# Lecture 10 - **Multivariate Normal and Principal Component Analysis**

BIOENG-210 Course Notes
Prof. Gioele La Manno

April 2025

## Contents

# 1 The Multivariate Normal Distribution

## 1.1 Introduction: From Bivariate to Multivariate Analysis

In our journey through statistical methods, we have progressively increased the complexity of our models. We began with univariate distributions, describing single variables like gene expression levels or cell sizes. We then moved to bivariate distributions, examining relationships between pairs of variables. In regression analysis, we explored how multiple predictors could inform a single outcome variable.

Now we take a bold step further into the realm of multivariate analysis, where we consider many variables simultaneously without designating any specific variable as an outcome. This shift marks the beginning of our exploration into unsupervised learning methods - approaches that help us discover patterns and structure in data without predefined target variables.

Imagine a single-cell RNA sequencing experiment where thousands of genes are measured across hundreds of cells. While regression might help us predict cell type based on gene expression profiles, multivariate analysis asks a different question: "What is the intrinsic structure of this high-dimensional data?" The answer to this question can reveal cellular states, developmental trajectories, and functional groupings that we might not have known to look for.

The multivariate normal distribution serves as our theoretical foundation for this exploration, just as the normal distribution formed the basis for much of our previous work.

## 1.2 Definition and Form

The multivariate normal distribution extends the familiar normal distribution to multiple dimensions. Just as the normal distribution is characterized by its mean and variance, the bivariate normal distribution is defined by its mean vector, two variances, and $\rho$ (the correlation coefficient). the multivariate normal is defined by two parameters: a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$.

**Definition 1.1** (Multivariate Normal Distribution)**.** A random vector $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$ follows a $p$-dimensional multivariate normal distribution, denoted $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if its probability density function is:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

where:

- $\boldsymbol{\mu} = E[\mathbf{X}]$ is the $p$-dimensional mean vector
- $\boldsymbol{\Sigma} = E[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T]$ is the $p \times p$ covariance matrix
- $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$

This formulation may appear intimidating, but we can build intuition by examining special cases. When $p = 1$, this reduces to the familiar univariate normal distribution. For $p = 2$, we recover the bivariate normal distribution we studied earlier, where the covariance matrix encodes the correlation between two variables.

Let's show how the familiar bivariate normal formula with correlation coefficient $\rho$ can be written in the multivariate normal form:

**Example 1.2** (From Standard to Matrix Form)**.** The bivariate normal distribution is often written as:

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}\right]\right)$$

We can derive the matrix form by starting with the standard form and working backward:

$$\exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2}+\frac{(x_2-\mu_2)^2}{\sigma_2^2}-\frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}\right]\right)$$

This exponent can be rearranged as:

$$-\frac{1}{2}\begin{pmatrix}x_1-\mu_1 & x_2-\mu_2\end{pmatrix}\begin{pmatrix}\frac{1}{(1-\rho^2)\sigma_1^2} & \frac{-\rho}{(1-\rho^2)\sigma_1\sigma_2}\\ \frac{-\rho}{(1-\rho^2)\sigma_1\sigma_2} & \frac{1}{(1-\rho^2)\sigma_2^2}\end{pmatrix}\begin{pmatrix}x_1-\mu_1\\ x_2-\mu_2\end{pmatrix}$$

This is in the form $-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$, where:

$$\boldsymbol{\Sigma}^{-1}=\frac{1}{1-\rho^2}\begin{pmatrix}\frac{1}{\sigma_1^2} & \frac{-\rho}{\sigma_1\sigma_2}\\ \frac{-\rho}{\sigma_1\sigma_2} & \frac{1}{\sigma_2^2}\end{pmatrix}$$

We can find $\boldsymbol{\Sigma}$ by computing the inverse of $\boldsymbol{\Sigma}^{-1}$. First, let's calculate the determinant of $\boldsymbol{\Sigma}^{-1}$:

$$\det(\boldsymbol{\Sigma}^{-1})=\frac{1}{(1-\rho^2)^2}\left(\frac{1}{\sigma_1^2}\cdot\frac{1}{\sigma_2^2}-\frac{\rho^2}{\sigma_1^2\sigma_2^2}\right)=\frac{1}{(1-\rho^2)^2}\cdot\frac{1-\rho^2}{\sigma_1^2\sigma_2^2}=\frac{1}{(1-\rho^2)\sigma_1^2\sigma_2^2}$$

Now we can calculate $\boldsymbol{\Sigma}$ using the formula for the inverse of a $2\times 2$ matrix:

$$\boldsymbol{\Sigma}=(1-\rho^2)\sigma_1^2\sigma_2^2\begin{pmatrix}\frac{1}{\sigma_2^2} & \frac{\rho}{\sigma_1\sigma_2}\\ \frac{\rho}{\sigma_1\sigma_2} & \frac{1}{\sigma_1^2}\end{pmatrix}$$

Simplifying, we get:

$$\boldsymbol{\Sigma}=\begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}$$

This reveals that $\boldsymbol{\Sigma}$ is indeed the covariance matrix where:

- Diagonal elements are variances $\sigma_1^2$ and $\sigma_2^2$
- Off-diagonal elements are covariances, expressed as $\rho\sigma_1\sigma_2$

The determinant of this matrix is $|\boldsymbol{\Sigma}|=\sigma_1^2\sigma_2^2(1-\rho^2)$, which completes the transformation to the multivariate normal form:

$$f(\mathbf{x})=\frac{1}{2\pi|\boldsymbol{\Sigma}|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

We found an equivalent expression but in a more compact form where basically everything but the normalizing constant is identical to the multivariate case.

