Lecture 14
Principal Component Analysis

Jochen Braun

Otto-von-Guericke-Universität Magdeburg,
Cognitive Biology Group

Engineering Neuroscience / Computational Neuroscience II
SS 2020

Credits: Jonathan Shlens (2005) "A Tutorial on Principal Component Analysis"

14. Principal Component Analysis (PCA)

PCA is a useful way to summarize high-dimensional data (repeated observations of multiple variables). This lecture provides the underlying linear algebra needed for practical applications. It also emphasizes consistent notation. The central ideas of PCA are **orthonormal coordinate** *systems, the distinction between* **variance and covariance**, *and the possibility of choosing an orthonormal basis to* **eliminate covariance**. *Technically, PCA may be performed either by* **eigenvector analysis** *of the covariance matrix or by* **singular value decomposition** *of the original observation matrix. Both variants are described. We illustrate the method with a non-biological example (image-denoising) and with a biological example (multi-unit activity, see* **Exercise 8**).

Overview

Motivation

Principal component analysis (PCA) is one of the most valuable results of applied linear algebra.

It is widely used – from neuroscience to computer graphics – because it is an easy way to simplify confusing data sets.
With minimal effort, PCA dramatically reduces the dimensionality of a large data sent and potentially reveals a simpler structure hiding within.

During brain development, an (approximate) PCA of sensory inputs is performed by activity-dependent Hebbian plasticity.

Variables and observations

We distinguish between *variables* (types of measurements) and *observations* (times or conditions of measurements).
Think of variables as 'effects' (possibly mixed, redundant, or noisy) and observations as variations over 'causes' (due to differences in time, condition, or other factors).

For example, what kind of food people eat is a *variable*, in that it is an 'effect' of many causal factors such as climate, culture, economics, and so on.

When considering different countries, historical periods, or climate zones, these causal factors will presumably differ from case to case. The different cases being compared may be considered *observations*.

1. British food consumption

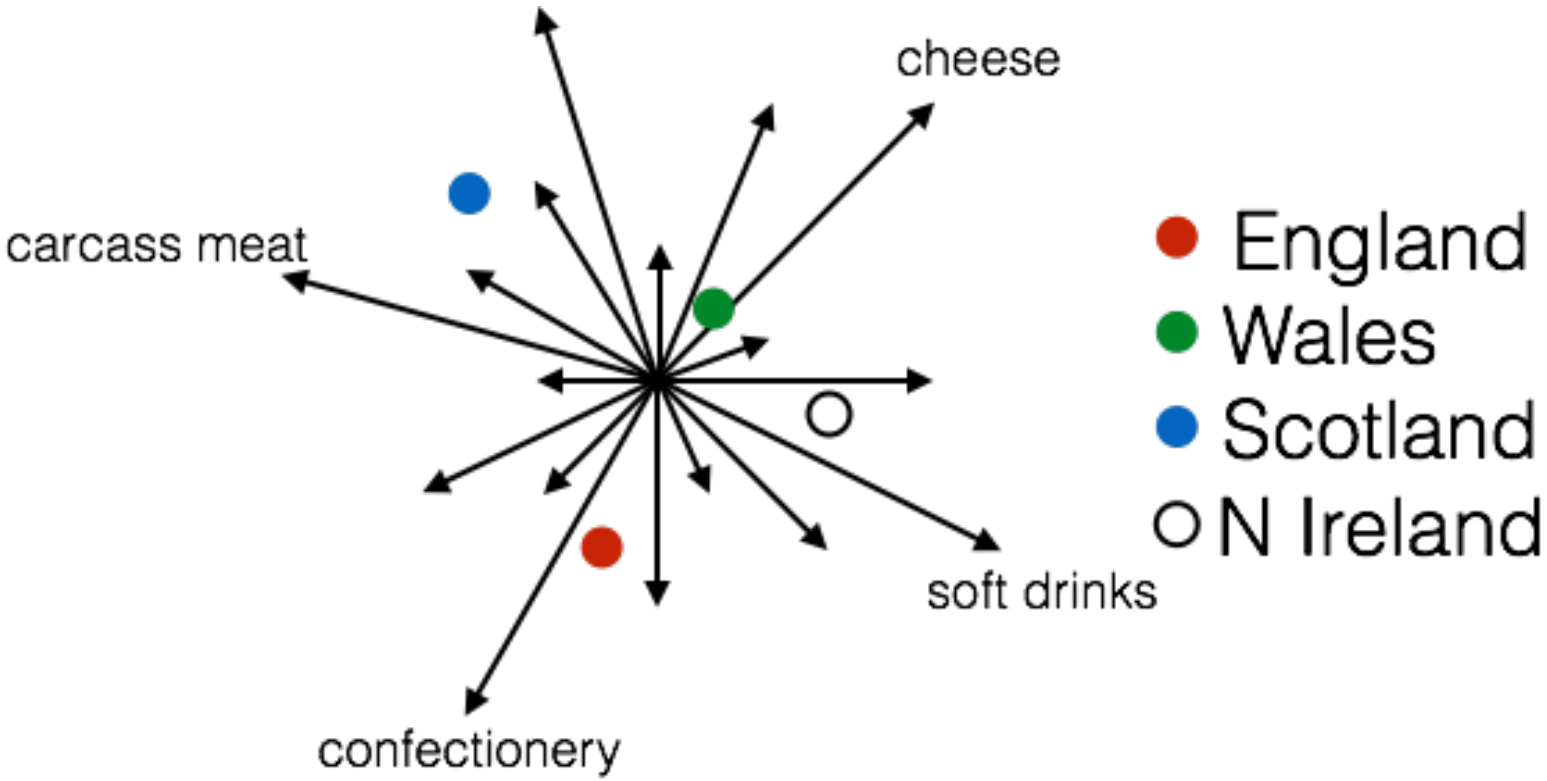
Consumption of different foods in Britain, in g/person/week:

	Eng	Wal	Scot	N Ire
cheese	105	103	103	66
carcass meat	245	227	242	267
other meat	685	803	750	586
fish	147	160	122	93
fats and oils	193	235	184	209
sugars	156	175	147	139
fresh potatoes	720	874	566	1033
fresh veg	253	265	171	143
other veg	488	570	418	355
processed potatoes	198	203	220	187
processed veg	360	365	337	334
fresh fruit	1102	1137	957	674
cereals	1472	1582	1462	1494
hot beverages	57	73	53	47
soft drinks	1374	1256	1572	1506
alcoholic drinks	375	475	458	135
confectionery	54	64	62	41

We consider foods as *variables* and countries as *observations*! (Credits: Mark Richardson).

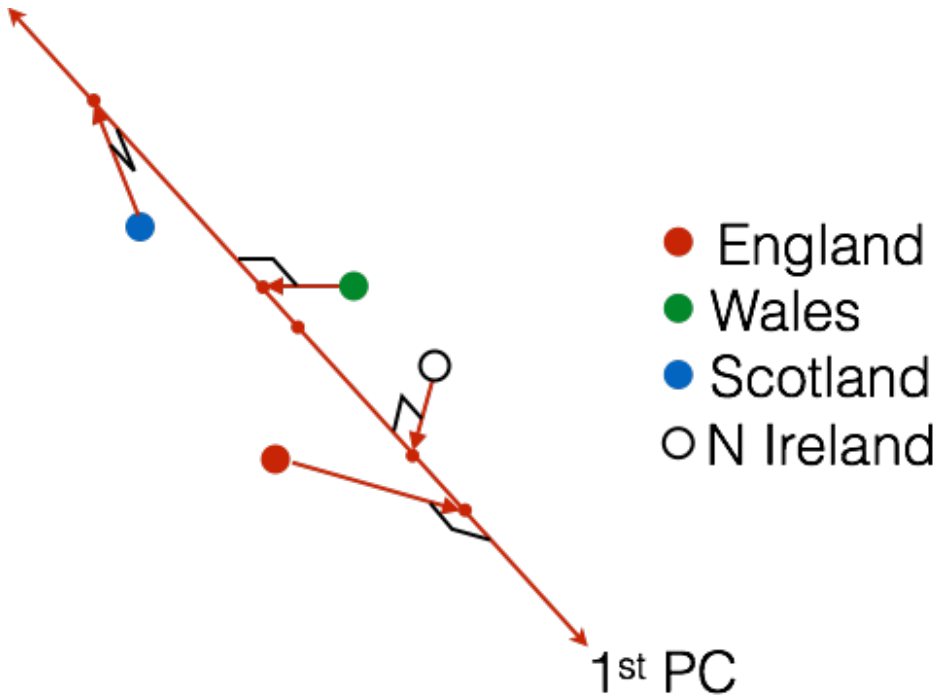
4 points in 17-dimensional space

We form an orthonormal basis of 17 food types and represent 4 observations of 17 variables as four points in this space (one point per observation):



Project to special line

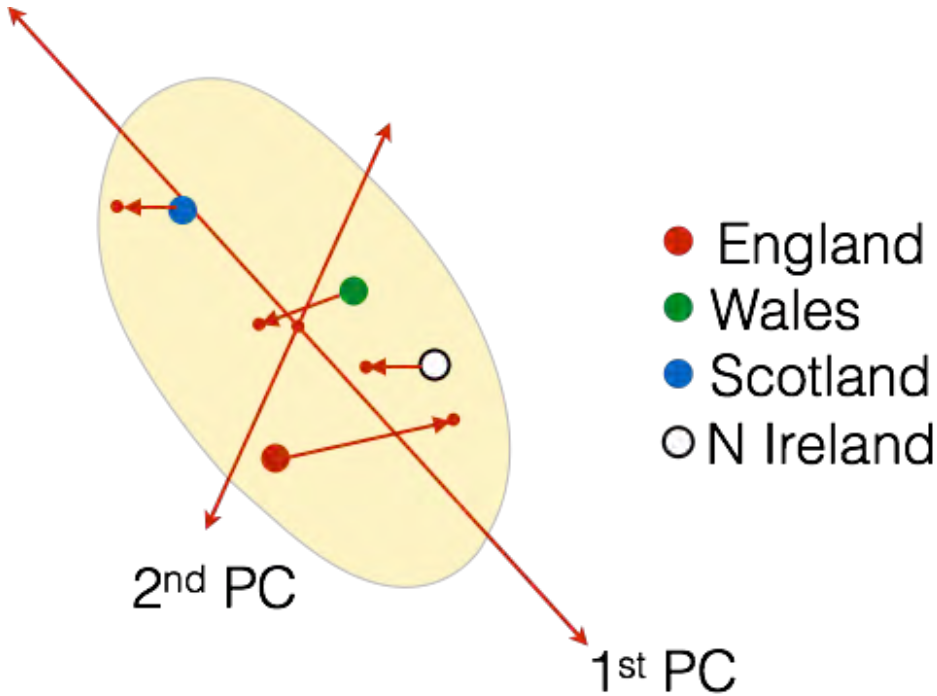
We may project our observations orthogonally on any line in 17D space. We seek the particular line along which the projections scatter most widely (show maximal variance).



We term this special line the *1st principal component* .

Project to special plane

Similarly, we may project orthogonally on any plane in 17D space. Again we seek a particular plane: the plane that spans the two orthogonal lines along which the projections scatter most widely and second most widely.



We term these orthogonal lines the *1st and 2nd principal component* .

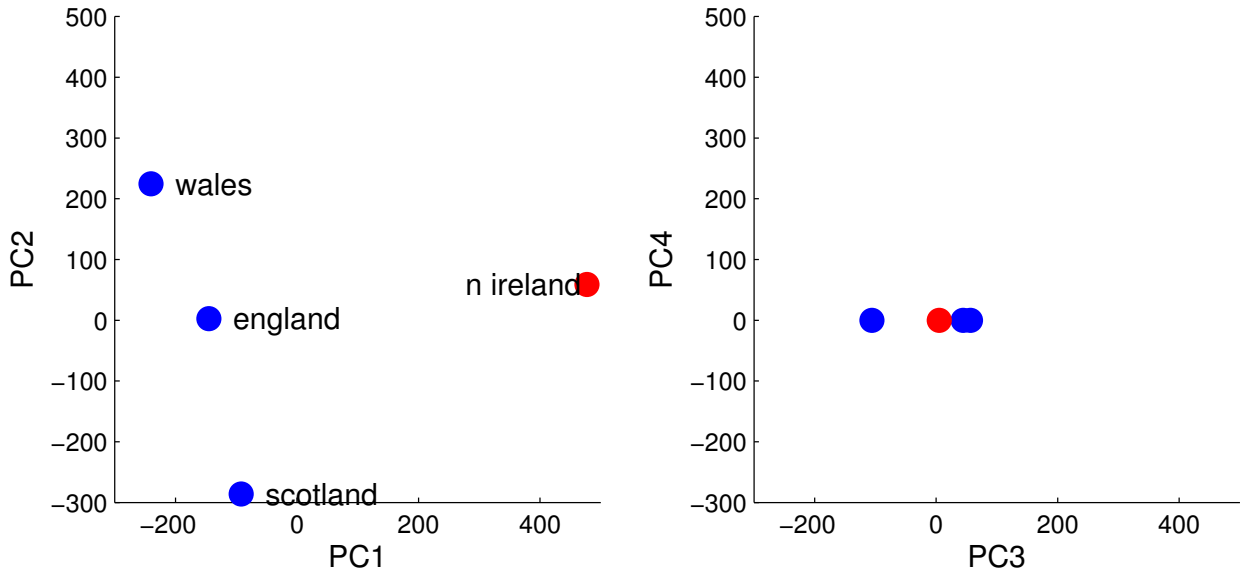
Maximal variance between observations (countries)

The axis of maximal variance (1st principal component) between countries is here shown horizontally

From this perspective, it is evident that England, Wales, and Scotland (blue) are comparatively similar, whereas N Ireland (red) is comparatively distinct.

