The spectral theorem

4433LALG3: Linear Algebra

Week 3, Lecture 9, Valente Ramírez

Mathematical & Statistical Methods group — Biometris, Wageningen University & Research

Overview

- Orthogonal diagonalization
- The covariance matrix
- Supplementary material

References:

- Nicholson §8.2

Section 1

Orthogonal diagonalization

Motivation

In this course we have made two claims regarding the choice of basis:

- When dealing with a linear transformation $T\colon \mathbb{R}^n \to \mathbb{R}^n$, the best choice of basis is one in which the associated matrix is **diagonal**
- In the context of projections/expansions, the best choice of basis is an **orthonormal** basis

Wouldn't it be great if we could have both simultaneously?

Recap: diagonalization

Recall that if a matrix A (i.e. a linear transformation T_A) has enough eigenvalues/eigenvectors, we can:

- Find a basis of eigenvectors $\{\mathbf{p}_1, \ldots, \mathbf{p}_n\}$, with associated eigenvalues $\lambda_1, \ldots, \lambda_n$
- Define the matrix $P = \begin{bmatrix} \mathbf{p}_1 & \ldots & \mathbf{p}_n \end{bmatrix}$
- Define the matrix $D = \mathrm{diag}\{\lambda_1, \ldots, \lambda_n\}$

In order to get:

$$P^{-1}AP = D, \quad \text{or equivalently,} \quad A = PDP^{-1}$$

Recap: orthogonal matrices

Recall that $\{\mathbf{p}_1, \ldots, \mathbf{p}_n\}$ is an *orthonormal* set if: $\mathbf{p}_i^\top \mathbf{p}_j = \delta_{ij}$.

In that case, the matrix $P = \begin{bmatrix} \mathbf{p}_1 & \ldots & \mathbf{p}_n \end{bmatrix}$ is called an *orthogonal matrix*.

Fundamental property

If P is orthogonal, then it is invertible and $P^{-1} = P^\top$.

The above fact should be so evident to you that you could explain it to a six-year old.

Orthogonal diagonalization

Question

For which matrices A is it possible to find an **orthonormal** matrix P such that $P^{-1}AP$ is *diagonal*?

Suppose A, P are as above.

Then we can write: $A = PDP^{-1}$, for a diagonal matrix D.

Notice that we can also write: $A = PDP^\top$.

Let's have a look at $A^\top$:

$$
\begin{aligned}
A^\top &= (PDP^\top)^\top \\
&= (P^\top)^\top D^\top P^\top && \text{(because } (XY)^\top = Y^\top X^\top\text{)} \\
&= PD^\top P^\top && \text{(because } (P^\top)^\top = P\text{)} \\
&= PDP^\top && \text{(because } D \text{ is symmetric)} \\
&= A && \text{(because } A = PDP^\top\text{)}
\end{aligned}
$$

Orthogonal diagonalization

Conclusion

Only **symmetric matrices** can be diagonalized by an orthogonal matrix!

The converse fact, that *all* symmetric matrices can be diagonalized by an orthogonal matrix is a fundamental result in linear algebra.

The spectral theorem

Theorem

A symmetric $n \times n$ matrix A can be diagonalized by an orthogonal matrix.

This means that:

- *A has n (real) eigenvalues $\lambda_1, \dots, \lambda_n$ (they may be repeated)*
- *It is possible to find an orthonormal set of eigenvectors $\{\mathbf{p}_1, \dots, \mathbf{p}_n\}$*
- *If we set $P = \begin{bmatrix} \mathbf{p}_1 & \dots & \mathbf{p}_n \end{bmatrix}$ and $D = \operatorname{diag}\{\lambda_1, \dots, \lambda_n\}$, then*

$$A = PDP^\top$$

The spectral theorem

Some remarks

- The name *spectral theorem* comes from the fact that the set of eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$ is called **the spectrum** of A
- Technically speaking, this is the finite-dimensional real spectral theorem
- Nicholson uses the name *Principal Axes Theorem* (c.f. Theorem 8.2.2)
- The factorization of A into $A = PDP^\top$ is called the **spectral decomposition** of A

The spectral theorem

Okay, so symmetric matrices have a very nice property: there is a special basis that is orthonormal and in which the transformation is diagonal.

We will see some important consequences of that.

But ...**who cares??**

Well ...**you!**

Symmetric matrices are among the most important objects in multivariate analysis.

Section 2

The covariance matrix

The covariance matrix

Suppose $X_1, \dots, X_p$ are random variables.