## 1.3 The Covariance Matrix: Encoding Relationships

The covariance matrix $\boldsymbol{\Sigma}$ is the multivariate analog of variance. It captures not only the spread of individual variables but also their interrelationships. For a $p$-dimensional random vector, $\boldsymbol{\Sigma}$ is a $p\times p$ symmetric positive semidefinite matrix:

$$\boldsymbol{\Sigma}=\begin{pmatrix}\sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p}\\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p}\\ \vdots & \vdots & \ddots & \vdots\\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2\end{pmatrix}$$

where:

- $\sigma_i^2 = \text{Var}(X_i)$ is the variance of the $i$-th variable (diagonal elements)
- $\sigma_{ij} = \sigma_{ji} = \text{Cov}(X_i, X_j)$ is the covariance between variables $i$ and $j$ (off-diagonal elements)

The covariance matrix contains all the information about the scale of variation in each dimension and the associations between dimensions. When variables are uncorrelated, all off-diagonal elements are zero, and the covariance matrix is diagonal.

### 1.3.1 Important Special Cases of Covariance Matrices

Several special forms of covariance matrices are particularly useful:

**Property 1.3** (Identity Covariance Matrix)**.** When $\mathbf{\Sigma} = \mathbf{I}$, we have:

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

This represents a standard multivariate normal distribution where all variables have unit variance and are uncorrelated (independent). The resulting density has spherical contours in $p$-dimensional space.

**Property 1.4** (Diagonal Covariance Matrix)**.** When $\mathbf{\Sigma} = \text{diag}(\boldsymbol{s})$, where $\boldsymbol{s} = (s_1, s_2, \ldots, s_p)$:

$$\mathbf{\Sigma} = \begin{pmatrix} s_1 & 0 & \cdots & 0 \\ 0 & s_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & s_p \end{pmatrix}$$

This represents uncorrelated (independent) variables with possibly different variances. The resulting density has ellipsoidal contours with axes aligned with the coordinate axes.

**Property 1.5** (Block Diagonal Covariance Matrix)**.** A block diagonal structure indicates groups of correlated variables that are uncorrelated with variables in other groups:

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{A} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{B} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{C} \end{pmatrix}$$

For example, with two blocks:

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{pmatrix}$$

This structure appears in biology when variables cluster into functional modules or pathways that operate independently.

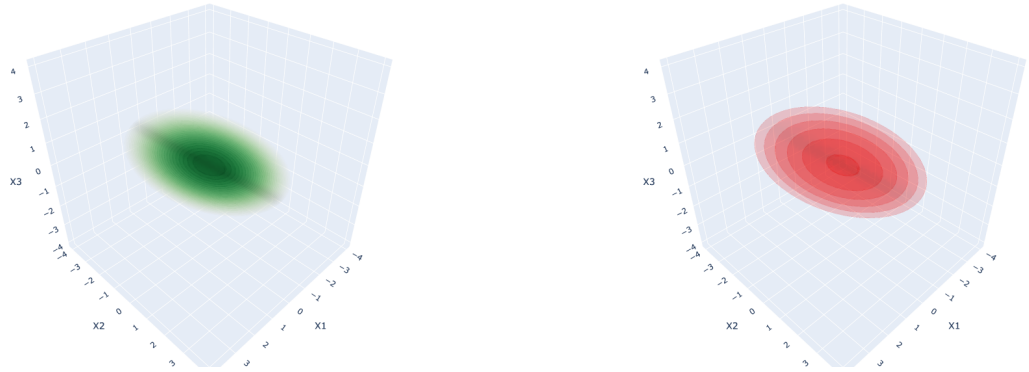
**Example 1.6** (Gene Expression Modules)**.** In single-cell RNA sequencing, we might observe a covariance structure like:

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{\text{cell cycle}} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{\text{metabolism}} \end{pmatrix}$$

This indicates that cell cycle genes covary with each other but not with metabolism genes, and vice versa. Such structure can reveal functional organization in biological systems.

## 1.4 Visualizing MN Distributions

Let's visualize the highest dimensional multivariate normal distribution we can visualize using contours and a density map: the trivariate normal distribution.




(a) Trivariate normal distribution visualization - density map

(b) Trivariate normal distribution visualization - contours (level sets)

Figure 1: **Trivariate normal distribution visualization**

As we move beyond three dimensions, direct visualization becomes challenging. However, the geometric principles remain the same.

## 1.5 MN Level Sets: Ellipsoids

You recall that level sets of the univariate normal distribution are simply points on a line, and for the bivariate case, they are ellipses in a plane. in the multivariate normal distribution creates ellipsoids in higher dimensions. Contours of equal density in $p$-dimensional space.

