$$\begin{aligned}
M_{\theta+BY}(u) &= \mathbb{E}\left[\exp\{u^\top(\theta+BY)\}\right] = \exp\left\{u^\top\theta\right\}\mathbb{E}\left[\exp\{(B^\top u)^\top Y\}\right] \\
&= \exp\left\{u^\top\theta\right\} M_Y(B^\top u) \\
&= \exp\left\{u^\top\theta\right\}\exp\left\{(B^\top u)^\top\mu + \frac{1}{2}u^\top B\Sigma B^\top u\right\} \\
&= \exp\left\{u^\top\theta + u^\top(B\mu) + \frac{1}{2}u^\top B\Sigma B^\top u\right\} \\
&= \exp\left\{u^\top(\theta+B\mu) + \frac{1}{2}u^\top B\Sigma B^\top u\right\}
\end{aligned}$$

And this last expression is the MGF of a $\mathcal{N}(\theta + B\mu, B\Sigma B^\top)$ distribution. □

Proposition (AY, BY indep $\iff$ $A\Sigma B^\top = 0$)

If $Y \sim \mathcal{N}(\mu_{p\times 1}, \Sigma_{p\times p})$, and $A_{m\times p}$, $B_{d\times p}$ be real matrices. Then,

$$AY \text{ independent of } BY \iff A\Sigma B^\top = 0.$$

Proof

It suffices to prove the result assuming $\mu = 0$ (and it simplifies the algebra).
First assume $A\Sigma B^\top = 0$. Let $W_{(m+d)\times 1} = \binom{AY}{BY}$ and $\theta_{(m+d)\times 1} = \binom{u_{m\times 1}}{v_{d\times 1}}$.

$$\begin{aligned}
M_W(\theta) &= \mathbb{E}[\exp\{W^\top\theta\}] = \mathbb{E}\left[\exp\left\{Y^\top A^\top u + Y^\top B^\top v\right\}\right] \\
&= \mathbb{E}\left[\exp\left\{Y^\top (A^\top u + B^\top v)\right\}\right] = M_Y(A^\top u + B^\top v) \\
&= \exp\left\{\frac{1}{2}(A^\top u + B^\top v)^\top \Sigma (A^\top u + B^\top v)\right\} \\
&= \exp\left\{\frac{1}{2}\left(u^\top A\Sigma A^\top u + v^\top B\Sigma B^\top v + u^\top \underbrace{A\Sigma B^\top}_{=0} v + v^\top \underbrace{B\Sigma A^\top}_{=0} u\right)\right\} \\
&= M_{AY}(u) M_{BY}(v),
\end{aligned}$$

i.e., the joint MGF is the product of the marginal MGFs, proving independence.

For the converse, assume that AY and BY are independent. Then, $\forall u, v$,

$$M_W(\theta) = M_{AY}(u) M_{BY}(v), \quad \forall u, v,$$

$$\implies \exp\left\{\frac{1}{2}\left(u^\top A\Sigma A^\top u + v^\top B\Sigma B^\top v + u^\top A\Sigma B^\top v + v^\top B\Sigma A^\top u\right)\right\}$$

$$= \exp\left\{\frac{1}{2} u^\top A\Sigma A^\top u\right\} \exp\left\{\frac{1}{2} v^\top B\Sigma B^\top v\right\}$$

$$\implies \exp\left\{\frac{1}{2} \times 2u^\top A\Sigma B^\top v\right\} = 1$$

$$\implies u^\top A\Sigma B^\top v = 0, \qquad \forall\ u \in \mathbb{R}^d, v \in \mathbb{R}^m,$$

$\implies$ the orthocomplement[a] of the column space of $A\Sigma B^\top$ is the whole of $\mathbb{R}^m$.

$$\implies A\Sigma B^\top = 0.$$

□

[a] recall that for $\mathrm{Q}_{m\times d}$ we have $\mathcal{M}^\perp(\mathrm{Q}) = \{y \in \mathbb{R}^m : y^\top \mathrm{Q} x = 0,\ \forall x \in \mathbb{R}^d\}$

Proposition (Density Function)

Let $\Sigma_{p\times p}$ be nonsingular. The density of $\mathcal{N}(\mu_{p\times 1}, \Sigma_{p\times p})$ is

$$f_Y(y) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left\{-\tfrac{1}{2}(y-\mu)^\top \Sigma^{-1}(y-\mu)\right\}$$

Proof.

Let $Z = (Z_1, \ldots, Z_p)^\top$ be a vector of iid $\mathcal{N}(0,1)$ random variables. Then, because of independence,

(a) the density of Z is

$$f_Z(z) = \prod_{i=1}^{p} f_{Z_i}(z_i) = \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z_i^2\right) = \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2}z^\top z\right).$$

(b) The MGF of Z is

$$M_Z(u) = \mathbb{E}\left\{\exp\left(\sum_{i=1}^{p} u_i Z_i\right)\right\} = \prod_{i=1}^{p} \mathbb{E}\{\exp(u_i Z_i)\} = \exp(u^\top u/2),$$

which is the MGF of a p-variate $\mathcal{N}(0, I)$ distribution.

proof continued

$\overset{(a)+(b)}{\Longrightarrow}$ the $\mathcal{N}(0, I)$ density is $f_Z(z) = \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2} z^\top z\right)$.

By the spectral theorem, Σ admits a square root, $\Sigma^{1/2}$. Furthermore, since Σ is non-singular, so is $\Sigma^{1/2}$.

Now observe that from our Property 2, we have $Y \overset{d}{=} \Sigma^{1/2} Z + \mu \sim \mathcal{N}(\mu, \Sigma)$.

By the change of variables formula,

$$\begin{aligned} f_Y(y) &= f_{\Sigma^{1/2} Z + \mu}(y) \\ &= |\Sigma^{-1/2}| f_Z\{\Sigma^{-1/2}(y - \mu)\} \\ &= \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(y - \mu)^\top \Sigma^{-1}(y - \mu)\right\}. \end{aligned}$$

[Recall that to obtain the density of $W = g(X)$ at w, we need to evaluate f_X at $g^{-1}(w)$ but also multiply by the Jacobian determinant of g^{-1} at w.]

□

Theorem (Karhunen-Loève)

Let $\mu \in \mathbb{R}^p$ and $\Sigma \in \mathbb{R}^{p \times p}$ be covariance with spectral decomposition

$$\Sigma = \mathsf{U}\Lambda\mathsf{U}^\top = \begin{pmatrix} u_1 & \dots & u_p \end{pmatrix} \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_p \end{pmatrix} \begin{pmatrix} u_1^\top \\ \vdots \\ u_p^\top \end{pmatrix} = \sum_{j=1}^{p} \lambda_j \, u_j \, u_j^\top,$$

where $\{u_j\}_{j=1}^p$ are the eigenvectors and $\{\lambda_j\}_{j=1}^p$ the eigenvalues. Then,

$$Y \sim N(\mu, \Sigma) \iff Y = \mu + \sum_{j=1}^{p} \lambda_j^{1/2} Z_j \, u_j = \sum_{j=1}^{p} (\lambda_j^{1/2} Z_j + u_j^\top \mu) u_j, \; Z_j \overset{iid}{\sim} N(0,1).$$

In words: if we do a change of basis and express $Y \sim N(\mu, \Sigma)$ in the basis of eigenvectors of Σ, the new coordinates become independent Gaussians with means $u_j^\top \mu$ (the coordinates of μ in the U basis) and variances λ_j.

Proof.

Recall that

$$Y \sim N(\mu, \Sigma) \iff Y = \Sigma^{1/2} W + \mu = \mathsf{U}\Lambda^{1/2}\mathsf{U}^\top W + \mu, \quad W \sim N(0, \mathsf{I}_{p\times p})$$

and defining $Z = \mathsf{U}^\top W$ we get $Z = \mathsf{U}^\top W \sim N(0, \mathsf{U}^\top\mathsf{U}) \equiv N(0, \mathsf{I}_{p\times p})$. So,

$$\Sigma^{1/2} W + \mu = \mathsf{U}\Lambda^{1/2} Z + \mu = \begin{pmatrix} u_1 & \dots & u_p \end{pmatrix} \begin{pmatrix} \lambda_1^{1/2} & & \\ & \ddots & \\ & & \lambda_p^{1/2} \end{pmatrix} \begin{pmatrix} Z_1 \\ \vdots \\ Z_p \end{pmatrix} + \mu$$

$$= \mu + \sum_{j=1}^{p} \lambda_j^{1/2} Z_j u_j, \quad Z_j \overset{iid}{\sim} N(0,1).$$

Finally, note that

$$\mu = \mathsf{U}^\top\mathsf{U}\mu = \begin{pmatrix} u_1 & \dots & u_p \end{pmatrix} \begin{pmatrix} u_1^\top \\ \vdots \\ u_p^\top \end{pmatrix} \mu = \sum_{j=1}^{n} (u_j^\top \mu) u_j .$$

□

Proposition

If $Y \sim \mathcal{N}(\mu, \Sigma)$, *then* $\text{supp}\{Y\} = \mathcal{R}(\Sigma) + \mu$

Proof.

Assume $\mu = 0$ wlog. When Σ is of full rank, $\mathcal{N}(\mu, \Sigma)$ admits a density, whose form shows that any open ball is assigned positive probability. For the possibly reduced rank case, let $r = \text{rank}(\Sigma)$, let $y \in \mathcal{R}(\Sigma)$ and $\epsilon > 0$ and observe that

$$\mathbb{P}\{\|Y - y\| < \epsilon\} \overset{(1)}{=} \mathbb{P}\{\|\mathsf{H}_Y Y - \mathsf{H}_Y y\| < \epsilon\} \overset{(2)}{=} \mathbb{P}\left\{\left\| \begin{pmatrix} \lambda_1^{1/2} Z_1 \\ \vdots \\ \lambda_r^{1/2} Z_r \end{pmatrix} - \begin{pmatrix} y^\top u_1 \\ \vdots \\ y^\top u_r \end{pmatrix} \right\| < \epsilon \right\}$$

$$\overset{(3)}{=} \mathbb{P}\{\|W - w\| < \epsilon\} > 0.$$

(1) is because $y \in \mathcal{R}(\Sigma)$ and we already know (support and covariance lemma) that $\text{supp}\{Y\} \subseteq \mathcal{R}(\Sigma)$, so $\mathsf{H}_Y Y = Y$ almost surely and $\mathsf{H}_Y y = y$.

(2) is by the Karhunen-Loève expansion.

(3) uses the fact that $W = (\lambda_1^{1/2} Z_1, \ldots, \lambda_r^{1/2} Z_r)^\top \sim N(0, \text{diag}\{\lambda_1, ..., \lambda_r\})$ is a Gaussian on $\mathbb{R}^r$ with non-singular covariance

So we have $y \in \mathcal{R}(\Sigma) \implies y \in \text{supp}(Y)$, so $\mathcal{R}(\Sigma) \subseteq \text{supp}(Y)$.

□

- When $\Sigma_{p\times p}$ is singular, the $N(\mu, \Sigma)$ distribution does not admit a density with respect to Lebesgue measure on $\mathbb{R}^p$.
- The support of $N(\mu, \Sigma)$ is $\mathcal{R}(\Sigma) + \mu$. This is an affine set of dimension $r = \text{rank}(\Sigma)$. So it admits an r-dimensional Lebesgue volume measure.
- Can we define a density w.r.t. this Lebesgue measure on the support?

Proposition

Let $X \sim N(\mu_{p\times 1}, \Sigma_{p\times p})$. Then, X admits a probability density with respect to Lebesgue measure on $\mu + \mathcal{R}(\Sigma^{1/2})$ given by

$$f_X(x) = \frac{1}{\prod_{j=1}^{r}(2\pi\lambda_j)^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^\top \Sigma^\dagger (x-\mu)\right\}, \quad x \in \mathcal{R}(\Sigma) + \mu,$$

where $r = \text{rank}(\Sigma) \leq p$ and $\{\lambda_1, \ldots, \lambda_r\}$ are the non-zero eigenvalues of Σ.

Proof.

The full rank case is already established, so we take $\mathrm{rank}(\Sigma) = r < p$. We start with the case $\mu = 0$ and

$$\Sigma = \mathrm{diag}\{\lambda_1, ..., \lambda_r, 0, ..., 0\}$$

for $\lambda_j > 0$, $j = 1, ..., r$. Given $y \in \mathcal{R}(\Sigma)$ let $N_\epsilon(y) = \prod_{i=1}^p (y_i - \epsilon, y_i + \epsilon)$ be the open rectangle of sidelength $2\epsilon > 0$ centred at y. Then $\mathbb{P}\{X \in N_\epsilon(y)\}$ equals

$$\mathbb{P}\left[\cap_{i=1}^p \{X_i \in (y_i - \epsilon, y_i + \epsilon)\}\right] = \underbrace{\prod_{i=1}^r \int_{y_i - \epsilon}^{y_i + \epsilon} \frac{e^{-x_i^2/(2\lambda_i)}}{\sqrt{2\pi\lambda_i}} dx_i}_{A(\epsilon)} \times \underbrace{\prod_{j=r+1}^p \mathbb{P}\{|\, Y_j - y_j\,| < \epsilon\}}_{B(\epsilon)}.$$

But for $j > r$ we have $y_j = 0$ and $Y_j = 0$ almost surely, so $B(\epsilon) = 1$. This establishes the form of the density in the mean zero and diagonal covariance case. For the general case $\mu \neq 0$ and $\Sigma = \mathsf{U}\Lambda\mathsf{U}^\top$, note that

$$\mathbb{P}\{\, Y \in A\} = \mathbb{P}\{\mathsf{U}^\top(\, Y - \mu) \in \mathsf{U}^\top(A - \mu)\} \text{ and}$$
$$\mathsf{U}^\top(\, Y - \mu) \sim N(0, \mathsf{U}^\top\Sigma\mathsf{U}) \equiv N(0, \mathrm{diag}\{\lambda_1, ..., \lambda_r, 0, ..., 0\}).$$

So the density is obtained by the change of variables $x \mapsto \mathsf{U}^\top(x - \mu)$ and observing that the term $\mathsf{U}\,\mathrm{diag}\{\lambda_1^{-1}, ..., \lambda_r^{-1}, 0, ..., 0\}\,\mathsf{U}^\top$ that will appear in the exponential's quadratic form equals $\Sigma^\dagger$ by definition of the pseudoinverse. □

Since linear dependence is the only kind of dependence between two jointly Gaussian random vectors X and Y, we may ask the question:

If we extract from X the part that is a perfect linear function of Y, is the "remainder" independent of Y?

Theorem (Regression Representation)

Let $\mu \in \mathbb{R}^{n+m}$ and Σ be a covariance on $\mathbb{R}^{n+m}$, expressed in block form as

$$\mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{XY}^\top & \Sigma_Y \end{pmatrix}.$$

If X and Y are random vectors in $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively, then the following two statements are equivalent:

- $\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}(\mu, \Sigma)$
- $X = \mu_X + \Sigma_{XY}\Sigma_Y^\dagger(Y - \mu_Y) + \varepsilon$, *with* $Y \sim \mathcal{N}(\mu_Y, \Sigma_Y)$, $\varepsilon \sim \mathcal{N}(0, \Sigma_X - \Sigma_{XY}\Sigma_Y^\dagger\Sigma_{XY}^\top)$, *and* $\varepsilon \perp\!\!\!\perp Y$

The theorem provides intuition as to how the pair $(X, Y)^\top$ arises:

- First, we affinely transform a realisation of $Y \sim \mathcal{N}(\mu_Y, \Sigma_Y)$ by a deterministic affine transformation.
- Then, we add an independent (of Y) zero mean Gaussian random variable ε.

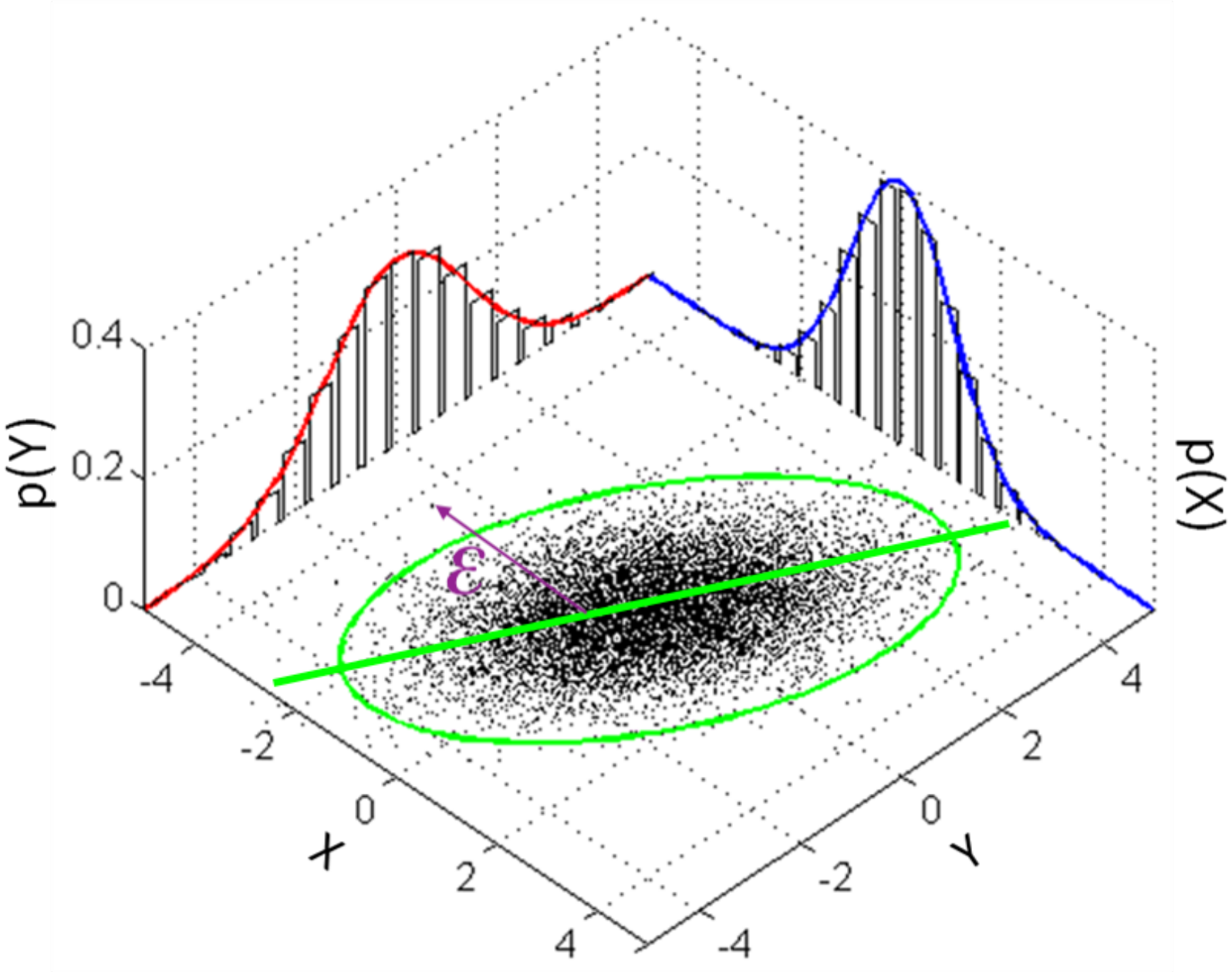
Consider the simplest case where X and Y are scalar, and $\mathrm{var}(Y) > 0$. Then the representation in the theorem reduces to the familiar expression

$$X = \beta_0 + \beta Y + \varepsilon, \qquad Y \sim \mathcal{N}(\mu_Y, \mathrm{var}(Y)) \text{ independent of } \varepsilon \sim \mathcal{N}(0, \mathrm{var}(\varepsilon))$$

where:

- $\beta_0 = \mu_X - \frac{\mathrm{cov}\{X,Y\}}{\mathrm{var}(Y)}\mu_Y$ is called the *intercept*
- $\beta = \mathrm{cov}\{X, Y\}/\mathrm{var}(Y)$ is called the *regression coefficient*
- ε is called the *error* or *innovation* that is *homoskedastic* in that $\mathrm{var}(\varepsilon) = \mathrm{var}(X) - \mathrm{cov}^2\{X, Y\}/\mathrm{var}(Y)$ does not vary with Y.

This explains why we call it the "regression representation".


0.4
0.2
0
p(Y)
p(X)
ε
-4
-2
0
2
4
X
Y

Proof when $\mu = 0$ and $\Sigma \succ 0$.

Since $\Sigma \succ 0$, we have $\Sigma_Y^\dagger = \Sigma_Y^{-1}$. We begin by proving the "$\Leftarrow$" direction. Note that if X is defined as stipulated by the representation, then,

$$\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \Sigma_{XY}\Sigma_Y^{-1} Y + \varepsilon \\ Y \end{pmatrix} = \begin{pmatrix} \mathsf{I}_{n\times n} & \Sigma_{XY}\Sigma_Y^{-1} \\ 0 & \mathsf{I}_{m\times m} \end{pmatrix} \begin{pmatrix} \varepsilon \\ Y \end{pmatrix},$$

where the conditions on ε and Y imply that

$$\begin{pmatrix} \varepsilon \\ Y \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} \Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^\top & 0 \\ 0 & \Sigma_Y \end{pmatrix}\right).$$

This implies that $(X^\top, Y^\top)^\top$ is jointly normally distributed with mean zero and covariance

$$\begin{pmatrix} \mathsf{I}_{n\times n} & \Sigma_{XY}\Sigma_Y^{-1} \\ 0 & \mathsf{I}_{m\times m} \end{pmatrix} \begin{pmatrix} \Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^\top & 0 \\ 0 & \Sigma_Y \end{pmatrix} \begin{pmatrix} \mathsf{I}_{n\times n} & 0 \\ \Sigma_Y^{-1}\Sigma_{XY}^\top & \mathsf{I}_{m\times m} \end{pmatrix}$$

$$= \begin{pmatrix} \mathsf{I}_{n\times n} & \Sigma_{XY}\Sigma_Y^{-1} \\ 0 & \mathsf{I}_{m\times m} \end{pmatrix} \begin{pmatrix} \Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^\top & 0 \\ \Sigma_{XY}^\top & \Sigma_Y \end{pmatrix} = \Sigma.$$

To prove the "$\Rightarrow$" direction we show that if $W = (X^\top, Y^\top)^\top \sim N(0, \Sigma)$, then:

1. $X - \Sigma_{XY}\Sigma_Y^{-1} Y \perp\!\!\!\perp Y$.
2. $X - \Sigma_{XY}\Sigma_Y^{-1} Y \sim \mathcal{N}(0, \Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^\top)$

To this aim we note that

$$X - \Sigma_{XY}\Sigma_Y^{-1} Y = \begin{pmatrix} \mathsf{I}_{n\times n} & -\Sigma_{XY}\Sigma_Y^{-1} \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} = \mathsf{Q}\,W \qquad \& \qquad Y = \begin{pmatrix} 0 & \mathsf{I}_{m\times m} \end{pmatrix} W = \mathsf{P}\,W.$$

Therefore:

- $X - \Sigma_{XY}\Sigma_Y^{-1} Y \perp\!\!\!\perp Y$ iff $\mathsf{P}\Sigma\mathsf{Q}^\top = 0$, which is verified because $\mathsf{P}\Sigma\mathsf{Q}^\top$ equals

$$\underbrace{\begin{pmatrix} 0 & \mathsf{I}_{m\times m} \end{pmatrix}}_{\mathsf{P}} \underbrace{\begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{XY}^\top & \Sigma_Y \end{pmatrix}}_{\Sigma} \underbrace{\begin{pmatrix} \mathsf{I}_{n\times n} \\ -\Sigma_Y^{-1}\Sigma_{XY}^\top \end{pmatrix}}_{\mathsf{Q}^\top} = \underbrace{\begin{pmatrix} 0 & \mathsf{I}_{m\times m} \end{pmatrix} \begin{pmatrix} \Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^\top \\ 0 \end{pmatrix}}_{=0}$$

- $\mathsf{Q}\,W \sim N(0, \mathsf{Q}\Sigma\mathsf{Q}^\top)$ and importing our previous calculation of $\Sigma\mathsf{Q}^\top$

$$\mathsf{Q}\Sigma\mathsf{Q}^\top = \begin{pmatrix} \mathsf{I}_{n\times n} & -\Sigma_{XY}\Sigma_Y^{-1} \end{pmatrix} \begin{pmatrix} \Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^\top \\ 0 \end{pmatrix} = \Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^\top.$$

□

Exercise: Use the "support and covariance" lemma to establish the general case.

Corollary (Gaussian Conditional Distributions)

Let $(X^\top \; Y^\top)^\top$ be a jointly Gaussian, comprised of concatenated random vectors in $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively, with mean and covariance

$$\mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{XY}^\top & \Sigma_Y \end{pmatrix}.$$

Then,

$$X|\{Y = y\} \sim \mathcal{N}\left(\mu_X + \Sigma_{XY}\Sigma_Y^\dagger(y - \mu_Y), \Sigma_X - \Sigma_{XY}\Sigma_Y^\dagger\Sigma_{XY}^\top\right).$$

Consequently, when Σ_Y is non-singular,

$$X|\{Y = y\} \sim \mathcal{N}\left(\mu_X + \Sigma_{XY}\Sigma_Y^{-1}(y - \mu_Y) \;,\; \Theta_X^{-1}\right). \qquad (*)$$

We highlight that when Σ_Y is non-singular, the conditional covariance of $X|Y$ is $\Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{XY}^\top = \Theta_X^{-1}$ where Θ_X is the top-left $n \times n$ block of the precision matrix $\Theta = \Sigma^{-1}$. This follows from our block inverse covariance Lemma. This observation will come in very handy in our next two Theorems. Call it $(*)$. In other words,

the covariance sub-block Σ_X of Σ is the covariance matrix of the marginal law of X.
the precision sub-block Θ_X of Θ is the precision matrix of the conditional law of $X|Y$

Theorem (Gaussian Conditional Independence)

Let $(X^\top, Y^\top, Z^\top)^\top$ *be an* $(n+m+p)$*-dimensional Gaussian vector with non-singular covariance matrix* $\Sigma \succ 0$ *and precision matrix* $\Theta = \Sigma^{-1}$*, expressed in block format as*

$$\Sigma = \begin{pmatrix} \Sigma_X & \Sigma_{XY} & \Sigma_{XZ} \\ \Sigma_{XY}^\top & \Sigma_Y & \Sigma_{YZ} \\ \Sigma_{XZ}^\top & \Sigma_{YZ}^\top & \Sigma_Z \end{pmatrix} \quad \& \quad \Sigma^{-1} = \Theta = \begin{pmatrix} \Theta_X & \Theta_{XY} & \Theta_{XZ} \\ \Theta_{XY}^\top & \Theta_Y & \Theta_{YZ} \\ \Theta_{XZ}^\top & \Theta_{YZ}^\top & \Theta_Z \end{pmatrix}$$

Then,

$$\{X \perp\!\!\!\perp Y\} \,\big|\, Z \iff \Theta_{XY} = 0.$$

Proof.

Set $W = (X^\top, Y^\top)^\top$. Since $(X^\top, Y^\top, Z^\top)^\top$ is Gaussian with non-singular covariance, we conclude that $F_{X,Y|Z} \equiv F_{W|Z}$ is Gaussian, with non-singular covariance $\Gamma = \Theta_W^{-1}$ (by $(*)$). In turn, this equals

$$\Gamma = \Theta_W^{-1} = \begin{pmatrix} \Theta_X & \Theta_{XY} \\ \Theta_{XY}^\top & \Theta_Y \end{pmatrix}^{-1}$$

i.e. the inverse of the top left $(n+m) \times (n+m)$ submatrix of Θ. Observe that:

- If $\Theta_{XY} = 0$, then

$$\Gamma = \begin{pmatrix} \Theta_X & 0 \\ 0 & \Theta_Y \end{pmatrix}^{-1} = \begin{pmatrix} \Theta_X^{-1} & 0 \\ 0 & \Theta_Y^{-1} \end{pmatrix}$$

  which implies that $F_{X,Y|Z}$ factorizes as $F_{X|Z} F_{Y|Z}$ by Property (5) of Gaussians, and the form of Gaussian conditionals.
- If $F_{X,Y|Z} = F_{X|Z} F_{Y|Z}$, then Property (5) of Gaussians implies that Γ is block-diagonal, and so its inverse is also block diagonal, implying that $\Theta_{XY} = 0$.

In the setting of our last theorem, using the Gaussian regression representation twice, we have:

- $X = \mu_X + \Sigma_{XZ}\Sigma_Z^\dagger(Z - \mu_Z) + \varepsilon_X$,
- $Y = \mu_Y + \Sigma_{YZ}\Sigma_Z^\dagger(Z - \mu_Z) + \varepsilon_Y$,

where

- $Z \sim \mathcal{N}(\mu_Z, \Sigma_Z)$
- $\varepsilon_X \sim \mathcal{N}(0, \Sigma_X - \Sigma_{XZ}\Sigma_Z^{-1}\Sigma_{XZ}^\top)$ with $\varepsilon_X \perp\!\!\!\perp Z$
- $\varepsilon_Y \sim \mathcal{N}(0, \Sigma_Y - \Sigma_{YZ}\Sigma_Z^{-1}\Sigma_{YZ}^\top)$ with $\varepsilon_Y \perp\!\!\!\perp Z$

Notice that $\varepsilon_X \not\perp\!\!\!\perp \varepsilon_Y$ in general. When are they independent?

Proposition (Regression and Gaussian Conditional Independence)

In the same context as above, we have

$$\begin{pmatrix} \varepsilon_X \\ \varepsilon_Y \end{pmatrix} \sim N\left(0, \begin{pmatrix} \Theta_X & \Theta_{XY} \\ \Theta_{XY}^\top & \Theta_Y \end{pmatrix}^{-1}\right)$$

Therefore,

$$\varepsilon_X \perp\!\!\!\perp \varepsilon_Y \iff \Theta_{XY} = 0 \iff \{X \perp\!\!\!\perp Y\} \,\big|\, Z$$

Proof.

Assume wlog that the mean is zero. By the regression representation,

$$\begin{pmatrix} \varepsilon_X \\ \varepsilon_Y \end{pmatrix} = \begin{pmatrix} X - \Sigma_{XZ}\Sigma_Z^{-1} Z \\ Y - \Sigma_{YZ}\Sigma_Z^{-1} Z \end{pmatrix} = \begin{pmatrix} \mathsf{I}_{n\times n} & 0 & -\Sigma_{XZ}\Sigma_Z^{-1} \\ 0 & \mathsf{I}_{m\times m} & -\Sigma_{YZ}\Sigma_Z^{-1} \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}.$$

Thus $(\varepsilon_X^\top, \varepsilon_Y^\top)^\top$ is jointly Gaussian. The form of $\mathrm{cov}\{\varepsilon_X\}$, $\mathrm{cov}\{\varepsilon_Y\}$, has already been established. The cross-covariance, is $\mathrm{cov}\{\varepsilon_X, \varepsilon_Y\} = \mathbb{E}[\varepsilon_X \varepsilon_Y^\top]$ which equals

$$\begin{aligned}
& \mathbb{E}[(X - \Sigma_{XZ}\Sigma_Z^{-1} Z)(Y - \Sigma_{YZ}\Sigma_Z^{-1} Z_Z)^\top] \\
= \; & \mathbb{E}[XY^\top] - \mathbb{E}[X(\Sigma_{YZ}\Sigma_Z^{-1} Z)^\top] - \mathbb{E}[\Sigma_{XZ}\Sigma_Z^{-1} ZY^\top] + \mathbb{E}[\Sigma_{XZ}\Sigma_Z^{-1} Z(\Sigma_{YZ}\Sigma_Z^{-1} Z)^\top] \\
= \; & \mathbb{E}[XY^\top] - \mathbb{E}[XZ^\top]\Sigma_Z^{-1}\Sigma_{YZ}^\top - \Sigma_{XZ}\Sigma_Z^{-1}\mathbb{E}[ZY^\top] + \Sigma_{XZ}\Sigma_Z^{-1}\mathbb{E}[ZZ^\top]\Sigma_Z^{-1}\Sigma_{YZ}^\top \\
= \; & \Sigma_{XY} - \Sigma_{XZ}\Sigma_Z^{-1}\Sigma_{YZ}^\top - \Sigma_{XZ}\Sigma_Z^{-1}\Sigma_{YZ}^\top + \Sigma_{XZ}\Sigma_Z^{-1}\Sigma_Z\Sigma_Z^{-1}\Sigma_{YZ}^\top \\
= \; & \Sigma_{XY} - \Sigma_{XZ}\Sigma_Z^{-1}\Sigma_{YZ}^\top
\end{aligned}$$

In summary,

$$\mathrm{cov}\left\{\begin{pmatrix}\varepsilon_X \\ \varepsilon_Y\end{pmatrix}\right\} = \begin{pmatrix} \Sigma_X - \Sigma_{XZ}\Sigma_Z^{-1}\Sigma_{XZ}^\top & \Sigma_{XY} - \Sigma_{XZ}\Sigma_Z^{-1}\Sigma_{YZ}^\top \\ \Sigma_{XY}^\top - \Sigma_{YZ}\Sigma_Z^{-1}\Sigma_{XZ}^\top\Sigma_{XZ}^\top & \Sigma_Y - \Sigma_{YZ}\Sigma_Z^{-1}\Sigma_{YZ}^\top \end{pmatrix} \overset{(?)}{=} \begin{pmatrix} \Theta_X & \Theta_{XY} \\ \Theta_{XY}^\top & \Theta_Y \end{pmatrix}^{-1}$$

It remains to ascertain whether "$\overset{(?)}{=}$" is true.

