Dimensionality Reduction

PCA

sparse PCA

Non-Negative matrix factorization

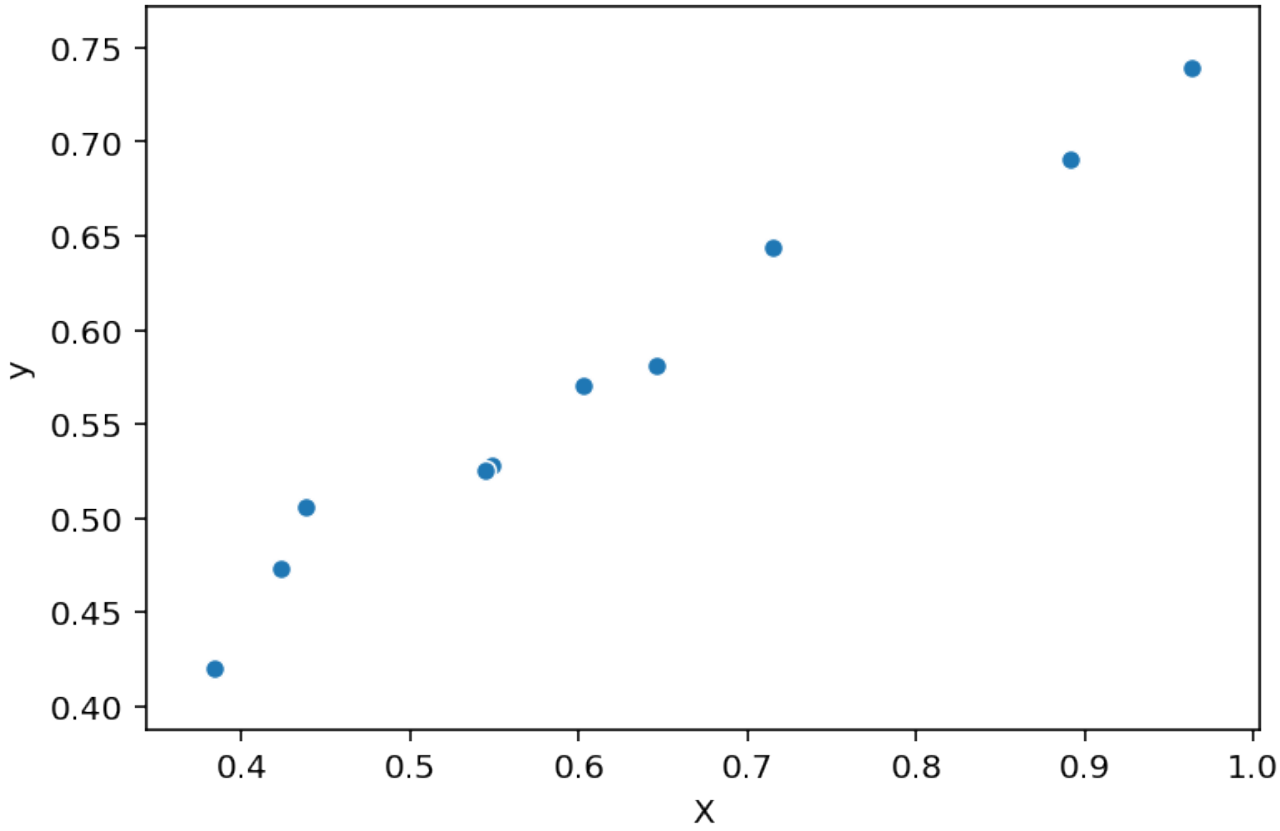
Centering

Choosing a technique

Compression

x	y
0.55	0.53
0.72	0.64
0.60	0.57
…	…

Suppose we want to compress these data.