If you remember the lecture about the bivariate normal distribution, you probably can guess, even if you never saw this formula that the formula for an ellipsoid can be expressed in linear algebra as:

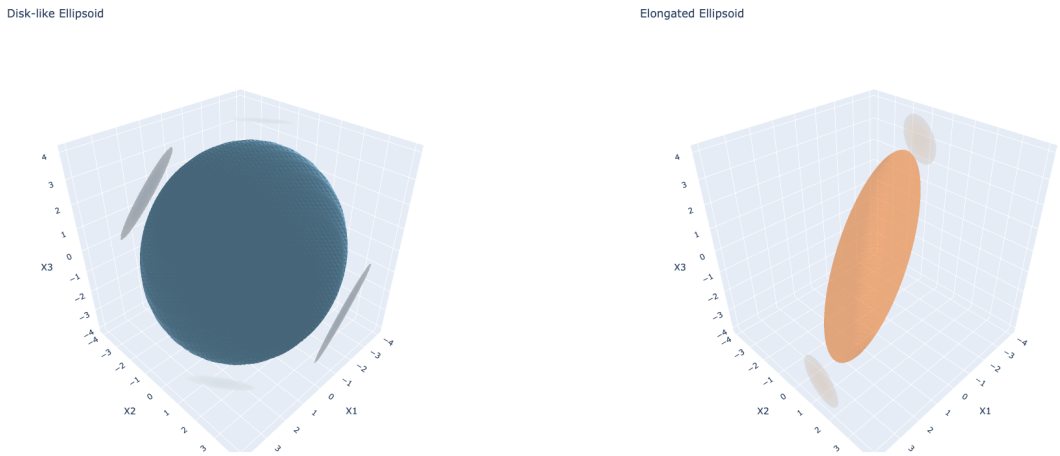
$$(\mathbf{x} - \mathbf{x}_c)^T \mathbf{M}^{-1} (\mathbf{x} - \mathbf{x}_c) = 1$$

with $\mathbf{x}_c$ being the center of the ellipsoid and $\mathbf{M}$ determines the shape and orientation of the ellipsoid. Note that M needs to be positive definite. Let's remember the definition:

**Definition 1.7** (Positive Definite Matrix)**.** A symmetric matrix $\mathbf{M}$ is positive definite if for all non-zero vectors $\mathbf{x}$, the quadratic form $\mathbf{x}^T \mathbf{M} \mathbf{x} > 0$.

In other words if it does not have negative eigenvalues. Geometrically: it represents a transformation that does not "flip the space".

The basic correspondence is that $\mu$ is the center of the ellipsoid, and $\mathbf{\Sigma}^{-1}$ determines the shape and orientation.




(a) Ellipsoid with disk-like shape

(b) Ellipsoid with elongated shape

Figure 2: **Examples of Ellipsoids in 3D**

# 2 Estimation in the Multivariate Normal

## 2.1 Sample Mean Vector and Covariance Matrix

Let's consider a dataset of $n$ observations, each a $p$-dimensional vector. Let's assume a Multivariate Normal distribution as a good model for our data. We want to estimate the parameters of this distribution, specifically the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$. The sample mean vector and covariance matrix are the natural estimators for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively.

- The sample mean vector is:

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$$

- The sample covariance matrix is:

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T$$

  If we indicate the centered matrix $\mathbf{X}_c = \mathbf{X} - \hat{\boldsymbol{\mu}}$, we can write it as:

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\mathbf{X}_c^T\mathbf{X}_c$$

These estimators are unbiased and, when the data truly follow a multivariate normal distribution, they are the maximum likelihood estimators (after adjusting the denominator in the covariance formula from $n-1$ to $n$).

## 2.2 Challenges in High Dimensions

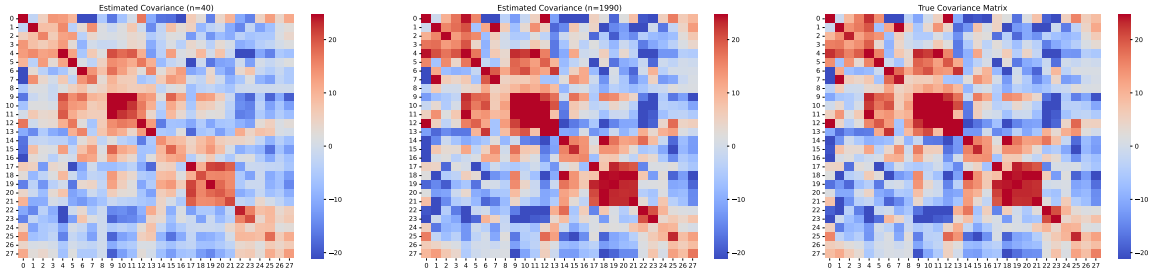
While these estimators work well in low dimensions with ample data, they face significant challenges in high-dimensional settings common in modern biology.

When the number of variables $p$ approaches or exceeds the sample size $n$, the sample covariance matrix becomes ill-conditioned or singular, meaning its inverse (required for density calculations) does not exist or is numerically unstable.

This phenomenon reflects the "curse of dimensionality" - as dimensionality increases, the volume of the space increases so rapidly that the available data becomes sparse. Definition:

**Definition 2.1** (Curse of Dimensionality)**.** The curse of dimensionality refers to various phenomena that arise when analyzing and organizing data in high-dimensional spaces that do not occur in low-dimensional settings. In particular, it describes the exponential increase in volume associated with adding extra dimensions to a mathematical space, leading to sparsity of data points.

Estimating a full $p \times p$ covariance matrix correspond to estimating $p(p+1)/2$ parameters (due to symmetry), which means that a sample size of $n$ must be larger than $p(p+1)/2$ to ensure a well-defined estimate.




(a) Estimated w/ 50 points (b) Estimated w/ 2000 points (c) True covariance matrix

Figure 3: **Covariance estimation in high dimensions**

## 2.3 Flipping the matrix

Let's take a step back and look at the data matrix $\mathbf{X}$ and question a bit the orientation we are using it in. Say that $X \in \mathbb{R}^{n \times p}$, with $n > p$, the data matrix from a single-cell RNA-seq experiment, where $n$ is the number of cells (the number of observations) and $p$ is the number of genes (the number of variables). We could consider its transpose $\mathbf{X}^T \in \mathbb{R}^{p \times n}$, where now which are the observations and which are the variables is flipped and, also here, we could chose a Multivariate Normal distribution to model the data $\mathbf{X}^T$. In other words, instead of studying the covariance of the genes, we could study the covariance of the cells.