Fortunately, we may recall $(*)$ to notice that the RHS equals Θ_W^{-1}, i.e. the covariance of $W = (X^\top, Y^\top)^\top$ given Z. And can use $(*)$ to calculate it explicitly. Namely, we can write

$$W \,|\{Z = z\} \sim N(\Sigma_{WZ}\Sigma_Z z, \Theta_W^{-1}) \equiv N(\Sigma_{WZ}\Sigma_Z z, \Sigma_W - \Sigma_{WZ}\Sigma_Z^{-1}\Sigma_{WZ}^\top)$$

We can now write out

$$\begin{aligned}
\Sigma_W - \Sigma_{WZ}\Sigma_Z^{-1}\Sigma_{WZ}^\top &= \begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{XY}^\top & \Sigma_Y \end{pmatrix} - \mathbb{E}\left[\begin{pmatrix} XZ^\top \\ YZ^\top \end{pmatrix}\right] \Sigma_Z^{-1} \mathbb{E}\left[(ZX^\top \; ZY^\top)\right] \\
&= \begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{XY}^\top & \Sigma_Y \end{pmatrix} - \begin{pmatrix} \Sigma_{XZ} \\ \Sigma_{YZ} \end{pmatrix} \Sigma_Z^{-1} (\Sigma_{XZ}^\top \; \Sigma_{YZ}^\top) \\
&= \begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{XY}^\top & \Sigma_Y \end{pmatrix} - \begin{pmatrix} \Sigma_{XZ}\Sigma_Z^{-1}\Sigma_{XZ}^\top & \Sigma_{XZ}\Sigma_Z^{-1}\Sigma_{YZ}^\top \\ \Sigma_{YZ}\Sigma_Z^{-1}\Sigma_{XZ}^\top & \Sigma_{YZ}\Sigma_Z^{-1}\Sigma_{YZ}^\top \end{pmatrix} \\
&= \begin{pmatrix} \Sigma_X - \Sigma_{XZ}\Sigma_Z^{-1}\Sigma_{XZ}^\top & \Sigma_{XY} - \Sigma_{XZ}\Sigma_Z^{-1}\Sigma_{YZ}^\top \\ \Sigma_{XY}^\top - \Sigma_{YZ}\Sigma_Z^{-1}\Sigma_{XZ}^\top & \Sigma_Y - \Sigma_{YZ}\Sigma_Z^{-1}\Sigma_{YZ}^\top \end{pmatrix}
\end{aligned}$$

which establishes that "$\overset{(?)}{=}$" is true indeed.
The second part of the Proposition follows from the fact that $\varepsilon_X \perp\!\!\!\perp \varepsilon_Y$ if and only if the covariance of $(\varepsilon_X^\top, \varepsilon_Y^\top)^\top$ is block-diagonal (Property 5 of Gaussians). $\square$

Let X and Y be random vectors in $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively and write

$$\mathrm{cov}\{X, Y\} = \begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{XY}^\top & \Sigma_Y \end{pmatrix}.$$

How can we best predict Y using an affine function of X?

Formally, we seek an affine map $\mathbb{R}^n \ni x \mapsto \mathsf{B}_* x + \beta_* \in \mathbb{R}^m$, such that

$$\mathbb{E}\| Y - \beta_* - \mathsf{B}_* X \|^2 \leq \mathbb{E}\| Y - \beta - \mathsf{B} X \|^2$$

is minimal over all choices of $\beta \in \mathbb{R}^m$ and $\mathsf{B} \in \mathbb{R}^{m \times n}$. Such a map is called a best linear predictor.

- When X and Y are jointly Gaussian, we saw that the conditional expectation of $Y|X$ is an affine transformation of X – hence it is the *de facto* best linear predictor, seeing as it is the best predictor (linear or otherwise):

$$\beta_* = \mu_Y - \Sigma_{YX}\Sigma_X^\dagger \mu_X \quad \& \quad \mathsf{B}_* = \Sigma_{YX}\Sigma_X^\dagger$$

- What can we say beyond jointly Gaussian vectors?

Theorem (Best Linear Prediction)

Regardless of the joint law of X and Y, the best linear predictor of $Y|X$ is the same as in the jointly Gaussian case.

Proof.

Writing $\mathbb{E}\|Y - \beta - \mathrm{B}X\|^2 = \sum_{i=1}^m \mathbb{E}\big[(Y_i - \beta_i - \mathrm{B}_i X)^2\big]$ for $\mathrm{B}_i \in \mathbb{R}^{1\times n}$ the ith row of B, we see that it suffices to consider a scalar Y and $\mathrm{B} \in \mathbb{R}^{1\times n}$.

The key step is to show that X and $\varepsilon := Y - \beta_* - \mathrm{B}_* X$ are uncorrelated. But we have done this already in the proof of Gaussian regression representation[a]!

Now $\mathbb{E}(Y - \beta - \mathrm{B}X)^2$ can be written as

$$\mathrm{var}\{Y - \beta - \mathrm{B}X\} + [\mathbb{E}(Y - \beta - \mathrm{B}X)]^2 = \mathrm{var}\{Y - \mathrm{B}X\} + [\mu_Y - \beta - \mathrm{B}\mu_X]^2$$

We can immediately check that $(\beta, \mathrm{B}) = (\beta_*, \mathrm{B}_*)$ minimises the second term (yielding zero). Let's also check that this choice also minimizes the first term.

$$\begin{aligned}
\mathrm{var}\{Y - \mathrm{B}X\} &= \mathrm{var}\{Y - \beta_* - \mathrm{B}_* X + \beta_* + \mathrm{B}_* X - \mathrm{B}X\} = \mathrm{var}\{\varepsilon + \beta_* + (\mathrm{B}_* - \mathrm{B})X\} \\
&= \mathrm{var}\{\varepsilon + (\mathrm{B}_* - B)X\} = \mathrm{var}\{\varepsilon\} + \mathrm{var}\{(\mathrm{B}_* - \mathrm{B})X\} \\
&= \mathrm{var}\{\varepsilon\} + (\mathrm{B}_* - \mathrm{B})\Sigma_X(\mathrm{B}_* - \mathrm{B})^\top
\end{aligned}$$

The first term does not depend on (β, B) and the second is clearly minimised at $\mathrm{B} = \mathrm{B}_*$ since $(\mathrm{B}_* - \mathrm{B})\Sigma_X(\mathrm{B}_* - \mathrm{B})^\top \succeq 0$ for all B. □

[a] because to establish independence in the Gaussian context, we established uncorrelatedness

Let $X = (X_1, \ldots, X_p)^\top$ be random p-vector, with covariance $\Sigma \succ 0$ and precision matrix $\Theta = \{\theta_{ij}\} = \Sigma^{-1}$. The partial correlation of X_i and X_j given $\{X_k\}_{k \neq i,j}$ is defined as

$$\rho_{ij|k \neq i,j} := -\frac{\theta_{ij}}{\sqrt{\theta_{ii}\theta_{jj}}}$$

It expresses the correlation between X_i and X_j when controlling for the linear effects of the remaining variables $\{X_k\}_{k \neq i,j}$ on X_i and X_j.

When the partial correlation vanishes, the corresponding variables are called partially uncorrelated[2].

- when $X = (X_1^\top, X_2^\top, X_3^\top)^\top$ is a partitioned Gaussian, X_1 and X_2 are partially uncorrelated given X_3 if and only if $X_1 \perp\!\!\!\perp X_2 | X_3$
- when $X = (X_1^\top, X_2^\top, X_3^\top)^\top$ is possibly non-Gaussian, X_1 and X_2 are partially uncorrelated given X_3 if and only if $X_1 - X_{1|3}^*$ and $X_2 - X_{2|3}^*$ are uncorrelated, where $X_{k|3}^*$ is the best linear predictor of X_k given X_3, $k = 1, 2$.
- The last sentence could be given the heading "regression and conditional uncorrelatedness" (compare to the "regression and Gaussian conditional independence" theorem)

[2] Notice that this is always in reference to another set of variables!

Gaussian Quadratic Forms and Concentration

If Z is standard Gaussian, what is the law of $Z^\top Z$?

Starting with $Z \sim \mathcal{N}(0,1)$. Note that $F_{Z^2}(y) = 0$ if $y < 0$, and for $y \geq 0$ the CDF is

$$\begin{aligned} F_{Z^2}(y) &= \mathbb{P}[Z^2 \leq y] = \mathbb{P}[|Z| \leq \sqrt{y}] = \mathbb{P}[-\sqrt{y} \leq Z \leq \sqrt{y}] \\ &= \Phi(\sqrt{y}) - \Phi(-\sqrt{y}) = \Phi(\sqrt{y}) - (1 - \Phi(\sqrt{y})) = 2\Phi(\sqrt{y}) - 1. \end{aligned}$$

Differentiating yields the PDF

$$f_{Z^2}(y) = 2\frac{d}{dy}\Phi(\sqrt{y}) = 2\frac{d}{d\sqrt{y}}\Phi(\sqrt{y})\frac{d}{dy}\sqrt{y} = [\ldots] = \frac{1}{\sqrt{2}\sqrt{\pi}} e^{-y/2} y^{-1/2}.$$

The MGF is

$$M_{Z^2}(t) = \frac{1}{\sqrt{2\pi}} \int_0^\infty e^{ty} y^{-1/2} e^{-y/2}\, dy = \frac{1}{\sqrt{2\pi}} \int_0^\infty y^{-1/2} e^{-(1-2t)y/2}\, dy,$$

and, provided that $1 - 2t > 0$, the integral is finite and we can substitute $u = (1-2t)y$ to get

$$M_{Z^2}(t) = (1-2t)^{-1/2} \int_0^\infty \frac{1}{\sqrt{2\pi}} u^{-1/2} e^{-u/2}\, du = (1-2t)^{-1/2}, \quad t < \frac{1}{2}.$$

The corresponding distribution is called the χ^2_1 distribution.

Proposition (Gaussian Sum of Squares and Chi-Square Distribution)

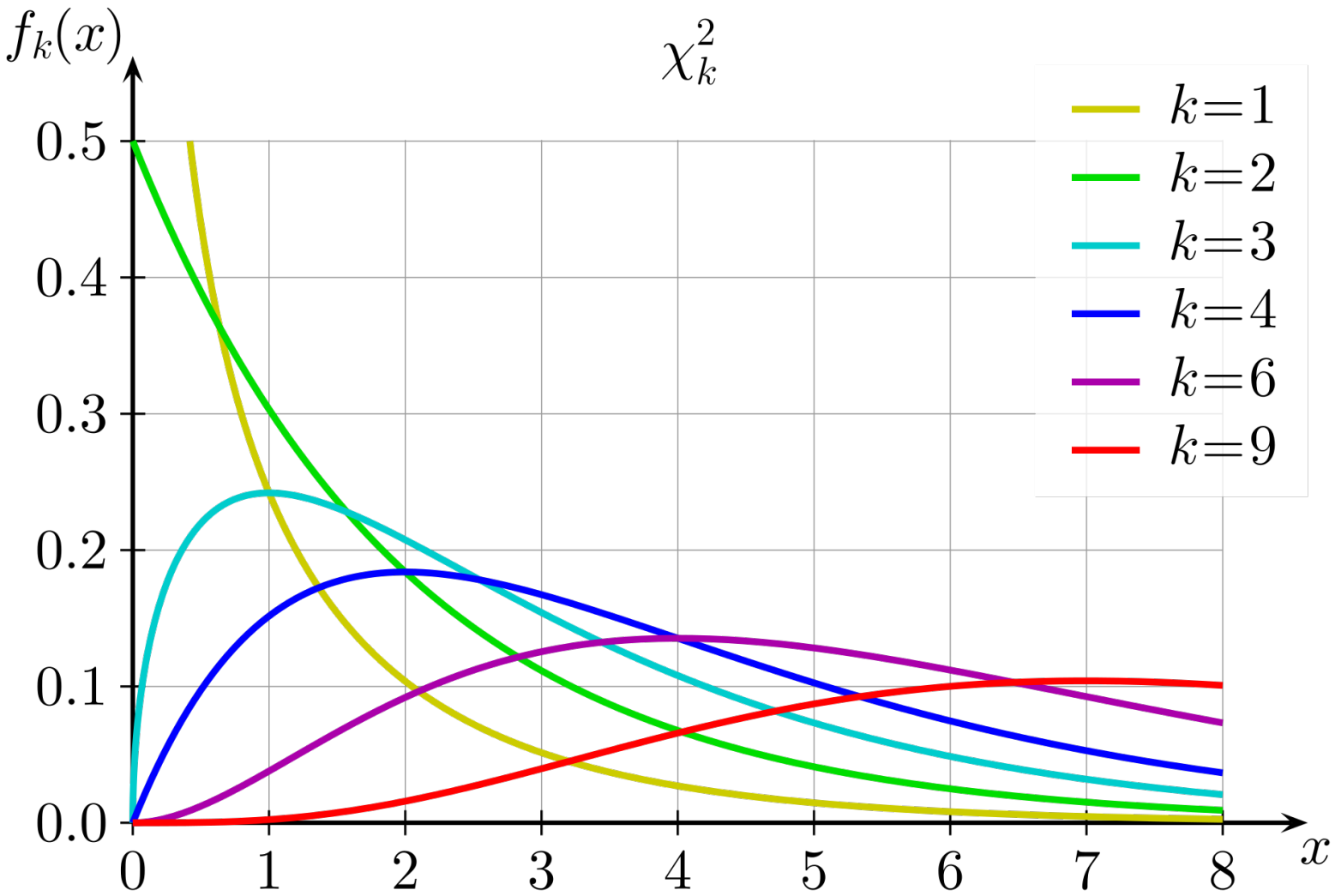
If $Z \sim N(0, \mathsf{I}_{k\times k})$, *then the moment generating function of the random variable* $Z^\top Z$ *is given by*

$$M_{Z^\top Z}(t) = (1-2t)^{-k/2}, \qquad t < \frac{1}{2}.$$

The proof is an easy exercise. The corresponding distribution is called the χ^2_k distribution. This law is completely specified by the parameter k, called the degrees of freedom.

Can easily establish establish the PDF, but it's not of much use to us.

Can easily verify $Z \sim N(0, \mathsf{I}_{k\times k})$ satisfies $\mathbb{E}[Z^\top Z] = k$ and $\text{var}\{Z^\top Z\} = 2k$.


$f_k(x)$
χ^2_k
$k=1$
$k=2$
$k=3$
$k=4$
$k=6$
$k=9$
0.5
0.4
0.3
0.2
0.1
0.0
0
1
2
3
4
5
6
7
8
x

Exercise:

- $X \sim N(0, \Sigma_{p\times p})$ and Σ invertible, then $X^\top \Sigma^{-1} X \sim \chi^2_p$.
- $X \sim N(0, \mathsf{I}_p)$ and H a projection, then $X^\top \mathsf{H} X \sim \chi^2_{\mathtt{rank}(\mathsf{H})}$.
- $X \sim N(0, \Sigma_{p\times p})$, then $X^\top \Sigma^\dagger X \sim \chi^2_{\mathtt{rank}(\Sigma)}$.

Lemma (Chernoff Bound for χ^2_k)

Given $Q \sim \chi^2_p$, one has

$$\mathbb{P}\left\{\left|\frac{Q-p}{p}\right| > u\right\} \leq 2e^{-pu^2/8}, \qquad \forall\, u \in (0,1).$$

The proof is left as an exercise.

The *plat principal* is the following:

Theorem (Gaussian Concentration of Measure)

Let $Z \sim \mathrm{N}(0, \mathrm{I}_{p\times p})$ and $X = \Sigma^{1/2} Z + \mu$ with Σ non-singular. Then, $\forall \epsilon \in (0,1)$,

$$\begin{aligned}\mathbb{P}\left\{||\Sigma^{-1/2}(X-\mu)||^2 \notin [(1-\epsilon)p, (1+\epsilon)\,p]\right\} &= \mathbb{P}\left\{||Z||^2 \notin [(1-\epsilon)p, (1+\epsilon)\,p]\right\} \\ &\leq 2e^{-\epsilon^2 p/8}\end{aligned}$$

Proof of the lemma.

A Chernoff bound combinines the MGF with Markov's inequality:

$$\mathbb{P}[Q - p > pu] = \mathbb{P}[e^{\lambda(Q-p)} > e^{\lambda pu}] \quad \leq \quad e^{-\lambda pu}\mathbb{E}[e^{\lambda(Q-p)}] = e^{-\lambda pu}e^{-\lambda p}M_Q(\lambda)$$

$$= \quad e^{-\lambda pu}\left(\frac{e^{-\lambda}}{\sqrt{1-2\lambda}}\right)^p \leq \exp\Big\{-\lambda pu + 2\lambda^2 p\Big\}$$

for all $|\lambda| < 1/4$ (in the last step we have used the inequality $\frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \leq e^{2\lambda^2}$ which is valid for $|\lambda| < 1/4$). Now we optimise the upper bound with respect to λ. It can be checked that $\exp\Big\{-\lambda pu + 2\lambda^2 p\Big\}$ has a global minimum at the value $\lambda^* = u/4$. Since $u \in (0,1)$, we have $|\lambda^*| < 1/4$, and we can plug λ^* into our inequality to obtain $\mathbb{P}[Q - p > pu] \leq e^{-pu^2/8}$.
Now we go in the other direction, and consider

$$\mathbb{P}[Q - p < -pu] = \mathbb{P}[-\theta(Q-p) > \theta pu] = \mathbb{P}[e^{-\theta(Q-p)} > e^{\theta pu}]$$

$$\leq e^{-\theta pu}\mathbb{E}[e^{-\theta(Q-p)}] = e^{-\theta pu}e^{-\theta p}M_Q(-\theta)$$

Provided $|\theta| < 1/4$ we may define $\lambda = -\theta$ and repeat the exact same steps as before to obtain $\mathbb{P}[Q - p < -pu] \leq e^{-pu^2/8}$. To complete the proof, we put the two inequalities together to observe that

$$\mathbb{P}\left\{\left|\frac{Q-p}{p}\right| > u\right\} = \mathbb{P}\left\{\frac{Q-p}{p} > u\right\} + \mathbb{P}\left\{\frac{Q-p}{p} < -u\right\} \leq 2e^{-pu^2/8}.$$

□

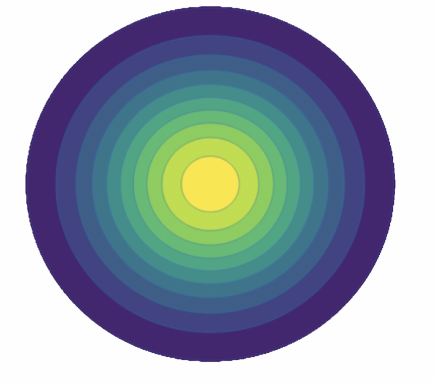
Upon closer inspection, the result is striking:

- given $\epsilon \in (0,1)$, the bound $2e^{-\epsilon^2 p/8}$ rapidly approaches 0 as p grows
- morally this says that:

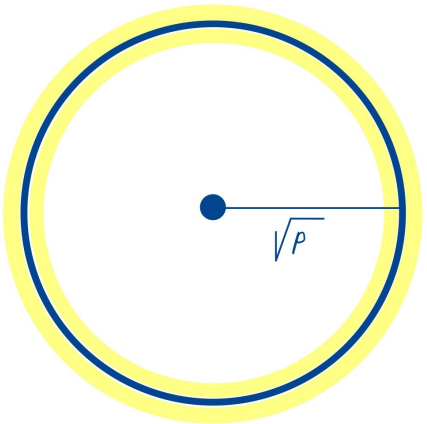
  in "high dimensions" (p large), the realisations of $Z \sim N(0, I_{p\times p})$ highly concentrate near the surface of the sphere of radius $\sqrt{p}$.

- In other words, the standard normal distribution in high dimension p is close to the uniform distribution on the sphere of radius $\sqrt{p}$:

$$N(0, I_p) \overset{p \text{ large}}{\approx} \text{Uniform}(\sqrt{p}\mathbb{S}^{p-1})$$



$$p \text{ large} \approx$$



- This may seem surprising given that the mode of the pdf is always at zero.

In the case $X \sim N(\mu, \Sigma_{p\times p})$ with $\Sigma = \mathsf{U}\,\mathrm{diag}\{\lambda, \ldots, \lambda_p\}\,\mathsf{U}^\top$ invertible, we have a similar concentration, but this time around an ellipsoid:

- centred at μ
- with the eigenvectors u_i of Σ as principle axes
- with principle axis lengths $2\lambda_i^{-1/2}\sqrt{p}$.

Proof of the theorem.

Using the Chernoff bound for the χ^2_k distribution, we get

$$
\begin{aligned}
2e^{-pu^2/8} \geq \mathbb{P}\left\{\left|\frac{||Z||^2 - p}{p}\right| > u\right\} &= \mathbb{P}\Big\{\{||Z||^2 - p > pu\} \cup \{||Z||^2 - p < -pu\}\Big\} \\
&= 1 - \mathbb{P}\Big\{\{||Z||^2 - p \leq pu\} \cap \{||Z||^2 - p \geq -pu\}\Big\} \\
&= 1 - \mathbb{P}\Big\{\{||Z||^2 \leq (1+u)p\} \cap \{||Z||^2 \geq (1-u)p\}\Big\} \\
&= 1 - \mathbb{P}\Big\{(1-u)p \leq ||Z||^2 \leq (1+u)p\Big\}
\end{aligned}
$$

and the result follows taking $\epsilon = u \in (0,1)$. □

Spherical and Elliptical Distributions

A key feature of a Gaussian vector $X \sim N(\mu, \Sigma)$ is that the representation

$$X \stackrel{d}{=} \mu + \Sigma^{1/2} Z, \qquad Z \sim N(0, \mathrm{I}).$$

- What if we replace $Z \sim N(0, \mathrm{I})$ by some other "spherical" random vector W
- Spherical means that

$$\mathrm{U}\, W \stackrel{d}{=} W \text{ for all orthogonal U.}$$

  - Equivalently (exercise) that $W = \xi U$ where $U \sim \mathrm{Unif}\{x : ||x|| = 1\}$, $\xi > 0$ is a random scalar (called the radial part), and $U \perp\!\!\!\perp \xi$
  - Equivalently (exercise) $v^\top W \stackrel{d}{=} ||v||\, W_1$, for all v (where $W = (W_1, \ldots, W_p)^\top$)

- If a spherical law has density f, then necessarily $f(x) = f(\mathrm{U}x)$ for all orthogonal U. Hence $f(x) = g(||x||^2)$ for some $g : [0, \infty) \to [0, \infty)$.

Definition (Elliptical Distributions)

An random vector $X = (X_1, \ldots, X_p)^\top$ is called *elliptical* with *location* μ and *dispersion* $\mathsf{A}\mathsf{A}^\top = \Sigma$ if and only if

$$X \overset{d}{=} \mu + \mathsf{A}\, W,$$

for W a spherical random vector, and $\mathsf{A}_{p \times d}$ with $p \leq d$.

- Elliptical distributions are affine transformations of spherical distributions, just like Gaussians are affine transformations of standard Gaussians.

Since any spherical random vector is represented as $\xi\, U$ where $\xi \perp\!\!\!\perp U$,

$$X \text{ is elliptical} \iff X \overset{d}{=} \mu + \xi \mathsf{A}\, U, \qquad U \sim \text{Unif}\{x : \|x\| = 1\}.$$

Notice that in the elliptical case (contrary to the spherical case) the radial part ξ is unique up to rescaling, since $\xi \mathsf{A} = (\xi / c)(c\mathsf{A})$ for any $c \neq 0$.

Let $X = \mu + \mathrm{A}W$ for a spherical random vector W in $\mathbb{R}^d$. If A is nonsingular,

$$W = A^{-1}(X - \mu)$$

which implies that if W has density $\psi(u) = g_\psi(\|u\|^2)$, then X has density

$$f_X(x) = \frac{1}{|\Sigma|^{1/2}}\,\psi\big(\Sigma^{-1/2}(x-\mu)\big) = \frac{1}{|\Sigma|^{1/2}}\,g_\psi\big((x-\mu)^\top\Sigma^{-1}(x-\mu)\big),$$

where $\Sigma = \mathrm{A}\mathrm{A}^\top$. Call ψ the generating density of f_X.

Comments

- The density depends on x only via $(x-\mu)^\top\Sigma^{-1}(x-\mu)$
- Hence it is constant on ellipsoids, i.e. has *elliptical contours* (hence the name)
- Since $X = \mu + c\mathrm{A}c^{-1}W$ for any $c \neq 0$, the dispersion matrix Σ is not unique (it is unique only up to rescaling).
- We have not assumed existence of second (or even first) moments.
  - If a first moment exists, then μ is the expectation.
  - If second moments exist, then some rescaled version of Σ is the covariance.

- Evidently, all spherical distributions, and all Gaussians are also elliptical.
- Gaussian variance mixtures, $Y = \mu + \sqrt{\zeta}AZ$, where $0 < \zeta \perp\!\!\!\perp Z \sim N(0, I)$ (exercise: show that $\text{cov}(Y) = \mathbb{E}[\zeta]\text{cov}(AZ)$ and $\text{corr}(Y) = \text{corr}\{AZ\}$.)
- A special case of Gaussian variance mixture (with $\nu/\zeta \sim \chi^2_\nu$ is the multivariate t distribution $t(\nu, \mu, \Sigma)$, with density

$$f(x) = \frac{\Gamma((\nu+p)/2)}{\Gamma(\nu/2)(\nu\pi)^{p/2}|\Sigma|^{1/2}} \left(1 + \frac{(x-\mu)^\top \Sigma^{-1}(x-\mu)}{\nu}\right)^{-\frac{\nu+p}{2}}, x \in \mathbb{R}^p,$$

where $\mu \in \mathbb{R}^p$, $\Sigma_{p\times p} \succ 0$, and $\nu \in \mathbb{N}$ are the degrees of freedom. Note that $\nu > 1$ is required for first moments to exist, and $\nu > 3$ for second moments.

Elliptical distributions are important because:

1. they retain some of the nice properties of Gaussians

   while

2. they allow for greater generality, heavier tails, and extremal dependence – the Gaussian does not:

   if $(X_1, X_2)^\top \sim N(0, \Sigma)$ with $\text{corr}(X_1, X_2) = \rho$ and $\text{var}(X_1) = \text{var}(X_2) = 1$, then $X_2 | X_1 \sim N(\rho x, 1 - \rho^2)$ and so

$$\mathbb{P}[X_2 > x | X_1 = x] = 1 - \Phi\left(x\sqrt{\frac{1-\rho}{1+\rho}}\right) \overset{x \to \infty}{\to} 0.$$

Which properties do they retain?

- Closure under marginalisation (exercise)
- Ellipticity is preserved under
  - affine transformations (exercise)
  - conditioning (possibly with a different generating density)

Closure Under Marginalization

If $W = (X^\top, Y^\top)^\top$ is jointly elliptical with location/dispersion,

$$\mu_W = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} \qquad \& \qquad \Sigma_W = \begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{XY}^\top & \Sigma_Y \end{pmatrix}$$

then:

1. X (resp. Y) is also elliptical with location μ_X (resp. μ_Y) and dispersion Σ_X (resp. Σ_Y). Furthermore, if W has generating density ψ, so do X and Y.
2. $Q = \mathrm{B}W + \theta$ is also elliptical, with location $\mathrm{B}\mu + \theta$ and dispersion $\mathrm{B}\Sigma\mathrm{B}^\top$. When they exist, the generating densities of W and Q coincide.
3. $X \,|\, \{Y = y\}$ is also elliptical, with location $\mu_X - \Sigma_{XY}\Sigma_Y^\dagger(y - \mu_Y)$ and dispersion $\Sigma_X - \Sigma_{XY}\Sigma_Y^\dagger\Sigma_{XY}^\top$. Should they exist, the generating densities of W and $X \,|\, Y$ need not coincide.

Sampling Theory: Gaussian and Approximate

An i.i.d. Gaussian sample is a collection $X_1, ..., X_n \overset{iid}{\sim} N(\mu_{p\times 1}, \Sigma_{p\times p})$ in $\mathbb{R}^p$. We can stack then row-wise to build what is known as a (Gaussian) data matrix:

$$\mathrm{X} = \begin{pmatrix} X_{11} & \dots & X_{1p} \\ X_{21} & \dots & X_{2p} \\ \vdots & & \vdots \\ X_{n1} & \dots & X_{np} \end{pmatrix} = \begin{pmatrix} X_1^\top \\ X_2^\top \\ \vdots \\ X_n^\top \end{pmatrix} = (X_1 \ X_2 \ \dots \ X_n)^\top.$$

In a data matrix:

- The n rows represent observations/cases.
- The p columns represent variables/features.
- Think of an $n \times p$ design matrix in linear models.

When sampling a Gaussian data matrix X, we wish to know sampling laws of:

- Linear transformations of X, i.e. matrices of the form

$$\mathrm{AXB}, \qquad \mathrm{A} \in \mathbb{R}^{m\times n}, \ B \in \mathbb{R}^{p\times q}.$$

- Quadratic forms of X, i.e. matrices of the form

$$\mathrm{X}^\top \mathrm{CX}, \qquad \mathrm{C} = \mathrm{C}^\top \in \mathbb{R}^{n\times n}$$

Why?

- Sample mean $\bar{X}$ is a linear form

$$\bar{X}^\top = \frac{1}{n}\mathbf{1}_n^\top \mathsf{X} = \mathsf{AXB},$$

with $\mathsf{A} = \frac{1}{n}\mathbf{1}_n^\top$ and $\mathsf{B} = \mathsf{I}_{p\times p}$.

- Sample covariance is a quadratic form

$$\widehat{\Sigma} = \frac{1}{n}\mathsf{X}^\top \mathsf{H}\mathsf{X}$$

where the centring matrix H_n of dimension n is defined as

$$\mathsf{H}_n = \mathsf{I}_{n\times n} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top.$$

Note that $\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top = \mathbf{1}_n(\mathbf{1}_n^\top\mathbf{1}_n)^{-1}\mathbf{1}_n^\top$ is the projection onto $\mathrm{span}\{\mathbf{1}_n\}$.Hence H_n is the projection onto $\mathrm{span}^\perp\{\mathbf{1}_n\}$.

- The sample mean/covariance are sufficient[3] for their population counterparts.

[3] under some conditions

Theorem (Gaussian Sampling)

Let $X_1, ..., X_n \overset{iid}{\sim} N(\mu, \Sigma)$ be an random sample of size n of Gaussian d-vectors. Then,

- *The sample mean is a Gaussian vector: $\bar{X} = \frac{1}{n}\mathsf{X}^\top \mathbf{1}_n \sim N(\mu, n^{-1}\Sigma)$*
- *The sample covariance is a Wishart matrix: $n\widehat{\Sigma} = \mathsf{X}^\top \mathsf{H}\mathsf{X} \sim W_p(\Sigma, n-1)$.*
- *The sample mean and sample covariance are independent $\bar{X} \perp\!\!\!\perp \widehat{\Sigma}$*

The results could follow from the behaviour of Gaussians under linear/quadratic transformaton[4]

But it might be cleaner/simpler to restate such theorems in terms of data matrices:

- Properties of linear forms involving Gaussian data matrices
- Properties of quadratic forms involving Gaussian data matrices

[4] vectorising row-wise and using Kronecker products

As a first question when is a linear form AXB of a Gaussian data matrix X also a Gaussian data matrix?

- Clearly AXB is always a Gaussian matrix (entries jointly Gaussian)
- But to call it a *Gaussian data matrix* it must have i.i.d. rows.

Theorem (Linear Forms of Gaussian Data Matrices)

If X *is a Gaussian* $n \times p$ *data matrix from* $N(\mu, \Sigma)$*, then* $\mathrm{A}_{m\times n}\mathrm{X}\mathrm{B}_{p\times q}$ *is an* $m \times q$ *Gaussian data matrix if and only if the following two conditions hold true:*

1. $\boxed{\mathrm{A}\mathbf{1}_n = \alpha \mathbf{1}_m \text{ for some } \alpha \in \mathbb{R}}$ *OR* $\boxed{\mathrm{B}^\top \mu = 0}$.
2. $\boxed{\mathrm{A}\mathrm{A}^\top = \beta \mathrm{I}_{m\times m} \text{ for some } \beta \in \mathbb{R}}$ *OR* $\boxed{\mathrm{B}^\top \Sigma \mathrm{B} = 0}$.