The **covariance matrix** of $\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix}$ is the **symmetric** matrix:

$$\mathrm{Cov}\left(\mathbf{X}\right) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_p^2 \end{bmatrix},$$

where $\sigma_i^2 = \mathrm{Var}(X_i)$, and $\sigma_{ij} = \mathrm{Cov}(X_i, X_j)$.

The covariance matrix is extremely important whenever you are interested in *linear relationships* between the variables.

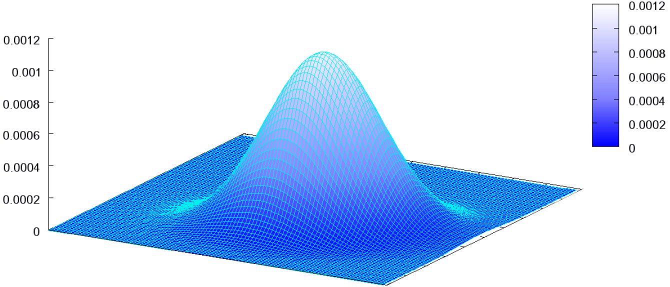
The multivariate normal distribution

Definition

A random vector $\mathbf{X} = \begin{bmatrix} X_1 & \dots & X_p \end{bmatrix}^\top$ is said to follow a **multivariate normal distribution** whenever its probability density function is given by

$$f_{\mathbf{X}}(x_1, \dots, x_p) = \frac{1}{\sqrt{(2\pi)^p \det(\Sigma)}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})},$$

where $\boldsymbol{\mu} = \mathbb{E}(\mathbf{X})$ and $\Sigma = \mathrm{Cov}(\mathbf{X})$. In such case we write $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$.



The multivariate normal distribution

In order to make our arguments more clear, today we assume $\boldsymbol{\mu} = \mathbf{0}$.

Set $C = \frac{1}{\sqrt{(2\pi)^p \det(\Sigma)}}$, and let $Q = \Sigma^{-1}$ (called the *precision matrix*).

The *pdf*, $f_{\mathbf{X}} = Ce^{-\frac{1}{2}\mathbf{x}^\top Q\mathbf{x}}$, is a composition of two transformations:

$$\begin{array}{ccccc} \mathbb{R}^p & \xrightarrow{\quad q \quad} & \mathbb{R} & \xrightarrow{\quad h \quad} & \mathbb{R} \\ \mathbf{x} & \longmapsto & \mathbf{x}^\top Q\mathbf{x} & \longmapsto & Ce^{-\frac{1}{2}\mathbf{x}^\top Q\mathbf{x}} \end{array}$$

The function $q(\mathbf{x}) = \mathbf{x}^\top Q\mathbf{x}$ takes the variables $(x_1, \ldots, x_p)$ and compresses them into a single number u.

The function $h(u) = Ce^{-\frac{1}{2}u}$ computes the density out of the "one-number summary" u.

Quadratic forms

Definition

A **quadratic form** on $\mathbb{R}^m$ is a transformation $q_A\colon \mathbb{R}^m \to \mathbb{R}$ given by

$$q_A(\mathbf{x}) = \mathbf{x}^\top A\mathbf{x},$$

for some symmetric matrix A.

Example

The transformation $q(x_1, x_2) = 4x_1^2 + 2x_1x_2 - 3x_2^2$ is a quadratic form. Indeed,

$$q(x_1, x_2) = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 4 & 1 \\ 1 & -3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

The geometry of the quadratic form

A fundamental fact regarding multivariate analysis is the following:

Fact

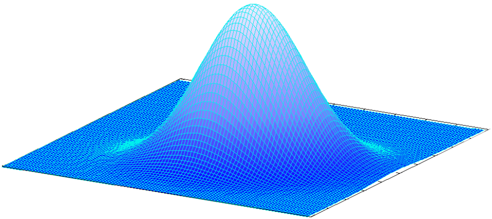
The level sets of the quadratic form Q are:

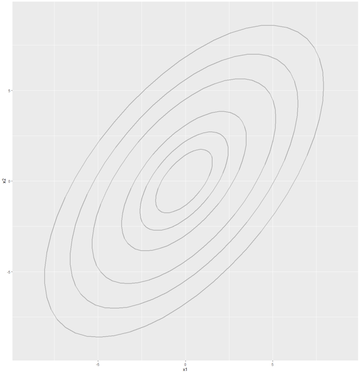
- ellipses, when $p = 2$,
- ellipsoids, when $p = 3$,
- hyperellipsoids, when $p \geq 4$.

Thus we can say that the geometry of Q is **elliptical**.

The geometry of the quadratic form

The level sets of $q(\mathbf{x})$ (and so also of $f_{\mathbf{x}}$) are ellipses.