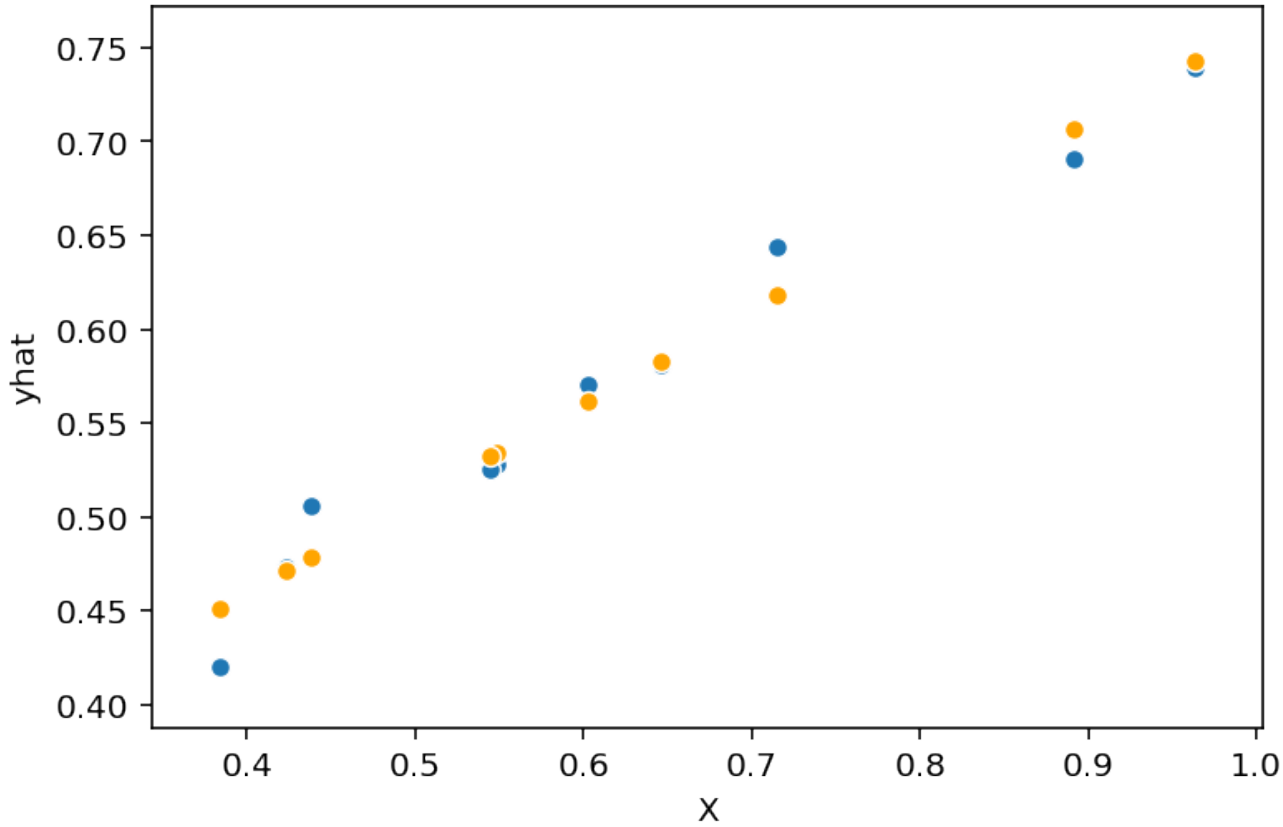


Compression

x	y
0.55	
0.72	
0.60	
…	

$$y \approx 0.26x + 0.5$$

Store x and
the formula.