Clearly, when AXB *is a Gaussian data matrix, it is from a* $N(\alpha \mathrm{B}^\top \mu, \beta \mathrm{B}^\top \Sigma \mathrm{B})$.

Proof.

Exercise. Note that post-multiplication of X involves adding weighted *variables*. Hence the rows of XB remain independent. Thus rows of AXB) will be independent unless premultiplication by A introduces some interdependence (premultiplication of X *adds* weighted *observations*). □

Then we can ask when are two different linear forms AXB and CXD of a Gaussian data matrix independent?

Theorem (Independence Between Linear Forms of Gaussian Data Matrices)

Let X *be a Gaussian data matrix from* $N(\mu, \Sigma)$. *Then*

$$\mathrm{AXB} \perp\!\!\!\perp \mathrm{CXD} \iff \mathrm{AC}^\top = 0 \text{ or } \mathrm{B}^\top \Sigma \mathrm{D} = 0.$$

Exercise. Prove the theorem.

Now let's consider what happens if we "square" a Gaussian data matrix X:

Definition (Wishart Matrix)

Let $\mathsf{X}_{n\times p}$ be a Gaussian data matrix from a $N(0,\Sigma)$ distribution. The $p\times p$ random matrix $\mathsf{X}^\top\mathsf{X}$ is said to follow a p-dimensional Wishart distribution with scale Σ and n degrees of freedom,

$$\mathsf{X}^\top\mathsf{X} \sim W_p(\Sigma_{p\times p}, n).$$

When $\Sigma = \mathsf{I}_{p\times p}$ we speak of the *standard* p-Wishart distribution with n d.f.

Let's try to get our head around this definition:

- When $n=1$ and $\Sigma = \mathsf{I}_{p\times p}$ then we are looking at the distribution of $ZZ^\top$ for $Z\sim N(0,\mathsf{I}_{p\times p})$ — the *outer product* of a standard Gaussian w/ itself.
- Compare this with the *inner product* of a standard Gaussian $Z^\top Z$ w/ itself.

Some properties are as follows (exercise):

- $\mathsf{W}\sim W_p(\Sigma_{p\times p},n) \iff \mathsf{W}\overset{d}{=}\sum_{i=1}^n \mathsf{W}_i, \quad \mathsf{W}_i\overset{iid}{\sim} W_p(\Sigma_{p\times p},1)$
- $\mathsf{W}\sim W_p(\Sigma_{p\times p},n) \implies \mathbb{E}[\mathsf{W}] = n\Sigma.$
- The lower triangular part of $\mathsf{W}\sim W_p(\Sigma_{p\times p},n)$ has density if and only if $n\geq p$.
- $\mathsf{W}\sim W_p(\Sigma_{p\times p},n) \implies \theta^\top\mathsf{W}\theta/\theta^\top\Sigma\theta\sim\chi^2_n,\ \forall\theta\notin\ker(\Sigma).$

Proposition (Closure under conjugation)

$\mathsf{W} \sim W_p(\Sigma, n) \implies \mathsf{B}^\top \mathsf{W} \mathsf{B} \sim W_p(\mathsf{B}^\top \Sigma \mathsf{B}, n)$

Proof.

Writing $\mathsf{W} = \mathsf{X}^\top \mathsf{X}$ for an $n \times p$ data matrix from $N(0, \Sigma)$, we have

$$\mathsf{B}^\top \mathsf{W} = \mathsf{B}^\top \mathsf{X}^\top \mathsf{X} \mathsf{B} = \mathsf{Y}^\top \mathsf{Y}$$

where $\mathsf{Y} = \mathsf{I}_{p \times p} \mathsf{X} \mathsf{B}$ is also a Gaussian data matrix, from a $N(0, \mathsf{B}^\top \Sigma \mathsf{B})$, by the data matrix linear form theorem (evidently, $\mathsf{I}_{p \times p} \mathbf{1}_p = \mathbf{1}_p$ & $\mathsf{I}_{p \times p}^\top \mathsf{I}_{p \times p} = \mathsf{I}_{p \times p}$). □

Corollary (Standardisation)

The random matrix W *has a* $W_p(\Sigma, n)$ *distribution if and only if* $\Sigma^{\dagger/2} \mathsf{W} \Sigma^{\dagger/2}$ *has a* $W_q(\mathsf{H}, n)$ *distribution, where* $\mathsf{H} = \Sigma^\dagger \Sigma$ *is the projection onto the range of* Σ.

- To see why this is standardisation, just assume Σ is non-singular.

Recall that when $Z \sim N(0, \mathsf{I}_p)$, and H is a projection, then $Z^\top \mathsf{H} Z \sim \chi^2_{\mathtt{rank}(\mathsf{H})}$.

What is the analogue for Wisharts? Note that $Z^\top \mathsf{H} Z = \|\mathsf{H}\mathsf{Z}\|^2$, so the corresponding Wishart quantity would be

$$\mathsf{Z}^\top \mathsf{H} \mathsf{Z}$$

where $\mathsf{Z}_{n\times p} = (Z_1^\top, ..., Z_n^\top)^\top$.

Theorem (Cochran's Theorem)

Let $\mathsf{X}_{n\times p}$ *be a Gaussian data matrix from a* $N(0, \Sigma)$ *and* H *be symmetric. Then,*

$$\mathsf{X}^\top \mathsf{H} \mathsf{X} \text{ is a Wishart matrix of mean } \Sigma \iff \mathsf{H} \text{ is a projection.}$$

When H *is indeed a projection,* $\mathsf{X}^\top \mathsf{H} \mathsf{X} \sim W_p(\Sigma, \mathtt{rank}(\mathsf{H}))$.

Proof.

Let $\mathsf{H} = \mathsf{U}\Lambda\mathsf{U}^\top$ be the spectral decomposition of H. Then $\mathsf{X}^\top\mathsf{H}\mathsf{X} = \mathsf{X}^\top\mathsf{U}\Lambda\mathsf{U}^\top\mathsf{X}$ and $\mathsf{Y} = \mathsf{U}^\top\mathsf{X}\mathsf{I}$ is a $N(0, \Sigma)$ Gaussian data matrix, because $\mathsf{U}^\top\mathsf{U} = \mathsf{I}$ and $\mathsf{I}0 = 0$.

Thus, $\mathsf{X}^\top\mathsf{H}\mathsf{X} = \mathsf{Y}^\top\Lambda\mathsf{Y} = \sum_{i=1}^n \lambda_i Y_i Y_i^\top$, where $Y_i \overset{iid}{\sim} N(0, \Sigma)$. Now:

- if H is a projection, then λ_i is either 0 or 1. Hence
$$\sum_{i=1}^n \lambda_i Y_i Y_i^\top = \sum_{j=1}^{\mathrm{rank}(\mathsf{H})} Y_i Y_i^\top \sim W_p(\Sigma, \mathrm{rank}(\mathsf{H})).$$
- Now let $\mathrm{rank}(\mathsf{H}) = q \leq n$. Then only its first q eigenvalues λ_i are non-zero. Assume now, as in the statement, that for some (yet unspecified) d, $\sum_{i=1}^q \lambda_i Y_i Y_i^\top \sim W_p(\Sigma, d)$. Multiplying both sides from left by $e_1^\top$ (the first canonical vector) and from right by e_1, then dividing both sides by $e_1^\top\Sigma e_1$ and finally using the last exercise in slide 114 we arrive at equality in distribution of:
  1. On the LHS a weighted sum of q independent χ_1^2, with λ_i as weights.
  2. On the RHS, a single χ_d^2

  Equating the corresponding MGFs, yields for all t sufficiently small
$$\prod_{i=1}^q (1 - 2\lambda_i t)^{-1/2} = (1 - 2t)^{-d/2} \implies \prod_{i=1}^q (1 - 2\lambda_i t) = (1 - 2t)^d$$
  and the last equality can only happen over an interval of t's if both polynomials have same degree and same roots. Hence $d = q$ and $\lambda_i = 1$.

We are now ready to prove the sampling theorem.

Proof (Sampling from a Gaussian)

The distribution of the sample mean vector follows directly from the theorem on linear forms of Gaussian data matrices, because

$$\bar{X}^\top = \underbrace{n^{-1}\mathbf{1}_n^\top}_{\mathrm{A}}\mathrm{X}$$

with $\mathrm{A}\mathbf{1}_n = n^{-1}\mathbf{1}_n^\top\mathbf{1}_n = 1$ and $\mathrm{A}\mathrm{A}^\top = n^{-1}\mathbf{1}_n^\top n^{-1}\mathbf{1}_n = n^{-1}$. The distribution of the sample covariance follows directly from Cochran's theorem.

Independence of the sample mean and sample covariance follow from the theorem on independence of Gaussian data matrix linear forms, by considering $\mathrm{CX} := \mathrm{HX}$ and $\mathrm{AX} := n^{-1}\mathbf{1}_n^\top\mathrm{X}$ and noting that $\mathrm{AC}^\top = \mathrm{H}\mathbf{1}_n = 0$ since H projects onto $\mathrm{span}^\perp\{\mathbf{1}_n\}$.

A small parenthesis.

When doing testing, we will be interested in constructing statistics of the form

$$(\bar{X} - \mu)^\top \Sigma^{-1} (\bar{X} - \mu).$$

But, more often than not, Σ will be unknown (nuiscance parameter). So we will replace it by the sample version,

$$(\bar{X} - \mu)^\top \widehat{\Sigma}^{-1} (\bar{X} - \mu).$$

Obviously, this will yield a different sampling distribution for the statistic than when using Σ itself.

Definition (Hotelling T^2 distribution)

For $n \geq p$, let $\tau^2 = n Y^\top \mathrm{W}^{-1} Y$ where $Y_{p\times 1}$ and $\mathrm{W}_{p\times p}$ are independently distributed as $N(0, \mathrm{I})$ and $W_p(\mathrm{I}, n)$, respectively. Then τ^2 is said to follow the Hotelling T^2 distribution with parameters p and n, written $\tau^2 \sim T^2(p, n)$.

Lemma

If $X \sim N(\mu, \Sigma_{p\times p})$ *independently of* $\mathrm{W} \sim W_p(\Sigma, n)$ *with* Σ *non-singular and* $n \geq p$*, then*

$$n(X-\mu)^\top \mathrm{W}^{-1}(X-\mu) \sim T^2(p,m).$$

Corollary

Let $\bar{X}$ *and* $\widehat{\Sigma}$ *be the sample mean and covariance of a* $N(\mu, \Sigma)$ *iid sample. If* $n \geq p$ *and* Σ *is non-singular, then*

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

Exercise: prove the lemma and the corollary.

- Recall that the square of a Student t_m distribution yields a $F_{1,m}$ distribution
- Hence $T^2(1,m) \equiv F_{1,m} \equiv (\mathrm{t}_m)^2$
- More generally, we have

$$T^2(p,m) \equiv \frac{mp}{m-p+1} F_{p,m-p+1}.$$

- It may happen that X is a data matrix, albeit non-Gaussian.
- What can we say about $\bar{X}$ and $\widehat{\Sigma}$ then?
- In general answer depends on row distribution.
- We are instead looking for *universality*.
- For this we need to consider an approximate/asymptotic law.
- Essentially two parameters we can play with: n and p

Definition (Weak Convergence)

Let $\{X_n\}$ be a sequence of random vectors of $\mathbb{R}^d$, and X a random vector of $\mathbb{R}^d$. Let $F_n, F : \mathbb{R}^d \to [0,1]$ be the corresponding joint CDFs. We say that X_n converges in distribution to X as $n \to \infty$ (and write $X_n \xrightarrow{d} X$) if

$$F_{X_n}(x) \overset{n\to\infty}{\longrightarrow} F_X(x)$$

for every continuity point $x \in \mathbb{R}^d$ of F_X.

There is a link between univariate and multivariate weak convergence:

Theorem (Cramér-Wold Device)

Let $\{X_n\}$ be a sequence of random vectors of $\mathbb{R}^d$, and X a random vector of $\mathbb{R}^d$. Then,

$$X_n \xrightarrow{d} X \Leftrightarrow \theta^\top X_n \xrightarrow{d} \theta^\top X, \ \forall \theta \in \mathbb{R}^d.$$

Exercise: show by counterexample that separate weak convergence of each coordinate does not imply weak convergence of the random vector.

When it comes to matrices,

- the definition is essentially the same
- the Cramér-Wold device for $n \times p$ matrices reads

$$\mathsf{M}_n \xrightarrow{d} \mathsf{M} \iff \mathrm{trace}\{\mathsf{A}^\top \mathsf{M}\} \xrightarrow{d} \mathrm{trace}\{\mathsf{A}^\top \mathsf{M}\}, \ \forall \mathsf{A} \in \mathbb{R}^{n \times p}.$$

Exercise: verify that the latter is the right formulation indeed.

Theorem (Strong Law of Large Numbers)

Let $\{X_n\}$ *be pairwise iid random vectors with* $\mathbb{E}X_k = \mu$ *and* $\mathbb{E}\|X_k\| < \infty$, *for all* $k \geq 1$. *Then,*

$$\frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow{a.s.} \mu$$

- "Strong" is as opposed to the "weak" law which requires $\mathbb{E}X_k^2 < \infty$ instead of $\mathbb{E}|X_k| < \infty$ and gives "$\xrightarrow{p}$" instead of "$\xrightarrow{a.s.}$"

Theorem (Central Limit Theorem)

Let $\{X_m\}$ *be an iid sequence of random vectors in* $\mathbb{R}^d$ *with mean* μ *and covariance* Σ *with* $\text{trace}\{\Sigma\} < \infty$. *Let* $\bar{X}_n := \sum_{i=1}^{n} X_i/n$ *be their sample mean. Then,*

$$\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} N(0, \Sigma).$$

Exercise: prove this CLT using the 1D CLT and the Cramér-Wold device

- Law of Large Numbers: assuming finite variance, rate of $n^{-1/2}$
- What about the CLT? What is the quality of the approximation?

Theorem (Berry-Essen-Benktus)

Let $X_1, ..., X_n$ be iid random p-vectors with mean 0 and covariance $\mathsf{I}_{p\times p}$. Define,

$$S_n = \frac{1}{\sqrt{n}}(X_1 + \ldots + X_n).$$

If $\mathcal{A}_p$ denotes the class of convex subsets of $\mathbb{R}^p$, then for $Z \sim N_p(0, I_p)$,

$$\sup_{A\in\mathcal{A}_p} |\mathbb{P}[S_n \in A] - \mathbb{P}[Z \in A]| \leq c \times \frac{p^{1/4}\mathbb{E}||X_i||^3}{\sqrt{n}}$$

for some universal constant $c \leq 400$

- Notice the dependence on dimension.
- Can we let p grow too?

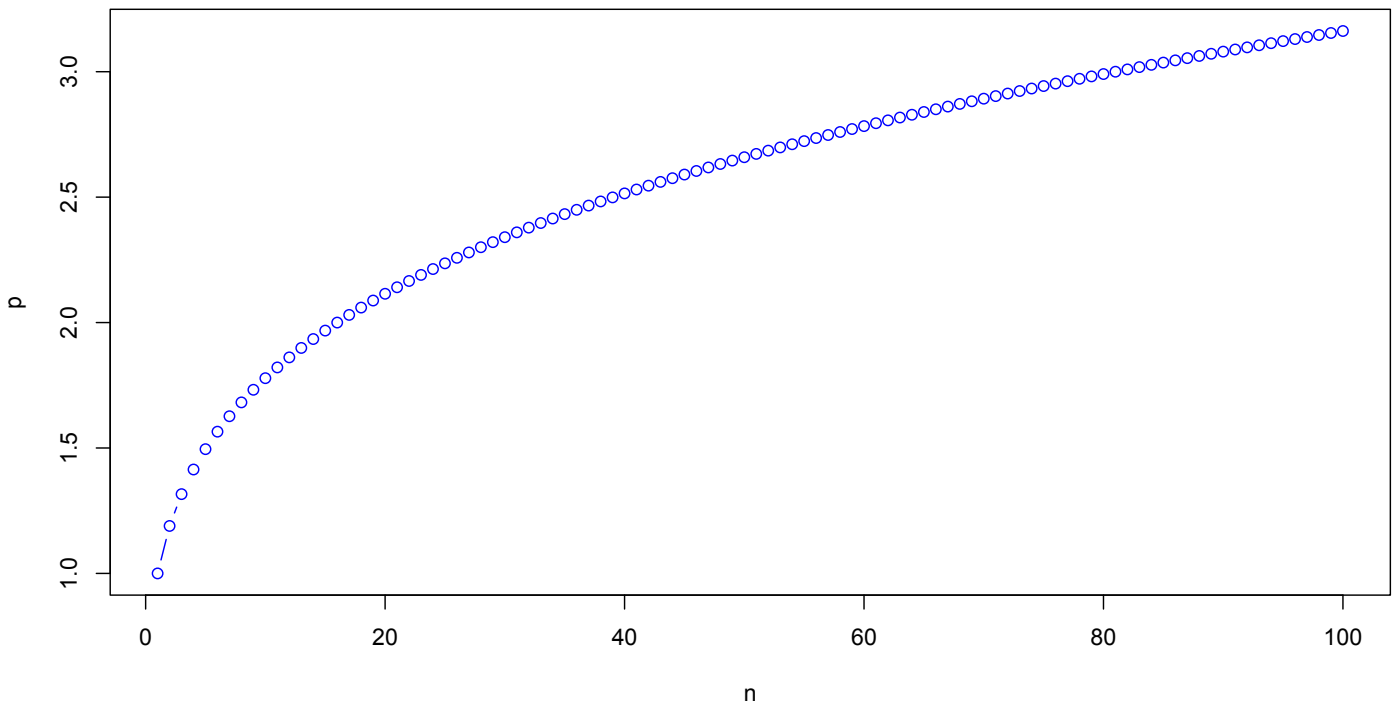
- By Jensen's inequality, when $\mathbb{E}[XX^\top] = \mathsf{I}_{p\times p}$, we have

$$\mathbb{E}\{\|X\|^3\} \geq \mathbb{E}\{\|X\|^2\}^{3/2} = p^{3/2}$$

- So to make the upper bound shrink to 0, it is necessary that $p = o(n^{2/7})$, i.e.

$$\frac{p_n}{n^{2/7}} \overset{n\to\infty}{\longrightarrow} 0.$$

- here's what (for example) $p \sim n^{2/8} \equiv n^{1/4}$ looks like



- What about cases when p is of same order or larger than n?
- In the context of the last theorem, assume that
  - $\mathbb{E}[X_{ij}^2] \geq b$ for all $j = 1, ..., p$.
  - $\mathbb{E}[|X_{ij}|^{2+k}] \leq B$ for $k = 1, 2$ and all $j = 1, ..., p$.
  - $\mathbb{E}[\exp\{|X_{ij}|/B\}] \leq 2$ for all $i \leq n$ and $j \leq p$.
- Then, if we focus only on rectangles $\mathcal{R}_p$ of $\mathbb{R}^p$, we have

Theorem (Chernozhukov, Chetverikov & Kato, 2017)

$$\sup_{R \in \mathcal{R}_p} |\mathbb{P}[S_n \in R] - \mathbb{P}[Z \in R]| \leq C \times \left(\frac{\log^7(pn)}{n} \right)^{1/6}$$

where $Z \sim N(0, \mathrm{I}_{p \times p})$ the constant C depends only on b and B.

- CCK simply requires $\frac{\log p_n}{n^{1/7}} \overset{n \to \infty}{\longrightarrow} 0$, i.e. $\log p = o(n^{1/7})$, allowing for $p \gg n$
- Compare to BEB necesary condition $p = o(n^{2/7})$
- Improvement comes at cost of smaller class of sets $\mathcal{R}_d \subset \mathcal{A}_d$
- Not a CLT in a traditional sense of "$\overset{d}{\to}$" (given approximation of probabilities without convergence to some fixed random vector – "moving target")

And what about the empirical covariance?

Observe that the empirical covariance is a sample average of the random matrices

$$X_1 X_1^\top, \ldots, X_n X_n^\top$$

- These are iid with mean Σ and some covariance. In the Gaussian (and elliptical) case this depends only on second moments but in general it will depend on 4th moments (see next two slides).
- They are elements of a real vector space of dimension $p(p+1)/2$.
- Therefore, the usual law of large numbers holds and the usual CLT hold
- And, provided we can standardise, the high dimensional CLT holds unchanged (i.e. with $\log[p(p+1)] \approx \log p^2 = 2 \log p = o(n^{1/7})$)

Let's specify, for the record, the covariance structure of $\mathrm{W} = XX^\top$, $X \sim N(0, \Sigma)$.

Covariance of Wishart matrix (a.k.a. Isserlis' Formula)

If $\mathrm{W} \sim W(\Sigma, 1)$, its covariance is, element-wise,

$$\begin{aligned} \mathrm{cov}\{w_{ij}, w_{kl}\} &= \mathrm{cov}\{X_i X_j, X_k X_l\} = \mathbb{E}[X_i X_j X_k X_l] - \mathbb{E}[X_i X_j]\mathbb{E}[X_k X_l] \\ &= (\sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk} + \sigma_{ij}\sigma_{kl}) - \sigma_{ij}\sigma_{kl} = \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk} \end{aligned}$$

The blue part is obtained by taking the mixed partial derivative of order 4 the MGF of a 4-dimensional Gaussian (a tedious but elementary calculation).

This can be arranged (exercise) in vectorised form as

$$\mathrm{cov}\{\mathrm{vec}(\mathrm{W})\} = (\Sigma \otimes \Sigma)(\mathrm{I}_{p^2 \times p^2} + \mathrm{K}_{pp})$$

where $\mathrm{K}_{pp} = \sum_{i=1}^p \sum_{i=1}^p \mathrm{H}_{ij} \otimes \mathrm{H}_{ij}^\top$ is the commutation matrix, with $\mathrm{H}_{ij} = e_i e_j^\top$. The commutation matrix "transposes vecs", i.e. $\mathrm{K}_{pp}\mathrm{vec}(\mathrm{W}) = \mathrm{vec}(\mathrm{W}^\top)$.

Exercise: Σ diagonal $\implies$ Wishart entries uncorrelated, ij element has variance

$$(1 + \mathbf{1}\{i = j\})\sigma_{ii}\sigma_{jj} \qquad \text{(recall notation } \sigma_{ii} \equiv \Sigma_{ii} = \sigma_i^2\text{)}$$

Exercise: $\Sigma \succ 0 \implies \mathrm{cov}[\{w_{ij}\}_{i \le j}] \succ 0$ (lower triangular part of W).

Warning: in non-Gaussian case, 2nd moments do not determine 4th moments:

$$\mathbb{E}[X_i X_j \, , \, X_k X_l] = m_4(i,j,k,l) \neq \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk} + \sigma_{ij}\sigma_{kl}$$

- So we can't use covariance structure of a Wishart matrix in the CLT limit.
- Instead, we need essentially all (mixed) 4th moments (which, are also *very* difficult to estimate)

A notable exception is for centred X with elliptical law of dispersion $\Psi \propto \Sigma$. The proportionality constant can be deduced to be $\mathbb{E}[X^\top \Psi^{-1} X]/p$.

Here, fourth moments enter only through a simple scalar parameter

$$\kappa = \frac{p}{(p+2)} \frac{\mathbb{E}\{[X^\top \Psi^{-1} X]^2\}}{\mathbb{E}^2[X^\top \Psi^{-1} X]} - 1$$

Indeed, one can calculate

$$\mathrm{cov}\{\mathrm{vec}(XX^\top)\} = (\kappa+1)(\Sigma \otimes \Sigma)(\mathsf{I}_{p^2 \times p^2} + \mathsf{K}_{pp}) + \kappa \mathrm{vec}(\Sigma)\mathrm{vec}(\Sigma)^\top.$$

or elementwise,

$$\mathbb{E}[X_i X_j X_k X_l] = (\kappa+1)\sigma_{ik}\sigma_{jl} + (\kappa+1)\sigma_{il}\sigma_{jk} + \kappa\sigma_{ij}\sigma_{lk}.$$

So if Σ diagonal, entries of $XX^\top$ are uncorrelated, as in Gaussian case.
In the Gaussian case, we can calculate $\kappa = 0$.

- Occasionally we need to studentise by an unknown but consistently estimatble quantity.
- Other times, we have a CLT for a quantity, but we are interested in some functional thereof.
- For such cases, statisticians rely on two essential tools:

Lemma (Slutsky)

Let X be a random vector in $\mathbb{R}^p$, $\theta \in \mathbb{R}^q$, and $g : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ be continuous on $\text{supp}\{X\} \times \{\theta\}$. *If $X_n \xrightarrow{d} X$ in $\mathbb{R}^p$ and $Y_n \xrightarrow{d} \theta$, then $g(X_n, Y_n) \xrightarrow{d} g(X, c)$.*

Theorem (Delta Method)

Let $Z_n := a_n(X_n - \theta) \xrightarrow{d} Z$ where $0 < a_n \uparrow \infty$ and $\theta \in \mathbb{R}^p$. Let $g(\cdot) : \mathbb{R}^p \to \mathbb{R}^q$ be differentiable at θ. Then, $a_n(g(X_n) - g(\theta)) \xrightarrow{d} [(\nabla g)(\theta)]Z$.

There is a fundamental difference between the low/high dimensional cases.

To see this,

- Consider the n iid p-vectors $\{X_i\}$ with mean 0 and covariance I_p.
- Whether in low/high dimension, when it holds, the CLT heuristically says

$$\sqrt{n}\bar{X} \overset{d}{\approx} N(0, \mathsf{I}_p), \qquad \text{for large } n.$$

$\Longrightarrow$ In the low dimensional case, $\bar{X}$ collapses to the true mean as $n \to \infty$

$\Longrightarrow$ In the high dimensional case, $\bar{X}$ concentrates on the sphere of radius $\sqrt{\frac{p}{n}}$.

(recall the concentration of measure phenomenon)

So depending on the ratio p/n we might very well not have a LLN.

This fundamental difference has immediate consequences when doing statistics at low vs high dimensions – they are two distinct and very different regimes.

Inference

Proposition (Gaussian Likelihood)

Let $X_1, ..., X_n \overset{iid}{\sim} N(\mu, \Sigma)$ *be a sample of Gaussian* p*-vectors. The likelihood* $L(\mu, \Sigma)$ *of* (μ, Σ) *is given by*

$$\frac{\mathbf{1}\{\mathcal{R}\left(\hat{\Sigma} + (\bar{x} - \mu)(\bar{x} - \mu)^\top\right) \subseteq \mathcal{R}(\Sigma)\}}{\left(\prod_{j=1}^{r}(2\pi\lambda_j(\Sigma))^{1/2}\right)^n} \exp\left\{-\frac{n}{2}(\bar{x} - \mu)^\top \Sigma^\dagger (\bar{x} - \mu)\right\} \times \exp\left\{-\frac{n}{2}\mathrm{trace}\left\{\Sigma^\dagger \hat{\Sigma}\right\}\right\}$$

where $r = \mathrm{rank}(\Sigma) \leq p$ *and* $\{\lambda_1(\Sigma), \ldots, \lambda_r(\Sigma)\}$ *are the positive eigenvalues of* Σ*. Consequently, the likelihood depends on the data only through* $(\bar{X}, \hat{\Sigma})$*.*

Some comments

- If/when they exist, the maximum likelihood estimators of (μ, Σ) will be functions of $(\bar{X}, \hat{\Sigma})$.
- When $\Sigma \succ 0$, we immediately conclude that $(\bar{X}, \hat{\Sigma})$ is sufficient for (μ, Σ), by the Fisher-Neyman factorisation theorem.
- Without restrictions on the support of the Gaussian law, the model is non-regular (no common dominating measure).

Proof.

We will make use of the following identity (exercise)

$$\sum_{i=1}^n (x_i - \mu)(x_i - \mu)^\top \overset{\star}{=} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^\top + n(\bar{x} - \mu)(\bar{x} - \mu)^\top.$$

Recall that when $X \sim \mathsf{N}(\mu_{p\times 1}, \Sigma_{p\times p})$, X admits a density on $\mu + \mathcal{R}(\Sigma)$ given by

$$f_X(x) = \frac{1}{\prod_{j=1}^r (2\pi\lambda_j(\Sigma))^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^\top \Sigma^\dagger (x-\mu)\right\} \mathbf{1}\{x \in \mathcal{R}(\Sigma) + \mu\},$$

where $r = \mathrm{rank}(\Sigma) \leq p$ and $\{\lambda_1(\Sigma), \ldots, \lambda_r(\Sigma)\}$ are the positive eigenvalues. Therefore, we obtain the joint density w.r.t. Lebesgue measure on $\mu + \mathcal{R}(\Sigma)$ as

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \frac{\prod_{i=1}^n \mathbf{1}\{x_i \in \mathcal{R}(\Sigma) + \mu\}}{\left(\prod_{j=1}^r (2\pi\lambda_j(\Sigma))^{1/2}\right)^n} \exp\left\{-\frac{1}{2}\sum_{i=1}^n (x_i - \mu)^\top \Sigma^\dagger (x_i - \mu)\right\}$$

Let's first focus on the "red" factor, and then the "green" factor. The "blue" factor needs no more work.

Re-writing the quadratic form as a trace, and using the identity $\star$, get red factor=

$$
\begin{aligned}
&= \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\mathrm{tr}\left(\Sigma^{\dagger}(x_i-\mu)(x_i-\mu)^{\top}\right)\right\} = \exp\left\{-\frac{1}{2}\mathrm{tr}\left(\Sigma^{\dagger}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^{\top}\right)\right\} \\
&= \exp\left\{-\frac{n}{2}\mathrm{tr}\left(\Sigma^{\dagger}(\bar{x}-\mu)(\bar{x}-\mu)^{\top}\right)\right\} \times \exp\left\{-\frac{1}{2}\mathrm{trace}\left\{\Sigma^{\dagger}\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})^{\top}\right\}\right\} \\
&= \exp\left\{-\frac{n}{2}(\bar{x}-\mu)^{\top}\Sigma^{\dagger}(\bar{x}-\mu)\right\} \times \exp\left\{-\frac{n}{2}\mathrm{trace}\left\{\Sigma^{\dagger}\hat{\Sigma}\right\}\right\}
\end{aligned}
$$

which is exactly the form sought.

As for the green factor, we need to to show that $\prod_{i=1}^{n}\mathbf{1}\{x_i \in \mathcal{R}(\Sigma)+\mu\}$, or equivalently $\prod_{i=1}^{n}\mathbf{1}\{x_i-\mu \in \mathcal{R}(\Sigma)\}$, is as stipulated. To do this we will use the identity $\star$ again, combined with the claim (old exercise, reminder after the proof) that for any $\mathsf{Q} \succeq 0$,

$$
v_1, \ldots, v_k \in \mathcal{R}(\mathsf{Q}) \overset{\heartsuit}{\iff} \mathcal{R}\left(\sum_{i=1}^{k} v_k v_k^{\top}\right) \subseteq \mathcal{R}(\mathsf{Q})
$$

Assuming the claim ♡ to be true, we immediately have

$$x_1 - \mu, ..., x_n - \mu \in \mathcal{R}(\Sigma) \overset{\heartsuit}{\iff} \mathcal{R}\left(\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^\top\right) \subseteq \mathcal{R}(\Sigma)$$

$$\overset{\star}{\iff} \mathcal{R}\left(n\hat{\Sigma} + n(\bar{x} - \mu)(\bar{x} - \mu)^\top\right) \subseteq \mathcal{R}(\Sigma)$$

Since the factor n has no bearing on the inclusion, the proof will be complete as soon as we establish our claim. This we do separately below. □

Lemma (Ranges and Spans)

$$v_1, ..., v_k \in \mathcal{R}(Q) \overset{\heartsuit}{\iff} \mathcal{R}\left(\sum_{i=1}^{k} v_k v_k^\top\right) \subseteq \mathcal{R}(Q)$$

(we have seen this early in the course)

The log-likelihood (up to constants) for μ and Σ based on a Gaussian data matrix $\mathsf{X}_{n\times p}$, $n > 1$, equals $-\infty$ when $\mathcal{R}\left(n\hat{\Sigma} + n(\bar{x}-\mu)(\bar{x}-\mu)^\top\right) \not\subseteq \mathcal{R}(\Sigma)$, and otherwise equals

$$\ell(\mu,\Sigma) = -\frac{n}{2}\sum_{j=1}^{\mathtt{rank}(\Sigma)} \log\lambda_j(\Sigma) - \frac{n}{2}(\bar{x}-\mu)^\top\Sigma^\dagger(\bar{x}-\mu) - \frac{n}{2}\mathtt{trace}\left\{\Sigma^\dagger\hat{\Sigma}\right\}$$

which is finite for all $\bar{x}-\mu \in \mathcal{R}(\Sigma)$. When $\Sigma \succ 0$ the log-likelihood is positive for all $\bar{x}-\mu$ and equals

$$\ell(\mu,\Sigma) = -\frac{n}{2}\log|\Sigma| - \frac{n}{2}(\bar{x}-\mu)^\top\Sigma^{-1}(\bar{x}-\mu) - \frac{n}{2}\mathtt{trace}\left\{\Sigma^{-1}\hat{\Sigma}\right\}.$$

- Imposing an assumption on the range of Σ (equivalently the support of the random vector) represents imposing a restriction on the model $\mathcal{N}_p = \{N(\mu,\Sigma) : \mu \in \mathbb{R}^p, \Sigma_{p\times p} \succeq 0\}$.
- Because the model $\mathcal{N}_p$ is non-regular, maximum likelihood estimation depends heavily on such restrictions.