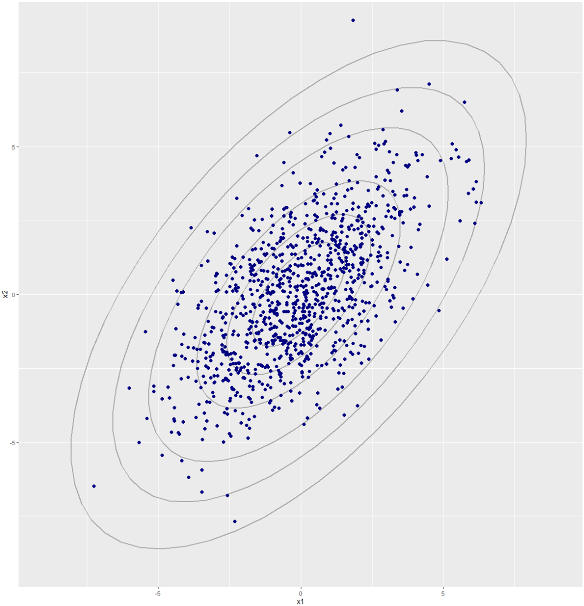




Sampling from the multivariate normal

Level sets of $q(\mathbf{x})$ define regions of iso-density.

$$n = 1000$$



Sampling from the multivariate normal

In this example,

$$\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \Sigma),$$

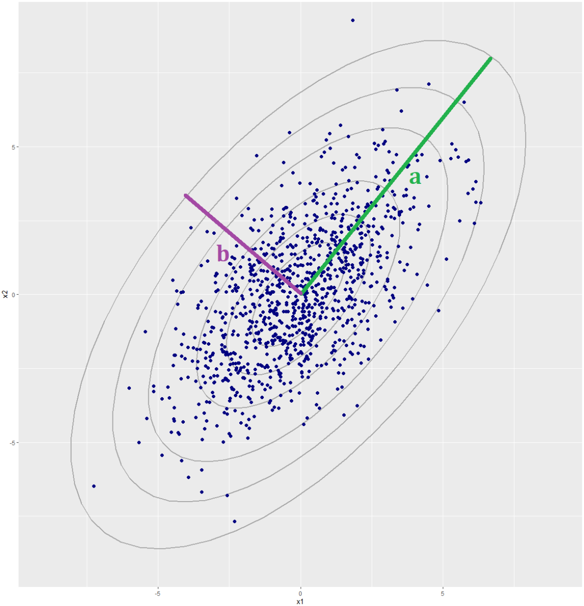
with

$$\Sigma = \begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix}.$$

How can we figure out the ellipses from Σ?

Want to know:

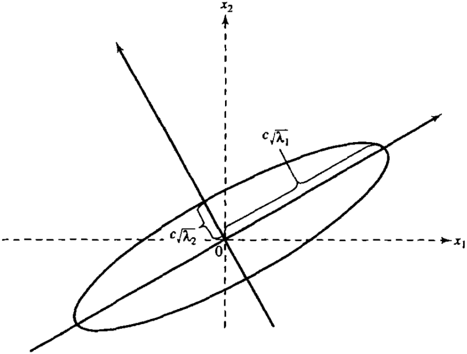
- Direction of the axes
- Aspect ratio $\frac{a}{b}$



Iso-density ellipses

The iso-density ellipse corresponding to the level curve $\mathbf{x}^\top Q\mathbf{x} = c^2$ has:

- Major axis in the direction of $\mathbf{v}_1$
- Major semi-axis length of $c\sqrt{\lambda_1}$
- Minor axis in the direction of $\mathbf{v}_2$
- Minor semi-axis length of $c\sqrt{\lambda_2}$



Source: *Johnson, Wichern – Applied Multivariate Statistical Analysis*, 6^{th} *ed.*

Above, $\mathbf{v}_1, \mathbf{v}_2$ are **eigenvectors** of Σ, and λ_1, λ_2 the corresponding **eigenvalues**. The eigenvalues are chosen so that $\lambda_1 > \lambda_2$.

Principal components

In the previous slide, the eigenvectors $\mathbf{v}_i$ are visualized in the (x_1, x_2)–plane.

Suppose $\mathbf{v}_1 = \begin{bmatrix} v_{11} \\ v_{21} \end{bmatrix}$, and $\mathbf{v}_2 = \begin{bmatrix} v_{12} \\ v_{22} \end{bmatrix}$.