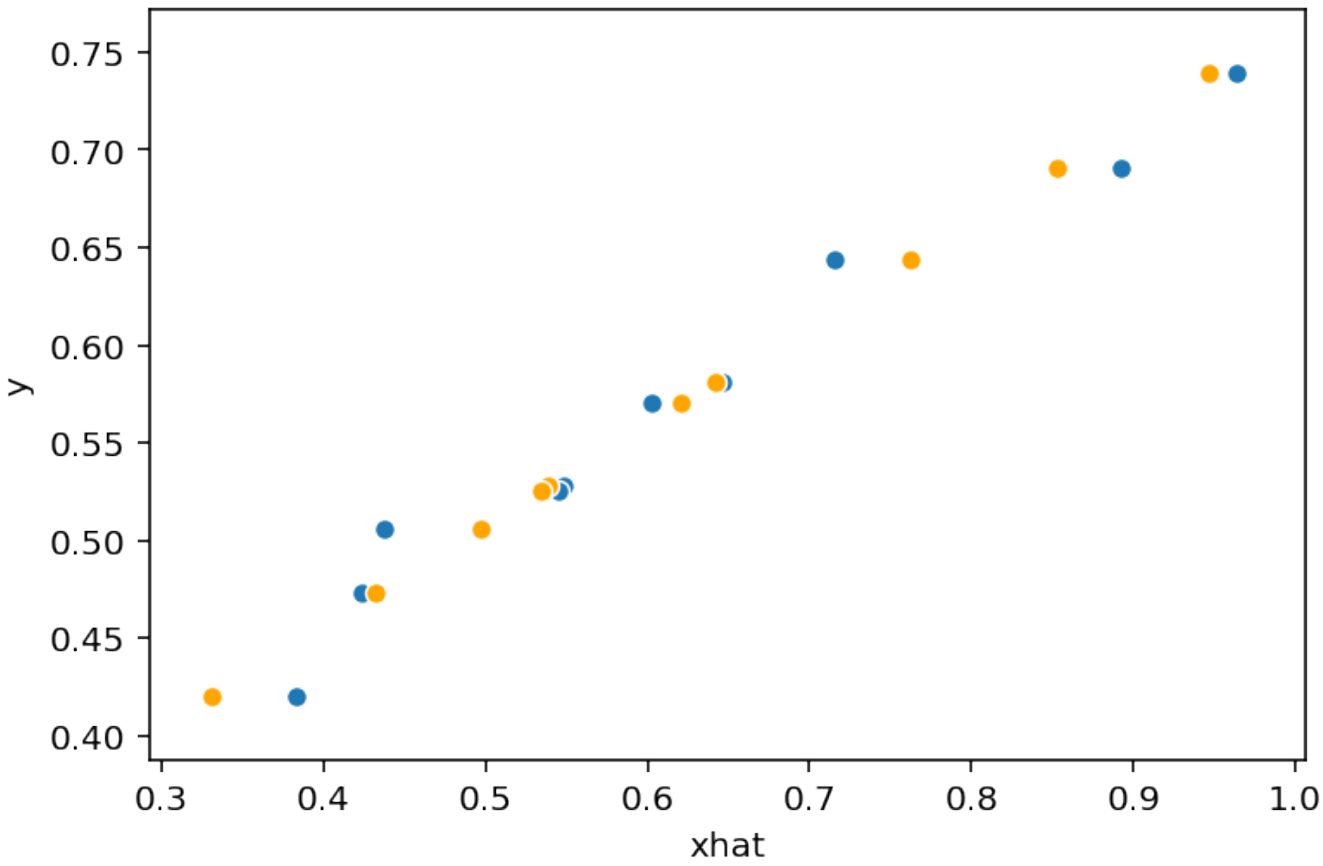
x is exactly right, y has some error: SSE=0.0028