Theorem (Gaussian Mean/Covariance MLE)

Let $X_1, ..., X_n \overset{iid}{\sim} N(\mu, \Sigma)$ be a sample of Gaussian p-vectors. Then,

1. *the unique MLE of μ is always $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.*
2. *Under the restriction $\mathcal{R}(\Sigma) = \mathcal{R}(\hat{\Sigma})$, the unique MLE of (μ, Σ) is $(\bar{X}, \hat{\Sigma})$.*
3. *Without any range restriction,*
   - *if $\hat{\Sigma}$ is non-singular, the unique MLE of (μ, Σ) is $(\bar{X}, \hat{\Sigma})$.*
   - *if $\hat{\Sigma}$ is singular, the MLE of μ is $\bar{X}$ but the MLE of Σ does not exist.*

Corollary (The non-singular and low-dimensional case.)

If $X_1, ..., X_n \overset{iid}{\sim} N(\mu, \Sigma)$ with $n > p$ and $\Sigma \succ 0$, then the unique MLE of μ and Σ are $\bar{X}$ and $\hat{\Sigma}$, respectively. Furthermore, $\bar{X}$ and $n\hat{\Sigma}/(n-1)$ are minimum variance unbiased estimators of μ and Σ.

Exercise: Prove the corollary (use projections for the second part).

Proof.

Notice that, unless $\Sigma \succ 0$, we need to be careful about the support of the joint density. The joint density vanishes unless

$$\mathcal{R}\left(n\hat{\Sigma} + n(\bar{x}-\mu)(\bar{x}-\mu)^\top\right) \subseteq \mathcal{R}(\Sigma),$$

or equivalently

$$x_i - \mu \in \mathcal{R}(\Sigma) \qquad \forall\ i \leq n.$$

When considering the joint density this is seen as a condition on the observations. But when considering the likelihood, this is seen as a condition on Σ and μ. When (μ, Σ) fail to satisfy it, the likelihood becomes zero (and the loglikelihood negatively infinite). Call this the "support condition".

Now consider estimation of μ first. The support condition compels us to only consider μ that satisfy $\bar{x} - \mu \in \mathcal{R}(\Sigma)$. Now, regardless of the choice of Σ, the middle term in the log-likelihood (the only term depending on μ)

$$-\frac{n}{2}(\bar{x}-\mu)^\top \Sigma^\dagger (\bar{x}-\mu)$$

attains its maximum of 0 when $\mu = \bar{x}$. This choice of μ trivially satisfies the support condition $\bar{x} - \mu \in \mathcal{R}(\Sigma)$, regardless of choice of Σ. So $\bar{X}$ is an MLE of μ. For uniqueness, let y be a candidate estimator satisfying the support condition, i.e. $\bar{x} - y \in \mathcal{R}(\Sigma)$. Since $\Sigma^\dagger \succ 0$ on $\mathcal{R}(\Sigma)$, we have $(\bar{x}-y)^\top \Sigma^\dagger (\bar{x}-y) = 0 \implies y = \bar{x}$.

To prove (2), note that wlog we can take $\mathcal{R}(\hat{\Sigma}) = \mathbb{R}^p$ (otherwise we rotate the space). Equivalently, we take Σ non-singular. Therefore any candidate Σ is also non-singular, so our search space is that of strictly positive definite matrices. When we plug in the MLE of μ, the loglikelihood reduces to

$$-\frac{n}{2}\log|\Sigma| - \frac{n}{2}\text{trace}\left\{\Sigma^{-1}\hat{\Sigma}\right\} = \frac{n}{2}\log|\Theta| - \frac{n}{2}\text{trace}\left\{\Theta\hat{\Sigma}\right\}$$

where $\Theta = \Sigma^{-1}$. Now let $\hat{V} = \hat{\Sigma}^{1/2} \succ 0$. Define $\Psi = \hat{V}\Theta\hat{V}$, and note that

$$\log|\Psi| = \log\left\{|\hat{V}|\,|\Theta|\,|\hat{V}|\right\} = \log\{|\hat{V}|\,|\hat{V}|\} + \log|\Theta| = \text{constant} + \log|\Theta|$$

while

$$\text{trace}\{\Theta\hat{\Sigma}\} = \text{trace}\{\hat{V}^{-1}\Psi\hat{V}^{-1}\hat{V}\hat{V}\} = \text{trace}\{\Psi\}.$$

Hence, up to constants, the loglikelihood can be expressed as

$$\log|\Psi| - \text{trace}\{\Psi\} = \sum_{i=1}^{p}\log\lambda_i(\Psi) - \sum_{i=1}^{p}\lambda_i(\Psi) = \sum_{i=1}^{p}(\log\lambda_i(\Psi) - \lambda_i(\Psi))$$

which can be optimised for each λ_i separately. Noting that 1 is the unique maximum of $\log x - x$ over $x > 0$ (check by differentiating), we get the unique MLE for $\lambda_i(\Psi) = 1$ for all i, i.e. $\Psi = \mathsf{I}_{p\times p}$, i.e. at $\Theta = \hat{\Sigma}^{-1}$ i.e. at $\Sigma = \hat{\Sigma}$.

To prove (3), plug in the MLE for the mean in the general form of the loglikelihood to get (up to constants)

$$\ell(\bar{x}, \Sigma) = \begin{cases} -\sum_{j=1}^{\mathrm{rank}(\Sigma)} \log \lambda_j(\Sigma) - \mathrm{trace}\left\{\Sigma^{\dagger}\hat{\Sigma}\right\} & \text{when } \mathcal{R}(\Sigma) \supseteq \mathcal{R}(\hat{\Sigma}), \\ -\infty & \text{otherwise.} \end{cases}$$

consider the sets

$$C_1 = \{\Sigma : \mathcal{R}(\Sigma) \supseteq \mathcal{R}(\hat{\Sigma})\}^c, \quad C_2 = \{\mathcal{R}(\hat{\Sigma}) = \mathcal{R}(\Sigma)\}, \quad C_3 = \{\mathcal{R}(\Sigma) \supset \mathcal{R}(\hat{\Sigma})\}.$$

Plugging in the MLE for the mean, we see that:

- Over C_1, the likelihood is zero.
- Over C_2, the maximal loglikelihood is finite and is attained uniquely at $\hat{\Sigma}$.
- Finally, over C_3, we can obtain a sequence with loglikelihood diverging to ∞ as follows. Take $\Sigma_m = \hat{\Sigma} + \alpha_n v v^{\top}$, where $v \in \mathcal{R}^{\perp}(\hat{\Sigma})$ is a unit vector (think of it as the "$(\hat{r}+1)$th eigenvector of $\hat{\Sigma}$") and $0 < \alpha_m < \lambda_{\hat{r}}(\hat{\Sigma})$ with $\alpha_m \downarrow 0$. For all α_m, the trace term yields the same value $-\hat{r}$. The "logdet" term on the other hand, equals $-\left(\sum_{j=1}^{\hat{r}} \log \lambda_j(\hat{\Sigma})\right) - \log(\alpha_m)$ which diverges. □

Exercise: If μ is known, the MLE of $\hat{\Sigma}$ (if it exists) becomes $\frac{1}{n}\sum(X_i - \mu)(X_i - \mu)^{\top}$.

Recall that the MLE is parametrisation equivariant: for any transformation g

$$\hat{\theta} \text{ is MLE of } \theta \Rightarrow g\left(\hat{\theta}\right) \text{ is MLE of } g(\theta)$$

If g is additionally 1-1, then uniqueness is also inherited when present.

Exercise: Show this. Note that we can find the MLE of $\phi = g(\theta)$ by maximising $\phi \mapsto \sup_{\theta \in g^{-1}(\phi)} L(\theta)$ where $L(\theta)$ is the likelihood for θ.

Thus, whenever $\Sigma \succ 0$ and $n > p$, we obtain the immediate corollaries:

- The MLE of the precision matrix $\Theta = \Sigma^{-1}$ is given by[5] $\hat{\Theta} := \hat{\Sigma}^{-1}$.
- The MLE of the correlation matrix R is given by

$$\hat{\mathrm{R}} = \mathrm{diag}\{\hat{\sigma}_1^{-1}, \ldots, \hat{\sigma}_p^{-1}\}\, \hat{\Sigma}\, \mathrm{diag}\{\hat{\sigma}_1^{-1}, \ldots, \hat{\sigma}_p^{-1}\}$$

  where $\hat{\sigma}_j$ is the jth diagonal element of $\hat{\Sigma}$.

[5] actually, our method of proof first showed $\hat{\Sigma}^{-1}$ to be the MLE of $\hat{\Sigma}^{-1}$, but anyway.

- Unless we have very special hypothesis structure (e.g. simple vs simple, or one-sided concerning a single coordinate of the mean), there will generally be no unequivocal choice of test (no optimal test).
- In special cases related to the mean, e.g. when Σ is known to be diagonal, one can do separate univariate tests, and combine them with a careful correction.
- A general (and usually sensible) general method is based on the likelihood ratio.

  (not the only approach, and other approaches can occasionally have advantages)
- Often we have a one-sample or a two-sample (or multi-sample) setting:
  - One sample: $X_1, ..., X_n \overset{iid}{\sim} N(\mu, \Sigma)$
  - Two sample: $X_1, ..., X_n \overset{iid}{\sim} N(\mu_X, \Sigma_X)$ and $Y_1, ..., Y_m \overset{iid}{\sim} N(\mu_Y, \Sigma_Y)$

Natural one-sample hypothesis pairs:

- $\{H_0 : \mu = \mu_0\}$ vs $\{H_1 : \mu \neq \mu_0\}$ (either when Σ known or Σ unknown)
- $\{H_0 : \Sigma = \Sigma_0\}$ vs $\{H_1 : \Sigma \neq \Sigma_0\}$.
- $\{H_0 : \Sigma \propto \mathsf{I}\}$ vs $\{H_1 : \Sigma \not\propto \mathsf{I}\}$ (sphericity test)
- $\{H_0 : \rho_{ij} = 0\}$ vs $\{H_1 : \rho_{ij} \neq 0\}$
- $\{H_0 : \rho_{ij|\text{rest}} = 0\}$ vs $\{H_1 : \rho_{ij|\text{rest}} \neq 0\}$
  (must be interpreted w/ care, partial corr is *always* wrt to a set of variables)

Natural two-sample hypothesis pairs (can be generalised to multi-sample case)

- $\{H_0 : \mu_X = \mu_Y\}$ vs $\{H_1 : \mu_X \neq \mu_Y\}$ with $\Sigma_X = \Sigma_Y$ (known or unknown)
- $\{H_0 : \Sigma_X = \Sigma_Y\}$ vs $\{H_1 : \Sigma_X \neq \Sigma_Y\}$
- $\{H_0 : \mu_X = \mu_Y\}$ vs $\{H_1 : \mu_X \neq \mu_Y\}$ with $\Sigma_X \neq \Sigma_Y$ (Behrens-Fisher problem)

Let's work out some cases to get the hang of it.

All the hypotheses previously formulated fall in the following general framework:

- $\mathbf{X} = (X_1^\top, ..., X_n^\top)^\top$ random vectors, w/ likelihood $L(\vartheta) = f_\vartheta(X_1, ..., X_n)$
- $\vartheta \in \Theta \subseteq \mathbb{R}^d$ where $\Theta = \Theta_0 \cup \Theta_1$ and $\Theta_0 \cap \Theta_1 = \emptyset$

Definition (Likelihood Ratio Test Statistic)

The *likelihood ratio statistic* for $H_0 : \vartheta \in \Theta_0$ vs $H_1 : \vartheta \in \Theta_1$ is

$$\Lambda = \sup_{\vartheta \in \Theta} L(\vartheta) / \sup_{\vartheta \in \Theta_0} L(\vartheta)$$

- Intuition: how much better do we do if we do if we do not restrict the maximisation of the likelihood to be over the subset Θ_0?
- We reject H_0 for large Λ or of some monotone increasing cts function of Λ.
- Which precise function depends on convenience (ease of calibration)
- Often the following increasing function is easy to calibrate (perhaps asymptotically)

$$2 \log \Lambda = 2 \left(\sup_{\vartheta \in \Theta} \log L(\vartheta) - \sup_{\vartheta \in \Theta_0} \log L(\vartheta) \right) = 2(\ell^* - \ell_0^*)$$

Implementing a likelihood ratio test requires two steps:

1. Determining the test statistic
2. Calibrating the test statistic (or a monotone transformation $\tau(\Lambda)$ theoref).

Calibration refers to finding the distribution of $\tau(\Lambda)$ under H_0, so that we can choose an appropriate quantile to define the critical region for rejection:

$$\tau(\Lambda) \overset{H_0}{\sim} F \implies \text{reject } H_0 \text{ at level } \alpha \text{ whenever } \tau(\Lambda) > q_{1-\alpha}(F)$$

- Often, especially in exponential families (like the Gaussian) we can find the exact sampling law under H_0.
- But more often it is not tractable, and we need an asymptotic approximation.
- A general such result for $\tau(\Lambda) = 2 \log \Lambda$ is given by Wilks' theorem.

Theorem (Wilk's theorem)

Let $X_1, ..., X_n$ be iid random vectors with density (frequency) depending on $\vartheta \in \mathbb{R}^p$ and satisfying the "usual regularity conditions" (see next slide). If the MLE sequence $\hat{\vartheta}_n$ is consistent for ϑ, then the likelihood ratio statistic Λ_n for $H_0 : \{\vartheta_j = \vartheta_{j,0}\}_{j=1}^s$, $s \leq p$, satisfies $2 \log \Lambda_n \xrightarrow{d} V \sim \chi^2_s$ when H_0 is true.

Note that Wilks' theorem applies for a simple null (not composite), though this does not need to fix all the parameters ($s < p$ is allowed).

Hypotheses of the form $H_0 : \{g_j(\vartheta) = a_j\}_{j=1}^s$, for g_j differentiable real functions, can also be handled by Wilks' theorem:

- Define $(\phi_1, ..., \phi_p) = g(\vartheta) = (g_1(\vartheta), ..., g_p(\vartheta))$
- $g_{s+1}, ..., g_p$ defined so that $\vartheta \mapsto g(\vartheta)$ is 1-1
- Apply theorem with parameter ϕ

The "usual regularity conditions" are as follows:

(A1) The parameter space $\Theta \in \mathbb{R}^p$ is open.

(A2) The support of f_ϑ is invariant w.r.t. ϑ

(A3) All mixed partial derivatives of ℓ w.r.t. ϑ up to degree 3 exist and are continuous.

(A4) $\mathbb{E}[\nabla_\vartheta \ell(X_i;\vartheta)] = 0 \ \forall \vartheta$ and $\text{cov}[\nabla_\vartheta \ell(X_i;\vartheta)] =: \boldsymbol{I}(\vartheta) \succ 0 \ \forall \vartheta$.

(A5) $-\mathbb{E}[\nabla^2_\vartheta \ell(X_i;\vartheta)] =: \boldsymbol{J}(\vartheta) \succ 0 \ \forall \vartheta$.

(A6) $\exists\, \delta > 0$ s.t. $\forall\, \vartheta \in \Theta$ and for all $1 \leq j,k,l \leq p$,

$$\left| \frac{\partial}{\partial \vartheta_j \partial \vartheta_k \partial \vartheta_l} \ell(x;u) \right| \leq M_{jkl}(x)$$

for $||\vartheta - u|| \leq \delta$ with M_{jkl} such that $\mathbb{E}[M_{jkl}(X_i)] < \infty$.

For a proof, see the "Statistical Inference" course, or Serfling, "Approximation Theorems of Mathematical Statistics" (Sec. 4.4.4.)

Let $X_1, \ldots, X_n$, be a random sample from $N(\mu, \Sigma_{p\times p})$ with $\Sigma \succ 0$ known. As Σ is known, H_0 is simple, and hence the maximal log-likelihood under H_0 is

$$\ell_0^* = \ell(\mu_0, \Sigma) = -\frac{n}{2}\log\det(2\pi\Sigma) - \frac{n}{2}\mathrm{tr}\left(\Sigma^{-1}\hat{\Sigma}\right) - \frac{n}{2}(\bar{X} - \mu_0)^\top\Sigma^{-1}(\bar{X} - \mu_0).$$

The unrestricted maximal loglikelihood occurs at the unrestricted MLE,

$$\ell^* = \ell(\bar{X}, \Sigma) = -\frac{1}{2}n\log\det(2\pi\Sigma) - \frac{1}{2}n\mathrm{tr}\left(\Sigma^{-1}\hat{\Sigma}\right).$$

Hence

$$2\log\Lambda = 2(\ell^* - \ell_0^*) = n(\bar{X} - \mu_0)^\top\Sigma^{-1}(\bar{X} - \mu_0),$$

which, under H_0, follows the χ^2_p distribution. Thus we reject H_0 at the level α iff

$$2\log\Lambda = n(\bar{X} - \mu_0)^\top\Sigma^{-1}(\bar{X} - \mu_0) > q_{1-\alpha}(\chi^2_p).$$

Note that we can invert this test to get a 95% confidence region for μ in the form

$$\{\mu_0 \in \mathbb{R}^p : n(\bar{X} - \mu_0)^\top\Sigma^{-1}(\bar{X} - \mu_0) \leq q_{1-\alpha}(\chi^2_p)\}$$

i.e. all the possible μ_0 for which H_0 isn't rejected.

One sample, $\{H_0 : \mu = \mu_0\}$ vs $\{H_1 : \mu \neq \mu_0\}$ (Σ unknown)

Let $X_1, \ldots, X_n$ be a random sample from $N(\mu, \Sigma_{p\times p})$ with $\Sigma \succ 0$ unknown and $n > p$. Σ must be estimated under H_0 and also under H_1. Therefore both hypotheses are composite. Based on our results,

- Under H_0 the MLE for μ and Σ are

$$\mu_0 \quad \text{and} \quad \hat{\Sigma}_0 = n^{-1}\sum_{i=1}^{n}(X_i - \mu_0)(X_i - \mu_0)^\top = \hat{\Sigma} + \delta\delta^\top,$$

  where $\delta = \bar{X} - \mu_0$.
- Under H_1 the MLE for μ and Σ are

$$\bar{X} \quad \text{and} \quad \hat{\Sigma}_1 = \hat{\Sigma}.$$

Plugging into the loglikelihood expression, we obtain $\ell_0^* = \ell(\mu_0, \hat{\Sigma} + \delta\delta^\top) =$

$$= -\frac{n}{2}\left[p\log(2\pi) + \log|\hat{\Sigma} + \delta\delta^\top| + \delta^\top\hat{\Sigma}^{-1}\delta + \mathrm{tr}\{(\hat{\Sigma} + \delta\delta^\top)^{-1}\hat{\Sigma})\}\right].$$

As $\hat{\Sigma} \succ 0$ a.s. when $n > p$, its rank 1 perturbed determinant is (exercise)

$$|\hat{\Sigma} + \delta\delta^\top| = |\hat{\Sigma}|(1 + \delta^\top\hat{\Sigma}^{-1}\delta),$$

which yields (via the Sherman-Morrison formula, exercise, and some algebra)

$$\ell_0^* = -\frac{n}{2}\left[p\log(2\pi) + \log|\hat{\Sigma}| + \log(1 + \delta^\top\hat{\Sigma}^{-1}\delta) + p\right].$$

In the unrestricted case, we calculate

$$\ell^* = \ell(\bar{X}, \hat{\Sigma}) = -\frac{n}{2}\left[p\log(2\pi) + \log|\hat{\Sigma}| + p\right],$$

and thus

$$2\log\Lambda = 2(\ell^* - \ell_0^*) = n\log(1 + \delta^\top\hat{\Sigma}^{-1}\delta) = n\log\left(1 + \frac{1}{n-1}(n-1)\delta^\top\hat{\Sigma}^{-1}\delta\right)$$

So this statistic depends upon $(n-1)\delta^\top\hat{\Sigma}^{-1}\delta$, which follows the $T^2(p, n-1)$ distribution (hence often referred to as the Hotelling one-sample T^2 statistic).

As $2\log\Lambda$ is a strictly increasing function of the Hotelling statistic, we reject H_0 at the level α iff

$$(n-1)\delta^\top\hat{\Sigma}^{-1}\delta > q_{1-\alpha}(T^2(p, n-1)).$$

Exercise: Establish the matrix determinant and Sherman-Morrison formulas:

- If $\Sigma \succ 0$, then $|\Sigma + uu^\top| = |\Sigma|(1 + u^\top\Sigma^{-1}u)$
- If $\Sigma \succ 0$, then $(\Sigma + uu^\top)^{-1} = \Sigma^{-1} - \frac{1}{1+u^\top\Sigma^{-1}u}\Sigma^{-1}uu^\top\Sigma^{-1}$

Two sample, $\{H_0 : \mu_X = \mu_Y\}$ vs $\{H_1 : \mu_X \neq \mu_Y\}$ with $\Sigma_X = \Sigma_Y = \Sigma$ known

Suppose that for $\Sigma_{p\times p} \succ 0$ known we have independent samples,

$$X_1, ..., X_n \overset{iid}{\sim} N(\mu_X, \Sigma) \quad \& \quad Y_1, ..., Y_m \overset{iid}{\sim} N(\mu_Y, \Sigma)$$

and we wish to discern whether they share the same mean or not.

- The two samples are not identically distributed under both hypotheses.
- Still a global likelihood w.r.t. $(\mu_X^\top, \mu_Y^\top)^\top \in \mathbb{R}^{2p}$ is obtained by multiplication.

$$\ell(\mu_X, \mu_Y) = -\frac{n+m}{2}\log|\Sigma| - \frac{n}{2}(\bar{x}-\mu_X)^\top\Sigma^{-1}(\bar{x}-\mu_X) - \frac{m}{2}(\bar{y}-\mu_Y)^\top\Sigma^{-1}(\bar{y}-\mu_Y)$$
$$-\frac{n}{2}\text{trace}\left\{\Sigma^{-1}\hat{\Sigma}_X\right\} - \frac{m}{2}\text{trace}\left\{\Sigma^{-1}\hat{\Sigma}_Y\right\}.$$

- The null $\{H_0 : \mu_X = \mu_Y\}$ corresponds to a restriction on the parameter space
- The corresponding restricted loglikelihood is

$$\ell(\mu_X) = -\frac{n+m}{2}\log|\Sigma| - \frac{n}{2}(\bar{x}-\mu_X)^\top\Sigma^{-1}(\bar{x}-\mu_X) - \frac{m}{2}(\bar{y}-\mu_X)^\top\Sigma^{-1}(\bar{y}-\mu_X)$$
$$-\frac{n}{2}\text{trace}\left\{\Sigma^{-1}\hat{\Sigma}_X\right\} - \frac{m}{2}\text{trace}\left\{\Sigma^{-1}\hat{\Sigma}_Y\right\}.$$

- The two loglikelihoods differ only through the quadratic forms in blue.

- The restricted likelihood corresponds to that of the parameter μ_X in a $N(\mu_X, \Sigma)$ model, based on an i.i.d. sample of size $n+m$. This is maximised w.r.t. μ_X at the *pooled sample mean*,

$$M = \frac{1}{n+m}\left(\sum_{i=1}^{n} X_i + \sum_{j=1}^{m} Y_j\right) = \frac{n}{n+m}\bar{X} + \frac{m}{n+m}\bar{Y}$$

- The unrestricted likelihood is maximised at $(\hat{\mu}_X, \hat{\mu}_Y) = (\bar{X}, \bar{Y})$, at which the blue term vanishes.
- Thus, the difference $2(\ell^* - \ell_0^*)$ can be seen to be equal to

$$\begin{aligned}
2(\ell^* - \ell_0^*) &= n(\bar{X} - M)^\top \Sigma^{-1}(\bar{X} - M) + m(\bar{Y} - M)^\top \Sigma^{-1}(\bar{Y} - M) \\
&= n\frac{m}{n+m}(\bar{X} - \bar{Y})^\top \Sigma^{-1} \frac{m}{n+m}(\bar{X} - \bar{Y}) \\
&\qquad + m\frac{n}{n+m}(\bar{Y} - \bar{X})^\top \Sigma^{-1} \frac{n}{n+m}(\bar{Y} - \bar{X}) \\
&= \frac{nm}{n+m}(\bar{X} - \bar{Y})^\top \left(\frac{n}{n+m}\Sigma^{-1} + \frac{m}{n+m}\Sigma^{-1}\right)(\bar{X} - \bar{Y}) \\
&= \frac{nm}{n+m}(\bar{X} - \bar{Y})^\top \Sigma^{-1}(\bar{X} - \bar{Y})
\end{aligned}$$

Under H_0, $\sqrt{\frac{nm}{n+m}}(\bar{X} - \bar{Y}) \sim N(0, \Sigma)$, thus we reject H_0 at the level α iff

$$2\log\Lambda = \frac{nm}{n+m}(\bar{X} - \bar{Y})^\top \Sigma^{-1}(\bar{X} - \bar{Y}) > q_{1-\alpha}(\chi^2_p).$$

Two sample, $\{H_0 : \mu_X = \mu_Y\}$ vs $\{H_1 : \mu_X \neq \mu_Y\}$ with $\Sigma_X = \Sigma_Y = \Sigma$ unknown

Suppose that $\Sigma_{p\times p} \succ 0$ unknown, $n + m > p$, and we have independent samples,

$$X_1, ..., X_n \overset{iid}{\sim} N(\mu_X, \Sigma) \quad \& \quad Y_1, ..., Y_m \overset{iid}{\sim} N(\mu_Y, \Sigma)$$

and we wish to discern whether they share the same mean or not.

- In this setting, one has an a.s. invertible *pooled empirical covariance*,

$$\hat{\Sigma} = \frac{1}{n+m}\left(\sum_{i=1}^{n} X_i X_i^\top + \sum_{j=1}^{m} Y_j Y_j^\top\right) - MM^\top \overset{exercise}{=} \frac{n}{n+m}\hat{\Sigma}_X + \frac{m}{n+m}\hat{\Sigma}_Y$$

- So the loglikelihood of the pooled sample is

$$\ell(\mu_X, \mu_Y, \Sigma) = -\frac{n+m}{2}\log|\Sigma| - \frac{n}{2}(\bar{x} - \mu_X)^\top\Sigma^{-1}(\bar{x} - \mu_X) - \frac{m}{2}(\bar{y} - \mu_Y)^\top\Sigma^{-1}(\bar{y} - \mu_Y)$$
$$- \frac{n}{2}\text{trace}\left\{\Sigma^{-1}\hat{\Sigma}_X\right\} - \frac{m}{2}\text{trace}\left\{\Sigma^{-1}\hat{\Sigma}_Y\right\}.$$
$$= -\frac{n+m}{2}\log|\Sigma| - \frac{n}{2}(\bar{x} - \mu_X)^\top\Sigma^{-1}(\bar{x} - \mu_X) - \frac{m}{2}(\bar{y} - \mu_Y)^\top\Sigma^{-1}(\bar{y} - \mu_Y)$$
$$- \frac{n+m}{2}\text{trace}\left\{\Sigma^{-1}\underbrace{\left(\frac{n}{n+m}\hat{\Sigma}_X + \frac{m}{n+m}\hat{\Sigma}_Y\right)}_{=\hat{\Sigma}}\right\}.$$

With this form in mind, we can now easily verify that:

- In the unconstrained case, the MLEs are the pooled mean and pooled covariance.
- In the constrained case, the MLEs are the two separate sample means and the pooled covariance.
- Repeating similar calculations as when Σ was known, we arrive at:

$$2(\ell^* - \ell_0^*) = \frac{nm}{n+m}(\bar{X} - \bar{Y})^\top \hat{\Sigma}^{-1}(\bar{X} - \bar{Y}) = \underbrace{nm(\bar{X} - \bar{Y})^\top (n\hat{\Sigma}_X + m\hat{\Sigma}_Y)^{-1}(\bar{X} - \bar{Y})}_{:=Q}$$

Which (when suitably rescaled) is known as the Hotelling's two-sample T^2.

Proposition

For $m + n > p > 1$, let $\mathrm{X}_{n\times p}$ and $\mathrm{Y}_{m\times p}$ be independent data matrices from $N(\mu_X, \Sigma_X)$ and $N(\mu_Y, \Sigma_Y)$, respectively. If $\mu_X = \mu_Y$ and $\Sigma_X = \Sigma_Y$, then

$$\left(1 - \frac{2}{n+m}\right) Q \sim T^2(p, n+m-2)$$

This, under H_0,we reject H_0 at the level α iff

$$\left(1 - \frac{2}{n+m}\right) Q > q_{1-\alpha}(T^2(p, n+m-2)).$$

Proof of the proposition.

From the Gaussian Sampling theorem and independence of X and Y, we have

$$\delta = \sqrt{\frac{nm}{n+m}}(\bar{X} - \bar{Y}) \sim N_p\left(\mu_X - \mu_Y, \frac{nm\Sigma_X}{n(n+m)} + \frac{nm\Sigma_Y}{m(n+m)}\right) \overset{H_0}{\equiv} N(0, \Sigma).$$

$$n\hat{\Sigma}_X \sim W_p(\Sigma_X, n-1) \overset{H_0}{\equiv} W_p(\Sigma, n-1)$$

$$m\hat{\Sigma}_Y \sim W_p(\Sigma_Y, m-1) \overset{H_0}{\equiv} W_p(\Sigma, m-1)$$

where $\{H_0 : \mu_X = \mu_Y \,\&\, \Sigma_X = \Sigma_X = \Sigma\}$. By independence, and the additivity property of the Wishart, we thus have under H_0 that

$$n\hat{\Sigma}_X + m\hat{\Sigma}_Y \sim W_p(\Sigma, n-1+m-1) \equiv W_p(\Sigma, n+m-2).$$

Moreover, the Gaussian sampling theorem states that $\hat{\Sigma}_X \perp\!\!\!\perp \bar{X}$ and $\hat{\Sigma}_Y \perp\!\!\!\perp \bar{Y}$. Moreover, as X $\perp\!\!\!\perp$ Y so $(\bar{X} - \bar{Y}) \perp\!\!\!\perp (n\hat{\Sigma}_X + m\hat{\Sigma}_Y)$.