We can use these coefficients to define new variables:

$$W_1 = v_{11}X_1 + v_{21}X_2 = \mathbf{v}_1^\top \mathbf{X},$$
$$W_2 = v_{12}X_1 + v_{22}X_2 = \mathbf{v}_2^\top \mathbf{X}.$$

These are called the (population) **principal components** of $\mathbf{X}$.

Principal components

The first principal component W_1 satisfies:

- W_1 is a linear combination of the X_i: $W_1 = \mathbf{v}^\top \mathbf{X}$,
- The coefficient vector $\mathbf{v}$ is such that:
  - It maximizes $\mathrm{Var}(\mathbf{v}^\top \mathbf{X})$,
  - subject to the constraint $\mathbf{v}^\top \mathbf{v} = 1$ (e.g. $\|\mathbf{v}\| = 1$) .

Subsequent components satisfy:

- W_k is a linear combination of the X_i: $W_k = \mathbf{v}^\top \mathbf{X}$,
- The coefficient vector $\mathbf{v}$ is such that:
  - It maximizes $\mathrm{Var}(\mathbf{v}^\top \mathbf{X})$,
  - subject to the constraint that W_k is uncorrelated to $W_1, \ldots, W_{k-1}$,
  - and subject to $\mathbf{v}^\top \mathbf{v} = 1$ (e.g. $\|\mathbf{v}\| = 1$) .

Principal components

By the way, it is also true that the <u>last</u> principal component W_p satisfies:

- W_p is a linear combination of the X_i: $W_p = \mathbf{v}^\top \mathbf{X}$,
- The coefficient vector $\mathbf{v}$ is such that:
  - It <u>minimizes</u> $\mathrm{Var}(\mathbf{v}^\top \mathbf{X})$,
  - subject to $\mathbf{v}^\top \mathbf{v} = 1$ (e.g. $\|\mathbf{v}\| = 1$) .

Principal components: visualization

Projection onto the line

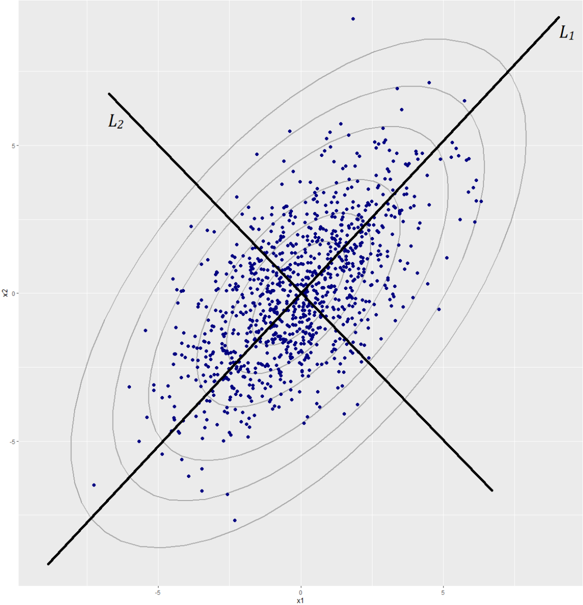
$$L_1 = \operatorname{span}(\mathbf{v}_1)$$

maximizes variance.

Projection onto the line

$$L_2 = \operatorname{span}(\mathbf{v}_2)$$

minimizes variance.



Section 3

Supplementary material

The sample covariance matrix

Suppose we have n independent observations of the variables $X_1, \ldots, X_p$.

The **sample covariance matrix** is the **symmetric** matrix:

$$S = \begin{bmatrix} s_1^2 & s_{12} & \ldots & s_{1p} \\ s_{21} & s_2^2 & \ldots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \ldots & s_p^2 \end{bmatrix},$$

where s_i^2 is the sample variance of the i^{th} variable, and s_{ij} is the sample covariance between the i^{th} and j^{th} variables.

Covariance computations in matrix notation

Formulas

Let $\mathbf{X}$ be a random vector of length p. Then $\mathrm{Cov}(\mathbf{X})$ is the $p \times p$ matrix given by:

$$\mathrm{Cov}(\mathbf{X}) = \mathbb{E}(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^\top.$$

Let X be a $n \times p$ matrix containing n observations on p variables. The sample covariance matrix of X is the $p \times p$ matrix given by:

$$S = \frac{1}{n-1}(C_n X)^\top (C_n X),$$

where C_n is the *centering matrix*, $C_n = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$.

Note: Because C_n is a projection operator, it is symmetric and idempotent. This means the last formula can be simplified to $S = \frac{1}{n-1}X^\top C_n X$.

Note

You don't need to memorize these formulas. It is just good to know that the covariance matrices can be directly defined from $\mathbf{X}$ or X, depending the case. Just look at the formulas and think how they relate to the univariate case.