Compression

x	y
	0.53
	0.64
	0.57
	…

$$x \approx 1.93y - 0.48$$

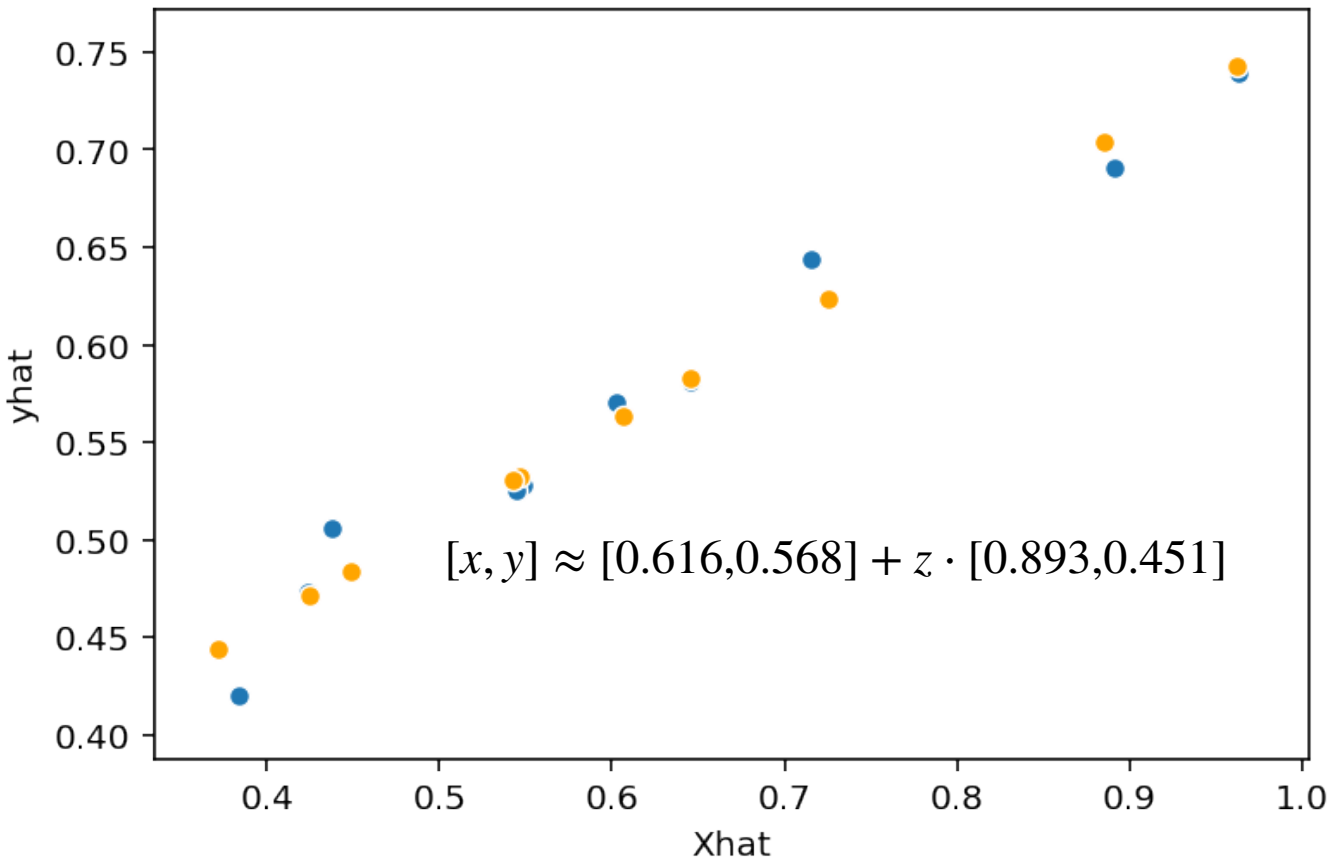
Store y and
the formula.



y is exactly right, x has some error: SSE=0.0109

Compression

z
-0.078
0.123
-0.010
…



Store z and the formula.

Both x and y have error, but SSE: 0.0023

Variance & Covariance

- **Variance** – A measure of the spread of the data in a dataset with mean $\bar{X}$
- **Covariance** – a measure of how much each of the dimensions varies from the mean with respect to each other.
- Covariance is measured between 2 dimensions to see if there is a relationship between the 2 dimensions, e.g., number of hours studied and grade obtained.
- The covariance between one dimension and itself is the variance