Principal components compared

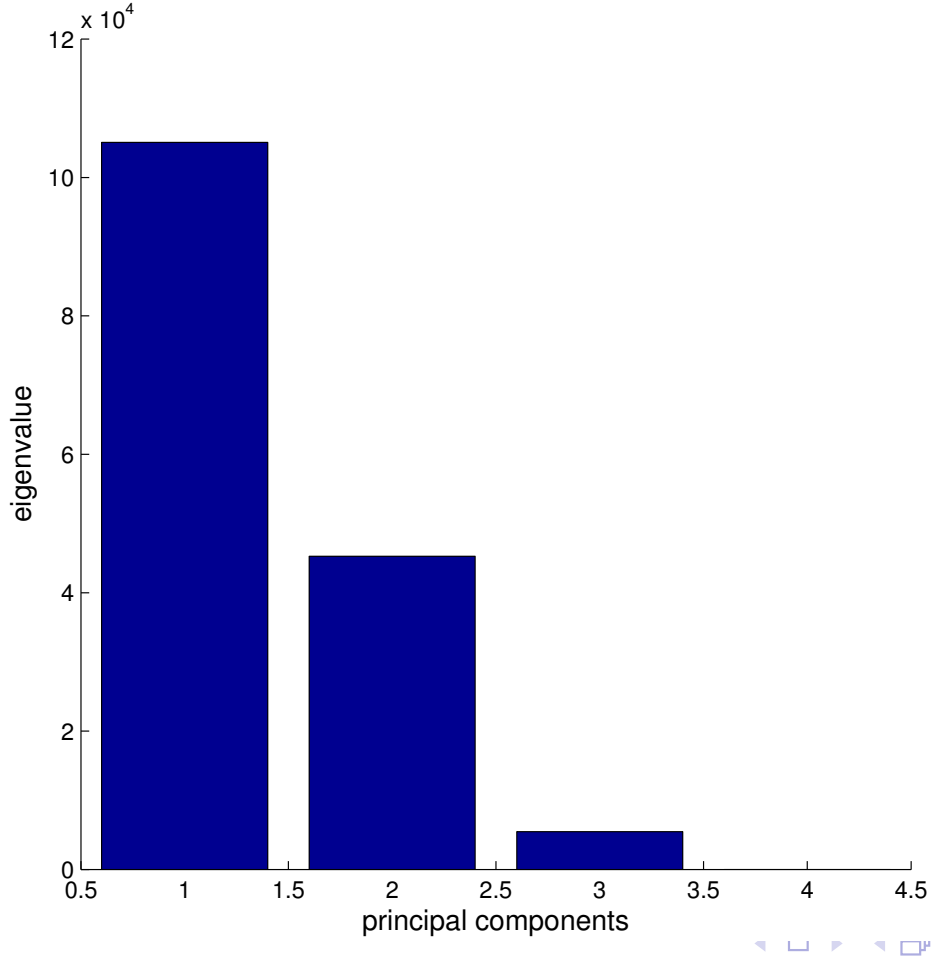
In total, there are 17 'principal components'. Together, they form an alternative orthonormal basis for our space. The 1^{st} and 2^{nd} principal components are shown on the left, the 3^{rd} and 4^{th} on the right:



Note that the first two PCs capture most of the variance between observations (countries).

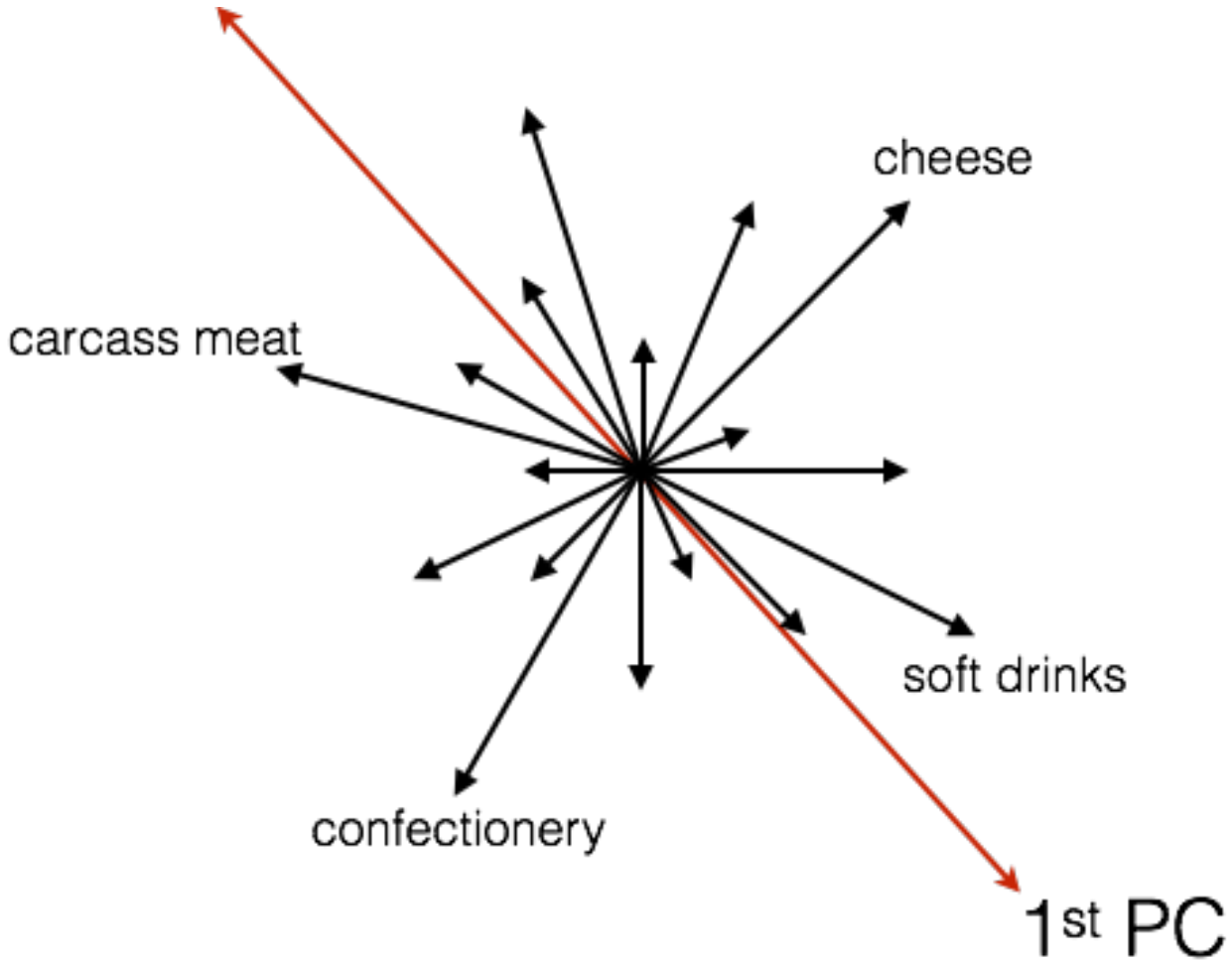
Variance captured

The relative importance of different principal components is measured by the variance captured. In our example, only four principal components capture variance. (This is because there were only 4 observations).



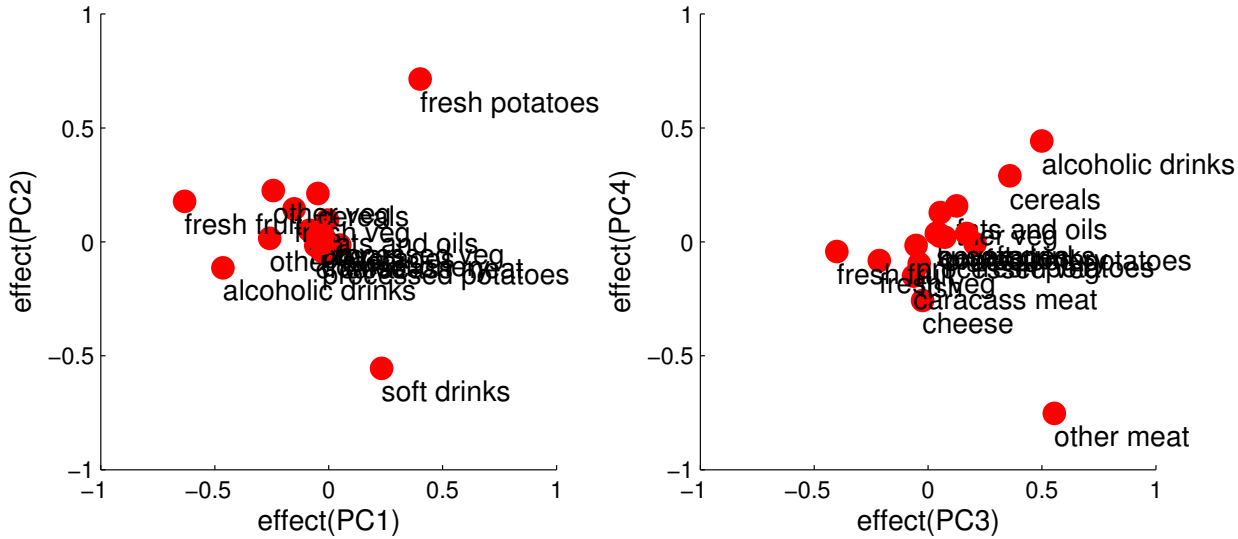
Contributions to principal component

A ‘principal component’ is linear combination of variables. In our example, it is a direction in the 17D space of food types.



Visualizing effects

To visualize the contributions ('effects') of different variables (foods) to a principal component, we can project the original 17 orthonormal basis vectors onto the PC of interest:



The largest effect on PC1 have fresh fruit and alcoholic drinks (both of which are consumed less in N Ireland). The largest effect on PC2 have fresh potatoes and soft drinks.

Summary

- In many fields (including neuroscience), we observe a multitude of variables under different conditions or at different times.
- It often makes sense to seek a simpler way to describe of the total variance (sum of variances over condition or time).
- PCA considers observations (conditions or times) as points in a high-dimensional space, the dimensionality of which is determined by the number of variables.
- PCA computes an alternative orthonormal basis for this space and it ranks the basis vectors ('principal components') by the variance captured.
- In effect, it identifies the points of view (directions of projection) from which observations appear most variable.

www.models.life.ku.dk
https://www.youtube.com/user/QualityAndTechnology/playlists
Principal Component Analysis – PCA

https://www.youtube.com/watch?v=K-F19DORO1w&list=PLBC24FD8C389FE9E4

8:20 to 10:20

2 Orthonormal bases

Consider m variables, **each with zero mean** and observed n times, with i^{th} observations grouped into **column** vectors (for now!)

$$\hat{\boldsymbol{x}}_i = \begin{pmatrix} x_{1i} \\ \vdots \\ x_{mi} \end{pmatrix} \qquad \sum_{i=1}^{n} x_{ji} = 0, \qquad i = 1 \ldots n, \qquad j = 1 \ldots m$$

and all n observations collected into a matrix with N columns and M rows

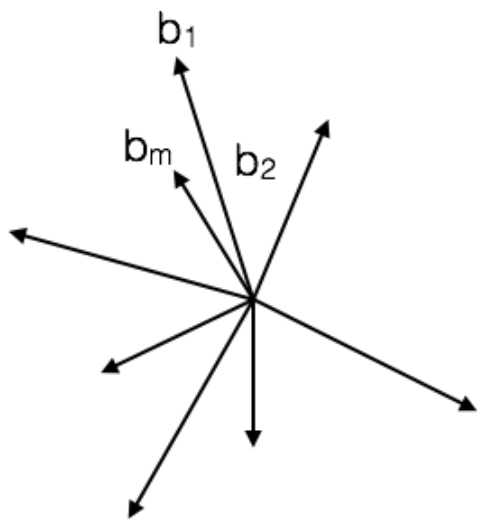
$$\boldsymbol{X} = (\hat{\boldsymbol{x}}_1 \quad \hat{\boldsymbol{x}}_2 \quad \ldots \quad \hat{\boldsymbol{x}}_n) = \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{pmatrix} \in \mathbb{R}^{m \times n}$$

Note the indexing with $x_{row,column}$ or $x_{variable,observation}$:
rows $\simeq$ **variables** and **columns** $\simeq$ **observations**

Orthonormal basis

Implictly, this assumes a basis of m orthonormal **row vectors** $\boldsymbol{b}_j = (b_{11} \ldots b_{1m})$, each with m elements, $\mathbf{B} \in \mathbb{R}^{m \times m}$:

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \\ \vdots \\ \boldsymbol{b}_m \end{pmatrix} = \begin{pmatrix} b_{11} & b_{12} & \ldots & b_{1m} \\ b_{21} & b_{22} & \ldots & b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1} & b_{m2} & \ldots & b_{mm} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix} = \boldsymbol{I}$$



'Orthonormal' means of unit length and pairwise orthogonal

$$\boldsymbol{b}_j \cdot \boldsymbol{b}_k = \delta_{jk} \qquad \boldsymbol{B}\boldsymbol{B}^T = \boldsymbol{B}^T\boldsymbol{B} = \boldsymbol{I}$$

Project **columns** $\hat{\boldsymbol{x}}_i \in \mathbb{R}^{m\times 1}$ of *individual* observations onto **rows** $\boldsymbol{b}_j$ of basis

$$\hat{\boldsymbol{x}}_i = \boldsymbol{B}\hat{\boldsymbol{x}}_i = \begin{pmatrix} \boldsymbol{b}_1 \cdot \hat{\boldsymbol{x}}_i \\ \boldsymbol{b}_2 \cdot \hat{\boldsymbol{x}}_i \\ \vdots \\ \boldsymbol{b}_m \cdot \hat{\boldsymbol{x}}_i \end{pmatrix} = \begin{pmatrix} b_{11}x_{1i} + \ldots + b_{1m}x_{mi} \\ b_{21}x_{1i} + \ldots + b_{2m}x_{mi} \\ \vdots \\ b_{m1}x_{1i} + \ldots + b_{mm}x_{mi} \end{pmatrix} = \begin{pmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{mi} \end{pmatrix}$$

Project entire matrix $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ of *all* observations onto basis

$$\boldsymbol{X} = \boldsymbol{B}\boldsymbol{X} = (\boldsymbol{B}\hat{\boldsymbol{x}}_1 \ldots \boldsymbol{B}\hat{\boldsymbol{x}}_n) = \begin{pmatrix} \boldsymbol{b}_1 \cdot \hat{\boldsymbol{x}}_1 & \ldots & \boldsymbol{b}_1 \cdot \hat{\boldsymbol{x}}_n \\ \boldsymbol{b}_2 \cdot \hat{\boldsymbol{x}}_1 & \ldots & \boldsymbol{b}_2 \cdot \hat{\boldsymbol{x}}_n \\ \vdots & \ddots & \ldots \\ \boldsymbol{b}_m \cdot \hat{\boldsymbol{x}}_1 & \ldots & \boldsymbol{b}_m \cdot \hat{\boldsymbol{x}}_n \end{pmatrix} \in \mathbb{R}^{m\times n}$$

Many orthonormal bases are possible

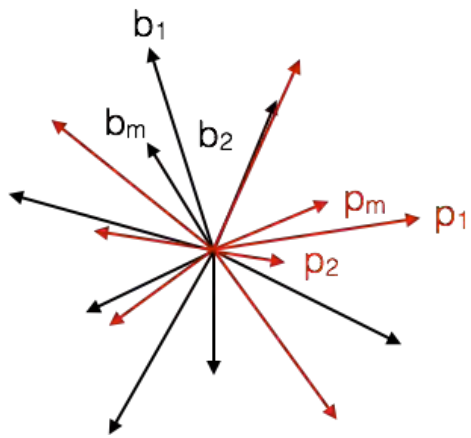
- The orthonormal basis $\boldsymbol{B} \in \mathbb{R}^{m \times m}$ reflects how we happened to collect the observations.
- Many other orthonormal bases $\boldsymbol{P} \in \mathbb{R}^{m \times m}$ may be obtained as linear combinations of $\boldsymbol{B}$.
- Perhaps in terms of other orthonormal bases, we can re-express our observations more simply or more meaningfully?