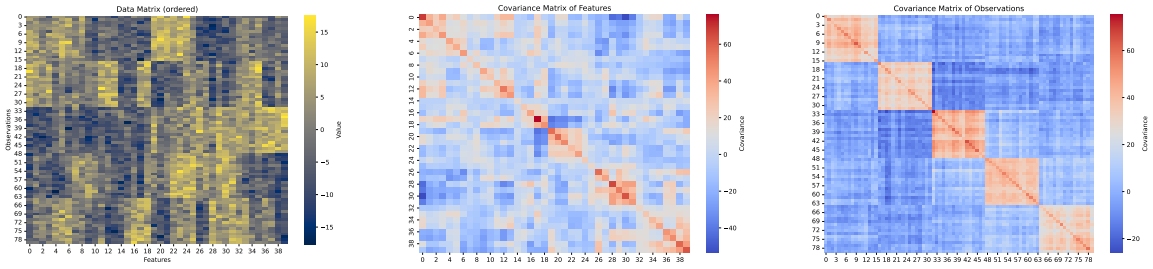
Therefore, there are two ways to look at the data matrix $\mathbf{X}$:

$$\hat{\mathbf{\Sigma}} = \frac{1}{n-1}\mathbf{X}_c^T\mathbf{X}_c$$

$$\hat{\mathbf{\Sigma}}_{\text{obs}} = \frac{1}{p-1}\mathbf{X}_{c'}\mathbf{X}_{c'}^T$$

Where we have:

- $\mathbf{X}_c$ is the centered data matrix where each column is centered by subtracting the mean of that column.
- $\mathbf{X}_{c'}$ is the centered data matrix where each row is centered by subtracting the mean of that row.




(a) Data matrix (b) Covariance of features (c) Covariance of observations

Figure 4: **Covariance estimation in high dimensions**

## 2.4 Geometry of the Multivariate Normal

We discussed level sets, let us study the level set with $c = 1$.

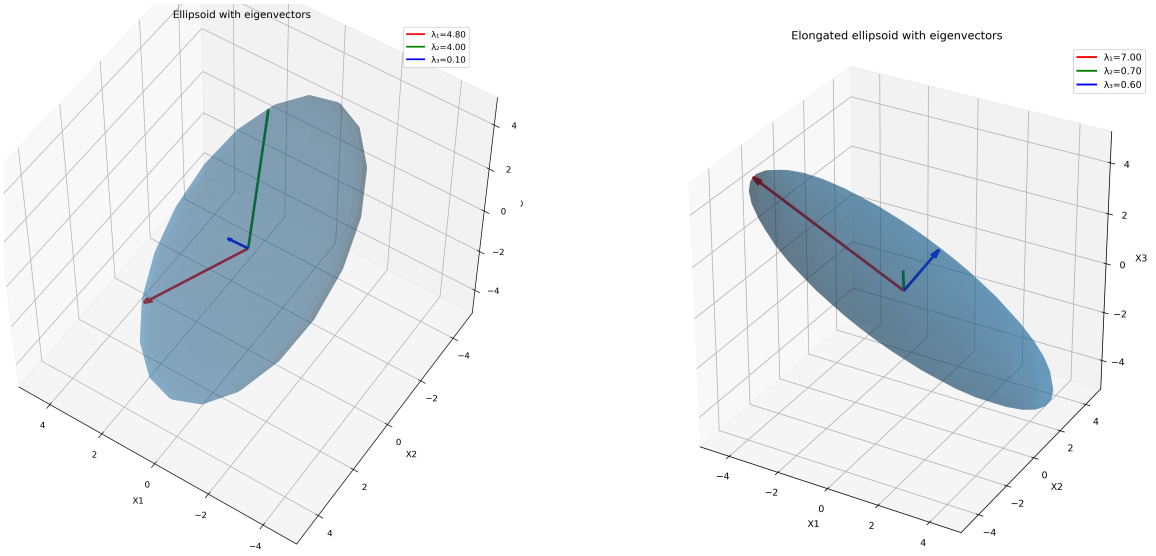
It is an ellipsoid with volume $V = \sqrt{|\mathbf{\Sigma}|}V_{ball}$ where $V_{ball}$ is the volume of a unit ball in $p$ dimensions. But which ellipsoid? A simple way to characterize it is to be able to say something about the major and minor axes of the ellipsoid.

Let's consider the equation $\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x} = 1$. Recall that a semiaxis of the ellipsoid is defined as the segment from the center to a point on the surface, along a direction that is preserved (only scaled, not rotated) when transforming the ellipsoid back into a sphere.

From linear algebra this sounds familiar and is expressed by the equation $Av = \lambda v$, which is the eigenvalue equation. Recall also that the eigenvalues $\lambda_i$ of $\mathbf{\Sigma}^{-1}$ are the reciprocals of the eigenvalues of $\mathbf{\Sigma}$, and the eigenvectors $v_i$ of $\mathbf{\Sigma}^{-1}$ are the same as those of $\mathbf{\Sigma}$.