Thus, by the "Hotelling Lemma" (slide 120)

$$\frac{n+m-2}{n+m} \underbrace{nm(\bar{X} - \bar{Y})^\top (n\hat{\Sigma}_X + m\hat{\Sigma}_Y)^{-1}(\bar{X} - \bar{Y})}_{Q} \sim T^2(p, n+m-2)$$

□

For $n > 2$ and $\Sigma \succ 0$, it suffices to consider the setting

- $\binom{X_1}{Y_1}, ..., \binom{X_n}{Y_n} \overset{iid}{\sim} N\left(\binom{\mu_X}{\mu_Y}, \begin{pmatrix} \sigma_X^2 & \sigma_X\sigma_Y\rho \\ \sigma_Y\sigma_X\rho & \sigma_Y^2 \end{pmatrix}\right)$
- Where $\text{var}\{X_1\} = \sigma_X^2$, $\text{var}\{Y_1\} = \sigma_Y^2$, $|\rho| < 1$
- $\{H_0 : \rho = 0\}$ vs $\{H_1 : \rho \neq 0\}$

The unrestricted loglikelihood occurs at the unrestricted MLE,

$$\ell^* = \ell(\bar{X}, \hat{\Sigma}) = -\frac{1}{2} n \log\det(2\pi\hat{\Sigma}) - \frac{n}{2}\text{tr}\left(\hat{\Sigma}^{-1}\hat{\Sigma}\right) = -\frac{n}{2}\log[2\pi] - \frac{n}{2}\log(\hat{\sigma}_X^2\hat{\sigma}_Y^2 - \hat{\rho}^2\hat{\sigma}_X^2\hat{\sigma}_Y^2) - n.$$

The loglikelihood is always maximised w.r.t. the mean at the sample mean. So for the restricted log-likelihood (under H_0) it suffices to consider the function

$$\ell(\bar{X}, \sigma_X^2, \sigma_Y^2) = -\frac{n}{2}\log[2\pi\sigma_X^2\sigma_Y^2] - \frac{n}{2}\text{tr}\left(\text{diag}\{\sigma_X^{-2}, \sigma_Y^{-2}\}\hat{\Sigma}\right)$$

$$= -\frac{n}{2}\log[2\pi] - \frac{n}{2}\log[\sigma_X^2] - \frac{n}{2}\log[\sigma_Y^2] - \frac{n}{2}\frac{\hat{\sigma}_X^2}{\sigma_X^2} - \frac{n}{2}\frac{\hat{\sigma}_Y^2}{\sigma_Y^2}.$$

with unique maximum[6] at $\sigma_X^2 = \hat{\sigma}_X^2$ and $\sigma_Y^2 = \hat{\sigma}_Y^2$, equalling

$$\ell_0^* = -\frac{n}{2}\log[2\pi] - \frac{n}{2}\log[\hat{\sigma}_X^2] - \frac{n}{2}\log[\hat{\sigma}_Y^2] - n$$

[6] recall that 1 is the unique maximum of $\log x - x$ over $x > 0$

It follows that the likelihood ratio

$$2\log\Lambda = 2(\ell^* - \ell_0^*) = -\frac{n}{2}\log(1-\hat{\rho}^2)$$

is a monotone function of the squared sample correlation $\hat{\rho}^2$. This is in turn a monotone function of

$$\kappa = \frac{\hat{\rho}^2}{1-\hat{\rho}^2}$$

Thus we reject when κ is large. In fact, $(n-2)\kappa \overset{H_0}{\sim} T^2(1, n-2)$, so we reject H_0 at level α iff

$$(n-2)\frac{\hat{\rho}^2}{1-\hat{\rho}^2} > q_{1-\alpha}(T^2(1, n-2)).$$

Theorem (Empirical Correlation Under Independence)

In the context of slide 158, we have

$$(n-2)\frac{\hat{\rho}^2}{1-\hat{\rho}^2} \overset{H_0}{\sim} T^2(1, n-2).$$

Proof.

Let $X = (X_1, ..., X_n)^\top \sim N(\mu_X \mathbf{1}_n, \sigma_X^2 \mathsf{I}_n)$, $Y = (Y_1, ..., Y_n)^\top \sim N(\mu_Y \mathbf{1}_n, \sigma_Y^2 \mathsf{I}_n)$, and $\mathsf{H}_n = \mathsf{I}_{n\times n} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$ the centring matrix (projecting onto $\mathbf{span}^\perp\{\mathbf{1}_n\}$),

$$\mathsf{H}_n = \mathsf{U}\Omega\mathsf{U}^\top, \quad \Omega = \mathrm{diag}_{n\times n}\{1, ..., 1, 0\} \ \& \ \mathsf{U}^\top\mathsf{U} = \mathsf{I}_{n\times n}.$$

Then $\Omega\mathsf{U}^\top X \sim N(0, \sigma_X^2\Omega)$, $\Omega\mathsf{U}^\top Y \sim N(0, \sigma_Y^2\Omega)$ and obviousy $\Omega^2 = \Omega$, so

$$\hat{\rho} = \frac{X^\top \mathsf{H}_n Y}{\sqrt{X^\top \mathsf{H}_n X Y^\top \mathsf{H}_n Y}} = \frac{X^\top \mathsf{U}\Omega\Omega\mathsf{U}^\top Y}{\sqrt{X^\top \mathsf{U}\Omega\Omega\mathsf{U}^\top X Y^\top \mathsf{U}\Omega\Omega\mathsf{U}^\top Y}} \overset{d}{=} \frac{W}{\|W\|}^\top \frac{V}{\|V\|}$$

where $W, V \overset{iid}{\sim} N(0, \mathsf{I}_{(n-1)\times(n-1)})$ (independence comes from H_0). Consequently we have the following collection of facts:

- $\frac{W}{\|W\|}, \frac{V}{\|V\|} \overset{iid}{\sim} \mathrm{Unif}$(on the surface of the unit sphere in $\mathbb{R}^{n-1}$).
- $\frac{V}{\|V\|} = \mathsf{V}e_1$ where $\mathsf{V} = \left(\frac{V}{\|V\|}, V_2, \ldots, V_{n-1}\right)$ is a random orthogonal matrix obtained by randomly extending $\frac{V}{\|V\|}$ to an orthonormal basis.
- $\frac{W}{\|W\|}^\top \frac{V}{\|V\|} = \frac{(\mathsf{V}^\top W)}{\|W\|}^\top e_1 = \frac{(\mathsf{V}^\top W)}{\|\mathsf{V}^\top W\|}^\top e_1 \overset{d}{=} \frac{W}{\|W\|}^\top e_1$ by independence

Therefore $\hat{\rho}^2 \overset{d}{=} \left(\frac{W^\top e_1}{\|W\|}\right)^2 \overset{d}{=} \frac{Z_1^2}{Z_1^2+\ldots+Z_{n-1}^2}$ for Z_i iid standard normal variables, so

$$\frac{\hat{\rho}^2}{1-\hat{\rho}^2} = \frac{Z_1^2}{Z_2^2+\ldots+Z_{n-1}^2}$$

which is the ratio of two independent χ^2 random variables. When each is renormalised by their respective degrees of freedom, we get

$$(n-2)\frac{\hat{\rho}^2}{1-\hat{\rho}^2} \sim F_{1,n-2} \equiv T^2(1,n-2).$$

(recall the definition of $F_{p,q}$ are ratio of independent χ_p^2/p by χ_q^2/q) □

Some comments:

- The ratio $\frac{\hat{\rho}^2}{1-\hat{\rho}^2}$ (and the test) can be arrived at using the regression representation (and can thus be interpreted in the same vein).
- Assuming $\Sigma \succ 0$ implies that $|\rho_{ij}| < 1$
- Thus, the derived test does not apply in the "boundary case".
- This is not an issue: if $\rho_{ij} = \pm 1$, we will see it immediately in the data (the two coordinates will realise perfectly along a line)

Partial correlation test $\{H_0 : \rho_{ij|k} = 0\}$ vs $\{H_1 : \rho_{ij|k} \neq 0\}$

Say we have a sample of size $n > 3$, $(X_i, Y_i, Z_i)^\top \overset{iid}{\sim} N(\mu, \Sigma_{3\times 3})$, with $\Sigma \succ 0$. We wish to test whether X is partially correlated with Y given Z,

$$\{H_0 : \rho_{XY|Z} = 0\} \quad \text{vs} \quad \{H_0 : \rho_{XY|Z} \neq 0\}$$

where we recall

$$\rho_{XY|Z} = -\theta_{XY}/\sqrt{\theta_{XX}\theta_{YY}}$$

Luckily, we don't have to do likelihood calculations again.

- To this aim, we will use the *regression representation (slide 85)*:

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \sim N\left(\mu, \begin{pmatrix} \Sigma_X & \Sigma_{XY} & \Sigma_{XZ} \\ \Sigma_{XY}^\top & \Sigma_Y & \Sigma_{YZ} \\ \Sigma_{XZ}^\top & \Sigma_{YZ}^\top & \Sigma_Z \end{pmatrix}\right) \iff \begin{pmatrix} \epsilon_X \\ \epsilon_Y \\ Z \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ \mu_Z \end{pmatrix}, \begin{pmatrix} \begin{pmatrix} \Theta_X & \Theta_{XY} \\ \Theta_{XY}^\top & \Theta_Y \end{pmatrix}^{-1} & 0 \\ 0 & \Sigma_Z \end{pmatrix}\right)$$

with $\epsilon_X = X - \Sigma_{XZ}\Sigma_Z^\dagger(Z - \mu_Z)$ and $\epsilon_Y = Y - \Sigma_{YZ}\Sigma_Z^\dagger(Z - \mu_Z)$

- In the 2×2 case the inverse has the explicit formula

$$\begin{pmatrix} \theta_X & \theta_{XY} \\ \theta_{XY} & \theta_Y \end{pmatrix}^{-1} = |\Theta|^{-1} \begin{pmatrix} \theta_Y & -\theta_{XY} \\ -\theta_{XY} & \theta_X \end{pmatrix}$$

- So we have reduced the problem to testing whether the correlation is zero in a 2×2 Gaussian setting!

- Therefore, the LRT will reject for large values of $\frac{\hat{\rho}^2_{\epsilon_X,\epsilon_Y}}{1-\hat{\rho}^2_{\epsilon_X\epsilon_Y}}$ for $\hat{\rho}_{\epsilon_X,\epsilon_Y}$ the MLE of the correlation of ϵ_X and ϵ_Y.
- We don't actually observe the sample of $(\epsilon_X, \epsilon_Y, Z)^\top$ but the induced likelihood is equivalent to that induced by the observable sample (i.e. it gives the same values at the same parameter choices)
- Thus, using equivariance, since

$$\rho_{\epsilon_X,\epsilon_Y} = \frac{-|\Theta^{-1}|\theta_{XY}}{\sqrt{|\Theta^{-1}|\theta_{XX}|\Theta^{-1}|\theta_{YY}}} = -\frac{\theta_{XY}}{\sqrt{\theta_{XX}\theta_{YY}}}$$

  we get that the MLE $\hat{\rho}_{\epsilon_X,\epsilon_Y}$ equals $-\frac{\hat{\theta}_{XY}}{\sqrt{\hat{\theta}_{XX}\hat{\theta}_{YY}}}$
- It turns out that the distribution under the null is now $T^2(1, n-3)$, after re-scaling by $(n-3)$, so we reject H_0 iff

$$(n-3)\frac{\hat{\theta}^2_{XY}/(\hat{\theta}_X\hat{\theta}_Y)}{1-\hat{\theta}^2_{XY}/(\hat{\theta}_X\hat{\theta}_Y)} > T^2(1, n-3).$$

- Exercise: Verify that $\hat{\rho}_{\epsilon_X,\epsilon_Y}$ is the sample correlation between the residuals obtained when regressing X on Z and those when regressing Y on Z. Use this to establish the null distribution, following the same steps when proving the theorem in slide 158, but using a projection matrix other than the centring matrix.

Let $X_1, \ldots, X_n$ be a random sample from $N(\mu, \Sigma_{p\times p})$ with $\Sigma \succ 0$. The restricted (under H_0) and unrestricted MLEs are, respectively:

- $\bar{X}$ and Σ_0.
- $\bar{X}$ and $\hat{\Sigma}$.

Thus,

$$\ell_0^* = \ell(\bar{X}, \Sigma_0) = -\frac{n}{2}\log|2\pi\Sigma_0| - \frac{n}{2}\mathrm{tr}\left(\Sigma_0^{-1}\hat{\Sigma}\right)$$

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

which yield

$$2\log\Lambda = 2(\ell^* - \ell_0^*) = -n\log\left(\frac{|\Sigma_0|}{|\hat{\Sigma}|}\right) + n\mathrm{tr}\left(\Sigma_0^{-1}\hat{\Sigma}\right) - np = n\log|\Sigma_0^{-1}\hat{\Sigma}| + n\mathrm{tr}\left(\Sigma_0^{-1}\hat{\Sigma}\right) - np.$$

If α and γ are the arithmetic and geometric means of the eigenvalues of $\Sigma_0^{-1}\hat{\Sigma}$, respectively, then $\mathrm{tr}(\Sigma_0^{-1}\hat{\Sigma}) = p\alpha$ and $|\Sigma_0^{-1}\hat{\Sigma}| = \gamma^p$. Thus,

$$2\log\Lambda = np(\alpha - \log\gamma - 1).$$

The exact distribution of this statistic is non-trivial to obtain, but Wilks' theorem applies so we can use the asymptotic approximation χ^2_m, where $m = p(p+1)/2$. Thus, if n is large enough, we reject H_0 at level α iff

$$2\log\Lambda > q_{1-\alpha}(\chi^2_{p(p+1)/2}).$$

Suppose that $\Sigma_X, \Sigma_Y \succ 0$ unknown and that we have two independent samples,

$$X_1, ..., X_n \overset{iid}{\sim} N(\mu_X, \Sigma_X) \quad \& \quad Y_1, ..., Y_m \overset{iid}{\sim} N(\mu_Y, \Sigma_Y), \qquad n, m > p$$

and wish to discern whether they share the same covariance or not.

- The loglikelihood of the pooled sample is

$$\ell(\mu_X, \mu_Y, \Sigma_X, \Sigma_Y) = -\frac{n}{2}\log|\Sigma_X| - \frac{m}{2}\log|\Sigma_Y| - \frac{n}{2}(\bar{x} - \mu_X)^\top \Sigma_X^{-1}(\bar{x} - \mu_X)$$
$$-\frac{m}{2}(\bar{y} - \mu_Y)^\top \Sigma^{-1}(\bar{y} - \mu_Y) - \frac{n}{2}\text{trace}\left\{\Sigma_X^{-1}\hat{\Sigma}_X\right\} - \frac{m}{2}\text{trace}\left\{\Sigma_Y^{-1}\hat{\Sigma}_Y\right\}.$$

- In the unconstrained case, this seprates into the sum of the two separate likelihoods corresponding to each sample

$$\ell(\mu_X, \mu_Y, \Sigma_X, \Sigma_Y) = \ell(\mu_X, \Sigma_X) + \ell(\mu_Y, \Sigma_Y)$$

which is maximised at the separate MLEs $(\bar{X}, \bar{Y}, \hat{\Sigma}_X, \hat{\Sigma}_Y)$ with maximum

$$\ell^* = -\frac{n}{2}\log|\hat{\Sigma}_X| - \frac{m}{2}\log|\hat{\Sigma}_Y| - \frac{(n+m)p}{2}.$$

- In the constrained case, the MLEs for the means are $(\bar{X}, \bar{Y})$ regardless of the choice of covariance.
- Plugging these in, the loglikelihood for the common covariance Σ becomes

$$\ell(\bar{X}, \bar{Y}, \Sigma) = -\frac{(n+m)}{2}\log|\Sigma| - \frac{n}{2}\text{trace}\left\{\Sigma^{-1}\hat{\Sigma}_X\right\} - \frac{m}{2}\text{trace}\left\{\Sigma^{-1}\hat{\Sigma}_Y\right\}$$
$$= -\frac{(n+m)}{2}\log|\Sigma| - \frac{(n+m)}{2}\text{trace}\left\{\Sigma^{-1}\hat{\Sigma}\right\}$$

  where $\hat{\Sigma} = \frac{n}{n+m}\hat{\Sigma}_X + \frac{m}{n+m}\hat{\Sigma}_Y$ is the weighted average of the two sample covariances (notice this is no longer equal to the pooled covariance because the two means are possibly different).
- Therefore $\hat{\Sigma}$ is the restricted MLE yielding $\ell_0^* = -\frac{(n+m)}{2}\log|\hat{\Sigma}| - \frac{(n+m)p}{2}$
- We arrive at

$$2(\ell^* - \ell_0^*) = -n\log|\hat{\Sigma}_X| - m\log|\hat{\Sigma}_Y| + (n+m)\log|\hat{\Sigma}|$$

- The exact distribution of this statistic is non-trivial to obtain, but Wilks' theorem applies so we can use the asymptotic approximation χ^2_m, where

$$m = \{\#\text{parameters} - \#\text{free parameters under } H_0\} = 2p + p(p+1) - 2p - \frac{p(p+1)}{2}$$

  Thus, if n is large enough, we reject H_0 at level α iff

$$2\log\Lambda > q_{1-\alpha}(\chi^2_{p(p+1)/2}).$$

Let $X_1, \ldots, X_n$ be a random sample from $N(\mu, \Sigma_{p\times p})$ with $\Sigma \succ 0$ and $n > p$. Consider the hypothesis pair,

$$\begin{cases} H_0: & \Sigma = \lambda \mathsf{I} \text{ for some } \lambda > 0, \\ H_1: & \Sigma \neq \lambda \mathsf{I} \text{ for all } \lambda > 0. \end{cases}$$

Both hypotheses are now *composite*. The LRT rejects H_0 for large values of

$$\left(\gamma(\hat{\Sigma})/\alpha(\hat{\Sigma})\right)^n$$

$\alpha(\hat{\Sigma})$ and $\gamma(\hat{\Sigma})$ are the arithmetic and geometric means of the eigenvalues of $\hat{\Sigma}$.

Exercise: verify this.

- As in the case $\{H_0 : \Sigma = \Sigma_0\}$ vs $\{H_1 : \Sigma \neq \Sigma_0\}$, the exact null sampling distribution of the LRT is not available in closed form.
- However Wilks' theorem does not apply because the null is composite.
- The asymptotic distribution *can* be obtained by different means, but is too convoluted to state.

- Notice that the tests related to the covariances globally generally depend only on the eigenvalues of the empirical covariance(s) and (when applicable) the null covariance.
- The Wilks χ^2 approximation will be valid only for simple null hypotheses, but not for composite hypotheses like sphericity.
- The LRT test statistics make sense more generally, when well defined, regardless of Gaussianity. In these cases, we can resort to asymptotic approximations e.g. Wilks (when applicable or by direct use of limit theorems).

Dimension Reduction

Dimension reduction is a means to introduce parsimony by way of projections or low-rank techniques.

The key principle of dimension reduction is, roughly speaking that

> *most of the "statistical action" is happening in some latent hyperplane of dimension far lower than the dimension p of the ambient space $\mathbb{R}^p$.*

The name of the game is looking for *good* linear functionals (projections) which:

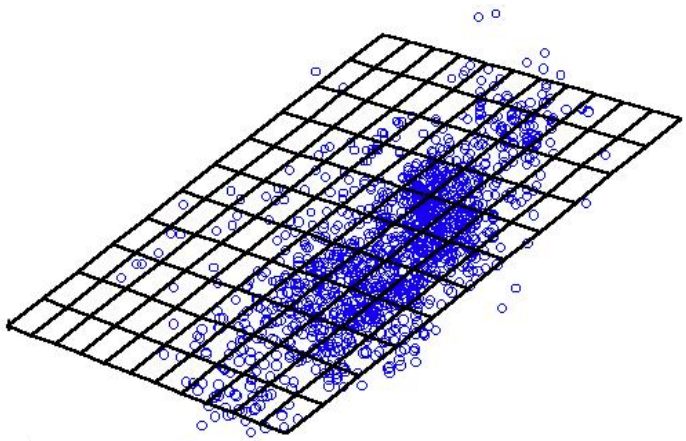
- capture most of the action, when all variables are treated equally
- distill most of the dependence, when variables are treated as input/output

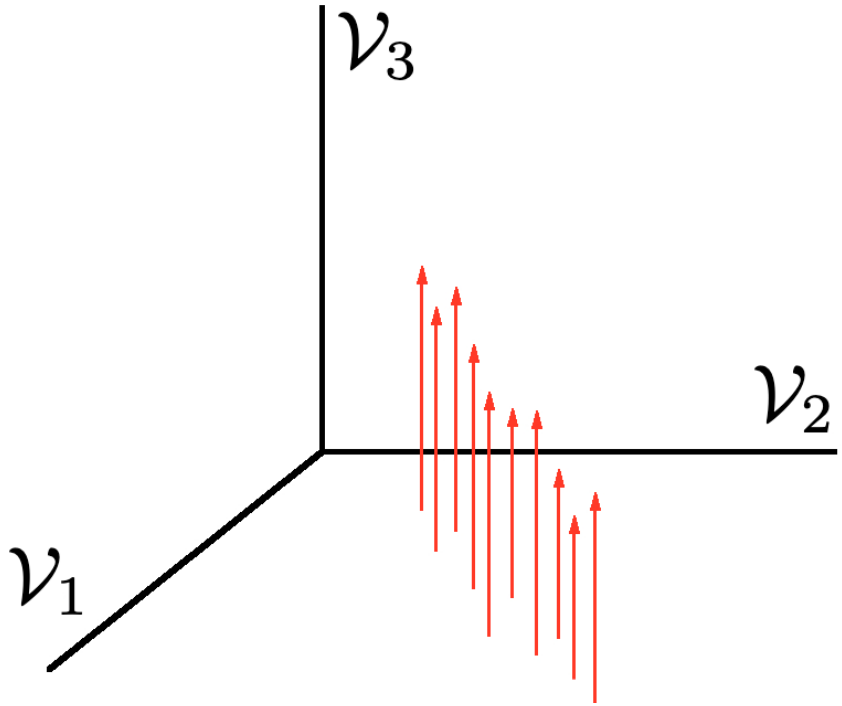
To this effect, we will see two types of analysis:

- Principal Component Analysis
- Canonical Correlation Analysis

Here are two caricatures to keep in mind:

- The "thin scatterplot" (for PCA)
- The "picket fence" (for CCA)





- Let X be a random vector in $\mathbb{R}^p$ with covariance matrix Σ.
- We seek $v_1 \in \mathbb{S}^{d-1}$ such that X's projection onto v_1 has maximal variance.
- And $j > 1$, we seek direction $v_j \in \text{span}^{\perp}\{v_1, \ldots, v_{j-1}\}$ such Y's projection onto v_j has maximal variance.

Solution: maximise $\text{var}(v_1^\top X) = v_1^\top \Sigma v_1$ over $\|v_1\| = 1$

$$v_1^\top \Sigma v_1 = v_1^\top \mathsf{U}\Lambda\mathsf{U}^\top v_1 = \|\Lambda^{1/2}\mathsf{U}^\top v_1\|^2 = \sum_{i=1}^{d} \lambda_i (u_i^\top v_1)^2 \qquad \text{[change of basis]}$$

Now $\sum_{i=1}^{d} (u_i^\top v_1)^2 = \|v_1\|^2 = 1$ so we have a convex combination of the $\{\lambda_j\}_{j=1}^{d}$,

$$\sum_{i=1}^{d} p_i \lambda_i, \qquad \sum_i p_i = 1, \quad p_i \geq 0, \quad i = 1, \ldots, d.$$

If $\lambda_1 \geq \lambda_i \geq 0$ so clearly this sum is maximised when $p_1 = 1$ and $p_j = 0 \; \forall j \neq 1$, i.e. $v_1 = \pm u_1$.

Iteratively, we find $v_j = \pm u_j$, i.e. the eigenvectors of Σ.

- The eigenvectors of Σ are called the *principal components* (of variation)
- The ratio $\lambda_j / \mathrm{tr}\{\Sigma\}$ gives the % of variance explained by the jth component.
- The actual projection $\langle u_1, X \rangle = u_1^\top X$ is called the *score* of X.
- Scores along different components are *uncorrelated*:

$$\mathrm{cov}\{u_i^\top X, u_j^\top X\} = u_i^\top \Sigma u_j = \lambda_j \mathbf{1}\{i = j\}.$$

- When $\sum_{j=1}^{k} \lambda_j / \mathrm{tr}\{\Sigma\}$ for $k \ll p$, PCA is useful for dimension reduction[7]
- PCA is always valid from a mathematical standpoint, but is most interesting from a statistical standpoint when
  - It helps reduce dimension considerably, and/or
  - When the principal components have a good interpretation as new variables.

[7] see next slide

Theorem (Optimal Linear Dimension Reduction Theorem)

Let X be a mean-zero random variable in $\mathbb{R}^p$ with $p \times p$ covariance Σ. Let H_k be the projection matrix onto the span of the first k eigenvectors of Σ. Then

$$\mathbb{E}\|X - \mathsf{H}_k X\|^2 \leq \mathbb{E}\|X - \mathsf{Q} X\|^2$$

for any $p \times p$ projection matrix Q with $\mathrm{rank}(\mathsf{Q}) \leq k$.

Intuitively: if you want to approximate a mean-zero random variable taking values $\mathbb{R}^p$ by a random variable that ranges over a subspace of dimension at most $k < p$, the optimal choice is the projection of the random variable onto the space spanned by its first k principal components.

"Optimal" is with respect to the mean squared error.

For the proof, recall that:

Q is a rank k projection if and only if $\mathsf{Q} = \sum_{j=1}^{k} v_j v_j^\top$ for orthonormal $\{v_j\}_{j=1}^{k}$.

Proof.

Write $\mathsf{Q} = \sum_{j=1}^{k} v_j v_j^\top$ for some orthonormal $\{v_j\}_{j=1}^{k}$. Then,

$$\begin{aligned}
\mathbb{E}\|X - \mathsf{Q}X\|^2 &= \mathbb{E}\left[X^\top(\mathsf{I}-\mathsf{Q})^\top(\mathsf{I}-\mathsf{Q})X\right] = \mathbb{E}\left[\mathrm{tr}\{(\mathsf{I}-\mathsf{Q})XX^\top(\mathsf{I}-\mathsf{Q})^\top\}\right] \\
&= \mathrm{tr}\{(\mathsf{I}-\mathsf{Q})\mathbb{E}\left[XX^\top\right](\mathsf{I}-\mathsf{Q})^\top\} = \mathrm{tr}\{(\mathsf{I}-\mathsf{Q})^\top(\mathsf{I}-\mathsf{Q})\Sigma\} \\
&= \mathrm{tr}\{(\mathsf{I}-\mathsf{Q})\Sigma\} = \mathrm{tr}\{\Sigma\} - \mathrm{tr}\{\mathsf{Q}\Sigma\} = \sum_{i=1}^{n}\lambda_i - \mathrm{tr}\left\{\sum_{j=1}^{k} v_j v_j^\top \Sigma\right\} \\
&= \sum_{i=1}^{n}\lambda_i - \sum_{j=1}^{k}\mathrm{tr}\left\{v_j v_j^\top \Sigma\right\} = \sum_{i=1}^{n}\lambda_i - \sum_{j=1}^{k} v_j^\top \Sigma v_j \\
&= \sum_{i=1}^{n}\lambda_i - \sum_{j=1}^{k}\mathrm{Var}[v_j^\top X]
\end{aligned}$$

If we can minimise this expression over all $\{v_j\}_{j=1}^{k}$ with $v_j^\top v_{j'} = \mathbf{1}\{j = j'\}$, then we're done. By PCA, this is done by choosing the top k eigenvectors of Σ. □

More generally, we can also show that:

Theorem (Eckhard-Young-Schmidt-Mirsky, Hilbert-Schmidt case)

Let $\Sigma_{p\times p} = \sum_{i=1}^{p} \lambda_i u_i u_i^\top \succeq 0$ *and* $\Sigma_k = \sum_{i=1}^{k} \lambda_i u_i u_i^\top$ *be its rank* k *spectral truncation. Then,*

$$\|\Sigma - \Sigma_k\|_{\mathbb{R}^{p\times p}} \leq \|\Sigma - \Gamma\|_{\mathbb{R}^{p\times p}}$$

for any Γ *of rank at most* k *(not necessarily non-negative definite). Here*

$$\|\mathsf{A}\|^2_{\mathbb{R}^{p\times p}} = \mathrm{tr}(\mathsf{A}^\top \mathsf{A}) = \|\mathrm{vec}(\mathsf{A})\|^2_{\mathbb{R}^{p^2}}.$$

Note that $\Sigma_k = H_k \Sigma \mathsf{H}_k$ *where* $\mathsf{H}_k = \sum_{i=1}^{k} u_i u_i^\top$ *projects onto* $\mathrm{span}\{u_1, \ldots, u_k\}$.

Shows that PCA can also be interpreted via the optimal low rank approximation of the covariance matrix. The theorem relies on Von Neumann's trace inequality

$$|\mathrm{tr}\{\mathsf{AB}\}| \leq \sum_i \sigma_i(A)\sigma_i(B)$$

(recall convention that singular values are always taken to be ≥ 0)

Proof (of the trace inequality)

By the SVD, the statement is equivalent to showing that

$$|\text{trace}\{\Lambda U \Omega V^\top\}| \leq \text{trace}\{\Lambda\Omega\}$$

for orthogonal matrices $\{U, V\}$ and (say) $p \times p$ diagonal $\{\Lambda, \Omega\}$. We express Λ and Ω as weighted averages of the projectors $P_k = \sum_{i=1}^{k} e_i e_i^\top$, with $\{e_i\}$ the canonical basis of $\mathbb{R}^p$:

$$\Lambda = (\lambda_1 - \lambda_2)P_1 + (\lambda_2 - \lambda_3)P_2 + \ldots + (\lambda_{p-1} - \lambda_p)P_{p-1} + \lambda_p P_p = \sum_{i=1}^{p} \alpha_i P_i$$

$$\Omega = (\omega_1 - \omega_2)P_1 + (\omega_2 - \omega_3)P_2 + \ldots + (\omega_{p-1} - \omega_p)P_{p-1} + \omega_p P_p = \sum_{i=1}^{p} \beta_i P_i$$

With this representation, our sought inequality becomes

$$\left|\sum_{i,j=1}^{p} \alpha_i \beta_j \text{trace}\{P_i U P_j V^\top\}\right| \leq \sum_{i,j=1}^{p} \alpha_i \beta_j \text{trace}\{P_i P_j\}.$$

This will follow by the triangle inequality if we can bound each term as

$$|\alpha_i \beta_j \text{trace}\{P_i U P_j V^\top\}| \leq \alpha_i \beta_j \text{trace}\{P_i P_j\},$$

For $i \geq j$, $P_i U P_j = (P_i u_1, ..., P_i u_j, 0, ..., 0)$ so we must show $\sum_{k=1}^{j} \langle P_i u_k, v_k \rangle \leq j$
This follows from the Cauchy-Schwarz inequality since $\|P_i u_k\| \leq \|u_k\| = 1$. □

Proof. (of the low rank approximation theorem).

Writing $\Gamma = W\Omega V^\top$ in SVD form, we open the square and use the trace inequality:

$$\begin{aligned}
\|\Sigma - \Gamma\|^2_{\mathbb{R}^{p\times p}} &= \|\Sigma\|^2_{\mathbb{R}^{p\times p}} + \|\Gamma\|^2_{\mathbb{R}^{p\times p}} - 2\mathrm{tr}(\Sigma\Gamma) \\
&\geq \sum_{j=1}^{p}\lambda_j^2 + \sum_{j=1}^{p}\omega_j^2 - 2\sum_{j=1}^{p}\underbrace{\lambda_j\omega_j}_{\geq 0} \\
&= \sum_{j=1}^{p}(\lambda_j - \omega_j)^2 \\
\implies \quad \inf_{\Gamma:\mathrm{rank}(\Gamma)\leq k} \|\Sigma - \Gamma\|^2_{\mathbb{R}^{p\times p}} &\geq \sum_{j=k+1}^{p}\lambda_j^2
\end{aligned}$$

Setting $\Gamma = \Sigma_k = H_k\Sigma H_k$ attains the lower bound on the RHS. □

- Note that positive definiteness does not play a role – using a truncated SVD gives similar result for the best low rank approximation of any matrix
- The result (with different proof) is valid for any unitarily invariant norm
- The "green inequality" is useful more generally.

If $X \sim N(\mu, \Sigma_{p\times p})$, it admits a Karhunen-Loève expansion:

$$X - \mu = \sum_{i=1}^{p} \xi_i u_i, \quad \xi_i \overset{iid}{\sim} N(0, \lambda_i)$$

where $\Sigma u_i = \lambda_i u_i$, $1 \leq i \leq p$, gives the spectrum of Σ, and $\xi_i = \langle X, u_i \rangle$.

Notice that:

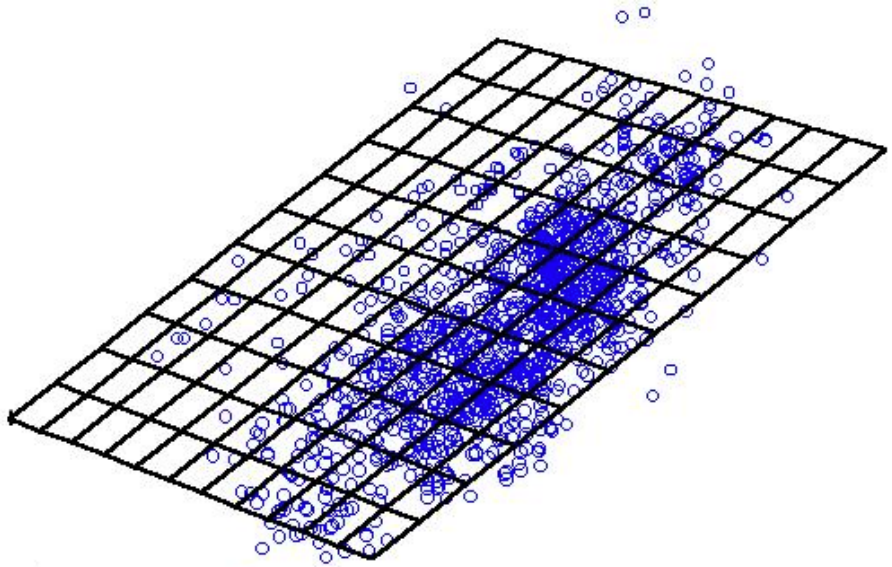
- u_i are precisely the principal components.
- ξ_i are precisely the scores.
- scores along different components are independent (not just uncorrelated).
- scores along different components are Gaussian.
- therefore, distinct component scores can be analysed completely separately

So, in the Gaussian case, PCA $\equiv$ KL, so we get independence/Gaussianity.