Alternative basis

Alternative basis $\boldsymbol{P} \in \mathbb{R}^{m \times m}$ of orthonormal **row** vectors $\boldsymbol{p}_j = (p_{j1} \dots p_{jm})$:

$$\boldsymbol{P} = \begin{pmatrix} \boldsymbol{p}_1 \\ \vdots \\ \boldsymbol{p}_m \end{pmatrix} = \begin{pmatrix} p_{11} & \cdots & p_{1m} \\ \vdots & \ddots & \vdots \\ p_{m1} & \cdots & p_{mm} \end{pmatrix} \in \mathbb{R}^{m \times m}$$

$$\boldsymbol{p}_j \cdot \boldsymbol{p}_k = \delta_{jk} \qquad \boldsymbol{P}\boldsymbol{P}^T = \boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{I}$$



Projection to alternative basis

Transform individual observations $\boldsymbol{x}_i \rightarrow \boldsymbol{y}_i$! Obtain y's as dot products (projections) of $\boldsymbol{x}_i$ on $\boldsymbol{p}_j$!

$$\hat{\boldsymbol{y}}_i = \boldsymbol{P}\hat{\boldsymbol{x}}_i = \begin{pmatrix} \boldsymbol{p}_1 \cdot \hat{\boldsymbol{x}}_i \\ \boldsymbol{p}_2 \cdot \hat{\boldsymbol{x}}_i \\ \vdots \\ \boldsymbol{p}_m \cdot \hat{\boldsymbol{x}}_i \end{pmatrix} = \begin{pmatrix} p_{11}x_{1i} + \ldots + p_{1m}x_{mi} \\ p_{21}x_{1i} + \ldots + p_{2m}x_{mi} \\ \vdots \\ p_{m1}x_{1i} + \ldots + p_{mm}x_{mi} \end{pmatrix} = \begin{pmatrix} y_{1i} \\ y_{2i} \\ \vdots \\ y_{mi} \end{pmatrix}$$

Transform all observations $\boldsymbol{X} \rightarrow \boldsymbol{Y}$

$$\boldsymbol{Y} = \boldsymbol{P}\boldsymbol{X} = (\boldsymbol{P}\hat{\boldsymbol{x}}_1 \ldots \boldsymbol{P}\hat{\boldsymbol{x}}_m) = \begin{pmatrix} \boldsymbol{p}_1 \cdot \hat{\boldsymbol{x}}_1 & \ldots & \boldsymbol{p}_1 \cdot \hat{\boldsymbol{x}}_m \\ \boldsymbol{p}_2 \cdot \hat{\boldsymbol{x}}_1 & \ldots & \boldsymbol{p}_2 \cdot \hat{\boldsymbol{x}}_m \\ \vdots & \ddots & \ldots \\ \boldsymbol{p}_m \cdot \hat{\boldsymbol{x}}_1 & \ldots & \boldsymbol{p}_m \cdot \hat{\boldsymbol{x}}_m \end{pmatrix} \in \mathbb{R}^{m \times n}$$

Keep in mind that $\hat{\boldsymbol{x}}_i$ and $\hat{\boldsymbol{y}}_i$ are **columns**, $\boldsymbol{b}_i$ and $\boldsymbol{p}_i$ are **rows**

Summary

- Observations $\boldsymbol{X}$ need not be expressed in native basis $\boldsymbol{B}$.
- Alternative orthonormal bases $\boldsymbol{P}$ are linear combinations of $\boldsymbol{B}$.
- Transformation to new basis is performed conveniently by matrix multiplication

$$\boldsymbol{Y} = \boldsymbol{P}\,\boldsymbol{X}$$

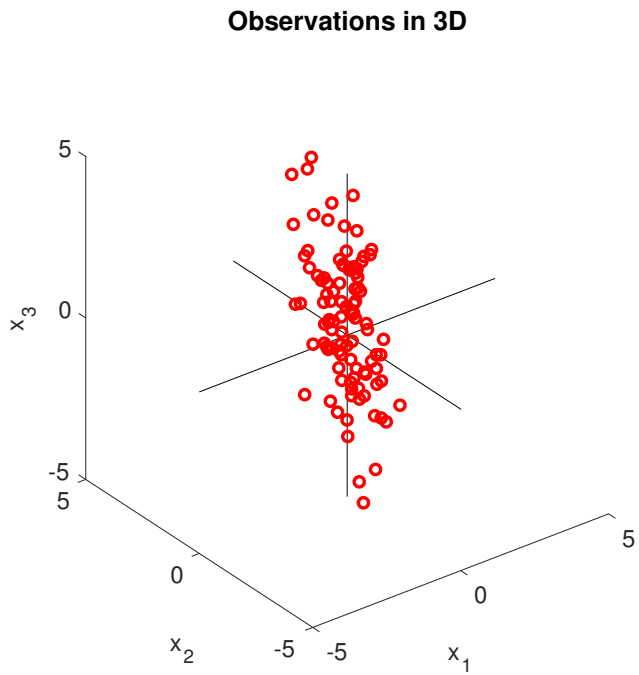
because dot products compute projections!

- Row vectors $\boldsymbol{p}_j$ will become *principal components* of $\boldsymbol{X}$.
- What would be a good choice for $\boldsymbol{P}$?
- In which respect could $\boldsymbol{Y}$ be an improvement over $\boldsymbol{X}$?

3 Variance

Consider the squared norm of individual observations $\hat{\boldsymbol{x}}_i$ (= square distance from origin) and the sum over all observations $\boldsymbol{X}$

$$|\hat{\boldsymbol{x}}_i|^2 = x_{1i}^2 + \ldots + x_{mi}^2 = \sum_{j=1}^{m} x_{ji}^2, \qquad \sum_i |\hat{\boldsymbol{x}}_i|^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} x_{ji}^2$$



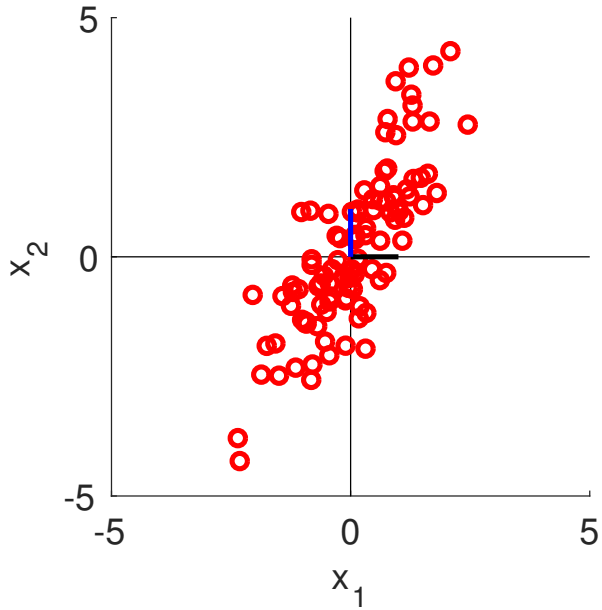
Neither the squared norm of individual observations nor their sum over all observations depend on the chosen basis!

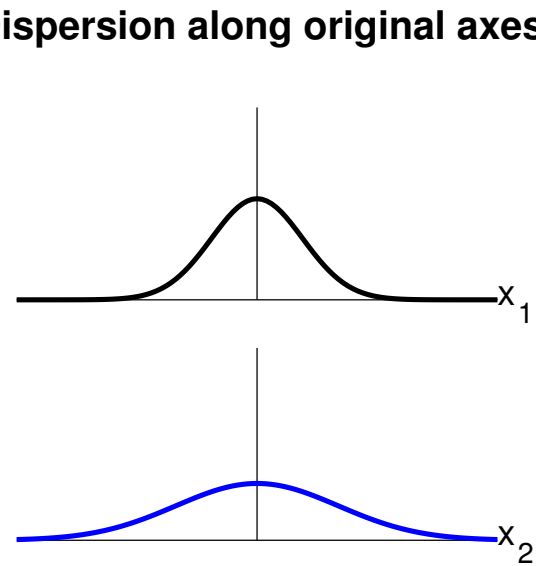
Variance within rows

Define $\boldsymbol{x}_j$ (distinct from $\hat{\mathbf{x}}_i$!) to represent **zero-mean rows** of x_{ji}

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_m \end{pmatrix} \qquad \begin{matrix} \sum_i x_{1i} = 0 \\ \vdots \\ \sum_i x_{mi} = 0 \end{matrix}$$

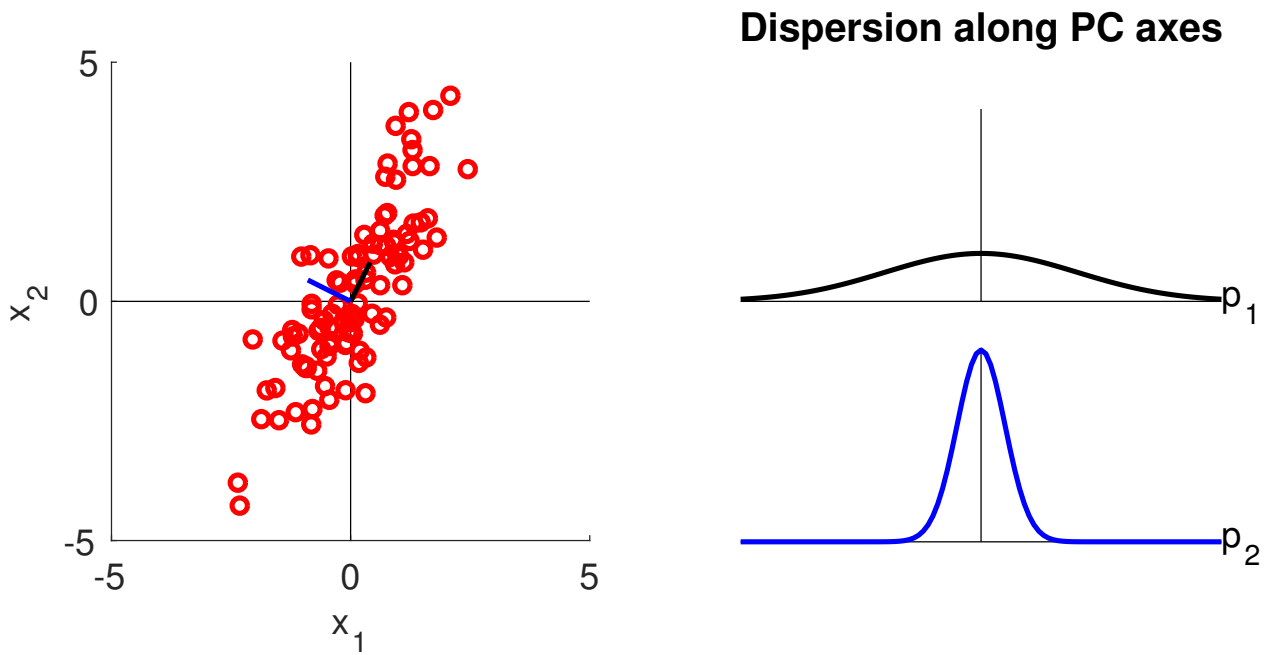
Typically, there is some, but not necessarily the same, variance within each of the different rows or dimensions, here x_1 and x_2:





Variance and orthonormal basis

The variance within each row or dimension typically changes with the basis. Importantly, the total variance remains always the same!



In the illustration, the rotated basis *maximizes* variance in one dimension (p_1) and minimizes it in the other (p_1).

Variance definition

The variance of **zero-mean** variable j over observations $i = 1, \ldots n$ is defined as

$$\begin{aligned}
\sigma_j^2 &= \frac{1}{n-1}\left[x_{j1}^2 + \ldots + x_{jn}^2\right] = \\
&= \frac{1}{n-1}\left(x_{j1} \ \ldots \ x_{jn}\right) \cdot \begin{pmatrix} x_{j1} \\ \vdots \\ x_{jn} \end{pmatrix} = \\
&= \frac{1}{n-1}\, \boldsymbol{x}_j \cdot \boldsymbol{x}_j
\end{aligned}$$

where $\boldsymbol{x}_j$ represents the **rows** of $\boldsymbol{X}$!