$$\text{var}(X) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})}{(n-1)}$$

$$\text{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{(n-1)}$$

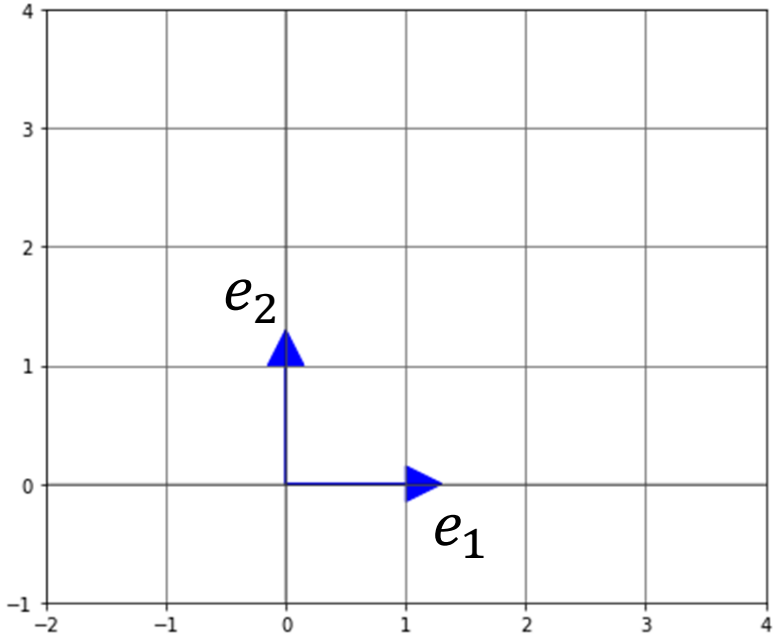
Basis

- A basis for $\mathbf{R}^n$ is a set of vectors which:
  - Spans $\mathbf{R}^n$, i.e. any vector in this *n*-dimensional space can be written as linear combination of these basis vectors.
  - Are linearly independent

Changing Basis

$$e_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

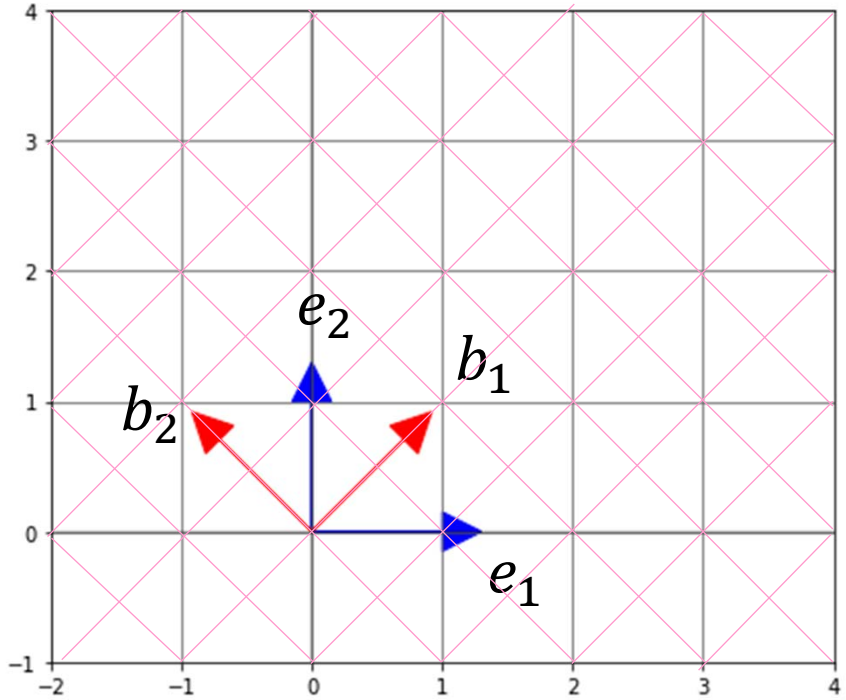
$$e_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$