From the perspective of the covariance remember that when $X \sim N(\mu, \Sigma_{p\times p})$,

$$\text{supp}\{X - \mu\} = \mathcal{R}(\Sigma).$$

Thus, by our low rank approximation theorem, PCA is equivalent to successive (and nested) dimension reductions of the support of X.



(in non-Gaussian case, we find the best fitting hyperplane of given dimension to the true support)

Why not just use principal components, no matter what?

- PCA basically represents a change-of-basis
- In the new basis, everything is mathematically simpler
- But our intuition/interest is in terms of original basis,
  - Coordinates in original basis correspond to variables/features. (age, weight, height,....)
  - Coordinates in PCA basis are linear combinations of variables/features: (e.g. $-0.3\times$ age $+0.275\times$ weight $-0.59\times$height $+\ldots$)
- Ideally, we find combintaions that are interpretable and/or sparse
- But there is no a priori guarantee that this may be the case.
- Motivates ℓ_1 penalised PCA:

$$u_\tau := \arg\max_{\|u\|=1} \left\{ u^\top \Sigma u - \tau \|u\|_1 \right\}$$

If we only have an iid sample, $X_1, ..., X_n$, we can define the sample principal components and the sample scores in precisely the same way as before, but replacing the mean/covariance (μ, Σ) with their sample versions $(\bar{X}, \hat{\Sigma})$.

- Caution: the observed (realised) sample scores have zero empirical correlation ...but...
  they are correlated as random variables, since they are based on empirical principal components (which are approximations to the true ones).
- In similar vein: in the Gaussian case, the sample scores will not be independent
- In the Gaussian case, MLE equivariance immediately establishes that:

Proposition

Let $X_1, ..., X_n \sim N(\mu, \Sigma_{p\times p})$, where Σ has spectrum $\{(\lambda_i, u_i)\}_{i=1}^p$, and assume that the MLE of (μ, Σ) exists (in which case it equals $(\bar{X}, \hat{\Sigma})$). Then, provided $\lambda_1 > ... > \lambda_p$, the MLE of $\{(\lambda_i, u_i)\}$ is given by the spectrum $\{(\hat{\lambda}_i, \hat{u}_i,)\}$ of $\hat{\Sigma}$.

- Strictly speaking, eigenvectors are unique up to sign, so we rather estimate the "eigenprojections" $u_i u_i^\top$ by their sample version $\hat{u}_i \hat{u}_i^\top$

In what sense does the optimal dimension reduction property of PCA hold at the sample level?

Corollary

Let $X_1, ..., X_n$ be iid random vectors in $\mathbb{R}^p$. The best approximating k-hyperplane to the points $\{X_1, ..., X_n\}$ is given by $\bar{X} + \mathcal{R}(\hat{\Sigma}_k)$, where $\hat{\Sigma}_k = \sum_{i=1}^k \hat{\lambda}_i \hat{u}_i \hat{u}_i^\top$ is the rank-k spectral truncation of $\hat{\Sigma}$. Equivalently, defining $\hat{\mathsf{H}}_k = \sum_{i=1}^k \hat{u}_i \hat{u}_i^\top$,

$$\sum_{j=1}^n \|(X_j - \bar{X}) - \hat{\mathsf{H}}_k(X_j - \bar{X})\|^2 \leq \sum_{j=1}^n \|X_j - v - \mathsf{Q}(X_j - v)\|^2$$

for any $v \in \mathbb{R}^p$ and $n \times n$ projection operator Q of rank at most k.

Exercise: prove the corollary. Hint: notice that randomness doesn't play a role, and do so for deterministic vectors. You can define a new random variable for which the (rescaled sums) correspond to expectations...

Viewing the spectrum $\{(\hat{\lambda}_i, \hat{u}_i,)\}$ of $\hat{\Sigma}$ as an estimator of the spectrum $\{(\lambda_i, u_i)\}$ of Σ, one naturally is led to the following questions:

1. (coarse) what performance guarantees (e.g. MSE) can we establish?
2. (refined) what is the (asymptotic) sampling distribution of $\{(\hat{\lambda}_i, \hat{u}_i,)\}$?

(1) is easier than (2), by way of what are known as perturbation bounds. Viewing $\hat{\Sigma}$ as a perturbation of Σ, we see how the spectrum is perturbed.

This is very easy to do at the level of eigenvalues:

Lemma (Eigenvalue Perturbation Bound)

$$\max_j |\hat{\lambda}_j - \lambda_j| \leq ||\hat{\Sigma} - \Sigma||_{\mathbb{R}^{p \times p}}$$

Exercise: check this (we've essentially already proven it!).

Eigenvectors require a little more work:

Theorem (Eigenvector Perturbation Bound)

Let $\Sigma \succ 0$ and $\hat{\Sigma} \succ 0$ have spectra (λ_j, u_j) and $(\hat{\lambda}_j, \hat{u}_j)$, respectively, both with distinct eigenvalues. Define $u_j^* = \mathrm{sign}\{\langle u_j, \hat{u}_j \rangle\} u_j$. Then,

$$\|\hat{u}_j - u_j^*\| \leq 2\sqrt{2}\alpha_j \|\hat{\Sigma} - \Sigma\|_{\mathbb{R}^{p\times p}},$$

where $\alpha_1 = (\lambda_1 - \lambda_2)^{-1}$ and $\alpha_j = \max\{(\lambda_{j-1} - \lambda_j)^{-1}, (\lambda_j - \lambda_{j+1})^{-1}\}$, $j \geq 2$.

- Distinct eigenvalues allow for individual eigendirections to be identifiable.
- But eigenvectors are unique only up to a sign change, hence the use of u_j^*

Proof

We will prove this by "wedging" the quantity $\|\Sigma\hat{u}_j - \lambda_j\hat{u}_j\|$ between the two terms in the sought inequality. Note that

$$\Sigma\hat{u}_j - \lambda_j\hat{u}_j = (\Sigma - \hat{\Sigma} + \hat{\Sigma})\hat{u}_j - (\lambda_j - \hat{\lambda}_j + \hat{\lambda}_j)\hat{u}_j = (\Sigma - \hat{\Sigma})\hat{u}_j + (\hat{\lambda}_j - \lambda_j)\hat{u}_j$$

Thus, the triangle inequality and the "green inequality" (slide 178) imply that

$$\|\Sigma\hat{u}_j - \lambda_j\hat{u}_j\| \leq \|(\Sigma - \hat{\Sigma})\hat{u}_j\| + \|(\hat{\lambda}_j - \lambda_j)\hat{u}_j\| \leq \|\Sigma - \hat{\Sigma}\|_\infty + \|\Sigma - \hat{\Sigma}\|_{\mathbb{R}^{p\times p}}$$

and since $\|\Sigma - \hat{\Sigma}\|_\infty \leq \|\Sigma - \hat{\Sigma}\|_{\mathbb{R}^{p\times p}}$ the RHS is majorised by $2\|\Sigma - \hat{\Sigma}\|_{\mathbb{R}^{p\times p}}$.

Now for all $1 \leq j \leq p$ we aim to lower bound $||\Sigma\hat{u}_j - \lambda_j \hat{u}_j||^2$ below by $(2\alpha_j^2)^{-1}||u_j^* - \hat{u}_j||^2$.

$$||\Sigma\hat{u}_j - \lambda_j \hat{u}_j||^2 = \sum_{k=1}^{p} \langle \Sigma\hat{u}_j - \lambda_j \hat{u}_j, u_k \rangle^2 = \sum_{k=1}^{p} (\langle \Sigma\hat{u}_j, u_k \rangle - \langle \lambda_j \hat{u}_j, u_k \rangle)^2$$

$$= \sum_{k=1}^{p} (\lambda_k - \lambda_j)^2 \langle \hat{u}_j, u_k \rangle^2 \geq \min_{k \neq j} (\lambda_k - \lambda_j)^2 \sum_{k \neq j} \langle \hat{u}_j, u_k \rangle^2 \geq \alpha_j^{-2} \sum_{k \neq j} \langle \hat{u}_j, u_k \rangle^2$$

Recalling that $u_j^* = \text{sign}\{\langle u_j, \hat{u}_j \rangle\} u_j$, observe that $||u_j^* - \hat{u}_j||^2$ can be written as

$$\sum_{k=1}^{p} \langle u_j^* - \hat{u}_j, u_k \rangle^2 = \{\text{sign}(\langle u_j^*, u_j \rangle) - \langle \hat{u}_j, u_j \rangle\}^2 + \sum_{k \neq j} \langle u_j^* - \hat{u}_j, u_k \rangle^2$$

$$= \{1 - |\langle \hat{u}_j, u_j \rangle|\}^2 + \sum_{k \neq j} (\langle u_j^*, u_k \rangle - \langle \hat{u}_j, u_k \rangle)^2 = \{1 - |\langle \hat{u}_j, u_j \rangle|\}^2 + \sum_{k \neq j} \langle \hat{u}_j, u_k \rangle^2$$

Since $\sum_{k=1} \langle \hat{u}_j, u_k \rangle^2 = 1$,

$$\{1 - |\langle \hat{u}_j, u_j \rangle|\}^2 = \sum_{k=1} \langle \hat{u}_j, u_k \rangle^2 - 2|\langle \hat{u}_j, u_j \rangle| + |\langle \hat{u}_j, u_j \rangle|^2$$

$$= \sum_{k \neq j} \langle \hat{u}_j, u_k \rangle^2 + 2\underbrace{\{|\langle \hat{u}_j, u_j \rangle|^2 - |\langle \hat{u}_j, u_j \rangle|\}}_{\leq 0} \leq \sum_{k \neq j} \langle \hat{u}_j, u_k \rangle^2$$

because $\langle \hat{u}_j, u_j \rangle \leq 1$. Thus $2\sum_{k \neq j} \langle \hat{u}_j, u_k \rangle^2 \geq ||u_j^* - \hat{u}_j||^2$.

Combining the inequalities in blue, and re-arranging the constant factors, we arrive at

$$4||\hat{\Sigma} - \Sigma||^2_{\mathbb{R}^{p \times p}} \geq ||\Sigma\hat{u}_j - \lambda_j \hat{u}_j||^2 \geq \alpha_j^{-2} \sum_{k \neq j} \langle \hat{u}_j, u_k \rangle^2 \geq (2\alpha_j^2)^{-1}||u_j^* - \hat{u}_j||^2 \qquad \square$$

Notice that one way to get rid of the use of u^* is to always use the convention that U (and $\hat{\mathrm{U}}$) are taken so that their diagonal elements are non-negative. This eradicates the sign ambiguity from all the eigenvectors. And will be useful in what comes next. Call this sign consistency.

Now we move on to (2) from our earlier list: distributional results on $\{(\hat{\lambda}_j, \hat{u}_j)\}$

- Exact sampling distribution is unwieldy, even in Gaussian case. (unless we have isotropy)
- Asymptotic distribution ($n \to \infty$, p fixed) easier to access (and arguably more useful/informative)

We will develop the asymptotic law of the empirical eigenvalues, and that of the empirical eigenvectors, and then see how they simplify when dealing with a Gaussian data matrix.

Recall that, as $n \to \infty$ with p fixed, $\sqrt{n}(\hat{\Sigma} - \Sigma) \xrightarrow{d} \mathsf{Z}$, where Z is a mean zero Gaussian random matrix. We can use this to obtain:

Theorem (CLT for Empirical Spectrum)

Let $X_1, ..., X_n$ be iid p-vectors whose covariance Σ has spectrum $\{(\lambda_i, u_i)\}_{i=1}^p$, with $\lambda_1 > ... > \lambda_p > 0$. Let $\{(\hat{\lambda}_i, \hat{u}_i,)\}$ be the spectrum of $\hat{\Sigma}$, and assume that $\{(\hat{u}_i, u_i)\}$ are chosen sign-consistently. Then,

1. $\{\sqrt{n}(\hat{\lambda}_j - \lambda_j)\}_{1 \le j \le p} \xrightarrow{d} N(0, \Phi)$, where $\Phi_{ij} = \mathbb{E}\left[\langle \mathsf{Z}u_i, u_i\rangle \langle \mathsf{Z}u_j, u_j\rangle\right]$.
2. $\sqrt{n}(\mathsf{U}^\top \hat{\mathsf{U}} - \mathsf{I}) \xrightarrow{d} \mathsf{W}$ for $\mathsf{W} = \{\mathsf{W}_{ij}\}$ a centred Gaussian matrix, such that

$$\operatorname{cov}\{\mathsf{W}_{ii'}, \mathsf{W}_{jj'}\} = \begin{cases} 0 & \text{if } i = i' \text{ or } j = j', \\ \mathbb{E}\left[\dfrac{\langle \mathsf{Z}u_i, u_{i'}\rangle}{\lambda_{i'} - \lambda_i} \dfrac{\langle \mathsf{Z}u_j, u_{j'}\rangle}{\lambda_{j'} - \lambda_j}\right] & \text{otherwise.} \end{cases}$$

- We can easily deduce from (2) that $\mathsf{UW}_n = \sqrt{n}(\hat{\mathsf{U}} - \mathsf{U}) \xrightarrow{d} \mathsf{UW}$ which is also a centred Gaussian limit.
- In fact, the proof shows that the sequences in (1) and (2) are jointly asymptotically Gaussian, for what it's worth.

Proof.

We will leverage the CLT for $\sqrt{n}(\hat{\Sigma} - \Sigma) \xrightarrow{d} Z$ in order to obtain the sought CLT. Assuming that U and $\hat{U}$ are defined sign-consistently, define

$$Q_n = U^\top \sqrt{n}(\hat{\Sigma} - \Sigma)U = \sqrt{n}(\underbrace{U^\top \hat{\Sigma} U}_{T_n} - U^\top \Sigma U) = \sqrt{n}(T_n - \Lambda)$$

$$D_n := \sqrt{n}(\hat{\Lambda} - \Lambda) \quad \& \quad W_n = \sqrt{n}(U^\top \hat{U} - I)$$

and observe that we may write

$$\underbrace{\Lambda + \frac{Q_n}{\sqrt{n}}}_{T_n} = \underbrace{\left(I + \frac{W_n}{\sqrt{n}}\right)}_{U^\top \hat{U}} \underbrace{\left(\Lambda + \frac{D_n}{\sqrt{n}}\right)}_{\hat{\Lambda}} \underbrace{\left(I + \frac{W_n}{\sqrt{n}}\right)^\top}_{\hat{U}^\top U}$$

or equivalently,

$$Q_n \overset{*}{=} W_n \Lambda + \Lambda W_n^\top + D_n + \frac{W_n D_n + W_n \Lambda W_n^\top + D_n W_n^\top}{\sqrt{n}} + \frac{W_n D_n W_n^\top}{n}.$$

We also note the constraint that

$$U^\top \hat{U} \text{ is orthogonal} \implies \left(I + \frac{W_n}{\sqrt{n}}\right) \text{ is orthogonal} \implies W_n + W_n^\top + \frac{W_n W_n^\top}{\sqrt{n}} \overset{**}{=} 0.$$

With all these definitions/relations in place, let us start to look at asymptotics:

1. $\mathsf{Q}_n \xrightarrow{d} \mathsf{Q} = \mathsf{U}^\top \mathsf{Z} \mathsf{U}$ where Z is the (centred Gaussian) weak limit of $\sqrt{n}(\hat{\Sigma} - \Sigma)$ and so Q is centred Gaussian itself.

2. All terms in $(*)$ scaled by $1/\sqrt{n}$ or $1/n$ converge to zero in probability, by submultiplicativity of the matrix norm $\|\cdot\|_\infty$ and our perturbation bounds:

$$\|\mathsf{W}_n\|_\infty \leq \|\sqrt{n}(\mathsf{U}^\top \hat{\mathsf{U}} - \mathsf{I})\|_{\mathbb{R}^{p\times p}} = \|\sqrt{n}(\hat{\mathsf{U}} - \mathsf{U})\|_{\mathbb{R}^{p\times p}} \leq c_1 \|\sqrt{n}(\hat{\Sigma} - \Sigma)\|_{\mathbb{R}^{p\times p}} \xrightarrow{d} c_1 \xi$$

$$\|\mathsf{D}_n\|_\infty \leq \|\sqrt{n}(\hat{\Lambda} - \Lambda)\|_{\mathbb{R}^{p\times p}} \leq \sqrt{p}\, \sup_{1\leq j\leq p} |\hat{\lambda}_j - \lambda_j| \leq c_2 \|\sqrt{n}(\hat{\Sigma} - \Sigma)\|_{\mathbb{R}^{p\times p}} \xrightarrow{d} c_2 \xi$$

   where $c_1, c_2 \in (0, \infty)$ and ξ is a scalar random variable, so dividing by a negative power of n kills off the last two terms of $(*)$ in the limit.

3. So by (1) and (2) combined with $(*)$, Slutsky's theorem implies that

$$\mathsf{W}_n \Lambda + \Lambda \mathsf{W}_n^\top + \mathsf{D}_n \xrightarrow{d} \mathsf{Q}$$

4. Additionally, (2) combined with $(**)$ implies $\mathsf{W}_n + \mathsf{W}_n^\top \xrightarrow{d} 0$, which means that the diagonal of W_n vanishes asymptotically, and consequently so does the diagonal of $\mathsf{W}_n \Lambda + \Lambda \mathsf{W}_n^\top$, seeing as Λ is a diagonal matrix.

5. On the other hand, D_n is – by definition– diagonal for all $n \geq 1$.

Consequently, letting $\mathcal{G} : \mathbb{R}^{p\times p} \to \mathbb{R}^{p\times p}$ be the projection onto diagonal matrices,

$$\begin{aligned} \mathsf{D}_n - \mathcal{G}\mathsf{Q}_n &= \mathcal{G}\mathsf{D}_n - \mathcal{G}\mathsf{Q}_n \\ &= \mathcal{G}\mathsf{D}_n + \mathcal{G}(\mathsf{W}_n\Lambda + \Lambda\mathsf{W}_n^\top) - \mathcal{G}\mathsf{Q}_n - \mathcal{G}(\mathsf{W}_n\Lambda + \Lambda\mathsf{W}_n^\top) \\ &= \mathcal{G}(\mathsf{W}_n\Lambda + \Lambda\mathsf{W}_n^\top + \mathsf{D}_n - \mathsf{Q}_n) - \mathcal{G}(\mathsf{W}_n\Lambda + \Lambda\mathsf{W}_n^\top) \\ &\xrightarrow{d} 0. \end{aligned}$$

This proves that $\mathsf{D}_n = \sqrt{n}(\hat{\Lambda} - \Lambda) \xrightarrow{d} \mathcal{G}\mathsf{Q} = \lim_{n\to\infty}\{\mathrm{law}(\mathcal{G}\mathsf{Q}_n)\}$, and so $\{\sqrt{n}(\hat{\lambda}_j - \lambda_j)\}_{1\le j\le p}$ has a centred Gaussian limit in distribution.

As for the limiting covariance of $\{\sqrt{n}(\hat{\lambda}_j - \lambda_j)\}_{1\le j\le p}$, this is simply the covariance of the diagonal elements of $\mathsf{Q} = \mathsf{U}^\top\mathsf{Z}\mathsf{U}$ (which coincide with the diagonal elements of $\mathcal{G}\mathsf{Q}$).

Noting that the latter is

$$\mathrm{cov}\{e_i^\top\mathsf{Q}e_i, e_j^\top\mathsf{Q}e_j\} = \mathrm{cov}\{e_i^\top\mathsf{U}^\top\mathsf{Z}\mathsf{U}e_i, e_j^\top\mathsf{U}^\top\mathsf{Z}\mathsf{U}e_j\} = \mathrm{cov}\{u_i^\top\mathsf{Z}u_i, u_j^\top\mathsf{Z}u_j\}$$

and since $\mathbb{E}\{\mathsf{Z}\} = 0$ the latter is $\mathbb{E}\left[\langle\mathsf{Z}u_i, u_i\rangle\langle\mathsf{Z}u_j, u_j\rangle\right]$, as claimed.

This settles the eigenvalues, and now we turn our attention to the eigenvectors.

Letting $\mathcal{G}^{\perp}$ be the projection onto matrices with zeros on the diagonal,

$$\mathsf{W}_n\Lambda - \Lambda\mathsf{W}_n - \frac{\mathsf{W}_n\mathsf{W}_n^{\top}}{\sqrt{n}} \overset{(**)}{\equiv} \mathsf{W}_n\Lambda + \Lambda\mathsf{W}_n^{\top} = \mathcal{G}^{\perp}(\mathsf{W}_n\Lambda + \Lambda\mathsf{W}_n^{\top} + \mathsf{D}_n) \xrightarrow{d} \mathcal{G}^{\perp}\mathsf{Q}$$

is asymptotically mean zero Gaussian. But $\frac{\mathsf{W}_n\mathsf{W}_n^{\top}}{\sqrt{n}} \xrightarrow{p} 0$ and we notice that the elements of $\mathsf{W}_n\Lambda - \Lambda\mathsf{W}_n$ are simply $w_{ij}(n)\lambda_j - \lambda_i w_{ij}(n) = (\lambda_j - \lambda_i)w_{ij}(n)$, so

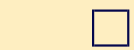
$$(\mathsf{W}_n)_{ij} = \tfrac{1}{\lambda_j - \lambda_i}(\mathsf{W}_n\Lambda - \Lambda\mathsf{W}_n)_{ij}\,.$$

Hence W_n itself has a centred Gaussian limit W, by Slutsky, with

$$\mathsf{W}_{ij} = \frac{\mathbf{1}\{i \neq j\}}{\lambda_j - \lambda_i}\mathsf{Q}_{ij}\,.$$

And, we can get the covariance between pairs of entries W by suitably rescaling the covariance of the corresponding pair of entries of $\mathsf{Q} = \mathsf{U}^{\top}\mathsf{Z}\mathsf{U}$:

$$\operatorname{cov}\{\mathsf{W}_{ii'}, \mathsf{W}_{jj'}\} = \begin{cases} 0 & \text{if } i = i' \text{ or } j = j', \\ \mathbb{E}\left[\frac{\langle \mathsf{Z}u_i, u_{i'}\rangle}{\lambda_{i'} - \lambda_i}\frac{\langle \mathsf{Z}u_j, u_{j'}\rangle}{\lambda_{j'} - \lambda_j}\right] & \text{otherwise.} \end{cases}$$



Corollary (Asymptotic Law of Wishart Spectrum)

Let $X_1, ..., X_n \sim N(\mu, \Sigma)$ be iid p-vectors whose covariance Σ has spectrum $\{(\lambda_i, u_i)\}_{i=1}^p$, with $\lambda_1 > ... > \lambda_p > 0$. Let $\{(\hat{\lambda}_i, \hat{u}_i,)\}$ be the spectrum of $\hat{\Sigma}$. Assume that $\{(\hat{u}_i, u_i)\}$ are chosen sign-consistently. Then,

1. $\{\sqrt{n}(\hat{\lambda}_j - \lambda_j)\}_{1 \leq j \leq p} \xrightarrow{d} N(0, \Phi)$, where $\Phi = \text{diag}\{2\lambda_1^2, ..., 2\lambda_p^2\}$.
2. $\mathsf{W}_n = \sqrt{n}(\mathsf{U}^\top \hat{\mathsf{U}} - \mathsf{I}) \xrightarrow{d} \mathsf{W}$ for W a centred Gaussian random matrix.
3. $\mathsf{W} \perp\!\!\!\perp \mathsf{D}$
4. Writing $\mathsf{W} = (W_1, \ldots, W_p)$ columnwise, we have:

$$\text{cov}\{W_i, W_i\} = \sum_{k \neq i} \frac{\lambda_i \lambda_k e_k e_k^\top}{(\lambda_i - \lambda_k)^2} \quad \& \quad \text{cov}\{W_i, W_j\} = -\frac{\lambda_i \lambda_j e_j e_i^\top}{(\lambda_i - \lambda_j)^2}, \ i < j.$$

Salient features in the Gaussian case:

- Eigenvalues asymptotically independent between them
- Eigenvectors asymptotically independent of eigenvalues.
- But eigenvectors not asymptotically independent between them (makes sense as they are orthogonal).
- Judging from the (Gaussian) asymptotic standard deviation of $\sqrt{2}\lambda_j$ we see that crossings will happen often even for well-spaced eigenvalues.
- We can easily deduce that $\Delta_n = \mathsf{U}\mathsf{W}_n = \sqrt{n}(\hat{\mathsf{U}} - \mathsf{U}) \xrightarrow{d} \mathsf{U}\mathsf{W} = \Delta$ with

$$\mathrm{cov}\{\Delta_i, \Delta_i\} = \sum_{k \neq i} \frac{\lambda_i \lambda_k u_k u_k^\top}{(\lambda_i - \lambda_k)^2} \quad \& \quad \mathrm{cov}\{\Delta_i, \Delta_j\} = -\frac{\lambda_i \lambda_j u_j u_i^\top}{(\lambda_i - \lambda_j)^2}, \ i \neq j.$$

Proof.

By our Gaussian assumption, we know that $n\hat{\Sigma} \sim W(\Sigma, n-1)$ and therefore $n\mathsf{U}^\top\hat{\Sigma}\mathsf{U} \equiv n\mathsf{T}_n \sim W(\Lambda, n-1)$. Consequently,

$$\mathsf{Q}_n = \sqrt{n}(\mathsf{T}_n - \Lambda) \xrightarrow{d} \mathsf{Q} \sim N(0, \mathsf{C}) \qquad \mathsf{C} \equiv \text{ covariance of } W(\Lambda, 1)$$

and so the entries of Q are uncorrelated (see slide 129 recalling that Λ is diagonal), and hence independent (as they are jointly Gaussian).

It follows that $\mathcal{G}\mathsf{Q}$ is independent of $\mathcal{G}^\perp\mathsf{Q}$ for any projection $\mathcal{G} : \mathbb{R}^{p\times p} \to \mathbb{R}^{p\times p}$ that "zeroes elements" and its complementary projection $\mathcal{G}^\perp$, in particular for $\mathcal{G}$ being the projection onto diagonal matrices. Recalling from our previous proof that

$$\sqrt{n}(\hat{\Lambda} - \Lambda) \xrightarrow{d} \mathcal{G}\mathsf{Q} \quad \& \quad \sqrt{n}(\mathsf{U}^\top\hat{\mathsf{U}} - \mathsf{I}) \xrightarrow{d} \mathcal{G}^\perp\mathsf{Q}$$

we establish (3), and the centred Gaussian limits claimed. The covariance in part (1) now follows by directly inspecting the corresponding entries of C from slide 129. As for the covariance in Part (4), recall from our last proof that

$$\mathsf{W}_{ij} = \frac{\mathbf{1}\{i\neq j\}}{\lambda_j - \lambda_i}\mathsf{Q}_{ij}$$

and thus scale the corresponding entries of C from slide 129 accordingly. □

We only used the Gaussian assumption to specify the asymptotic covariance.

In non-Gaussian settings, we can still specify the asymptotic covariance, but it depends on comprehensive (mixed) fourth moment structure, which is unwieldy.

However, we saw (slide 130) that dependence on fourth moments is "minimal" in ellpitical families. Indeed, we can straightforwardly deduce the extension below:

Theorem (Asymptotic Law of Elliptical Spectrum)

Let $X_1, ..., X_n$ be centred iid elliptical p-vectors whose covariance Σ has spectrum $\{(\lambda_i, u_i)\}_{i=1}^p$, with $\lambda_1 > ... > \lambda_p > 0$. Let $\{(\hat{\lambda}_i, \hat{u}_i,)\}$ be the spectrum of $\hat{\Sigma}$. Assume that $\{(\hat{u}_i, u_i)\}$ are sign-consistent. Letting κ be as in slide 130,

1. $\mathsf{D}_n = \sqrt{n}(\hat{\Lambda} - \Lambda) \xrightarrow{d} \mathsf{D}$ for D a diagonal and centred Gaussian random matrix.
2. $\mathsf{W}_n = \sqrt{n}(\mathsf{U}^\top \hat{\mathsf{U}} - \mathsf{I}) \xrightarrow{d} \mathsf{W}$ for W a centred Gaussian random matrix.
3. $\mathsf{W} \perp\!\!\!\perp \mathsf{D}$
4. Writing $\mathsf{W} = (W_1, \ldots, W_p)$ columnwise, we have:

$$\mathrm{cov}\{W_i, W_i\} = (1+\kappa)\sum_{k \neq i} \frac{\lambda_i \lambda_k e_k e_k^\top}{(\lambda_i - \lambda_k)^2} \quad \& \quad \mathrm{cov}\{W_i, W_j\} = -(1+\kappa)\frac{\lambda_i \lambda_j e_j e_i^\top}{(\lambda_i - \lambda_j)^2}, \ i < j.$$

High level summary: assuming $\lambda_1 > \lambda_2 > \ldots > \lambda_p > 0$ and sign-consistency.

No matter what,

- Sample eigenvectors and eigenvalues are jointly asymptotically Gaussian
- Sample eigenvectors of different index remain dependent, even asymptotically

In the Gaussian-case,

- Covariance structure for sample eigenvalues/vectors is tractable and depends only on second moments. This structure shows that:
  - Sample eigenvalues of different index are asymptotically mutually independent
  - Sample eigenvectors are asymptotically independent of sample eigenvalues

In the elliptical case,

- Covariance structure for sample eigenvalues/vectors is tractable, depends on second moments and the "4th-moment-parameter" κ. Structure shows that:
  - Sample eigenvalues of different index are asymptotically mutually independent
  - Sample eigenvectors are asymptotically independent of sample eigenvalues

In the non-elliptical case,

- Covariance structure for sample eigenvalues/eigenvectors is possibly intractable, depends on comprehensive mixed fourth moments.
- Evan asymptotically, sample eigenvalues may be dependent for different indices, and may be dependent with sample eigenvectors.

A priori, there is no unequivocal way to choose a truncation level k.

We can interpret k in various ways:

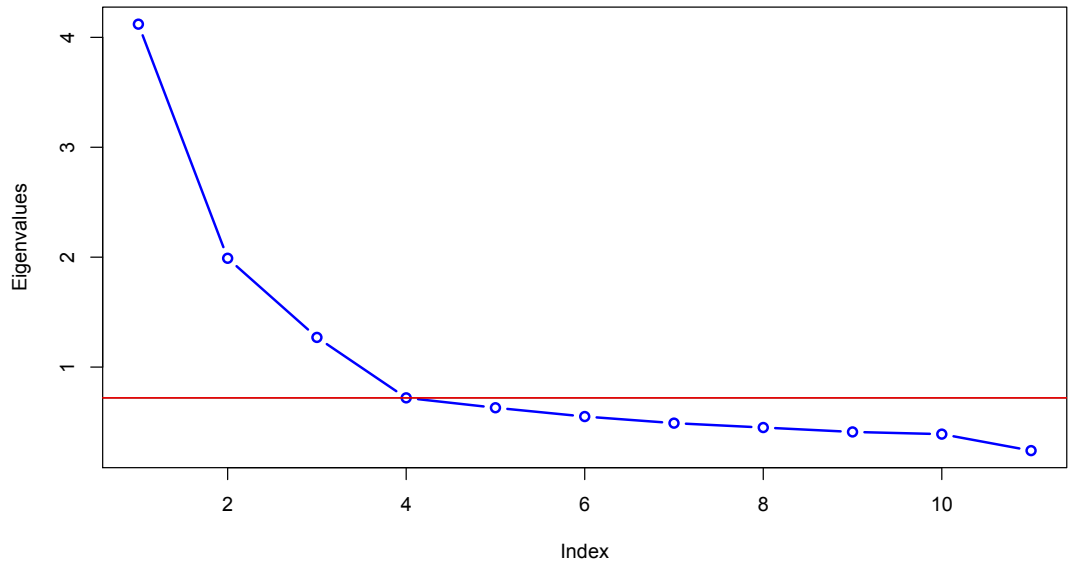
- As a tuning parameter in an approximation problem (% of variance explained)
- As a tuning parameter in an inverse problem (condition numbers, CV)
- As a model parameter to be inferred (testing/estimation)
- As a model index to be selected over (model selection)

Conversely, approaches to choosing k implicitly or explicitly represent a choice of interpretation. Sometimes different approaches give essentially same criterion. But not always. They often boil down to "eigenvalue decay" criteria.

No single approach is superior in all circumstances, and the choice of method is often guided by the specific data and problem at hand.

Combinations of methods can be employed (with careful calibration of significance levels if the testing approach is among them, to avoid data snooping bias).

No matter which method one chooses, the scree plot often shows up:



- Represents "derivative" of approximation error function.
- Leveling off suggests diminishing returns in terms of approximation.
- Often seek "elbows" if such are present.
- Rationale: past elbow, scree plot is essentially constant. No point in truncating to dimension past elbow point, you might as well not reduce at all, since all the remaining dimensions are virtually exchangeable.
- Can also add approximate error bars (CI) in Gaussian case.