So we have gathered enough information to list two of important properties:

- The eigenvalues of $\mathbf{\Sigma}$ determine the lengths of the semi-axes of the ellipsoid. Specifically the semi-axis $i$ of the level set with $c = 1$ has length $\sqrt{\lambda_i}$, where $\lambda_i$ is the $i$-th eigenvalue of $\mathbf{\Sigma}$.
- The eigenvectors of $\mathbf{\Sigma}$ determine the orientation of the ellipsoid in $p$-dimensional space. Specifically sorting the eigenvalues in descending order of eigenvalues gives the ordered semi-axes.




(a) Ellipsoid with eigenvectors and eigenvalues (b) Ellipsoid with eigenvectors and eigenvalues

Figure 5: **Ellipsoid with eigenvectors and eigenvalues**

This is a powerful result because it allows us to understand the geometry of the multivariate normal distribution in terms of its covariance structure. It is expecially useful when we have a dataset in multiple dimensions and we want to understand the relationships between the variables.

## 2.5 Interpretation: Principal Components

The eigenvalues and eigenvectors of the covariance matrix $\mathbf{\Sigma}$ can be interpreted in terms of principal components.

$$\mathbf{\Sigma}\mathbf{v}_i = \lambda_i \mathbf{v}_i \quad \text{where} \quad \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$$

The principal components are the directions in which the data varies the most. The first principal component corresponds to the direction of maximum variance, the second principal component is orthogonal to the first and captures the second most variance, and so on.

$$\mathbf{v}_1 = \arg\max_{\|\mathbf{v}\|=1} \text{Var}(\mathbf{v}^T \mathbf{X})$$

The eigenvalues of the covariance matrix represent the amount of variance explained by each principal component.

$$\text{Variance explained by } \mathbf{v}_i = \lambda_i$$

The eigenvectors represent the directions of these principal components in the original variable space.

$$\text{Total variance} = \sum_{i=1}^{p} \lambda_i \quad \text{and} \quad \text{Proportion explained by } \mathbf{v}_i = \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}$$

This interpretation is particularly useful in high-dimensional data analysis, where we often seek to reduce dimensionality while preserving as much variance as possible.

$$\mathbf{X}_k \approx \sum_{i=1}^{k} \mathbf{v}_i \mathbf{v}_i^T \mathbf{X} \quad \text{where} \quad k \ll p$$

## 2.6 Analysis of eigenvalues

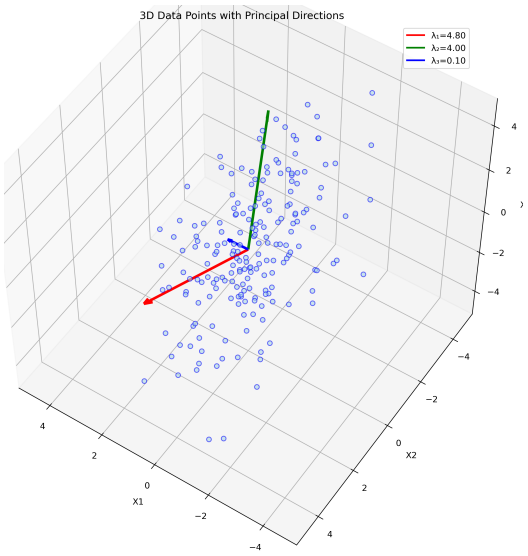
The eigenvalues of the covariance matrix $\mathbf{\Sigma}$ can be analyzed to understand the structure of the data. The eigenvalues $\lambda_i$ are non-negative and can be ordered as $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$.

The eigenvalues can be interpreted as the amount of variance explained by each principal component. The rank of the covariance matrix is equal to the number of non-zero eigenvalues.

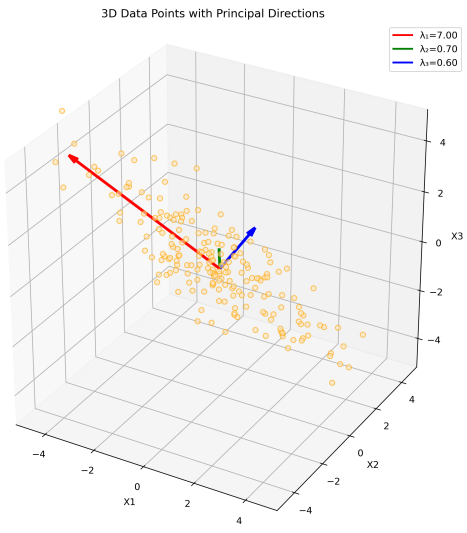
These plots can be used to visualize the rank of the covariance matrix and the amount of variance explained by each principal component. What does it mean to have a rank $k$ covariance matrix? It means that all our variables have a strong correlation-anticorrelation structure. Where basically there are $k$ groups of variables that are strongly correlated with each other.

In these case one talks about the presence of $k$ latent factors that explain the data.

Here, we can see that the eigenvalues decay rapidly, indicating that most of the variance is explained by a small number of principal components. The eigenvalues equal to 0 correspond to components that don't contribute to the variance of the data (they emerge from the existence of strictly correlated variables in the dataset). (Please note that these plots are generated from simulated data, in real data the eigenvalues "equal to zero" are not equal to exactly zero because of noise in the data.)