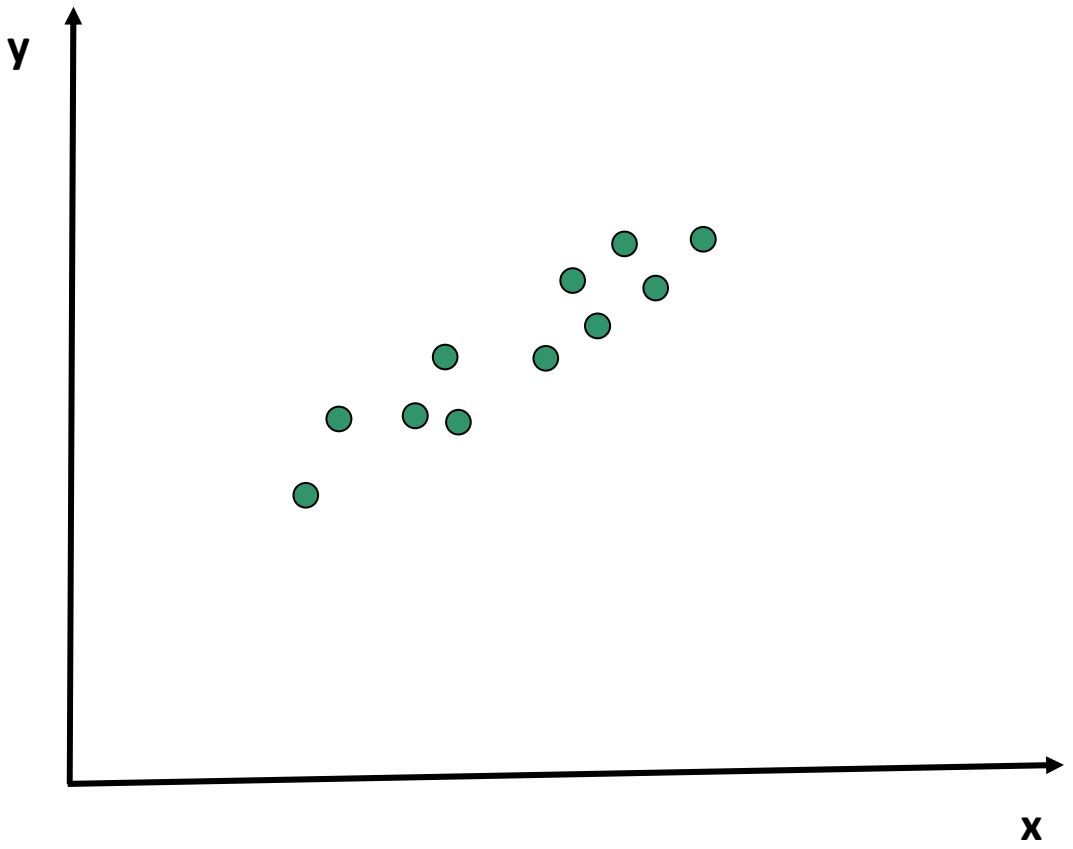
Changing Basis

$$e_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

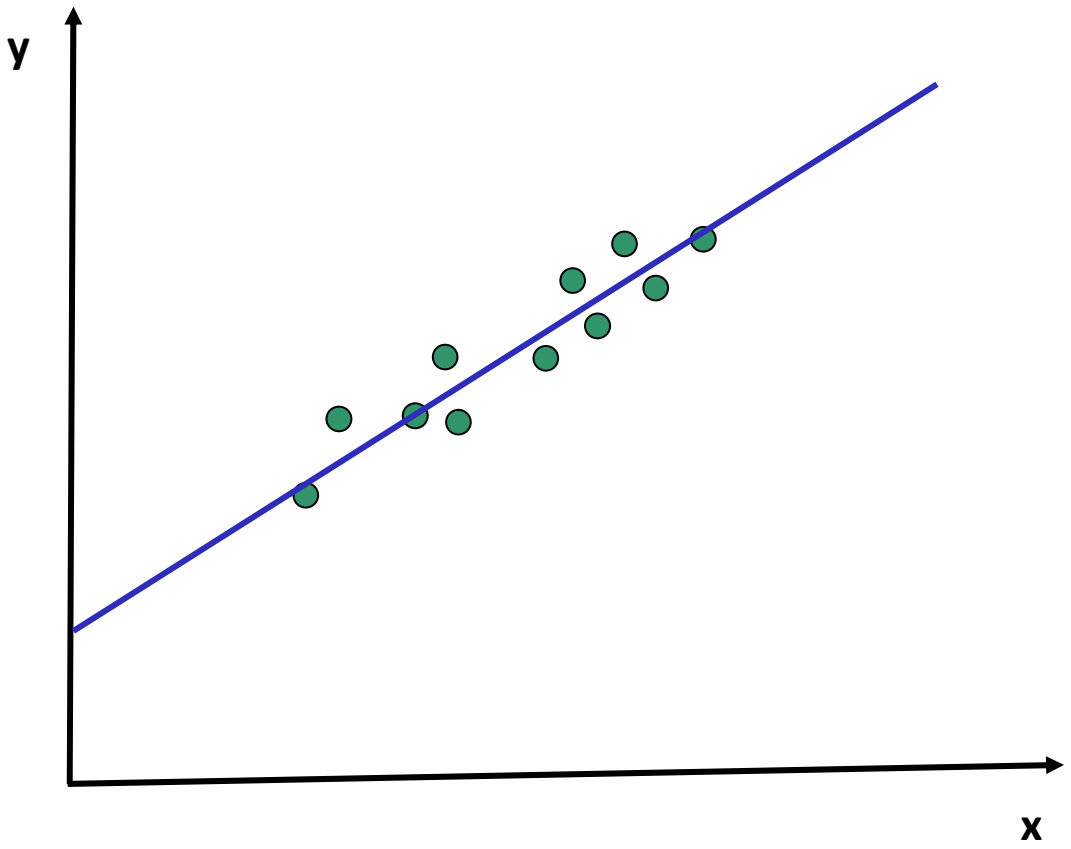
$$e_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$