The total variance over all rows j reflects the total squared norm

$$\sum_j \sigma_j^2 = \frac{1}{n-1} \sum_j \boldsymbol{x}_j \cdot \boldsymbol{x}_j = \frac{1}{n-1} \sum_{j=1}^{m} \sum_{i=1}^{n} x_{ji}^2$$

Covariance matrix

Covariance of dimensions or rows j and k

$$\sigma_{jk} = \frac{1}{n-1}\left[x_{j1}x_{k1} + \ldots + x_{jn}x_{kn}\right] = \frac{1}{n}\left(x_{j1} \ \ldots \ x_{jn}\right) \cdot \begin{pmatrix} x_{k1} \\ \vdots \\ x_{kn} \end{pmatrix}$$

$$\boldsymbol{C_X} = \frac{1}{n-1}\begin{pmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1m} \\ \sigma_{21} & \sigma_2^2 & \ldots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1} & \sigma_{m2} & \ldots & \sigma_m^2 \end{pmatrix} =$$

$$= \frac{1}{n-1}\begin{pmatrix} \boldsymbol{x}_1 \cdot \boldsymbol{x}_1 & \boldsymbol{x}_1 \cdot \boldsymbol{x}_2 & \ldots & \boldsymbol{x}_1 \cdot \boldsymbol{x}_m \\ \boldsymbol{x}_2 \cdot \boldsymbol{x}_1 & \boldsymbol{x}_2 \cdot \boldsymbol{x}_2 & \ldots & \boldsymbol{x}_2 \cdot \boldsymbol{x}_m \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{x}_m \cdot \boldsymbol{x}_1 & \boldsymbol{x}_m \cdot \boldsymbol{x}_2 & \ldots & \boldsymbol{x}_m \cdot \boldsymbol{x}_m \end{pmatrix} =$$

ctd

$$= \frac{1}{n-1} \begin{pmatrix} \boldsymbol{x}_1 \cdot \boldsymbol{x}_1 & \boldsymbol{x}_1 \cdot \boldsymbol{x}_2 & \dots & \boldsymbol{x}_1 \cdot \boldsymbol{x}_m \\ \boldsymbol{x}_2 \cdot \boldsymbol{x}_1 & \boldsymbol{x}_2 \cdot \boldsymbol{x}_2 & \dots & \boldsymbol{x}_2 \cdot \boldsymbol{x}_m \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{x}_m \cdot \boldsymbol{x}_1 & \boldsymbol{x}_m \cdot \boldsymbol{x}_2 & \dots & \boldsymbol{x}_m \cdot \boldsymbol{x}_m \end{pmatrix} =$$

$$= \frac{1}{n-1} \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_m \end{pmatrix} \begin{pmatrix} \boldsymbol{x}_1^T \boldsymbol{x}_2^T \dots \boldsymbol{x}_m^T \end{pmatrix} = \frac{1}{n-1} \boldsymbol{X} \boldsymbol{X}^T \in \mathbb{R}^{m \times m}$$

Recall that $\boldsymbol{x}_j \in \mathbb{R}^{1 \times n}$, $\boldsymbol{X} \in \mathbb{R}^{m \times n}$, $\boldsymbol{X}^T \in \mathbb{R}^{n \times m}$, and $\boldsymbol{C}_{\boldsymbol{X}} \in \mathbb{R}^{m \times m}$.

Interim summary

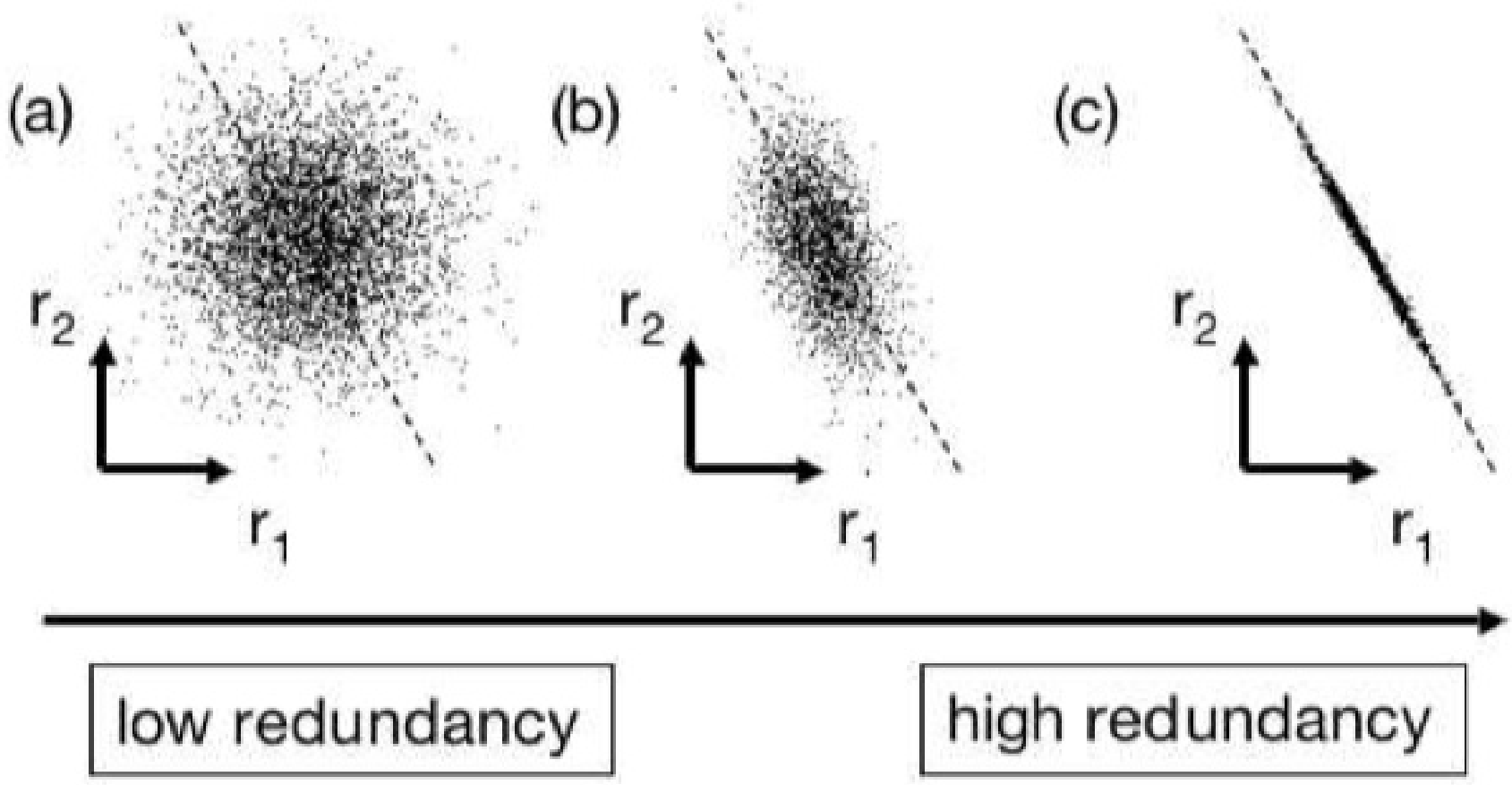
Properties of $\boldsymbol{C_X} = \frac{1}{n-1}\boldsymbol{X}\boldsymbol{X}^T$

- $\boldsymbol{C_X}$ is a square symmetric $m \times m$ matrix.
- Its diagonal terms are the *variances* of individual variables (rows of observations $\boldsymbol{X}$).
- Its off-diagonal terms are the *covariances* between different variables (two rows of observations $\boldsymbol{X}$).

$\boldsymbol{C_X}$ captures the correlations between all pairs of observations.

- In the diagonal terms, large (small) values indicate signal (noise).
- In the off-diagonal terms, large (small) values indicate high (low) redundancy.

Redundancy



Non-redundant and redundant variance

$$C_X = \frac{1}{n-1} \underbrace{\begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_m^2 \end{pmatrix}}_{non-redundant:signal} +$$

$$+ \frac{1}{n-1} \underbrace{\begin{pmatrix} 0 & \sigma_{12} & \dots & \sigma_{1m} \\ \sigma_{21} & 0 & \dots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1} & \sigma_{m2} & \dots & 0 \end{pmatrix}}_{redundant:noise}$$

Signal-to-noise ratio

The signal-to-noise ratio is defined as

$$SNR = \frac{\sigma^2_{signal}}{\sigma^2_{noise}}$$

It also specifies the fractional distribution of variance between signal and noise

$$f_{signal} = \frac{SNR}{1 + SNR} = \frac{\sigma^2_{signal}}{\sigma^2_{signal} + \sigma^2_{noise}}$$

$$f_{noise} = \frac{1}{1 + SNR} = \frac{\sigma^2_{noise}}{\sigma^2_{signal} + \sigma^2_{noise}}$$

Choosing transformation P

We can now formulate the criteria for choosing a transformation that helps reduce redundancy

$$\boldsymbol{Y} = \boldsymbol{P}\,\boldsymbol{X}, \qquad \boldsymbol{C_X} \rightarrow \boldsymbol{C_Y}$$

- Seek to maximize variance within as few dimensions of $\boldsymbol{Y}$ as possible.
- In identifying the most variable dimensions, we increase signal-to-noise ratio.
- In identifying redundant dimensions, we discover ways to reduce the dimensionality of our observations.

- Formally, we seek to maximize a few diagonal terms and minimize the other diagonal and all the off-diagnonal terms of $\boldsymbol{C_Y}$!
- A fully diagonal $\boldsymbol{C_Y}$ is always attainable! How variance distributes over the diagonal depends on our data.

4 Diagonalizing variance

There are many ways of choosing $\boldsymbol{P}$ such as to diagonalize $\boldsymbol{C_Y}$

$$\boldsymbol{Y} = \boldsymbol{P}\,\boldsymbol{X}, \qquad \boldsymbol{C_X} \rightarrow \boldsymbol{C_Y}$$

For example:

- Select the normalized direction in $\mathbb{R}^{m \times m}$ along which the variance in $\boldsymbol{X}$ is maximal. Save this vector as $\boldsymbol{p}_1$.
- Find a second, orthonormal direction along which the variance $\boldsymbol{X}$ is maximal. Save this vector as $\boldsymbol{p}_2$.
- Repeat until an orthonormal basis with m vectors has been selected.

Proceeding in this way does not take full benefit of linear algebra!

Orthogonal diagonalization

Every square symmetric $\boldsymbol{S}$ matrix is *orthonormally diagonizable*

$$\boldsymbol{S} = \boldsymbol{E}\,\boldsymbol{D}\boldsymbol{E}^T$$

with square matrix $\boldsymbol{E}$ collecting **column** eigenvectors of $\boldsymbol{S}$:

$$\boldsymbol{E} = (\boldsymbol{e}_1\,\boldsymbol{e}_2\,\ldots\,\boldsymbol{e}_m)\,, \qquad \boldsymbol{E}\boldsymbol{E}^T = \boldsymbol{E}^T\boldsymbol{E} = \boldsymbol{I}$$

and diagonal matrix $\boldsymbol{D}$ collecting eigenvalues of $\boldsymbol{S}$:

$$\boldsymbol{D} = \begin{pmatrix} \lambda_1 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \lambda_m \end{pmatrix} \qquad diag(\boldsymbol{D}) = (\lambda_1\,\lambda_2\,\ldots\,\lambda_m)$$

The relevance of this trick is that it applies to covariance matrices, which are both square and symmetric!

A matrix is what a matrix does!

$$\boldsymbol{S} = \underbrace{\boldsymbol{E}}_{\text{rotate back in } \mathbb{R}^{m\times m}} \underbrace{\boldsymbol{D}}_{\text{stretch}} \underbrace{\boldsymbol{E}^T}_{\text{rotate in } \mathbb{R}^{m\times m}} \in \mathbb{R}^{m\times m}$$

After rotation, j^{th} component is strechted by $\lambda_j = D_{jj}$

Application to covariance

Given orthonormal diagonalization $\boldsymbol{C_X} = \boldsymbol{E}\,\boldsymbol{D}\,\boldsymbol{E}^T$, the covariance of $\boldsymbol{Y} = \boldsymbol{P}\boldsymbol{X}$ may become diagonal:

$$\begin{aligned}
\boldsymbol{C_Y} &= \frac{1}{n-1}\boldsymbol{Y}\,\boldsymbol{Y}^T = \frac{1}{n-1}(\boldsymbol{P}\boldsymbol{X})(\boldsymbol{P}\boldsymbol{X})^T = \\
&= \frac{1}{n-1}(\boldsymbol{P}\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{P}^T) = \frac{1}{n-1}\boldsymbol{P}(\boldsymbol{X}\boldsymbol{X}^T)\boldsymbol{P}^T = \\
&= \boldsymbol{P}\,\boldsymbol{C_X}\,\boldsymbol{P}^T = \boldsymbol{P}\,\boldsymbol{E}\,\boldsymbol{D}\,\boldsymbol{E}^T\,\boldsymbol{P}^T = \boldsymbol{I}\,\boldsymbol{D}\,\boldsymbol{I} = \boldsymbol{D}
\end{aligned}$$

if only the **rows** of $\boldsymbol{P}$ are chosen as the **columns** of $\boldsymbol{E}$:

$$\boldsymbol{P} = \boldsymbol{E}^T \quad \Leftrightarrow \quad \boldsymbol{P}\boldsymbol{E} = \boldsymbol{I} \quad \Leftrightarrow \quad \boldsymbol{E}^T\boldsymbol{P}^T = \boldsymbol{I}$$

Sort principal components

By convention, we sort rows of $\boldsymbol{P}$ (columns of $\boldsymbol{E}$) from the largest to the smallest eigenvalue (as long as they are non-zero):

The *first principal component* $\boldsymbol{p}_1$ captures the largest variance $\sigma_1^2 = \lambda_1 = D_{1,1}$.