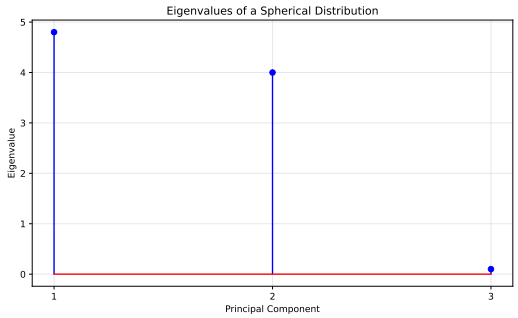



(a) A 3D dataset with principal directions

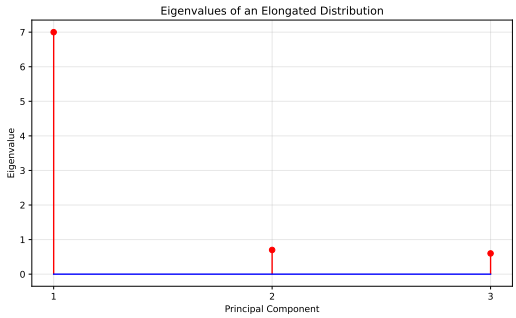



(b) Another 3D dataset with principal directions

Figure 6: **A 3D dataset with principal directions**

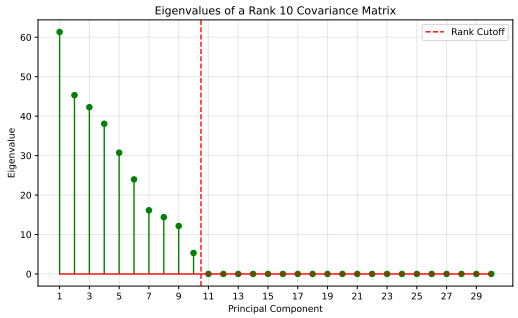



(a) Eigenvalues of a "disk" distribution

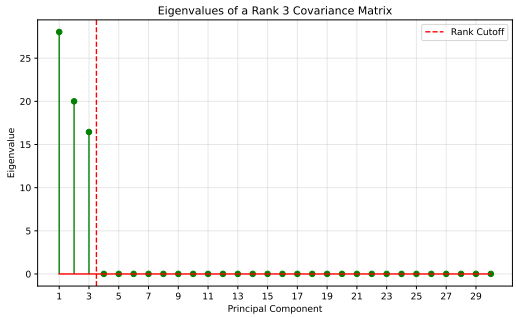



(b) Eigenvalues of an elongated distribution

Figure 7: **Eigenvalues of covariance matrices**




(a) Eigenvalues of a rank 10 covariance matrix




(b) Eigenvalues of a rank 3 covariance matrix

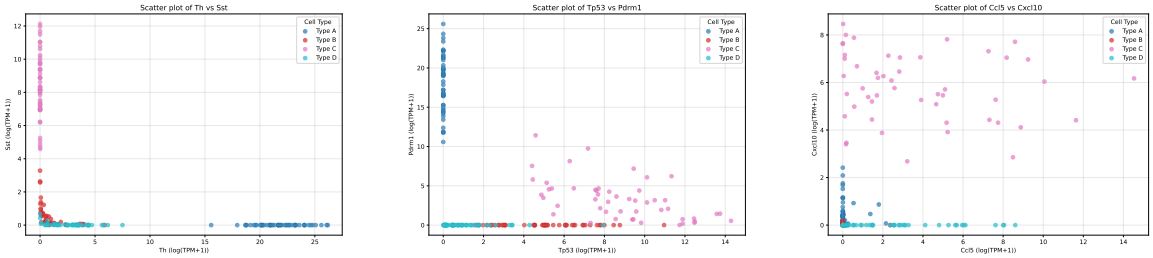
Figure 8: **Eigenvalues of low-rank covariance matrices**

# 3 Principal Component Analysis (PCA)

## 3.1 An anticipation

Before getting to lost in further analysis it is worth to anticipate that the theory of what we just discussed and what we will discuss will lead us to remarkable analytical "powers".

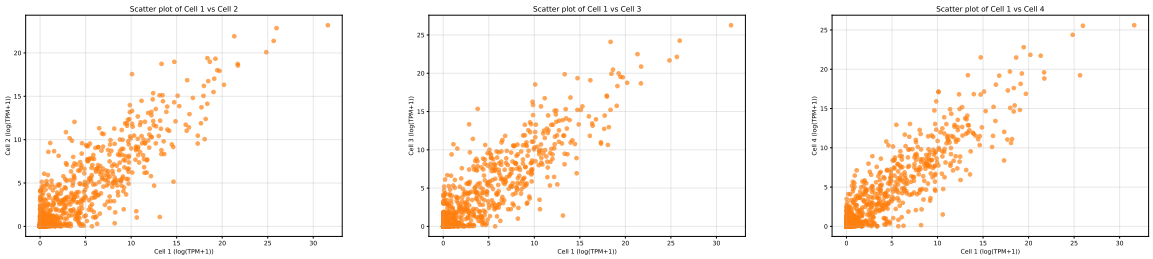
To see this, let's consider the following example: We have a dataset of $n$ observations, each a $p$-dimensional vector. Say, again, our single-cell RNA-seq experiment, where $n$ is the number of cells (the number of observations) and $p$ is the number of genes (the number of variables). With your previous knowledge, of analytical methods, if given this dataset you would probably approach "having a look at the data" by plotting scatters of pairs of genes against each other, where points are cells, maybe color-coded by some metadata (e.g. time point).




(a) Scatter plot of Th vs Sst (b) Scatter plot of Tp53 vs Pdrm1 (c) Scatter plot of Ccl5 vs Cxcl10

Figure 9: **Scatter plots of genes**

You could also consider plotting the other scatters, the one where you compare pairs of cells.