Covariance computations in matrix notation

Consider two random variables X_1 and X_2, and a linear combination of them:

$$aX_1 + bX_2.$$

We are interested in the variance of this new variable. Applying the usual formulas:[1]

$$\begin{aligned}\operatorname{Var}(aX_1 + bX_2) &= a^2 \operatorname{Var}(X_1) + 2ab \operatorname{Cov}(X_1, X_2) + b^2 \operatorname{Var}(X_2) \\ &= a^2\sigma_1^2 + 2ab\sigma_{21} + b^2\sigma_2^2.\end{aligned}$$

The last formula contains a^2, ab, b^2 ... looks like a quadratic form $q(a, b)$.

In fact, you can check that:

$$\operatorname{Var}(aX_1 + bX_2) = \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}.$$

[1] We've already done this in Lecture 7. See: "Application: the dot product".

Covariance computations in matrix notation

What we saw in the last slide is true for linear combinations in general.

Important fact

Let $\mathbf{X}$ be a random vector of length n, and $\Sigma = \mathrm{Cov}(\mathbf{X})$.

Given a linear combination $W = a_1 X_1 + \ldots + a_n X_n$ (which we can also write as $W = \mathbf{a}^\top \mathbf{X}$), its variance is given by:

$$\mathrm{Var}(W) = \mathbf{a}^\top \Sigma\, \mathbf{a}.$$

In particular, if $\mathbf{a} = \mathbf{v}_i$ is a (normalized) eigenvector of Σ, then $W_i = \mathbf{v}_i^\top \mathbf{X}$ is the i^{th}–principal component, and

$$\mathrm{Var}(W_i) = \mathbf{v}_i^\top \Sigma\, \mathbf{v}_i = \mathbf{v}_i^\top (\lambda_i \mathbf{v}_i) = \lambda_i.$$

Conclusion: λ_i equals the variance of the i^{th}–principal component.

Positive-definite matrices

Suppose A is either a (population) covariance matrix or a sample covariance matrix. Then A has a very special property:

Property

Suppose A is a covariance matrix with eigenvalues $\lambda_1, \ldots, \lambda_p$.

- If $\det(A) \neq 0$, then $\lambda_i > 0$ for all i,
- In the very exeptional case that $\det(A) = 0$, then $\lambda_i \geq 0$ for all i.

Definition

A matrix is called **positive-definite** if it is symmetric and all its eigenvalues are positive.

Definition

A matrix is called **positive semi-definite** if it is symmetric and all its eigenvalues are either positive or zero.

The inverse of a symmetric matrix

Let Σ be a symmetric matrix, and assume $\det \Sigma \neq 0$.

We want to compute $Q = \Sigma^{-1}$.

It is straightforward to check that if $\Sigma = PDP^\top$, then $Q = PD^{-1}P^\top$.

Therefore,

$$\Sigma = P \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p \end{bmatrix} P^\top \quad \Rightarrow \quad Q = P \begin{bmatrix} \lambda_1^{-1} & & & \\ & \lambda_2^{-1} & & \\ & & \ddots & \\ & & & \lambda_p^{-1} \end{bmatrix} P^\top.$$

Notice that Q is also symmetric. Moreover, if Σ is positive-definite, so is Q.

The square root matrix

When it comes to matrices, the concept of the square root is not well defined.

Exercise:

Consider the following matrices:

$$A = \begin{bmatrix} 2 & -3 \\ 0 & -1 \end{bmatrix}, \quad B = \begin{bmatrix} 2 & -1 \\ 0 & 1 \end{bmatrix}, \quad C = -A, \quad D = -B.$$

Verify, by doing the necessary multiplications, that $A^2 = B^2 = C^2 = D^2$.

Which of the above matrices has the right to be called the square root of A^2?

The square root matrix

In many statistical applications, it is useful to have a matrix A, such that $A^2 = \Sigma$, where Σ is a covariance matrix.

This can be done following the next convention.

Definition

Let Σ denote a *positive-definite* matrix, with spectral decomposition $\Sigma = PDP^\top$, where $D = \mathrm{diag}\{\lambda_1, \ldots, \lambda_p\}$.

We define **the square root** of Σ as the matrix $\Sigma^{1/2}$ with spectral decomposition:

$$\Sigma^{1/2} = P\widetilde{D}P^\top, \quad \text{where} \quad \widetilde{D} = \mathrm{diag}\{\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_p}\}.$$

In the above definition we always choose $\sqrt{\lambda_i} > 0$.

Exercise:

Verify that $\left(\Sigma^{1/2}\right)^2 = \Sigma$.