The *second principal component* $\boldsymbol{p}_2$, captures the next largest variance $\sigma_2^2 = \lambda_2 = D_{2,2}$.

$\vdots$

The m^{th} *principal component* $\boldsymbol{p}_m$, captures the smallest variance $\sigma_m^2 = \lambda_m = D_{m,m}$.

Different roads to Rome

Orthonormal diagonalization: requires covariance

- Form covariance of observations $\boldsymbol{C_X} = \frac{1}{n-1}\boldsymbol{X}\boldsymbol{X}^T \in \mathbb{R}^{m \times m}$
- Orthonormal diagonalization $\boldsymbol{C_X} = \boldsymbol{E}\boldsymbol{D}\boldsymbol{E}^T$ into eigenvectors $\boldsymbol{E} \in \mathbb{R}^{m \times m}$ and eigenvalues $diag(\boldsymbol{D})$.
- Obtain principal components as $\boldsymbol{P} = \boldsymbol{E}^T$ and variance contributions as $diag(\boldsymbol{D})$.

Singular-value decomposition (SVD): proceeds directly from observations

- Singular value decomposition of $\boldsymbol{X}^T = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T \in \mathbb{R}^{n \times m}$.
- Obtain principal components as $\boldsymbol{P} = \boldsymbol{V}^T$ and variance contributions as $diag(\boldsymbol{\Sigma}^T\boldsymbol{\Sigma})$.

Singular Value Decomposition

What the orthonormal diagonalization is for square and symmetric matrices, the *singular value decomposition* is for arbitrary rectangular matrices. Every rectangular matrix $\boldsymbol{A}$ has a singular value decomposition

$$\boldsymbol{A} = \boldsymbol{U}\,\boldsymbol{\Sigma}\,\boldsymbol{V}^T \in \mathbb{R}^{n\times m}$$

with *left singular vectors*

$$\boldsymbol{U} \in \mathbb{R}^{n\times n} \qquad \textit{orthonormal}$$

and *singular values*

$$\boldsymbol{\Sigma} \in \mathbb{R}^{n\times m} \qquad \textit{diagonal}$$

and *right singular vectors*

$$\boldsymbol{V} \in \mathbb{R}^{m\times m} \qquad \textit{orthonormal}$$

Orthonormal (and square)

$$\boldsymbol{U} = \begin{pmatrix} u_{11} & \dots & u_{1n} \\ \vdots & \ddots & \vdots \\ u_{n1} & \dots & u_{nn} \end{pmatrix} \in \mathbb{R}^{n \times n}, \quad \boldsymbol{V} = \begin{pmatrix} v_{11} & \dots & v_{1m} \\ \vdots & \ddots & \vdots \\ v_{m1} & \dots & v_{mm} \end{pmatrix} \in \mathbb{R}^{m \times m}$$

Diagonal (and rectangular)

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sqrt{\lambda_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sqrt{\lambda_m} \\ \vdots & \dots & \vdots \\ 0 & \dots & 0 \end{pmatrix} \in \mathbb{R}^{n \times m}$$

A matrix is what a matrix does!

$$\boldsymbol{A} = \underbrace{\boldsymbol{U}}_{\text{rotate in } \mathbb{R}^{n\times n}} \underbrace{\boldsymbol{\Sigma}}_{\text{stretch and pad}} \underbrace{\boldsymbol{V}^T}_{\text{rotate in } \mathbb{R}^{m\times m}} \qquad \in \mathbb{R}^{n\times m}$$

After first rotation, j^{th} component is stretched by $\sqrt{\lambda_j} = \Sigma_{jj}$

Application to PCA

Choosing

$$\boldsymbol{A} = \frac{1}{\sqrt{n-1}} \boldsymbol{X}^T \in \mathbb{R}^{n \times m}, \qquad \text{with} \qquad \boldsymbol{A} = \boldsymbol{U}\,\boldsymbol{\Sigma}\,\boldsymbol{V}^T$$

The desired orthonormal basis $\boldsymbol{P}$ is

$$\boldsymbol{P} = \boldsymbol{V}^T \in \mathbb{R}^{m \times m}$$

and the diagonal matrix of eigenvalues $\boldsymbol{D}$ is

$$\boldsymbol{D} = \begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_m \end{pmatrix} = \boldsymbol{\Sigma}\,\boldsymbol{\Sigma}^T = \begin{pmatrix} \sigma_1^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_m^2 \end{pmatrix}$$

Proof

The choice of $\boldsymbol{A} = \frac{1}{\sqrt{n-1}} \boldsymbol{X}^T = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ ensures that

$$\begin{aligned} \boldsymbol{C}_{\boldsymbol{X}} &= \frac{1}{n-1} \boldsymbol{X}\boldsymbol{X}^T = \boldsymbol{A}^T\boldsymbol{A} = \\ &= (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T)^T(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T) = \\ &= (\boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{U}^T)(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T) = \\ &= \boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{\Sigma}\boldsymbol{V}^T \end{aligned}$$

$$\Leftrightarrow \qquad \boldsymbol{C}_{\boldsymbol{X}}\boldsymbol{V} = \boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^T\boldsymbol{\Sigma}\boldsymbol{V} = \boldsymbol{D}\boldsymbol{V}$$

so that $\boldsymbol{V}$ are the eigenvectors and $\boldsymbol{\Sigma}^T\boldsymbol{\Sigma}$ the eigenvalues of $\boldsymbol{C}_{\boldsymbol{X}}$.

Summary diagonalizing variance

Seek linear transformation $\boldsymbol{Y} = \boldsymbol{P}\boldsymbol{X}$ such that covariance $\boldsymbol{C}_{\boldsymbol{Y}}$ is diagonal matrix $\boldsymbol{D}$!

SVD of rectangular matrix $\boldsymbol{A} = \frac{1}{n-1}\boldsymbol{X}^T$ into $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ gives us

$$\boldsymbol{P} = \boldsymbol{V}^T, \qquad \boldsymbol{D} = \boldsymbol{\Sigma}^T\boldsymbol{\Sigma}$$

Orthonormal diagonalization of square and symmetric matrix $\boldsymbol{C}_{\boldsymbol{X}}$ into $\boldsymbol{C}_{\boldsymbol{X}} = \boldsymbol{E}\boldsymbol{D}\boldsymbol{E}^T$ gives us

$$\boldsymbol{P} = \boldsymbol{E}^T, \qquad \boldsymbol{D} = \boldsymbol{D}$$

What have we achieved? By re-representing observations in PC coordinates, we have eliminated redundancy and revealed main sources of signal (variance):

$$\boldsymbol{C_X} = \frac{1}{n-1} \underbrace{\begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1m} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1} & \sigma_{m2} & \dots & \sigma_m^2 \end{pmatrix}}_{non-diagonal}$$

$$\sigma^2_{total} = \sigma_1^2 + \sigma_2^2 + \ldots + \sigma_m^2$$

$$\boldsymbol{C_Y} = \frac{1}{n-1} \underbrace{\begin{pmatrix} \rho_1^2 & 0 & \dots & 0 \\ 0 & \rho_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \rho_m^2 \end{pmatrix}}_{diagonal}$$

$$\sigma^2_{total} = \rho_1^2 + \rho_2^2 + \ldots + \rho_m^2$$

5. De-noising with PCA

Full projection matrix $\boldsymbol{V}^T \in \mathbb{R}^{m \times m}$

$$\boldsymbol{Y} = \boldsymbol{V}^T \boldsymbol{X}$$

$$\begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ \vdots & \ddots & \ddots & \vdots \\ y_{m1} & y_{m2} & \cdots & y_{mn} \end{pmatrix} = \begin{pmatrix} v_{11} & \cdots & v_{m1} \\ \vdots & \ddots & \vdots \\ v_{1m} & \cdots & v_{mm} \end{pmatrix} \begin{pmatrix} x_{11} & x_{1,2} & \cdots & x_{1n} \\ \vdots & \ddots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix}$$

Truncated projection matrix $\tilde{\boldsymbol{V}}^T \in \mathbb{R}^{r \times m}$, with $r < m$

$$\tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{V}}^T \boldsymbol{X}$$

$$\begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{r1} & y_{r2} & \cdots & y_{rn} \end{pmatrix} = \begin{pmatrix} v_{11} & \cdots & v_{m1} \\ v_{1r} & \cdots & v_{mr} \end{pmatrix} \begin{pmatrix} x_{11} & x_{1,2} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots & \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix}$$

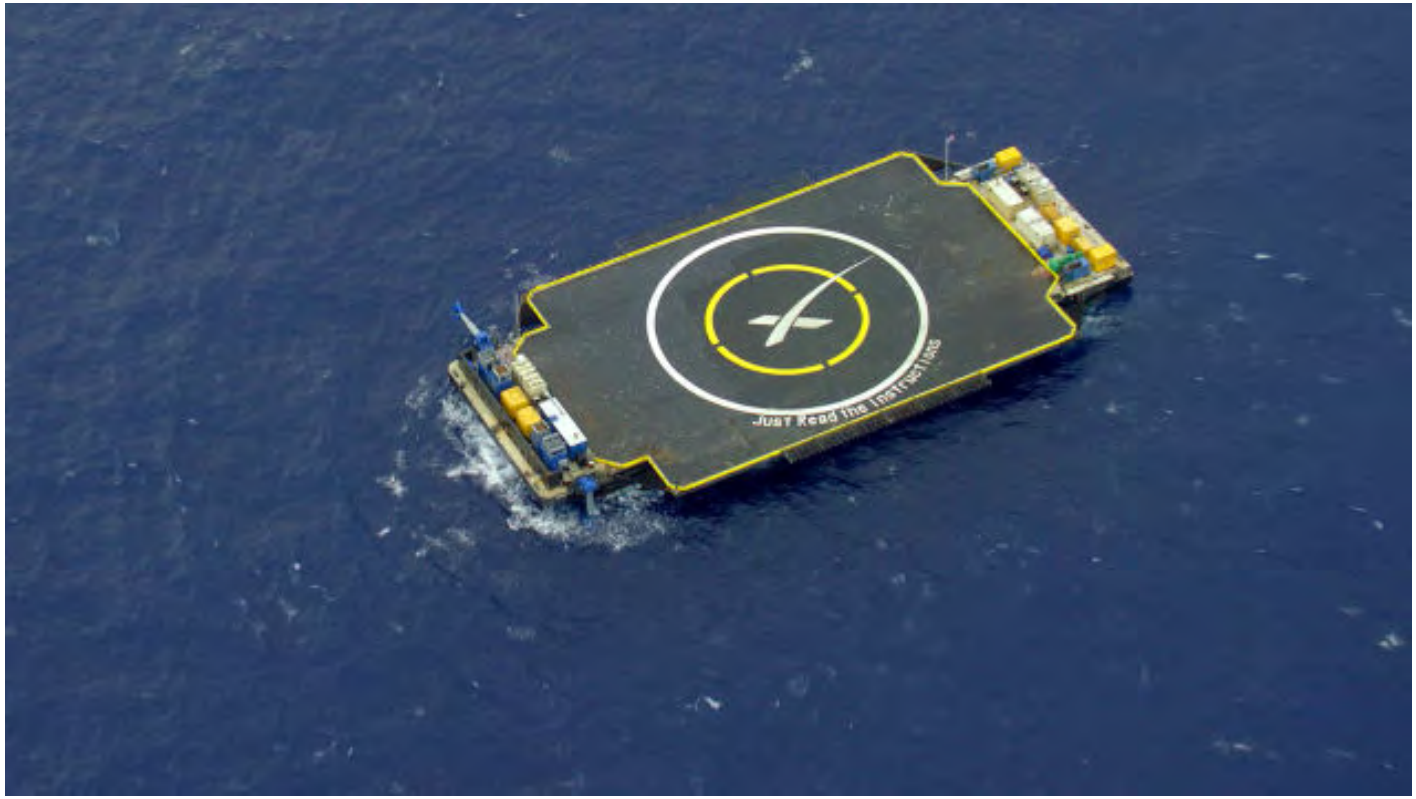
The truncated projection $\tilde{\boldsymbol{Y}}$ combines the first r *principal components* of the original observations $\boldsymbol{X}$. Being smaller, it constitutes a *compressed* version of $\boldsymbol{X}$.

It may be *uncompressed* into a matrix $\tilde{\boldsymbol{X}} \in \mathbb{R}^{n \times m}$ by performing the backprojection

$$\tilde{\boldsymbol{X}} = \left(\tilde{\boldsymbol{V}}^T\right)^{-1} \tilde{\boldsymbol{Y}}$$

$$\begin{pmatrix} x_{11} & x_{1,2} & \cdots & x_{1n} \\ \vdots & \ddots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix} = \begin{pmatrix} v_{11} & v_{r1} \\ \vdots & \vdots \\ v_{m1} & v_{mr} \end{pmatrix} \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{r1} & y_{r2} & \cdots & y_{rn} \end{pmatrix}$$

If higher principal components reflect noise, compressing and uncompressing with principal components also serves to *de-noise* the original data.