(a) Scatter plot of cell 1 vs cell 2 (b) Scatter plot of cell 1 vs cell 3 (c) Scatter plot of cell 1 vs cell 4

Figure 10: **Scatter plots of cells**

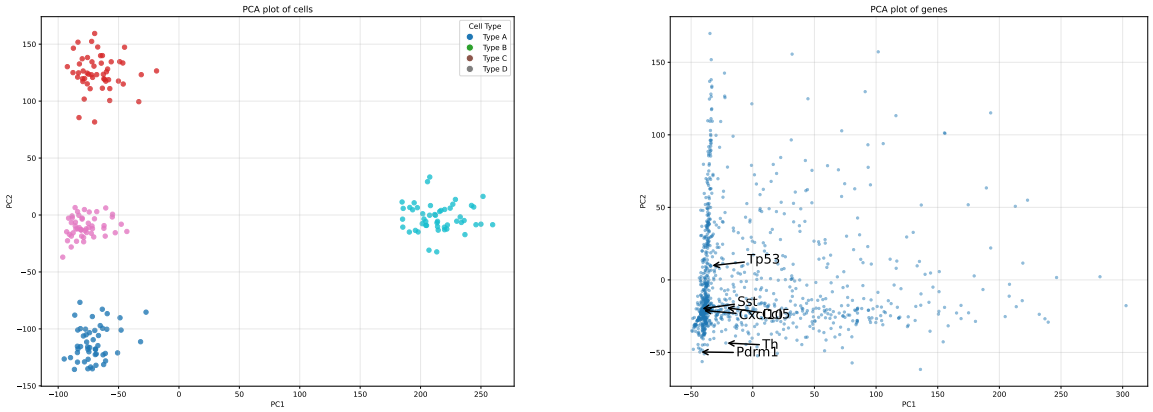
You might also consider studying mutual information matrices to understand pairwise relationships between genes, or evaluate meaningful measures of distance between cells.

But what if you could do something more powerful, to get a more global view of the data? I am going to anticipate it now and we are going to understand this in the next sections.

## 3.2 From Eigendecomposition to PCA

Building on our understanding of the multivariate normal distribution and its geometry, we can now explore a powerful technique for analyzing high-dimensional data: Principal Component Analysis (PCA). While we have discussed the eigendecomposition of the covariance matrix from a theoretical perspective, in practice there are computational and numerical considerations that lead us to a slightly different approach.

The eigendecomposition of the covariance matrix gives us valuable insight into the structure of our data, but computing the sample covariance matrix explicitly can be problematic for several reasons:




(a) PCA plot of cells

(b) PCA plot of genes

Figure 11: **PCA plot of cells and genes**

- For large datasets with many variables, computing and storing the full $p \times p$ covariance matrix can be memory-intensive
- When $p > n$ (more variables than observations), the sample covariance matrix is necessarily singular and has at most rank $n$
- Numerical stability issues can arise when computing eigenvalues and eigenvectors of ill-conditioned matrices
- In terms of computational complexity, the eigendecomposition of a $p \times p$ matrix is $O(p^3)$, while SVD is $O(np^2)$, with even more scalable "online" algorithms available.

These practical challenges motivate us to consider an alternative but equivalent approach based on the Singular Value Decomposition (SVD) of the data matrix itself.

## 3.3 Singular Value Decomposition (SVD)

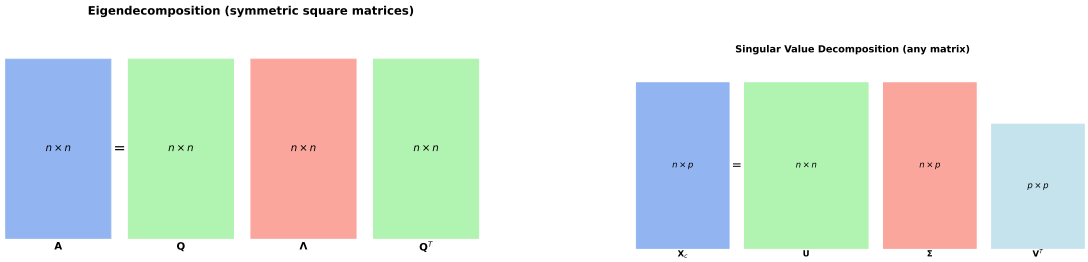
The SVD provides a powerful factorization of any matrix, not just data matrices. For our centered data matrix $\mathbf{X}_c \in \mathbb{R}^{n \times p}$, the SVD gives us:

$$\mathbf{X}_c = \mathbf{U}\mathbf{S}\mathbf{V}^T$$

where:

- $\mathbf{U} \in \mathbb{R}^{n \times n}$ is an orthogonal matrix whose columns are the left singular vectors
- $\mathbf{S} \in \mathbb{R}^{n \times p}$ is a rectangular diagonal matrix containing the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min(n,p)} \geq 0$
- $\mathbf{V} \in \mathbb{R}^{p \times p}$ is an orthogonal matrix whose columns are the right singular vectors

The beauty of this decomposition becomes apparent when we connect it to the eigendecomposition of the covariance matrix:




(a) Eigendecomposition of a matrix

(b) SVD decomposition of a matrix

Figure 12: **Eigendecomposition vs SVD**

$$\hat{\mathbf{\Sigma}} = \frac{1}{n-1}\mathbf{X}_c^T\mathbf{X}_c \tag{1}$$

$$= \frac{1}{n-1}\mathbf{V}\mathbf{S}^T\mathbf{U}^T\mathbf{U}\mathbf{S}\mathbf{V}^T \tag{2}$$

$$= \frac{1}{n-1}\mathbf{V}\mathbf{S}^T\mathbf{S}\mathbf{V}^T \tag{3}$$

$$= \mathbf{V}\left(\frac{1}{n-1}\mathbf{S}^T\mathbf{S}\right)\mathbf{V}^T \tag{4}$$

$$\tag{5}$$

Since $\mathbf{U}$ is orthogonal, $\mathbf{U}^T\mathbf{U} = \mathbf{I}$, and $\mathbf{S}^T\mathbf{S}$ is a diagonal matrix with entries $\sigma_i^2$.
This reveals that:

- The right singular vectors $\mathbf{V}$ from the SVD are exactly the eigenvectors of the covariance matrix
- The singular values $\sigma_i$ relate to the eigenvalues of the covariance matrix as $\lambda_i = \sigma_i^2/(n-1)$

The SVD gives us the principal components without explicitly forming the covariance matrix, offering computational advantages especially for high-dimensional data.

## 3.4 The Dual Nature of PCA: Variables and Observations

PCA has a fascinating duality, which we hinted at earlier when discussing the transposition of the data matrix. Let's explore this more deeply now.

When we compute PCA, we obtain two sets of vectors:

- The right singular vectors $\mathbf{V}$ (columns of $\mathbf{V}$), which are the eigenvectors of $\mathbf{X}_c^T\mathbf{X}_c$. These describe directions in the variable space.
- The left singular vectors $\mathbf{U}$ (columns of $\mathbf{U}$), which are the eigenvectors of $\mathbf{X}_c\mathbf{X}_c^T$. These describe directions in the observation space.

These two perspectives offer complementary insights into our data's structure.

### 3.4.1 Variable Space (Right Singular Vectors)

In the variable space, the principal components represent new axes that define a convenient coordinate system. Each axis (principal component) is a linear combination of the original variables.

For gene expression data, these new axes can be thought of as "metagenes" - patterns of gene expression that vary together across cells. These axes are chosen specifically to align with the directions of maximum variance in the data.

For instance, the first principal component axis might align with coordinated expression of cell cycle genes, while the second might align with stress response genes. These orthogonal axes reveal the underlying structure of gene co-expression.

The coefficients that define these new axes are called "loadings" and are given by the entries in the right singular vectors:

$$\text{Loading of variable } j \text{ on PC } k = V_{jk}$$

These loadings tell us how much each original variable contributes to defining each new axis.

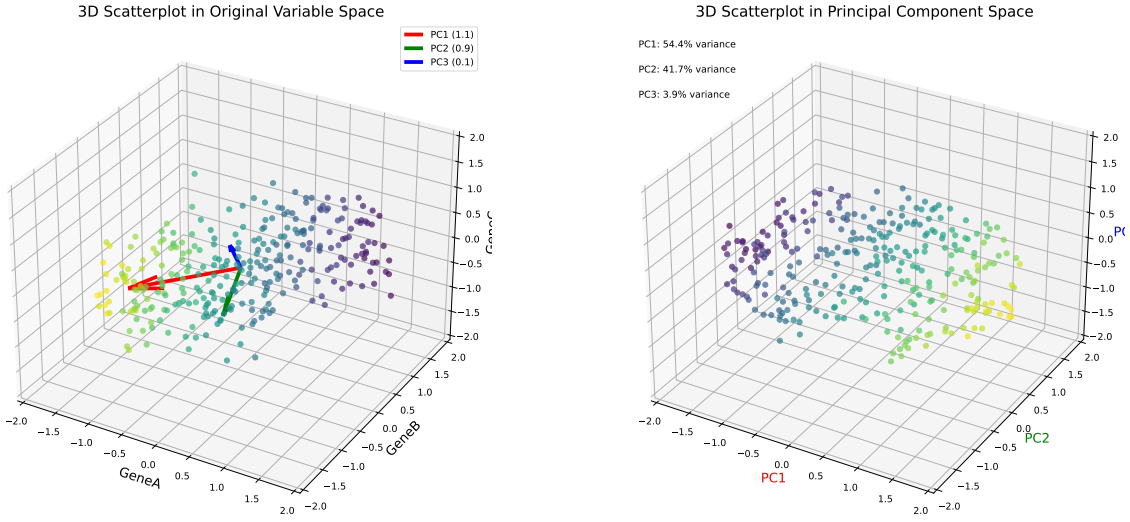
### 3.4.2 Observation Space (Left Singular Vectors)

In the observation space, PCA provides new coordinates for each observation along these principal component axes. Vectors (i.e. points in the visualization below) represent the observations (e.g., cells), and their coordinates are given by the left singular vectors. For example, in single-cell data, each cell gets a set of coordinates in this new system.

The coordinates of each observation in the principal component coordinate system are given by:

$$\text{PC score for observation } i \text{ on PC } k = U_{ik} \cdot \sigma_k$$

These "PC scores" are effectively the coordinates of each observation in the new principal component space. They allow us to visualize high-dimensional data in a lower-dimensional coordinate system where the axes are ordered by importance (variance explained), often revealing clusters or trajectories that were hidden in the original space.




(a) 3D scatterplot in original variable space

(b) 3D scatterplot where axes are PC1-2-3

Figure 13: **PC coordinates as a change of basis**

This geometric interpretation makes PCA powerful for visualization and analysis: it creates a new coordinate system where the axes are optimally aligned to capture the data's variance. The

first few axes often capture the most important structure, allowing effective dimensionality reduction while preserving relationships between observations.