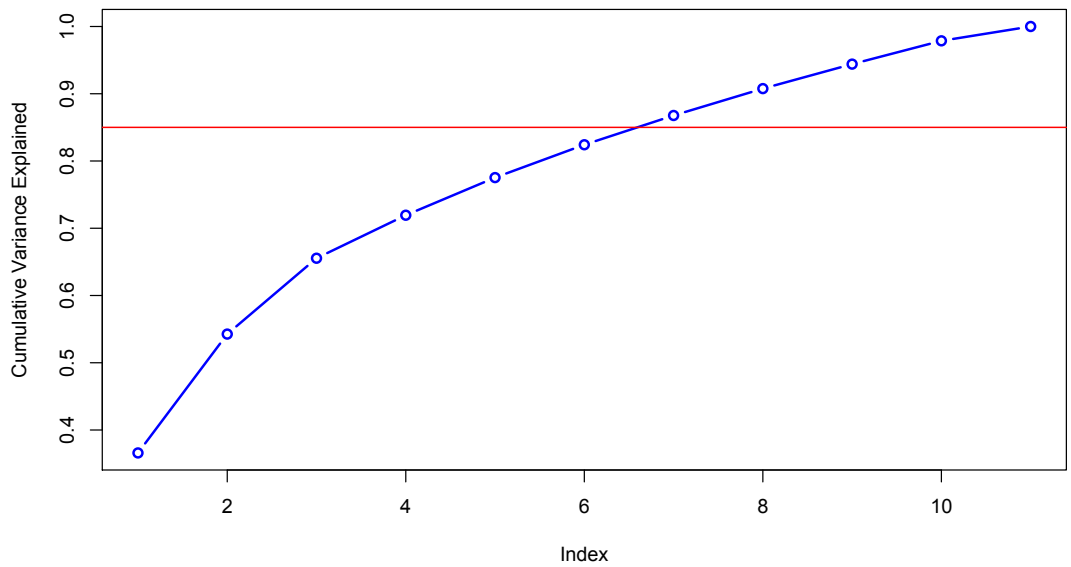
"Scree is a collection of broken rock fragments at the base of a cliff or other steep rocky mass that has accumulated through periodic rockfall" (Wikipedia)

Scree slope at the bottom of Yamnuska, Alberta, Canada (Wikipedia)

% of variance explained is simple enough in principle:

$$k = \min\{1 \leq j \leq p : (\hat{\lambda}_1 + \ldots + \hat{\lambda}_k)/\mathrm{tr}(\hat{\Sigma}) \geq 1 - \beta\}$$

- β can be chosen to some standard level, e.g. 0.15 or 0.1 (no gold standard)
- More often β is chosen to depend on the empirical eigenvalues , e.g. via simultaneous inspection of the scree plot and cumulative variance plot:



(here line is drawn at $\beta = 0.15$, corresponding to 85% variance explained)

- Doing so subconsciously corresponds to some form of penalized % of var

Often, the sample covariance is used as a device for a downstream task, usually through its (generalised) inverse, or that of its square root:

linear prediction, testing, classification

In this case, if the sample covariance is ill-conditioned,

$$\mathrm{CN} := \hat{\lambda}_1 / \hat{\lambda}_p \gg 1,$$

it can lead to wildly fluctuating outcomes even under small sampling variation.

Look at condition indices

$$\mathrm{CI}_j := \hat{\lambda}_1 / \hat{\lambda}_j, \quad j = 2, ..., p,$$

and truncate at first j where $\mathrm{CI}_j > c^*$ for some threshold c^*.

- Intuitively: you try to choose the maximal rank k truncation that still leads to a well-conditioned (according to the threshold) matrix.

(notice that this relates bijectively to the scree plot)

In the case of prediction, one can also use Cross Validation (CV).

Possible model for random vectors with "nearly rank k" covariance is

$$\{\text{lower dimensional signal}\} + \{\text{isotropic noise}\}$$

Specifically, we model X as

$$X = L + \epsilon,$$

i.e. X is a noisy measurement of the latent vector of interest L, where:

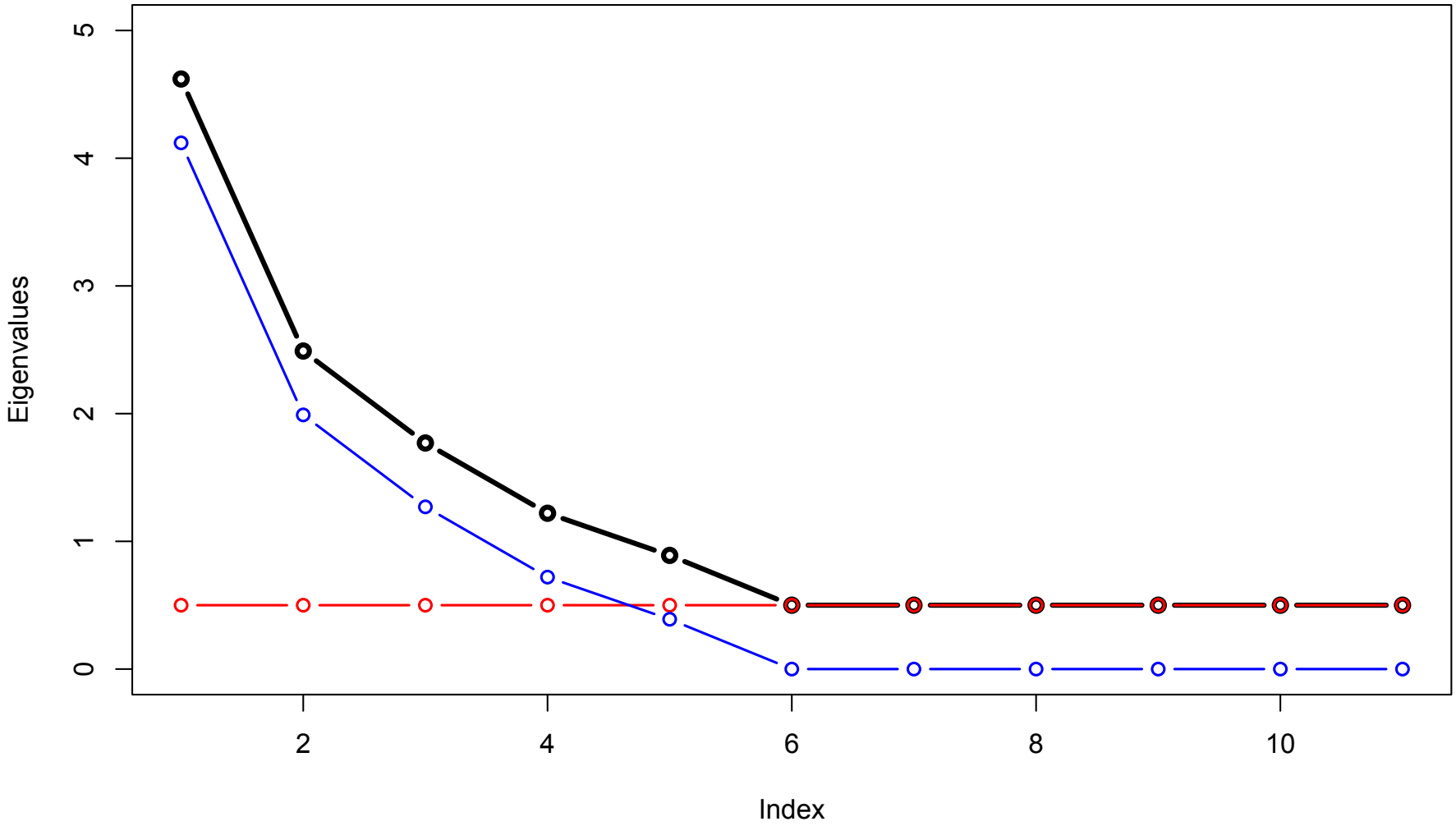
- L is a random vector in $\mathbb{R}^p$ with rank $k < p$ covariance $\Phi = \sum_{j=1}^{k} \phi_j u_i u_i^\top$.
- ϵ is a random vector in $\mathbb{R}^p$ with diagonal covariance $\theta \mathsf{I}_{p\times p}$, $\theta > 0$.
- $\text{cov}\{L, \epsilon\} = 0$ and $\Big[\{\phi_i \text{ distinct }\} \text{ OR } \{\phi_k > \theta\}\Big]$ (for identifiability)

a.k.a. "spiked covariance model". The covariance of X then becomes

$$\text{cov}\{X\} = \Phi + \theta \mathsf{I} = \sum_{i=1}^{k} (\phi_i + \theta) u_i u_i^\top + \sum_{i=k+1}^{p} \theta u_i u_i^\top$$

In this setting:

$$\text{choosing } k \iff \text{inferring as of where population scree plot becomes flat}$$



This is a population scree plot under the previous model. The sample version will not be as clear-cut!

Given a candidate $k \in \{0, ..., k_{max}\} = K$, where $k_{max} \leq p - 2$ (think why),

- we can test the last $p - k$ principal components of X for sphericity,
- equivalently, test the sphericity of $\Delta_k = \text{cov}\{\mathsf{Q}_k \mathsf{U}^\top X\}$, with Q_k being the $(p - k) \times p$ matrix obtained when deleting the first k rows of $\mathsf{I}_{p \times p}$,

$$\begin{cases} H_{k,0}: & \Delta_k = \theta \mathsf{I}_{(p-k)\times(p-k)} \text{ for some } \theta > 0, \\ H_{k,1}: & \Delta_k \neq \theta \mathsf{I}_{(p-k)\times(p-k)} \text{ for all } \theta > 0. \end{cases}$$

Whenever the hypothesised value k is chosen by scree plot inspection (data snooping) over the set K, we will need to adjust for multiple testing:

- Let p_k be the p-value corresponding to $H_{k,0}$
- Let $p_{(1)} \leq p_{(2)} \leq ...$ be the ordered p-values, from smallest to largest.
- Starting at $j = 1$ and going up,
  - If $p_{(j)} \leq \frac{\alpha}{(|K| - j + 1)}$, reject the hypothesis corresponding to $p_{(j)}$ and go to $j + 1$.
  - If $p_{(j)} > \frac{\alpha}{(|K| - j + 1)}$, "accept" the hypothesis corresponding to $p_{(j)}$, and all hypotheses corresponding to $p_{(j')}$ with $j' \geq j$, and terminate.

This is the Bonferroni-Holm adjustment, ensuring that the probability of falsely rejecting $H_K = \cup_{k \in K} H_{0,k}$ is at most α. Note that,

$$\cup_{k \in K} H_{0,k} = \{H_0 : \mathrm{rank}(\Phi) \leq k_{\max}\}$$

The least j (if any) for which $H_{j,0}$ is accepted is the *de facto* estimate of $\mathrm{rank}(\Phi)$

Some remarks on the testing perspective:

- Likelihood ratio test for $\{H_{0,j} \text{ vs } H_{1,j}\}$ similar to the "full case". Here, too, asymptotic distribution is convoluted to state, but available.
- Other test statistics are also possible, leading to approximately χ^2 sampling laws under the null.
- Criteria related to thresholds (% of variance, condition numbers) have a confirmatory (as opposed to exploratory) version via tests (that the population quantity satisfies the threshold). These are of limited interest in practice.

If we assume a specific distribution (e.g. Gaussian) we can employ model selection in the context of the low rank plus noise model:

- Then, the low rank plus noise covariance testing scheme can be seen as a sort of analysis of deviance for covariance:
  - Each such model is a restricted version of the general (unrestricted model) when $k = p$.
  - For $k_1 \leq k_2$ the corresponding models are *nested*.
  - Thus the test at step k can be seen as a likelihood ratio test for a submodel.
- More generally (and for different models) small k yield parsimonious models
- At the same time, smaller k will usually yield worse fit (lower max likelihood)
- thus can use an information criterion (AIC/BIC)
- Depending on the precise model, these will lead to threshold criteria.
- Exercise: in the low rank plus noise models, they take the form $\hat{\lambda}_k > \texttt{threshold}(n)$ which resemble a (sample-size dependent) condition number criterion.

PCA is neither invariant nor equivariant to re-scaling variables differentially:

- Changing scale (or units) in one variable changes the PCs.
- This change is not commensurate to the re-scaling (no equivariance).
- Concretely: if we have a trivariate context with heigh in m, weight in kg and age in years, we may want to switch to gr, cm, and months. But we get different results if we
  - Multiply the data by 100, 1000 and 1/10 and then perform a PCA.
  - Perform a PCA and multiply the coefficients of the three variables in the components by 100, 1000 and 1/10.
- Ideally all the variables have similar scales. Otherwise, changing to a very small scale in one variable will exaggerate its contribution to u_1.

Two often employed (but not definitive) solutions:

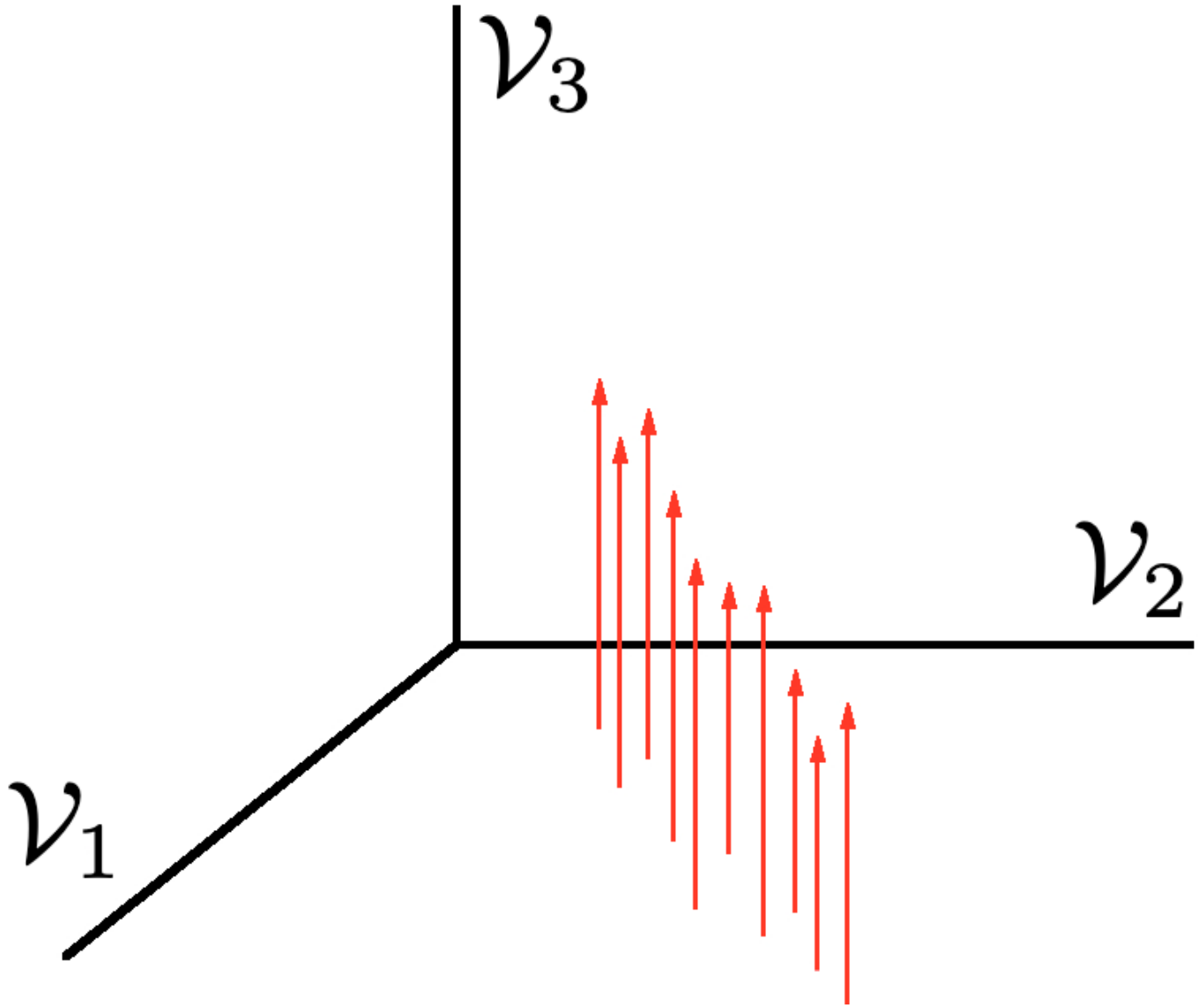
- Consider "natural" units. Hopefully the domain expert knows in precisely what scale they wish to discover dependencies. This relates to the notion of effect size: what changes are scientifically –as opposed to statistically– significant in the context of the problem.
- Standardize all variables (hence, the PCs are derived from the correlation matrix rather than the covariance matrix). This might seem the best, but it has problems of its own (scree plots, sphericity tests, testing for components all become dubious in terms of interpretation).

The context:

- Suppose that our variables can be naturally assigned into two groups:
  - "Inputs". For example, lifestyle/exercise variables.
  - "Outputs". For example, health indicators.
- Seek to understand associations between "inputs" and "outputs".
- We investigate the pairwise correlations between all input/output pairs.
- But such approach is arguably inefficient and ineffective :
  - If both groups have cardinality p there are $p(p-1)/2$ such pairs.
  - Possibly no single pair is too correlated, but the groups are as a whole

Canonical Correlation Analysis seeks to approximate/summarize the associations with relatively few statistical summaries.

- Each summary the correlation between some linear combination or input variables, and some other linear combination of output variables.
- It is in this sense that CCA can be thought of as an extension of regression (regression can be thought of as CCA with a singleton "output" group)
- Another way of thinking about it: CCA is to cross-covariance matrices what PCA is to covariance matrices ("between" vs "within" dependencies).


ν_3
ν_2
ν_1

Let X and Y be random vectors in $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively and write

$$\Sigma = \mathrm{cov}\{X, Y\} = \begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{XY}^\top & \Sigma_Y \end{pmatrix}.$$

Assume wlog that $p \leq q$. We seek $\alpha \in \mathbb{R}^p$ and $\beta \in \mathbb{R}^q$ to maximise

$$\mathrm{cov}\{\alpha^\top X, \beta^\top Y\} = \alpha^\top \Sigma_{XY} \beta.$$

Clearly, we need a constraint, or else the objective can grow without bound:

$$\mathrm{var}\{\alpha^\top X\} = \alpha^\top \Sigma_X \alpha = 1 \quad \& \quad \mathrm{var}\{\beta^\top Y\} = \beta^\top \Sigma_Y \beta = 1.$$

Such a pair (α_1, β_1) is called the first pair of canonical variables, and its covariance

$$\mathrm{cov}\{\alpha_1^\top X, \beta_1^\top Y\} = \mathrm{corr}\{\alpha_1^\top X, \beta_1^\top Y\} = \omega_1$$

is called the first canonical correlation.

The second pair of canonical variables (α_2, β_2) and the second canonical correlation is defined in similar way, but with the additional constraints:

$$\mathrm{cov}\{\alpha_1^\top X, \alpha_2^\top X\} = \alpha_1^\top \Sigma_X \alpha_2 = 0 \quad \& \quad \mathrm{cov}\{\beta_1^\top Y, \beta_2^\top Y\} = \beta_1^\top \Sigma_Y \beta_2 = 0.$$

Provided $\Sigma_X, \Sigma_Y \succ 0$, we will show that $k = \mathrm{rank}(\Sigma_{XY})$ canonical pairs exist.

Moreover, we will show that the canonical pairs are given by

$$\alpha_j = \Sigma_X^{-1/2} u_j \quad \& \quad \beta_j = \Sigma_Y^{-1/2} v_j, \qquad j \leq k,$$

where (u_j, v_j) are the singular vectors of the *canonical correlation matrix*[8]

$$\Phi = \Sigma_X^{-1/2} \Sigma_{XY} \Sigma_Y^{-1/2} = \mathsf{U}\Omega\mathsf{V}^\top,$$

and canonical correlations given by the corresponding singular values of Φ.

Exercise: Check that the canonical pairs $\{(\alpha_j, \beta_j)\}_{j=1}^k$ satisfy the sought constraints. To do this, write

$$\mathsf{A}_{p\times p} = \Sigma_X^{-1/2}\mathsf{U} \quad \& \quad \mathsf{B}_{q\times q} = \Sigma_Y^{-1/2}\mathsf{V}$$

where $\mathsf{U} = (u_1 \ldots u_p)$ and $\mathsf{V} = (v_1 \ldots v_q)$ and check that

$$\mathrm{cov}\left\{\begin{pmatrix}\mathsf{A}X\\ \mathsf{B}Y\end{pmatrix}\right\} = \begin{pmatrix}\mathsf{A} & 0\\ 0 & \mathsf{B}\end{pmatrix} \mathrm{cov}\{X, Y\} \begin{pmatrix}\mathsf{A}^\top & 0\\ 0 & \mathsf{B}^\top\end{pmatrix} = \begin{pmatrix}\mathsf{I}_{p\times p} & \Omega\\ \Omega^\top & \mathsf{I}_{q\times q}\end{pmatrix}.$$

[8] as distinct from the cross-correlation matrix!

Theorem (Canonical Correlation Analysis)

Let X and Y be random vectors in $\mathbb{R}^p$ and $\mathbb{R}^q$, with $p \leq q$, and let

$$\Sigma = \mathrm{cov}\{X, Y\} = \begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{XY}^\top & \Sigma_Y \end{pmatrix}$$

with $\Sigma_X, \Sigma_Y \succ 0$. Let $k = \mathrm{rank}(\Sigma_{XY})$ and let $\mathsf{U}\Omega\mathsf{V}^\top$ be the SVD of

$$\Phi = \Sigma_X^{-1/2} \Sigma_{XY} \Sigma_Y^{-1/2}.$$

where $\mathsf{U} = (u_1 \dots u_p)$ and $\mathsf{V} = (v_1 \dots v_q)$. Then, for $j = 1, \dots, k$,

$$\sup_{\binom{\alpha}{\beta} \in \mathcal{C}_j(X,Y)} \mathrm{cov}\{\alpha^\top X, \beta^\top Y\} = \mathrm{corr}\{(\Sigma_X^{-1/2} u_j)^\top X, (\Sigma_Y^{-1/2} v_j)^\top Y\} = \omega_j$$

where the constraint sets $\mathcal{C}_j(X, Y) \subset \mathbb{R}^{p+q}$ are defined as

$$\mathcal{C}_1(X, Y) = \left\{ \binom{\alpha}{\beta} \in \mathbb{R}^{p+q} : \mathrm{var}\{\alpha^\top X\} = \mathrm{var}\{\beta^\top Y\} = 1 \right\},$$

$$\mathcal{C}_j(X, Y) = \left\{ \binom{\alpha}{\beta} \in \mathcal{C}_1(X, Y) : \mathrm{cov}\{\alpha^\top X, \alpha_i^\top X\} = \mathrm{cov}\{\beta^\top Y, \beta_i^\top Y\} = 0 \ \forall\, i < j \right\}, \ j \geq 2.$$

Proof.

That $\alpha_j = \Sigma_X^{-1/2} u_j$ and $\beta_j = \Sigma_Y^{-1/2} v_j$ are feasible (satisfy the constraints) has already been established by our last exercise. To establish that the supremum is equal to the two quantities as stated, we proceed in two steps. First we notice that

$$\text{corr}\{(\Sigma_X^{-1/2} u_j)^\top X, (\Sigma_Y^{-1/2} v_j)^\top Y\} = u_j^\top \Sigma_X^{-1/2} \Sigma_{XY} \Sigma_Y^{-1/2} v_j = u_j^\top \mathsf{U}\Omega\mathsf{V}^\top v_j = \omega_j.$$

So the second equality is immediately true by the SVD. As for the first equality,

$$\text{cov}\{\alpha^\top X, \beta^\top Y\} = \alpha^\top \Sigma_{XY} \beta = (\Sigma_X^{1/2}\alpha)^\top \Sigma_X^{-1/2} \Sigma_{XY} \Sigma_Y^{-1/2} (\Sigma_Y^{1/2}\beta) = \gamma^\top \Phi \theta.$$

So we have the equivalences (with the analogous implications for β's and θ's)

$$\alpha^\top \Sigma_X \alpha = 1 \iff \gamma^\top \gamma = 1 \text{ and } \text{cov}\{\alpha^\top X, \alpha_i^\top X\} = 0 \iff (\underbrace{\Sigma_X^{1/2}\alpha_i}_{:=\gamma_i})^\top \gamma = 0.$$

Hence, as (α, β) range over the constraint sets $\mathcal{C}_j(X, Y)$, (γ, θ) range over

$$\mathcal{C}_1' = \{(\alpha^\top, \beta^\top)^\top \in \mathbb{R}^{p+q} : \gamma^\top \gamma = \theta^\top \theta = 1\},$$
$$\mathcal{C}_j' = \left\{(\alpha^\top, \beta^\top)^\top \in \mathcal{C}_1' : \gamma^\top \gamma_i = \theta^\top \theta_i = 0 \ \forall\, i < j\right\}, \quad j = 2, \ldots, k.$$

The result will now follow from Cauchy-Schwarz and the SVD of Φ.

By the Cauchy-Schwarz inequality, we have

$$|\gamma^\top \Phi\theta| \leq \|\gamma\|\|\Phi\theta\| = \|\gamma\|\sqrt{\theta^\top \Phi^\top \Phi\theta}.$$

With this in mind, we now note:

- The upper bound is attained when γ is collinear with $\Phi\theta$. So to maximise the expression, we seek feasible γ and θ such that γ is collinear with $\Phi\theta$.
- $\Phi^\top\Phi \succeq 0$. So, by PCA, the second term of the upper bound is maximised over the constraint sets $\mathcal{C}_1'$ and $\mathcal{C}_j'$ at the first k eigenvectors of $\Phi^\top\Phi$, respectively. Equivalently, over the first k *right singular vectors* $\{v_j\}_{j\leq k}$ of Φ.
- Once this choice is made, we note the choice of unit vectors γ in the constraint sets do not affect the value of the objective, so long as they are collinear to the corresponding $\Phi v_j \overset{SVD}{=} \omega_j u_j$. This forces us to choose the γ from the constraint sets $\mathcal{C}_1'$ and $\mathcal{C}_j'$ as the *left singular vectors* $\{u_j\}_{j\leq k}$ of Φ.

Backtransforming from γ's and θ's to α's and β's now completes the proof.

Exercise: show that when $p = 1$, the only non-trivial canonical correlation vector is the (standardised) leasts squares estimator of the regression coefficient vector.

Contrary to PCA, the nature of CCA constraints make it equivariant under standardisation – and invertible affine transformation more generally:

Theorem (Invariance/Equivariance of Canonical Correlations/Pairs)

In the same context as the previous theorem, let

$$f(x) = \mathsf{F}x + \phi \quad \textit{and} \quad g(y) = \mathsf{G}y + \gamma$$

be invertible affine transformations on $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively. Then,

1. *the canonical correlations of $\{f(X), g(Y)\}$ are the same as those of $\{X, Y\}$.*
2. *the canonical pairs of $\{f(X), g(Y)\}$ are the the inversely transformed canonical pairs of $\{X, Y\}$, via f^{-1} and g^{-1}, respectively.*

Proof.

Covariance is invariant to translations, so we may assume $\phi = \gamma = 0$. Let

$$\kappa_{X,Y}(\alpha, \beta) = \operatorname{cov}\{\alpha^\top X, \beta^\top Y\} \quad \& \quad \kappa_{f,g}(\alpha, \beta) = \operatorname{cov}\{\alpha^\top \mathsf{F}X, \beta^\top \mathsf{G}Y\}$$

be the original and transformed objectives.

Consider the (bijective) change of variables

$$\begin{pmatrix}\tilde{\alpha}\\ \tilde{\beta}\end{pmatrix} = \mathsf{D}\begin{pmatrix}\alpha\\ \beta\end{pmatrix} = \begin{pmatrix}\mathsf{F}^\top\alpha\\ \mathsf{G}^\top\beta\end{pmatrix}, \qquad \text{where} \quad \mathsf{D} = \begin{pmatrix}\mathsf{F}^\top & 0\\ 0 & \mathsf{G}^\top\end{pmatrix}.$$

In these variables:

$$\kappa_{f,g}(\alpha,\beta) = \mathrm{cov}\{\alpha^\top \mathsf{F}X, \beta^\top \mathsf{G}Y\} = \mathrm{cov}\{\tilde{\alpha}^\top X, \tilde{\beta}^\top Y\} = \kappa_{X,Y}(\tilde{\alpha},\tilde{\beta}).$$

$$(\alpha,\beta) \in \mathcal{C}_1(f,g) \Leftrightarrow \mathrm{var}\{\underbrace{\alpha^\top \mathsf{F}X}_{\tilde{\alpha}^\top X}\} = \mathrm{var}\{\underbrace{\beta^\top \mathsf{G}Y}_{\tilde{\beta}^\top Y}\} = 1 \Leftrightarrow (\tilde{\alpha},\tilde{\beta}) \in \mathcal{C}_1(X,Y)$$

And, given this equivalence, we have for $j = 2, ..., k$,

$$\begin{aligned}(\alpha,\beta) \in \mathcal{C}_j(f,g) &\Leftrightarrow (\alpha,\beta) \in \mathcal{C}_1(f,g) \,\&\, \mathrm{cov}\{\alpha^\top \mathsf{F}X, \alpha_i^\top \mathsf{F}X\} = \mathrm{cov}\{\beta^\top \mathsf{G}Y, \beta_i^\top \mathsf{G}X\} = 0 \;\forall\, i < j\\ &\iff (\tilde{\alpha},\tilde{\beta}) \in \mathcal{C}_1(X,Y) \quad \& \quad \mathrm{cov}\{\tilde{\alpha}^\top X, \tilde{\alpha}_i^\top X\} = \mathrm{cov}\{\tilde{\beta}^\top Y, \tilde{\beta}_i^\top X\} = 0 \;\forall\, i < j\\ &\iff (\tilde{\alpha},\tilde{\beta}) \in \mathcal{C}_j(X,Y)\end{aligned}$$

Letting $(\tilde{\alpha}_j^*, \tilde{\beta}_j^*)$ be the maximiser of $\kappa_{X,Y}$ over $\mathcal{C}_j(X,Y)$, and $\begin{pmatrix}\alpha^*\\ \beta^*\end{pmatrix} = \mathsf{D}^{-1}\begin{pmatrix}\tilde{\alpha}_j^*\\ \tilde{\beta}_j^*\end{pmatrix}$,

$$\kappa_{X,Y}(\tilde{\alpha}_j^*, \tilde{\beta}_j^*) \geq \kappa_{X,Y}(\tilde{\alpha}_j, \tilde{\beta}_j), \quad \forall (\tilde{\alpha}_j, \tilde{\beta}_j) \in \mathcal{C}_j(X,Y)$$

$$\implies \kappa_{f,g}(\alpha_j^*, \beta_j^*) \geq \kappa_{f,g}(\alpha_j, \beta_j), \quad \forall (\alpha_j, \beta_j) \in \mathcal{C}_j(f,g) \qquad \square$$

- At the level of sample, CCA can be carried out by replacing the covariance matrix by the sample covariance matrix. Since $(\hat{\Sigma}_X, \hat{\Sigma}_Y)$ are consistent for (Σ_X, Σ_Y), we have $\mathbb{P}\{\hat{\Sigma}_X, \hat{\Sigma}_Y \succ 0\} = 1$ for all n sufficiently large (exercise)
- So for n sufficiently large, when the singular values of the cross-covariance Σ_{XY} are distinct, the sample canonical pairs and canonical correlations are the MLE of their population versions.
- As for their asymptotic properties, notice that the sample canonical pairs/correlations can be related to the eigenvectors/values of $\hat{\Phi}\hat{\Phi}^\top$ and $\hat{\Phi}^\top\hat{\Phi}$, where $\hat{\Phi} := \hat{\Sigma}_X^{-1/2}\hat{\Sigma}_{XY}\hat{\Sigma}_Y^{-1/2}$
- So we can use our perturbation bounds once we can control the deviations $\|\hat{\Phi}\hat{\Phi}^\top - \Phi\Phi^\top\|_{\mathbb{R}^{q\times q}}$ and $\|\hat{\Phi}^\top\hat{\Phi} - \Phi^\top\Phi\|_{\mathbb{R}^{p\times p}}$
- In the presence of invertibilty, $\Sigma \mapsto \Phi$ and $\hat{\Sigma} \mapsto \hat{\Phi}$ are Lipschitz continuous, and so we can obtain such bounds. Similar arguments involving the differentiability of these maps can be used to obtain $\sqrt{n}$ asymptotic Gaussian limits (allowing inference) via the delta method. Exercise: Use the spectrum to show that $\Sigma \mapsto \Sigma^2$, $\Sigma \mapsto \Sigma^{-1}$, $\Sigma \mapsto \Sigma^{1/2}$ are C^1 at $\Sigma \succ 0$.