BW original BW denoised

BW original BW denoised

Summary de-noising

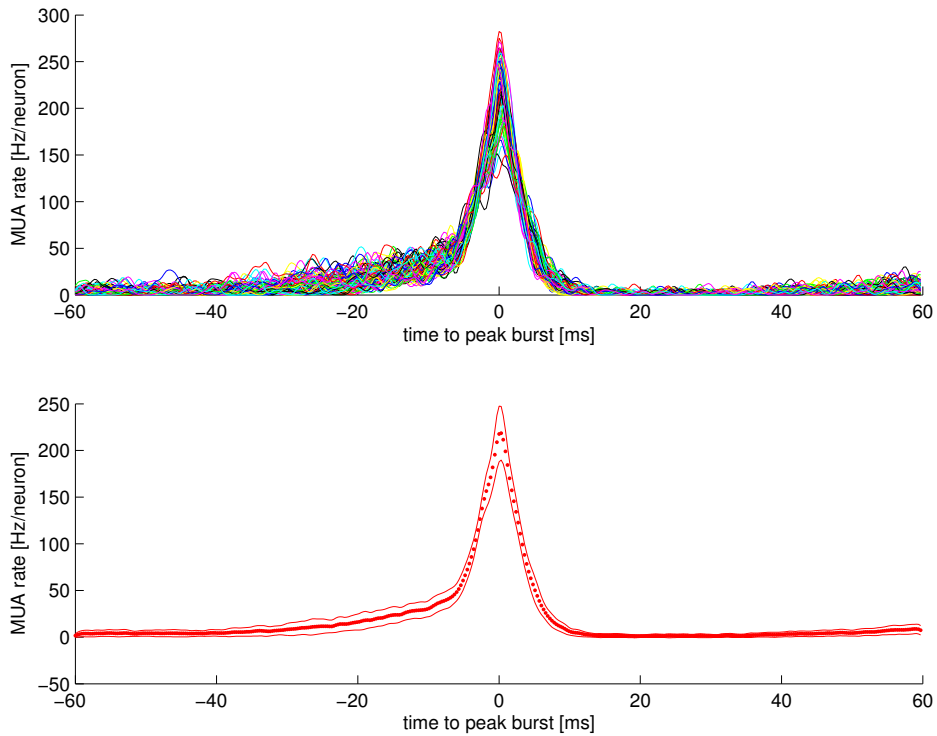
- Transformed observations $\boldsymbol{Y} = \boldsymbol{P}\boldsymbol{X}$ span the entire principal component space.
- It makes sense to consider only the first few dimensions, where most of the variance is concentrated.
- To this end, transformed observations $\boldsymbol{Y}$ may be truncated to $\tilde{\boldsymbol{Y}}$ by zeroing the unwanted dimensions.
- The results may be visualized by back-transforming the truncated observations into the original space

$$\tilde{\boldsymbol{X}} = \boldsymbol{P}^T \tilde{\boldsymbol{Y}}$$

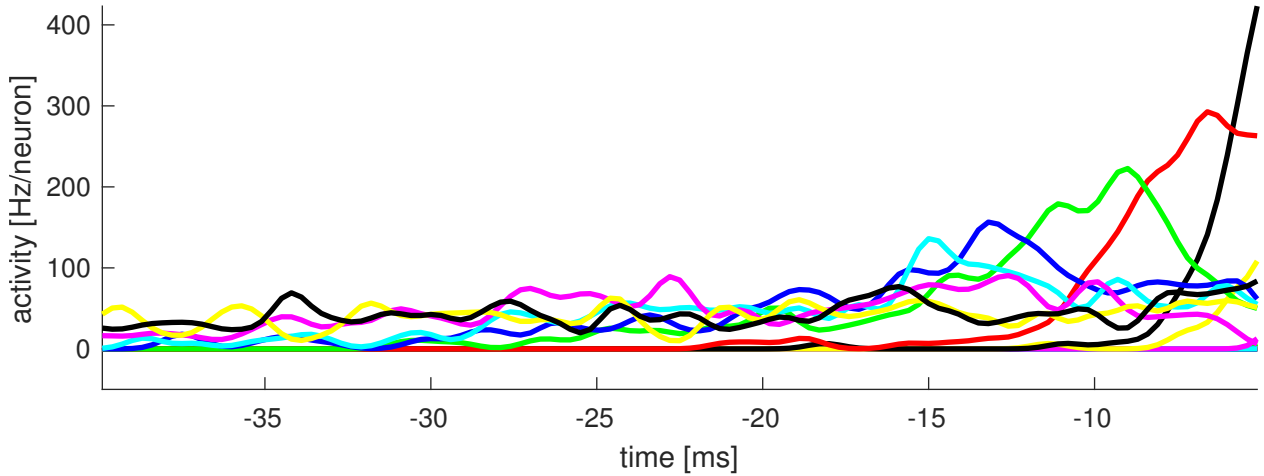
- The result is a 'de-noised' version of the original observations.

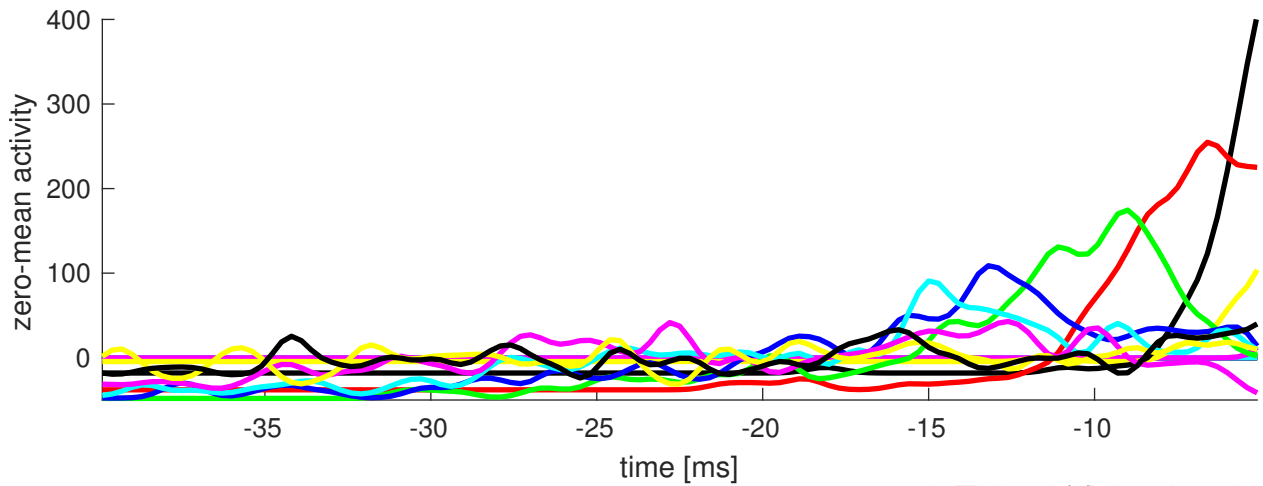
6. Multi-unit activity

Consider a network with intermittent population bursts. We have recorded the activity of 400 neurons over approximately 100 bursts, beginning 60 *ms* before and ending 60 *ms* after each burst. We have sorted neurons by overall activity and have saved the average MUA of 20 groups (of 20 neurons each) in 3 *ms*-wide time-bins, spaced 0.25 *ms* apart.



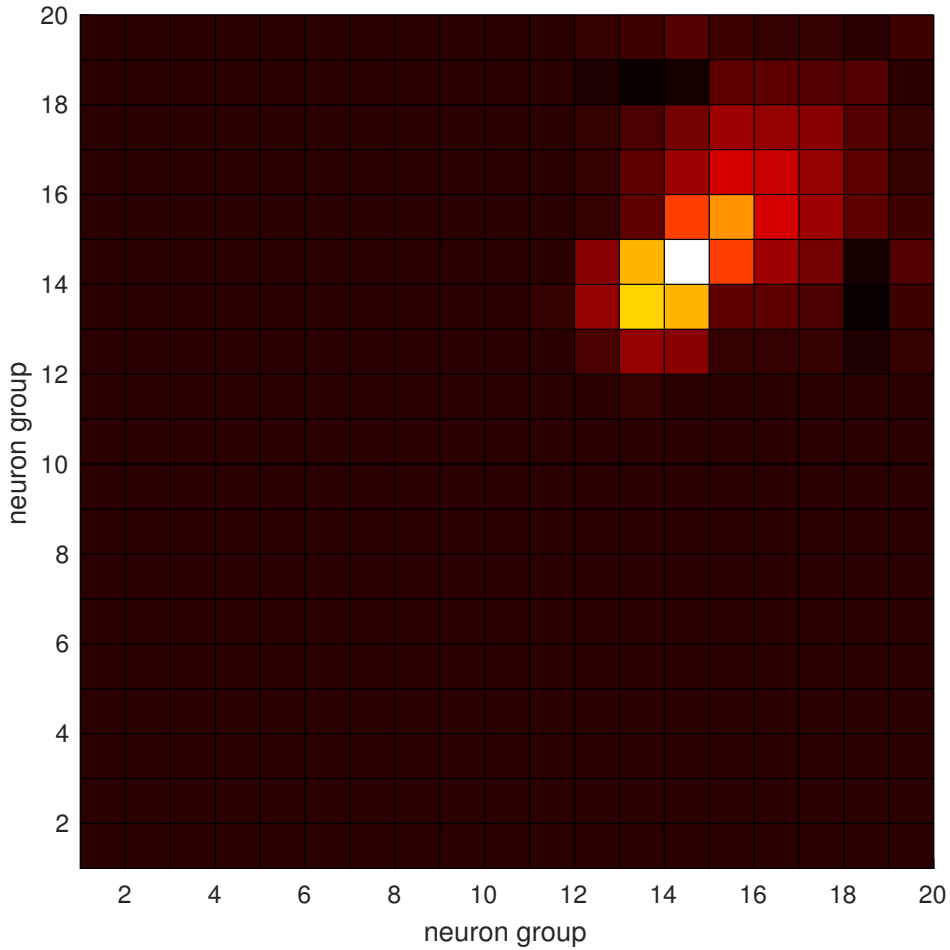
We are interested in the initial stages of each burst and decide to examine the period from 40 *ms* to 5 *ms* before the peak of each burst. Thus we have recordings from approximately 120 time points (spaced 0.25 *ms* apart over 35 *ms*). Notice that neuron groups differ in their level and time of their activity.





Covariance

Forming activity matrix $\boldsymbol{X}_0$ (rows = groups of neurons, columns = time points) and covariance $\boldsymbol{C}_{\boldsymbol{X}} = \boldsymbol{X}_0 \boldsymbol{X}_0^T / 120$, we note that most variance occurs in the more active groups of neurons:

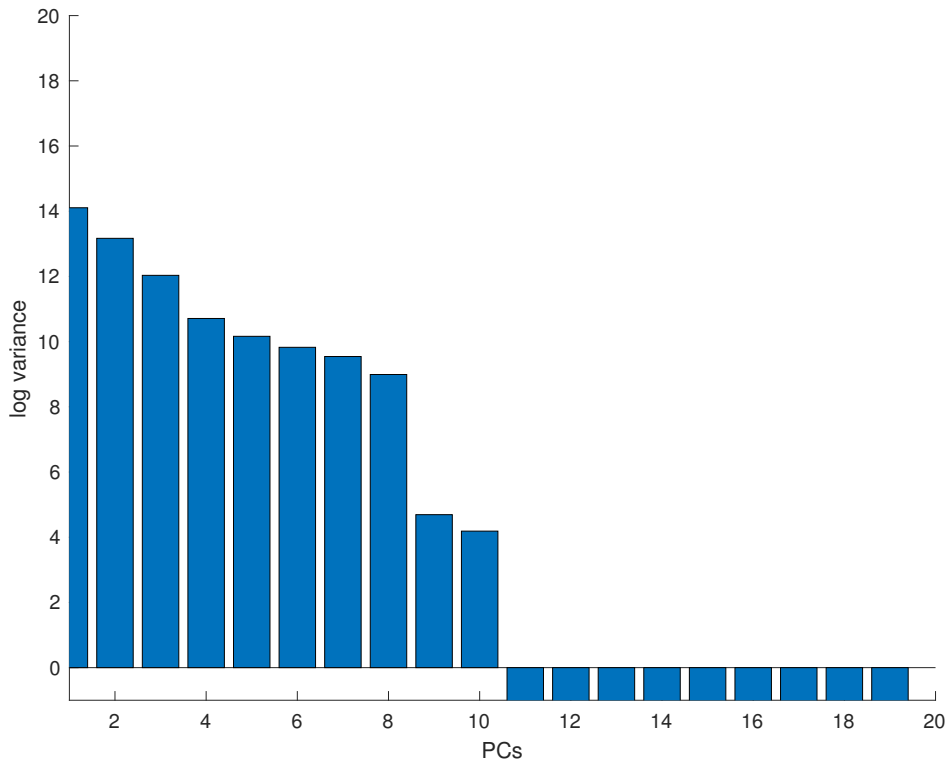


Variance captured by principal components

Using singular value decomposition

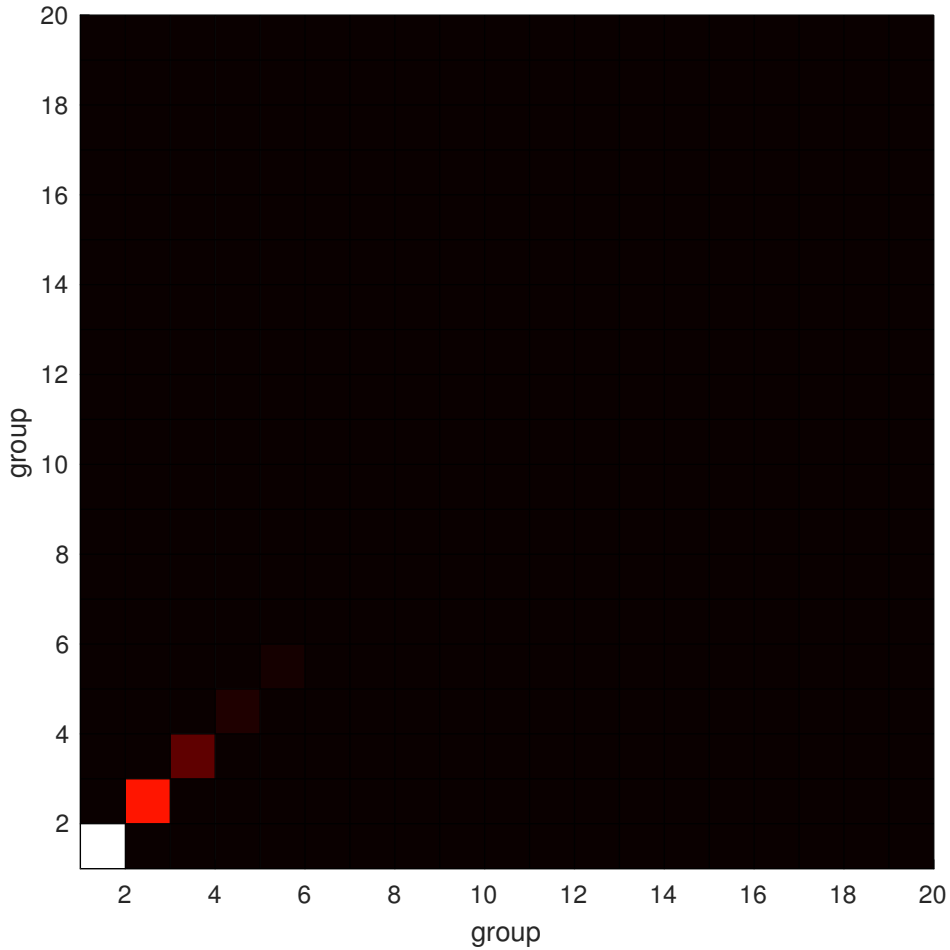
$$\boldsymbol{X}_0 = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$$

we obtain the principal components $\boldsymbol{P} = \boldsymbol{V}^T$ and the variance $diag(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T)$ captured by each:



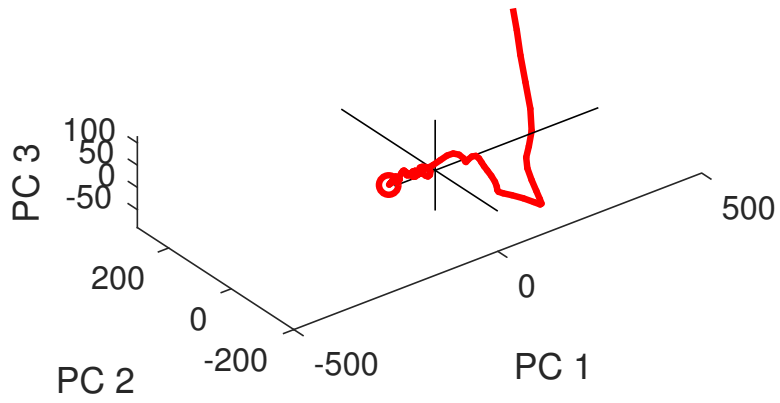
Covariance of transformed observations

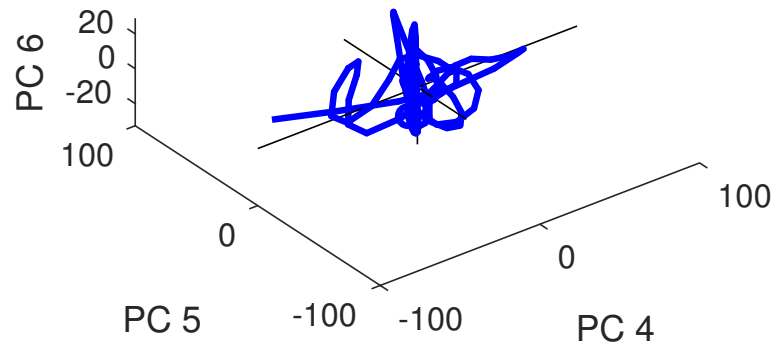
Transforming observations $\boldsymbol{Y} = \boldsymbol{P}\boldsymbol{X}_0$, we find that covariance $\boldsymbol{C}_{\boldsymbol{Y}} = \boldsymbol{Y}\,\boldsymbol{Y}^T/120$ is diagonalized, as expected:

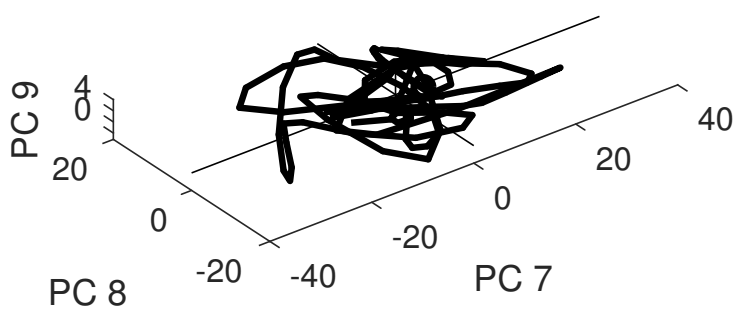


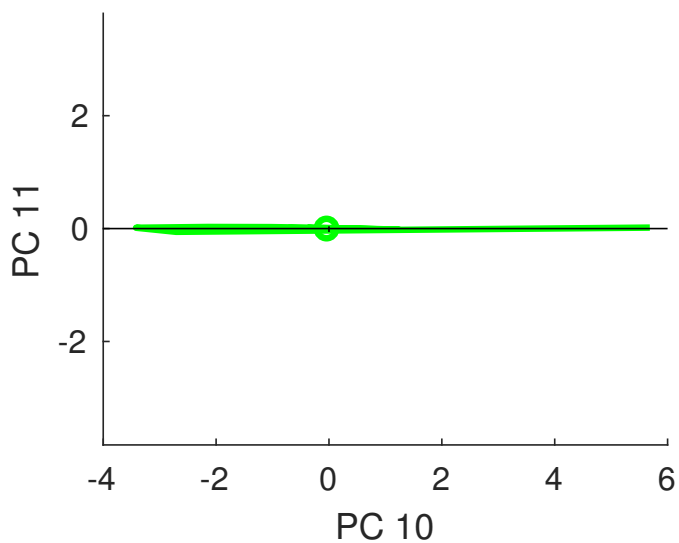
Visualizing observations in PC space

Having transformed observations, we can visualize the *average burst initiation* in the principal component space:



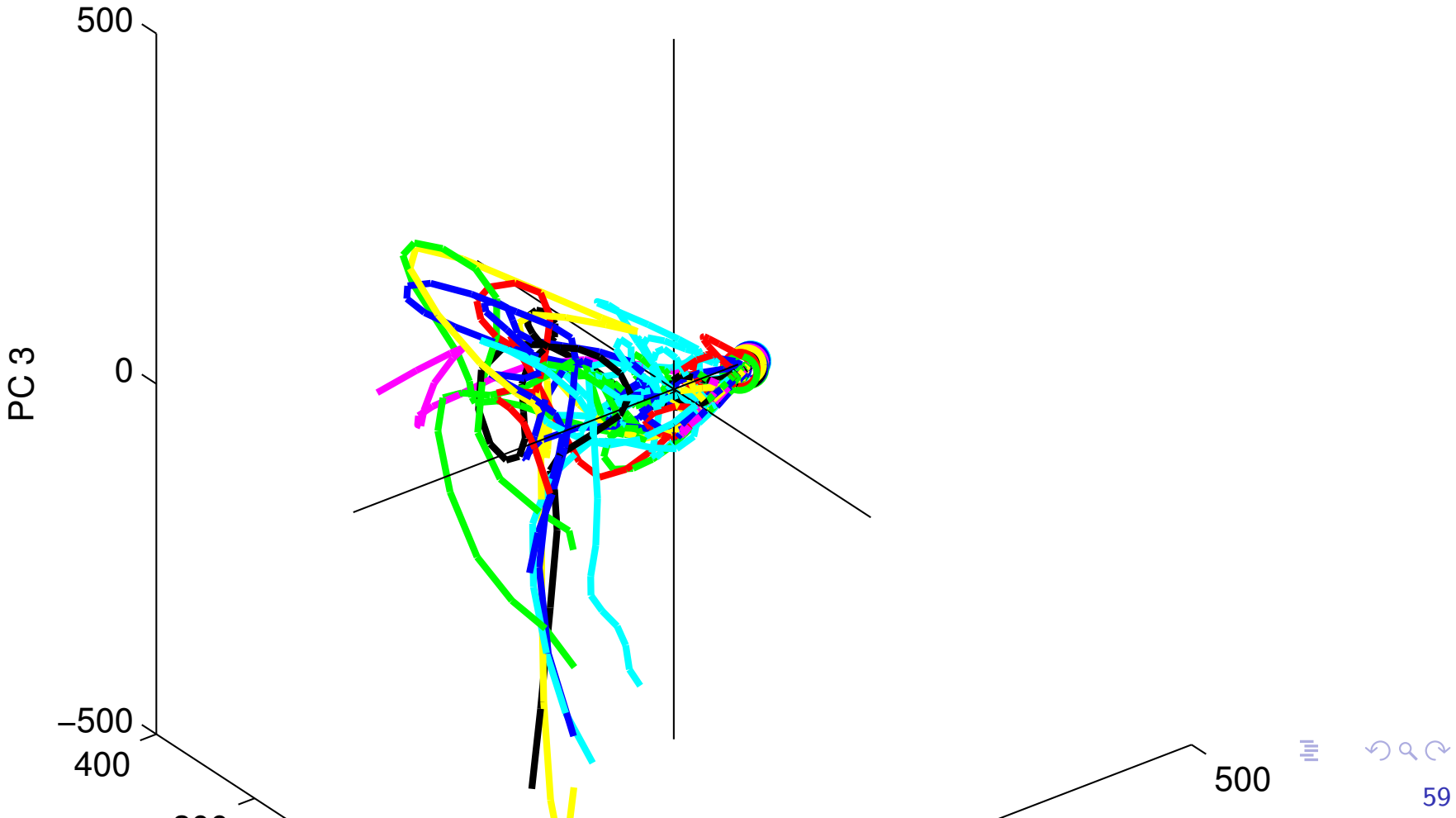






Individual burst initiations

In space of first three principal components:

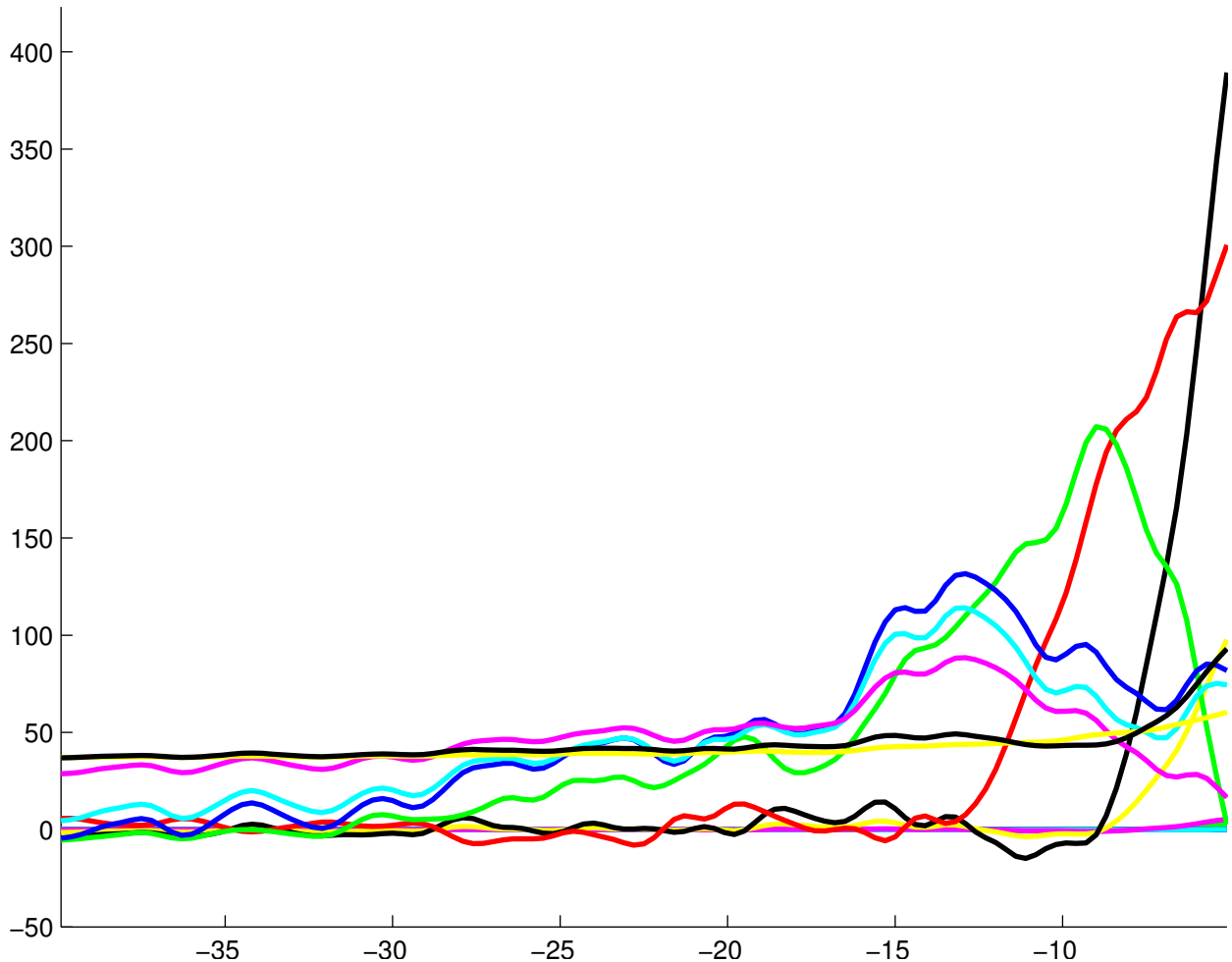


De-noised observations

Backprojecting the first three rows of $\boldsymbol{Y}$ into the original space

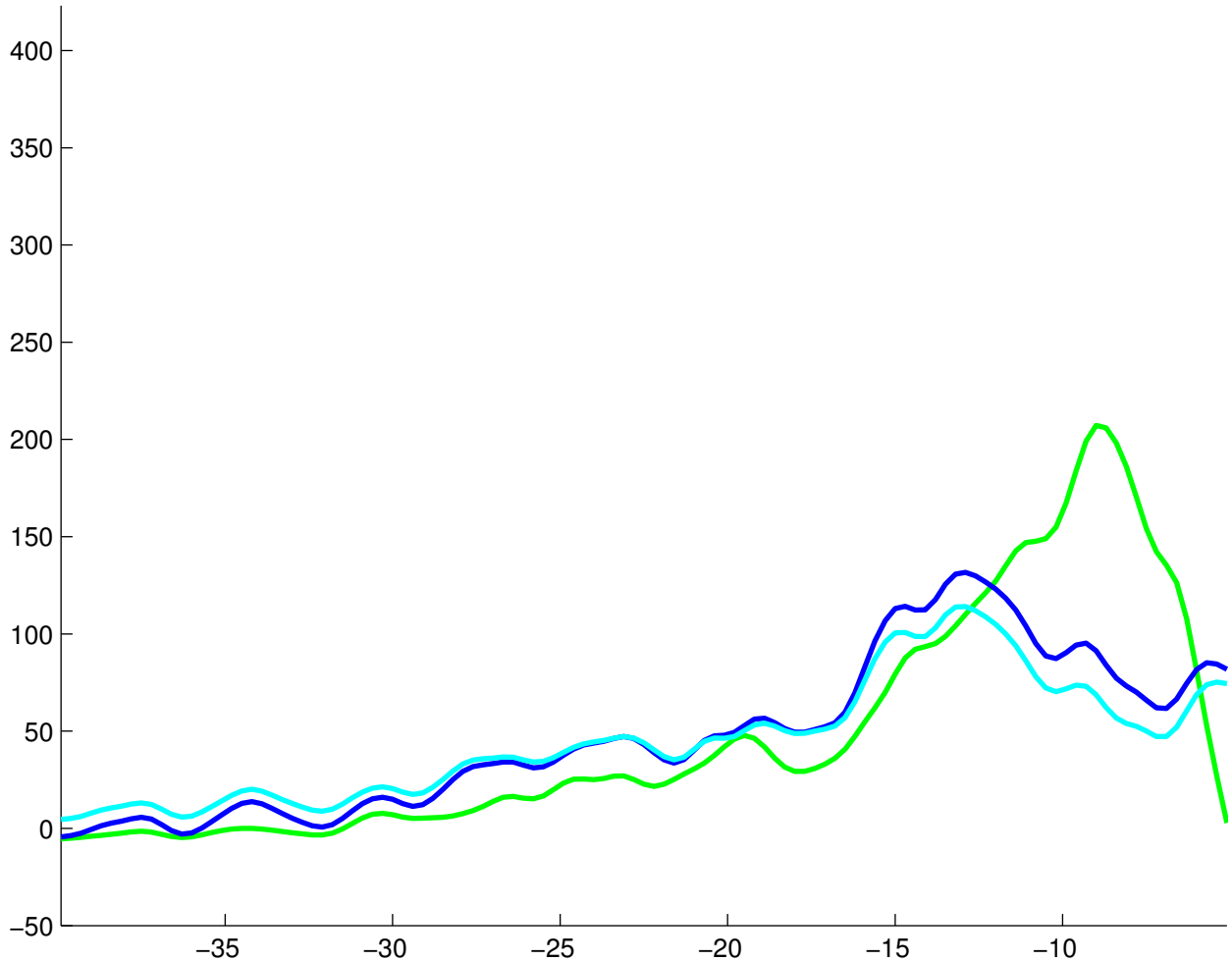
$$\boldsymbol{X}_0' = \boldsymbol{P}^T \boldsymbol{Y}'$$

we obtain the typical time-course of burst-initation:



Potentially important groups of neurons

This draws our attention to three groups of neurons (15, 16, 17) whose activity seems to initiate the burst. To test this hypothesis, we can target these groups with experimental manipulations.



Summary multi-unit activity

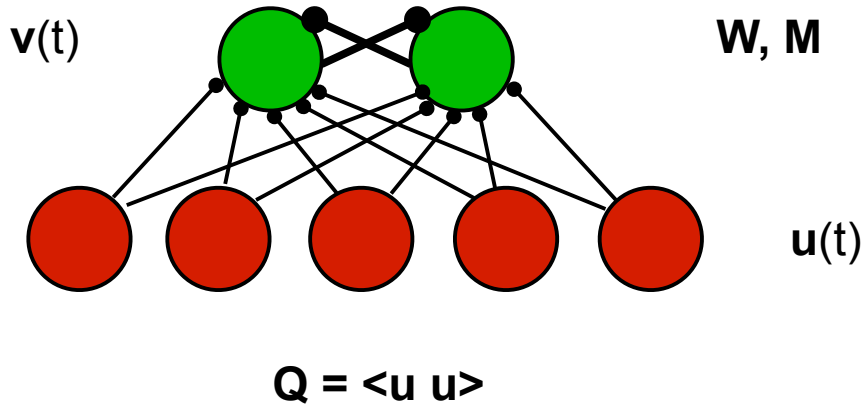
- We have used PCA to obtain survey a large data set (time-dependent activity of 20 groups of neurons).
- We find that variance is concentrated in a few groups of neurons.
- The de-noised time-course reveals the typical development of group activity.
- This helps identify the groups that initiate bursting activity.

Summary of PCA assumptions

1. *Linearity*
   We frame the problem as a change of basis $\mathbb{R}^{m \times m}$ space.
2. *Variance is a sufficient statistic*
   We assume the variance fully describes the distribution of our (zero-mean) variables.
3. *Large variances are important*
   We assume variables with high variance are 'signal', whereas those with low variance are 'noise'.
4. *Principal components are orthogonal*
   This lets us perform PCA with linear algebra techniques.

Link to representational learning

Recall "unsupervised learning" with competing output units.



Learning shaped by

- *correlations* between input neurons, specifically the eigenvectors of the correlation matrix: *principal components* of input.
- *competition* between output neurons, so that different neurons can represent different *linear combinations* of principal components.

Output representation in principal component space!

Next: Sparse Coding

Appendix: Singular value decomposition

A singular-value decomposition $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ decomposes the transformation (change of basis) performed by $\boldsymbol{A}$ into three distinct and intuitively intelligible steps. Consider a linear transformation $\boldsymbol{x} \to \boldsymbol{y}$

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{x}, \qquad \boldsymbol{y} \in \mathbb{R}^n, \qquad \boldsymbol{x} \in \mathbb{R}^m, \qquad \boldsymbol{A} \in \mathbb{R}^{n \times m}$$

Expressing $\boldsymbol{y}$ in the orthonormal basis $\boldsymbol{U}$ and $\boldsymbol{x}$ in the orthonormal basis $\boldsymbol{V}$, we have

$$\hat{\boldsymbol{y}} = \boldsymbol{U}\boldsymbol{y} \quad \Leftrightarrow \quad \boldsymbol{y} = \boldsymbol{U}^T\hat{\boldsymbol{y}}, \qquad \hat{\boldsymbol{x}} = \boldsymbol{V}\boldsymbol{x} \quad \Leftrightarrow \quad \boldsymbol{x} = \boldsymbol{V}^T\hat{\boldsymbol{x}}$$

By left-multiplying the SVD with $\boldsymbol{U}^T$, we can reformulate the transformation in terms of $\hat{\boldsymbol{x}} \to \hat{\boldsymbol{y}}$

$$\boldsymbol{U}^T \cdot \Big| \qquad \boldsymbol{y} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{x} \qquad \Leftrightarrow \qquad \hat{\boldsymbol{y}} = \boldsymbol{U}^T\boldsymbol{y} = \boldsymbol{\Sigma}\hat{\boldsymbol{x}}$$

Thus, in terms of the orthonormal bases $\boldsymbol{U}$ and $\boldsymbol{V}$, the transformation *rescales* components 1 to m with the singular values in $\boldsymbol{\Sigma}$ and zeroes all others:

$$\begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{x}_m \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} \sqrt{\lambda_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sqrt{\lambda_m} \\ \vdots & \dots & \vdots \\ 0 & \dots & 0 \end{pmatrix} \begin{pmatrix} \hat{x}_1 \\ \hat{x}_2 \\ \vdots \\ \hat{x}_m \end{pmatrix}$$

Note that $\boldsymbol{\Sigma}^T\boldsymbol{\Sigma}$ is square and diagonal:

$$\boldsymbol{\Sigma}^T\boldsymbol{\Sigma} = \boldsymbol{\Sigma}\boldsymbol{\Sigma}^T = \begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_m \end{pmatrix} \in \mathbb{R}^{m \times m}$$

Derivation of SVD

Consider a covariance matrix and its orthonormal eigenvectors and real positive eigenvalues

$$\boldsymbol{X}^T \boldsymbol{X}\, \boldsymbol{v}_j = \lambda_j \boldsymbol{v}_j, \qquad \boldsymbol{X} \in \mathbb{R}^{n\times m}, \boldsymbol{v}_j \in \mathbb{R}^{m\times 1},$$

Use $\boldsymbol{X}$ to transform $m \times 1$ eigenvectors $\boldsymbol{v}_j$ into $n \times 1$ vectors $\boldsymbol{u}_j$

$$\boldsymbol{u}_j = \frac{1}{\sqrt{\lambda_j}} \boldsymbol{X}\, \boldsymbol{v}_j, \qquad \boldsymbol{u}_j \in \mathbb{R}^{n\times 1}$$

Astonishingly, the transformed vectors also form an orthonormal basis! Their norm is unity! They are pairwise orthogonal!

$$||\boldsymbol{u}_j|| = 1, \qquad \boldsymbol{u}_j\, \boldsymbol{u}_k = \delta_{ik}$$

Proof

$$
\begin{aligned}
||\boldsymbol{u}_j||^2 &= \boldsymbol{u}_j \cdot \boldsymbol{u}_j = \boldsymbol{u}_j^T \boldsymbol{u}_j = \\
&= \frac{1}{\lambda_j}(\boldsymbol{X}\,\boldsymbol{v}_j)^T(\boldsymbol{X}\,\boldsymbol{v}_j) = \frac{1}{\lambda_j}(\boldsymbol{v}_j^T\,\boldsymbol{X}^T)\,(\boldsymbol{X}\,\boldsymbol{v}_j) = \\
&= \frac{1}{\lambda_j}\,\boldsymbol{v}_j^T\,\lambda_j\,\boldsymbol{v}_j = \boldsymbol{v}_j \cdot \boldsymbol{v}_j = ||\boldsymbol{v}_j^2|| = 1
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{u}_j \cdot \boldsymbol{u}_k &= \boldsymbol{u}_j^T \boldsymbol{u}_k = \\
&= \frac{1}{\sqrt{\lambda_j\lambda_k}}(\boldsymbol{X}\,\boldsymbol{v}_j)^T(\boldsymbol{X}\,\boldsymbol{v}_k) = \frac{1}{\sqrt{\lambda_j\lambda_k}}(\boldsymbol{v}_j^T\,\boldsymbol{X}^T)\,(\boldsymbol{X}\,\boldsymbol{v}_k) = \\
&= \frac{1}{\sqrt{\lambda_j\lambda_k}}\,\boldsymbol{v}_j^T\,\lambda_k\,\boldsymbol{v}_k = \sqrt{\frac{\lambda_k}{\lambda_j}}\,\boldsymbol{v}_j \cdot \boldsymbol{v}_k = \delta_{jk}
\end{aligned}
$$

Collect intro matrix

Collect $\boldsymbol{u}$'s into $n \times m$ matrix

$$\boldsymbol{U}' = (\boldsymbol{u}_1 \ldots \boldsymbol{u}_m) \in R^{n \times m}$$

'Pad' with further orthonormal vectors to complete square matrix

$$\boldsymbol{U} = \begin{pmatrix} \boldsymbol{u}_1 & \ldots & \boldsymbol{u}_m & \boldsymbol{u}'_{m+1} & \ldots & \boldsymbol{u}'_n \end{pmatrix} \in R^{n \times n}$$

Now have orthonormal basis

$$\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I} \qquad \in \mathbb{R}^{n \times n}$$

ctd

Collect the individual vector transformation

$$\boldsymbol{u}_j \sqrt{\lambda_j} = \boldsymbol{X}\, \boldsymbol{v}_j$$

into matrix transformation

$$\boldsymbol{U}\, \boldsymbol{\Sigma} = \boldsymbol{X}\, \boldsymbol{V} \qquad \in R^{n \times m}$$

$$\boldsymbol{U} \in R^{n \times n}, \qquad \boldsymbol{\Sigma} \in R^{n \times m}, \qquad \boldsymbol{X} \in R^{n \times m}, \qquad \boldsymbol{V} \in R^{m \times m},$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sqrt{\lambda_1} & 0 & \dots & 0 \\ 0 & \sqrt{\lambda_2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & \sqrt{\lambda_m} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \end{pmatrix} \in R^{n \times m}$$

ctd

Finally, right-multiply with $\boldsymbol{V}^T$ to obtain

$$\boldsymbol{U}\,\boldsymbol{\Sigma} \;=\; \boldsymbol{X}\,\boldsymbol{V} \qquad |\cdot \boldsymbol{V}^T$$

$$\boldsymbol{U}\,\boldsymbol{\Sigma}\,\boldsymbol{V}^T \;=\; \boldsymbol{X}$$

To recapitulate, for a rectangular matrix $\boldsymbol{X} \in R^{n\times m}$, we considered the eigenvectors and eigenvalues of the covariance matrix

$$\boldsymbol{X}^T\,\boldsymbol{X}\,\boldsymbol{v}_j = \lambda_j\,\boldsymbol{v}_j$$

and obtained the singular value decomposition of our matrix $\boldsymbol{X}$

$$\boldsymbol{U}\,\boldsymbol{\Sigma}\,\boldsymbol{V}^T \;=\; \boldsymbol{X}, \qquad \Sigma_{jj} = \sqrt{\lambda_j}, \qquad \boldsymbol{V} = \begin{pmatrix} \boldsymbol{v}_1 & \dots & \boldsymbol{v}_m \end{pmatrix}$$