(Gaussian) Graphical Models

As an important special case consider a stationary Markov Chain:

- A sequence of identically distributed random scalars $\xi_1, \xi_2, \xi_3, ...$
- Markov property: the past is independent of the future given the present:

$$\{\xi_i\}_{i<k} \perp\!\!\!\perp \{\xi_j\}_{j>k} | \xi_k, \qquad \forall k.$$

- by stationarity, the transition density $f_{\xi_{k+1}|\xi_k} = g$ is time-invariant.

Assuming that $X = (\xi_1, ..., \xi_p)^\top$ is jointly centred Gaussian, this implies that

$$\begin{aligned} \xi_1 &\sim N\left(0, \sigma^2/(1-\rho^2)\right), \qquad \xi_{k+1} = \rho\xi_k + \varepsilon_{k+1} \\ \rho &= \text{corr}\{\xi_k, \xi_{k+1}\}, \quad |\rho| < 1, \qquad \varepsilon_i \perp\!\!\!\perp \xi_i, \qquad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2). \end{aligned}$$

This is known as the Gaussian stationary AR(1) model (autoregressive of order 1)

Exercise: Show this via the regression representation of conditional independence.

The Markov property stipulates that the density factorises:

$$f(\xi_1, ..., \xi_n; \Theta) = f_{\xi_1}(\xi_1; \Theta) \prod_{j=1}^{p-1} f_{\xi_{j+1}|\xi_i}(\xi_{j+1}|\xi_j; \Theta)$$

where the conditional densities $f_{\xi_{j+1}|\xi_i}(\cdot|y; \Theta)$ are $N(\rho y, \sigma^2)$ pdf's. For a single realisation of the vector $X = (\xi_1, ..., \xi_p)^\top$, this yields a loglikelihood (up to constants)

$$\ell_1(\rho, \sigma^2) = -\frac{1}{2} \log\left(\frac{\sigma^2}{1-\rho^2}\right) - \frac{(p-1)\log\sigma^2}{2} - \frac{(1-\rho)^2\xi_1^2}{2\sigma^2} - \sum_{j=1}^{p-1} \frac{(\xi_{j+1} - \rho\xi_j)^2}{2\sigma^2}.$$

When n independent realisations $X_i = (\xi_{i,1}, ..., \xi_{i,p})^\top$ are available, we get

$$\ell_1(\rho, \sigma^2) = -\frac{n}{2} \log\left(\frac{\sigma^2}{1-\rho^2}\right) - \frac{n(p-1)}{2}\log\sigma^2 - \frac{(1-\rho)^2}{2\sigma^2}\sum_{i=1}^{n}\xi_{i,1}^2 - \sum_{i=1}^{n}\sum_{j=1}^{p-1}\frac{(\xi_{i,j+1} - \rho\xi_{i,j})^2}{2\sigma^2}.$$

Notice the information gain: get order np observations for 2 parameters! (instead of n observations to estimate order p^2 parameters)

What about the precision matrix?

The only pairs (ξ_i, ξ_j) that are not conditionally independent given $\{\xi_k : k \notin \{i, j\}\}$ are adjacent pairs, i.e. $|i - j| = 1$. Consequently:

the precision matrix Θ of a stationary Gaussian AR(1) model is tridiagonal.

Now notice that Σ is determined elementwise via

$$\text{cov}\{\xi_i, \xi_{i+k}\} = \sqrt{\text{var}\{\xi_i\}\text{var}\{\xi_{i+k}\}} \times \text{corr}\{\xi_i, \xi_{i+k}\} = \frac{\sigma^2}{(1-\rho^2)}\rho^k$$

And so we can directly verify that $\Theta = \Sigma^{-1} = \mathsf{L}^\top \mathsf{L}$ where

$$\mathsf{L} = \sigma^{-1} \times \begin{pmatrix} \sqrt{1-\rho^2} & 0 & \dots & \dots & 0 \\ -\rho & 1 & 0 & \dots & 0 \\ 0 & -\rho & 1 & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & -\rho & 1 \end{pmatrix}$$

With this decomposition, we can calculate the loglikelihood of L based on a single realisation of X using the expression of the multivariate Gaussian density:

$$\ell_1(\mathsf{L}) = \frac{1}{2}\log|\mathsf{L}^\top\mathsf{L}| - \frac{1}{2}X^\top\mathsf{L}^\top\mathsf{L}X$$

Obviously, we'll get the same expression via the Markov factorisation (exercise).

In summary, the Markov property

- yields a factorisation of the joint density by suitable conditioning.
- leads to a sparse precision matrix.
- substantially increases statistical efficiency.

Is there a more general structure underlying all this?

After all, recall that for jointly Gaussian $(X^\top, Z^\top, Y^\top)^\top$ with $\Sigma \succ 0$,

$$X \perp\!\!\!\perp Y \,|\, Z \iff f_{XZY} = f_{X|Z} \times f_{Y|Z} \times f_Z \iff \Theta_{XY} = 0$$

Recall that, in the AR(1) model, the only pairs (ξ_i, ξ_j) that are not conditionally independent given all else are adjacent pairs, i.e. $|i-j| = 1$.

Define a graph $G = (V, E)$ with, vertex/edge set, respectively

$$V = [p] = \{1, ..., p\} \quad \& \quad E = \{(i,j) \in [p]^2 : |i-j| = 1\} \subset V^2$$

We do not allow loops (i.e. self-edges). Then,

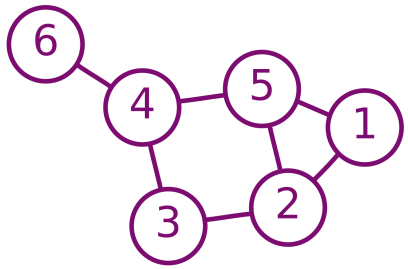
$$\xi_i \perp\!\!\!\perp \xi_j | \{\xi_k : k \in [p] \setminus \{i,j\}\} \iff (i,j) \notin E$$

. In other words:

- absence of edge (i,j) indicates conditional independence given all else.
- absence of edge (i,j) indicates that dependence between ξ_i and ξ_j is indirect
- presence edge (i,j) indicates direct dependence between ξ_i and ξ_j – dependence that is not undone by any conditioning.

A graphical model generalises the chain-like dependence structure to more general combinatorial dependence structures encoded by a more general graph.

A Gaussian Graphical Model encodes the conditional independencies amongst the coordinates $V := \{1, ..., p\}$ of a Gaussian vector via the edges in a graph on V.



There are three ways a graph $G = (E, V)$ could define a Markov property:

Pairwise, Local and Global Markov Property

1. Pairwise Markov. No edge between X_i and X_j implies their conditional independence given remaining variables: $(i,j) \notin E \implies X_i \perp\!\!\!\perp X_j | \{X_k\}_{k \neq i,j}$
2. Local Markov. Conditional on its graph neighbours, ith variable is independent of all other variables: $X_i \perp\!\!\!\perp \{X_j : (i,j) \notin E\} | \{X_k : (i,k) \in E\}$
3. Global Markov. Two subvectors are conditionally independent given a subvector that separates[a] them in G:

$$S \subset V \text{ separates } A \subset V \text{ from } B \subset V \text{ in } G \implies X_A \perp\!\!\!\perp X_B | X_S$$

[a] $S \subset V$ separates $A, B \subset V$ in G if removing S from V disconnects A from B

In full generality, regardless of Gaussianity, it's not hard to see (exercise) that:

$$\text{Global Markov} \implies \text{Local Markov} \implies \text{Pairwise Markov}$$

To go the other way around, we need to exclude "perfect dependence":

Theorem (Equivalence of Markov Properties – Gaussian case)

The three Markov properties are equivalent for $N(\mu, \Sigma)$ on $\mathbb{R}^p$ with $\Sigma \succ 0$.

Remark: In the non-Gaussian case the theorem is valid provided we replace non-singularity of the covariance, with everywhere positivity of the joint density.

Proof.

Write $V = \{1, ..., p\}$ and $\Theta = \Sigma^{-1}$. Given the exercise above, it suffices to prove that when $\Sigma \succ 0$, the pairwise Markov property with respect to some graph $G = (E, V)$ implies the global Markov property with respect to G. Assume that G encodes the pairwise Markov property. In the Gaussian case, this happens iff

$$E = \{i \neq j : \Theta_{ij} \neq 0\}$$

by the Gaussian conditional independence theorem (slide 83).

Now we need to show that this graph structure also yields the global Markov property. To this aim, assume that $S \subset V$ separates $A, B \subset V$ in the graph G. If (A, B, S) partition V, i.e. $A \cup B \cup S = V$, then we are done. To see this:

- Since A and B are separated, there is no edge from $i \in A$ to $j \in B$. It follows that $\Theta_{ij} = 0$ for all $(i, j) \in A \times B$. Blockwise, this says $\Theta_{AB} = 0$.
- Hence, by the Gaussian conditional independence theorem (slide 83) we have $X_A \perp\!\!\!\perp X_B$ given all other variables.
- But "all other variables" coincides with X_S, since $V = A \cup B \cup S$.

Now consider the general case. Let $R = V \setminus [A \cup B \cup S]$ be the "remaining vertices". Partition $R = R_A \cup R_B \cup R_0$ where:

- R_A contains all the vertices in R that are path-connected with A
- R_B contains all the vertices in R that are path-connected with B
- $R_0 = R \setminus [R_A \cup R_B]$ are the remaining vertices in R.

We highlight that this is indeed a partition of R: it must be that $R_A \cap R_B = \emptyset$ or else S would not separate A and B. For the same reason, S necessarily separates R_A from R_B, R_A from B, and R_B from A (possibly trivially so). Finally, any $v \in R_0$ is disconnected from both B and R_B, or else v would be contained in R_B (similarly for A and R_A, but we won't need that). Now our trick will be to augment the sets A and B, use the first part of the proof, and finally marginalise.

Define $\tilde{A} = A \cup R_A \cup R_0$ and $\tilde{B} = B \cup R_B$. Then, from our preceding discussion, S separates $(\tilde{A}, \tilde{B})$, and furthermore $(\tilde{A}, S, \tilde{B})$ is a partition V, so by the first part of the proof

$$X_{\tilde{A}} \perp\!\!\!\perp X_{\tilde{B}} | X_S.$$

But since $A \subset \tilde{A}$ and $B \subset \tilde{B}$, this implies that

$$X_A \perp\!\!\!\perp X_B | X_S.$$

This completes the proof. □

Where did we specifically rely on Gaussianity in this proof?

- In the first bullet of the last page, we were able to go from pairwise conditional independence to blockwise conditional independence.
- This is a remarkable feature of Gaussians: no interactions involve more than pairs of variables. This can already be seen at the level of density:

$$\log f(x) = \text{const} - \frac{1}{2} \sum_{v \in V} \Theta_{vv} x_v^2 - \frac{1}{2} \sum_{(v,v') \in E} \Theta_{vv'} x_v x_{v'}, \quad x = (x_1, \ldots, x_p)^\top \in \mathbb{R}^p.$$

Now that we've clarified how the Markov property with respect to a graph relates to precision matrix sparsity, let's turn to the factorisation of the density.

Let's revisit the AR(1) example momentarily. The factorisation

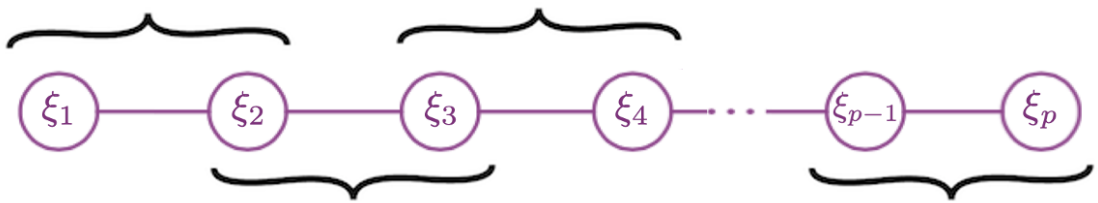
$$f(u_1, \ldots, u_n) = f_{\xi_1}(u_1) \prod_{j=1}^{p-1} f_{\xi_{j+1}|\xi_i}(u_{j+1}|u_j)$$

used the fact that the graph was "well-ordered", to arrive at a form

$$f(u_1, \ldots, u_n) \propto \prod_{j=1}^{p-1} \psi_j(u_j, u_{j+1}), \qquad \psi_i : \mathbb{R}^2 \to (0, \infty)$$

To see the general picture, where there is usually no ordering, we need some definitions related to a graph:

- A clique of $G = (V, E)$ is a fully connected subset of V.
- A maximal clique is a clique that is not a strict subset of another clique.



The AR(1) model joint density factorises over its maximal cliques.

Let $V = \{1, ..., p\}$, $G = (V, E)$ a graph, and $f(x_1, ..., x_p) > 0$ be an everywhere positive density on $\mathbb{R}^p$. We say that $f > 0$ factorises with respect to G if

$$f_{X_1,...,X_p}(x_1, ..., x_p) = \prod_{C \subseteq V} \psi_C(x_C),$$

for 2^p interaction functions $\psi_C > 0$ such that $\psi_C = 1$ unless C is a clique.

We use the shorthand notation $x_C = (x_{i_1}, ..., x_{i_k})$ for $C = \{i_1, ..., i_k\} \subseteq V$.

- Said differently, $f > 0$ factorises as product of positive functions ψ_C with C ranging over the collection $\mathcal{C}(G)$ of cliques of G.
- The reason we give the definition the way we do, is to give a "parsimony/reductive" intuition – removing terms from a larger product corresponding to non-cliques.
- This factorization implies that the global distribution can be understood in terms of local interactions.
- Conversely: provides a way to construct complex distributions from simpler building blocks, modeling local interactions.

In statistical physics terminology, one calls a density of the form

$$f_{X_1,\dots,X_p}(x_1,\dots,x_p) \propto \exp\left\{\sum_{C\in\mathcal{C}(G)} \phi_C(x_C)\right\}$$

with real-valued and non-identically-vanishing valued potential functions $\phi_C : \mathbb{R}^{|C|} \to \mathbb{R}$, a Gibbs distribution with respect to G.

(to see the relation consider potential $\phi_C = \log \psi_C$ with interaction ψ_C as above)

The fundamental result linking conditional independence and factorisation is:

Theorem (Hammersley-Clifford)

Let $f : \mathbb{R}^p \to (0,\infty)$ be an everywhere positive probability density function. Then,

f factorises w.r.t. G $\iff$ f satisfies the local Markov property w.r.t. G

Proof.

Assume $f > 0$ factorises w.r.t. G. Consider the conditional of X_i given all else,

$$f(x_i|x_{V\setminus\{i\}}) = \frac{\prod_{C\subseteq V}\psi_C(x_C)}{\int_{\mathbb{R}}\prod_{C\subseteq V}\psi_C(x_C)dx_i} = \frac{\prod_{C\not\ni i}\psi_C(x_C)\prod_{C\ni i}\psi_C(x_C)}{\prod_{C\not\ni i}\psi_C(x_C)\int_{\mathbb{R}}\prod_{C\ni i}\psi_C(x_C)dx_i}$$

In summary,

$$f(x_i|x_{V\setminus\{i\}}) = \prod_{C\ni i}\psi_C(x_C)/\int_{\mathbb{R}}\prod_{C\ni i}\psi_C(x_C)\,dx_i.$$

Therefore, given any $j \neq i$, the RHS depends on x_j only if $(i,j) \in E$:

- if $C \ni i$ and C is not a clique, then $\psi_C = 1$, so there is no dependence on x_j.
- if $C \ni i$ and C is a clique, and then $j \in C$ if and only if $(i,j) \in E$. So, in turn, ψ_C (and hence the RHS) depends on x_j if and only if $(i,j) \in E$.

In other words, f satisfies the local Markov property w.r.t. G.

In the other direction, assume $f > 0$ satisfies the local Markov property w.r.t. G. Define $\psi_\emptyset \equiv \psi_\emptyset(x_\emptyset) = f(\mu)$ for some fixed reference point $\mu \in \mathbb{R}^p$ (e.g. take $\mu = 0$). The argument $x_\emptyset$ simply corresponds to the function $\psi_\emptyset$ being a constant (does not depend on any coordinates). Define the remaining ψ's recursively, via

$$\psi_C(x_C) = \begin{cases} 1 & \text{if } C \text{ is not a clique,} \\ 1 & \text{if } x_j = \mu_j \text{ for some } j \in C \\ \dfrac{f(x_C, \mu_{V\setminus C})}{\prod_{B\subset C}\psi_B(x_B)} & \text{otherwise.} \end{cases}$$

Clearly the collection $\{\psi_C\}$ has the form we seek. It only remains to show that f factorises as $\prod_{C\subseteq V}\psi_C$ with these specific interaction functions.

Given any $x \in \mathbb{R}^p$, let $A = A(x)$ be the set of all coordinates where x and μ disagree: $A = \{j \in V : x_j \neq \mu_j\}$. Decompose

$$x = (x_A, x_{V \setminus A})^\top \equiv (x_A, \mu_{V \setminus A})^\top.$$

Suppose $A = \emptyset$. In this case $x = \mu$ and we need to show that $f(\mu)$ factorises as stipulated. Checking the definition of $\psi_C(\mu_C)$, we notice that there is no $C \subseteq V$ for which $\psi_C(\mu_C) \neq 1$ except $C = \emptyset$, for which we defined $\psi_\emptyset(x_\emptyset) \equiv f(\mu)$. Therefore the factorisation holds in the form

$$\prod_{C \subseteq V} \psi_C(\mu_C) = \psi_\emptyset = f(\mu).$$

Now suppose that $A \neq \emptyset$ and is a clique. Then,

$$\prod_{C \subseteq V} \psi_C(x_C) = \psi_A(x_A) \prod_{C \subset A} \psi_C(x_C) \prod_{C \not\subset A} \psi_C(x_C) = \frac{f(x_A, \mu_{V \setminus A})}{\prod_{C \subset A} \psi_C(x_C)} \prod_{C \subset A} \psi_C(x_C) \prod_{C \not\subset A} \psi_C(x_C)$$

and the terms in the last product equal 1 because $C \not\subseteq A$ means that $x_j = \mu_j$ for some $j \in C$, in which case our construction yields $\psi_C = 1$. Noting that $f(x_A, \mu_{V \setminus A}) = f(x_A, x_{V \setminus A})$, we once again get the sought factorisation.

Finally, suppose that $A \neq \emptyset$ and is not a clique. In this case, we will establish the factorisation by induction on the size of A. The base case is $|A| = 0$, where we have already established the factorisation. For $k \leq p$, assume the factorisation holds for $|A| = k - 1$, and let's show it holds for $|A| = k$.

Since A is not a clique, there exist $i, j \in A$ with $(i, j) \notin E$. We thus have

$$\begin{aligned} f(x_A, x_{V\setminus A}) \equiv f(x_A, \mu_{V\setminus A}) &= f(x_i | x_{A\setminus\{i\}}, \mu_{V\setminus A}) f(x_{A\setminus\{i\}}, \mu_{V\setminus A}) \\ &= f(x_i | x_{A\setminus\{i\}}, \mu_{V\setminus A}) \frac{f(x_{A\setminus\{i\}}, \mu_{V\setminus A}, \mu_i)}{f(\mu_i | x_{A\setminus\{i\}}, \mu_{V\setminus A})} \\ &= \frac{f(x_i | x_{A\setminus\{i,j\}}, \mu_j, \mu_{V\setminus A})}{f(\mu_i | x_{A\setminus\{i,j\}}, \mu_j, \mu_{V\setminus A})} f(x_{A\setminus\{i\}}, \mu_{V\setminus A}, \mu_i) \end{aligned}$$

using the local Markov prop to go in the last step: the green terms don't depend on x_j, because $(i, j) \notin E$, so we can fix x_j to whichever value we wish without changing the expression. So we fixed $x_j = \mu_j$. Using the inductive hypothesis,

$$= \frac{\frac{f(x_i, x_{A\setminus\{i,j\}}, \mu_j, \mu_{V\setminus A})}{f(x_{A\setminus\{i,j\}}, \mu_j, \mu_{V\setminus A})}}{\frac{f(\mu_i, x_{A\setminus\{i,j\}}, \mu_j, \mu_{V\setminus A})}{f(x_{A\setminus\{i,j\}}, \mu_j, \mu_{V\setminus A})}} f(x_{A\setminus\{i\}}, \mu_{V\setminus A}, \mu_i) = \frac{\prod_{C\subseteq A\setminus\{j\}} \psi_C(x_C)}{\prod_{C\subseteq A\setminus\{i,j\}} \psi_C(x_C)} \prod_{C\subseteq A\setminus\{i\}} \psi_C(x_C) = \prod_{C\subseteq A} \psi_C(x_C)$$

$= \prod_{C\subseteq V} \psi_C(x_C)$ by definition of ψ_C and fact that $(A \setminus \{j\}) \setminus (A \setminus \{i, j\}) = \{i\}$. □

Therefore, we get the immediate corollary:

Corollary (Hammersley-Clifford Theorem, Gaussian case)

Let $V = \{1, ..., p\}$, $G = (V, E)$ a graph, and $X \sim N(0, \Sigma)$ on $\mathbb{R}^p$ with $\Sigma \succ 0$. Then, the following statements are equivalent:

1. *the density of X factorises with respect to G*
2. *X satisfies a Markov property[a] with respect to G*
3. *given $i \neq j$, the (i, j) entry of $\Theta = \Sigma^{-1}$ is zero if and only if $(i, j) \notin E$.*

[a] we say 'a' Markov property, because all three Markov are equivalent when $\Sigma \succ 0$.

- When the graph is geometric, then one can make use of factorisation cleverly in order to carry out likelihood estimation conforming to a graphical model.
- The key here is that the graph is quite sparse (correspondingly, the precision matrix is very sparse, and one can see things through that lens).
- But, in general, the "list of cliques" is difficult to obtain – in fact NP-hard.

There are two (in a sense dual) problems that one might consider in this context, given $X_1, ..., X_n \sim N(\mu, \Sigma)$:

- Fitting a Gaussian distribution (estimating μ and Σ) subject to the constraint that $N(\mu, \Sigma)$ factorises with respect to a given graph G. This is also known as covariance selection.
- Estimating the graph G with respect to which $N(\mu, \Sigma)$ factorises when the parameters are unknown. This is also known as structure estimation.

In light of the Hammersley-Clifford theorem, when $\Sigma \succ 0$, these two problems reduce to:

- Estimate the model parameters under the constraint that $\Theta_{ij} = 0$ for known set E of pairs (i, j).
- Estimate the location of zeroes of Θ_{ij} amongst pairs (i, j).

Given a normal data matrix X, a (Gaussian) covariance selection model consists in the family

$$\{N(\mu, \Sigma) : \mu \in \mathbb{R}^p, \Sigma \in \mathcal{S}\}$$

for some subset $\mathcal{S}$ of the cone of $p \times p$ non-negative definite matrices.

- Depending on our choice of $\mathcal{S}$ an MLE may or may not exist.
- A standard choice is $\mathcal{S}$ being the set of strictly positive definite matrices.
- Even then, we saw the MLE does not exist when $\hat{\Sigma}$ is singular.
- So we expect that existence/uniqueness of the MLE is subtle for general $\mathcal{S}$

What might be reasonable choices?

- We may choose to impose linear constraints on Σ
- We may choose to impose linear constraints on Σ^{-1} (and assume it exists)

Focussing on the special case where constraints fix some elements to be zero:

- Imposing this on Σ leads to straightforward estimation: when MLE exists, we can annhilate the corresponding entries by equivariance.
- Imposing this on Σ^{-1} leads to estimation under a graphical model.

Since the constraints apply to the covariance we can take $\mu = 0$. Now consider a graph $G = (V, E)$ on the index set $V = \{1, ..., p\}$. The model we wish to fit is

$$\{N(0, \Sigma) : \Sigma \succ 0 \,\&\, e_i^\top \Sigma^{-1} e_j = 0 \,\text{whenever}\, (i, j) \notin E\}.$$

When $\Sigma \succ 0$, the loglikelihood is (up to constants)

$$\ell(\Sigma) = -\log|\Sigma| - \text{trace}\{\Sigma^{-1}\hat{\Sigma}\}$$

which can be equivalently expressed via $\Theta = \Sigma^{-1}$ as

$$\ell(\Theta) = \log|\Theta| - \text{trace}\{\Theta\hat{\Sigma}\}.$$

The crucial observations now are that:

- The objective $\ell(\Theta)$ is strictly concave over the set $\Theta \succ 0$. (exercise)
- The constraint set $\{\Theta \succ 0 : e_i^\top \Theta e_j = 0 \,\text{whenever}\, (i, j) \notin E\}$ is convex.

In conclusion, maximising the loglikelihood under a graphical model constraint is equivalent to a strictly convex optimisation problem – so provided a maximiser exists, it will also be unique.

It turns out that loglikelihood maximisation under a graphical model constraint is equivalent to the entropy maximisation under second moment constraints:

Theorem (Graphical Modeling as Matrix Completion)

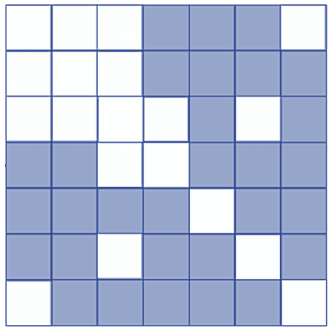
Maximising the loglikelihood $\ell(\Sigma) = -\log|\Sigma| - \mathrm{trace}\{\Sigma^{-1}\hat{\Sigma}\}$ *over the set*

$$\{\Sigma \succ 0 : e_i^\top(\Sigma^{-1})e_j = 0 \text{ whenever } (i,j) \notin E\}$$

is equivalent to maximising the entropy $H(\Sigma) \propto \log|\Sigma| + \mathrm{const}$ *over the set*

$$\{\Sigma \succ 0 : e_i^\top \Sigma e_j = e_i^\top \hat{\Sigma} e_j \text{ whenever } (i,j) \in E \text{ or } i = j\}.$$

Intuitively: delete all non-adjacent (w.r.t. G) off-diagonal entries of $\hat{\Sigma}$. Then complete the missing entries to maximise the determinant.



Proof.

The objective is negatively infinite at singular matrices, so we focus on interior points of the of the non-negative definite cone. We then have a differentiable strictly concave objective in Θ with equality constraints. We thus resort to the method of Lagrange multipliers. Define the Lagrangian

$$\mathcal{L}(\Theta, \varepsilon) = \log|\Theta| - \operatorname{tr}\{\Theta\hat{\Sigma}\} - \operatorname{tr}\{\Theta\varepsilon\} = \log|\Theta| - \operatorname{tr}\{\Theta(\hat{\Sigma} + \varepsilon)\}$$

for a symmetric $\varepsilon = \{\varepsilon_{ij}\}$ with $\varepsilon_{ij} = 0$ when $i = j$ or $(i, j) \in E$ (ε_{ij} are Lagrange multipliers corresponding to the equality constraints). Constrained optima must be saddlepoints of the Lagrangian. So if Θ_* is a constrained maximum of ℓ, then Θ_* is a critical point of $\tilde{\ell}(\Theta) = \log|\Theta| - \operatorname{tr}\{\Theta\tilde{\Sigma}\} = \log|\Theta| - \operatorname{tr}\{\Theta(\hat{\Sigma} + \varepsilon)\}$.
This is just a Gaussian loglikelihood corresponding to an empirical covariance $\tilde{\Sigma} = \hat{\Sigma} + \varepsilon$ instead of $\hat{\Sigma}$. When $\Theta \succ 0$ this can only have a unique critical point when $\tilde{\Sigma} \succ 0$, and that critical point is $\Theta_* = \tilde{\Sigma}^{-1} = (\hat{\Sigma} + \varepsilon)^{-1}$.
Now it remains to plug such a Θ_* back into the Lagrangian, and minimise over ε with $\varepsilon_{ij} = 0$ when $i = j$ or $(i, j) \in E$. Equivalently, it remains to choose ε to minimize $\ell\big((\hat{\Sigma} + \varepsilon)^{-1}\big) = -\log|\hat{\Sigma} + \varepsilon| - p$. Evidently, minimisation occurs at ε such that $\hat{\Sigma} + \varepsilon \succ 0$, compatibly with the requirement of the critical point Θ_*.
Finally, recall that entropy is minus the expected loglikelihood,

$$-\mathbb{E}\left[-\log|\Sigma| - \operatorname{tr}\{\Sigma^{-1}\hat{\Sigma}\} + \text{const}\right] = \log|\Sigma| + p + \text{const} \quad \square$$

If a feasible point exists, the graphically constrainted MLE exists uniquely.

- Existence of a feasible point $\equiv$ existence of a valid matrix completion
- Any completion must agree with the diagonal constraints – so the trace (and hence Frobenius norm) is (bounded by a) constant on the feasible set.
- Hence completion set is convex & compact, while objective is strictly concave.

Now notice a subtle distinction:

- At the level of (complete) data: If we start out with observation of the complete empirical matrix $\hat{\Sigma} \succ 0$, then it is clear that there exists at least one feasible point – namely the completion to $\hat{\Sigma}$ itself! So we can use gradient ascent to find the maximum.
- (Related to the last point) Small sample size or missing data if we have limited sample size $n < p$ (so $\hat{\Sigma}$ is singular), but $n \times n$ submatrices are non-singular, what graphs could we impose to get a graphically constrained MLE? (recall the Markov chain model, where the effective # of parameters is reduced). Similarly, if we have missing data (missing entries) what missingness patterns can he handle by graphical modeling?
- At the level of (local) modeling: What if we model some second moments but not the full covariance vector. If we arrange the partial moments into a matrix, can it admit a positive-definite completion?
(and hence a maximum entropy completion?)

Suppose we are interested in a Gaussian random vector $X = (X_1, X_2, X_3, X_4)^\top$...

...but we can only (or are only willing to) model the pairs (X_i, X_j), for

$$(i, j) \in \{(1,2), (2,3), (3,4), (1,4)\}$$

E.g. we prescribe corresponding covariances as follows:

$$\begin{pmatrix} 1 & \rho & ? & -\rho \\ \rho & 1 & \rho & ? \\ ? & \rho & 1 & \rho \\ -\rho & ? & \rho & 1 \end{pmatrix}$$

- Of course we must take care that all specified marginal covariances of all orders are positive definite (otherwise, it's a no-go from the start).
- This can be checked to be valid in our example.
- However, there is no valid positive-definite completion in this case!
- The missing entry pattern implicitly specifies a graph G: $i \neq j$ are adjacent if and only if their covariance is specified.
- It turns out that so long as subcovariances $\succ 0$, the problem of completion is entirely contingent on the structure of the graph (think Hammersley-Clifford)

Theorem (Grone, Johnson, Sá & Wolkowicz)

Let Σ_{partial} be a partial covariance with missing entry pattern graph G. Provided all specified subcovariances of Σ_{partial} are positive-definite:

$$\Sigma_{\text{partial}} \text{ admits a positive-definite completion} \iff G \text{ is chordal}^a$$

[a] every cycle of length 4 has a chord.

- Remarkably, the proof is constructive: it manifests a completion.
- It makes use of the fact that any chordal graph can be turned into a complete graph by adding one edge at a time in such a way, that the resulting graph remains chordal at each step.
- Following this ordering of edge additions, the partial matrix is completed entry by entry in such a way as to maximize the determinant of the largest complete submatrix that contains the missing entry
- Thus, the construction yields the maximum entropy completion, not just any completion!
- We will not prove the theorem in its full generality. But we will establish feasibility of completion in a special class of chordal graphs: *serrated* partial covariances, aka variable memory Markov chains(exercise)

Now let's consider the opposite direction:

> *Given $X_1, ..., X_n \sim N(\mu, \Sigma)$ with $\Sigma \succ 0$, estimate the corresponding Markov structure (equivalently the edge set of its graphical model).*

In light of our results, the problem reduces to estimating the zero pattern of Θ. When $n > p$ we have a natural approach based on thresholding:

- We can test for the presence of each possible edge, by testing for the corresponding partial correlation.
- The test statistic has the same null distribution, with same parameters, in each case.
- Selecting a significance level is in 1-1 correspondence with selecting a threshold for hard thresholding.
- The former can be chosen according to FDR considerations, and the latter via asymptotic considerations.
- A weakness of this approach is that it's unclear if the resulting matrix is $\succ 0$.

A different approach is to use a penalised loglikelihood, to promote graph sparsity:

$$\ell_{pen}(\Theta) = \log|\Theta| - \mathrm{tr}\{\Theta\hat{\Sigma}\} + \tau \sum_{i<j} |\Theta_{ij}|.$$

(requires tuning $\tau > 0$; we actually use partial correlation matrix instead of Θ to balance scales)