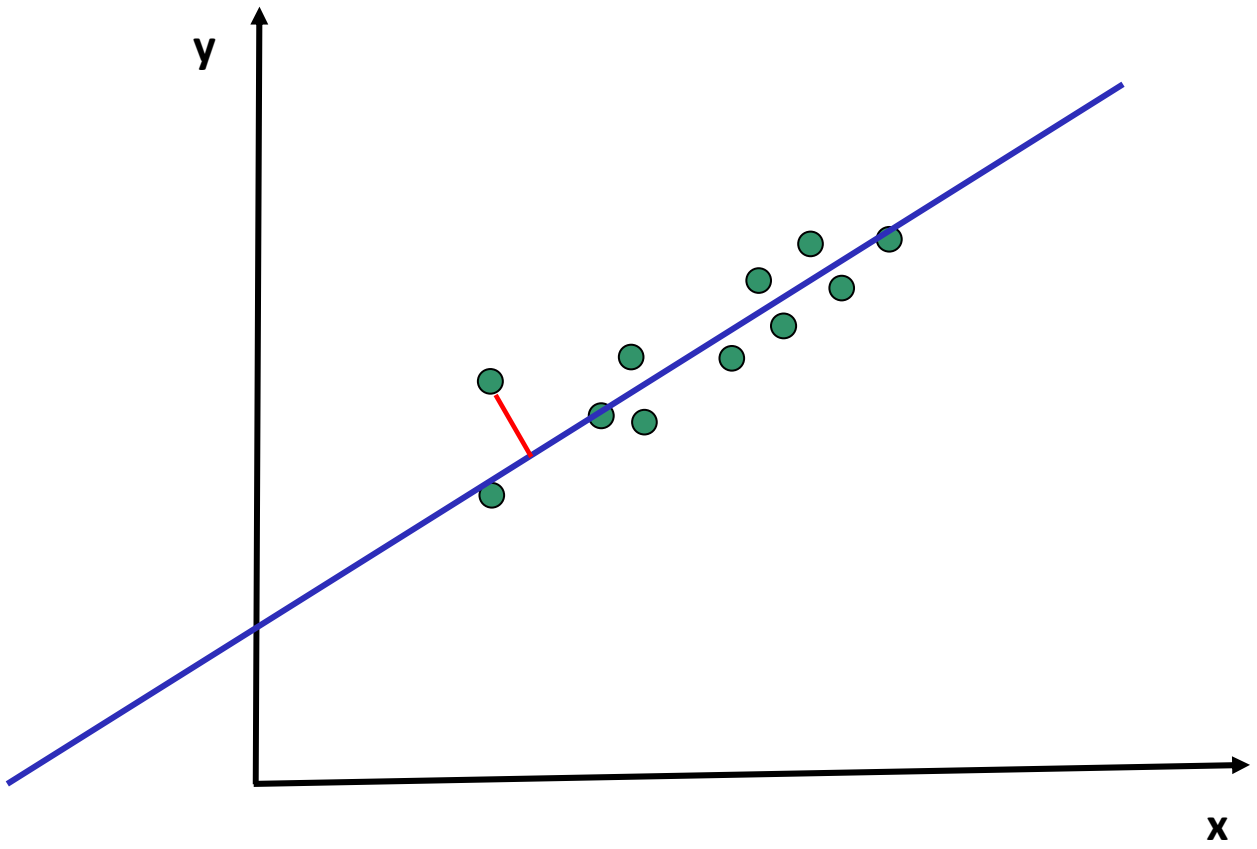
Application of changing basis



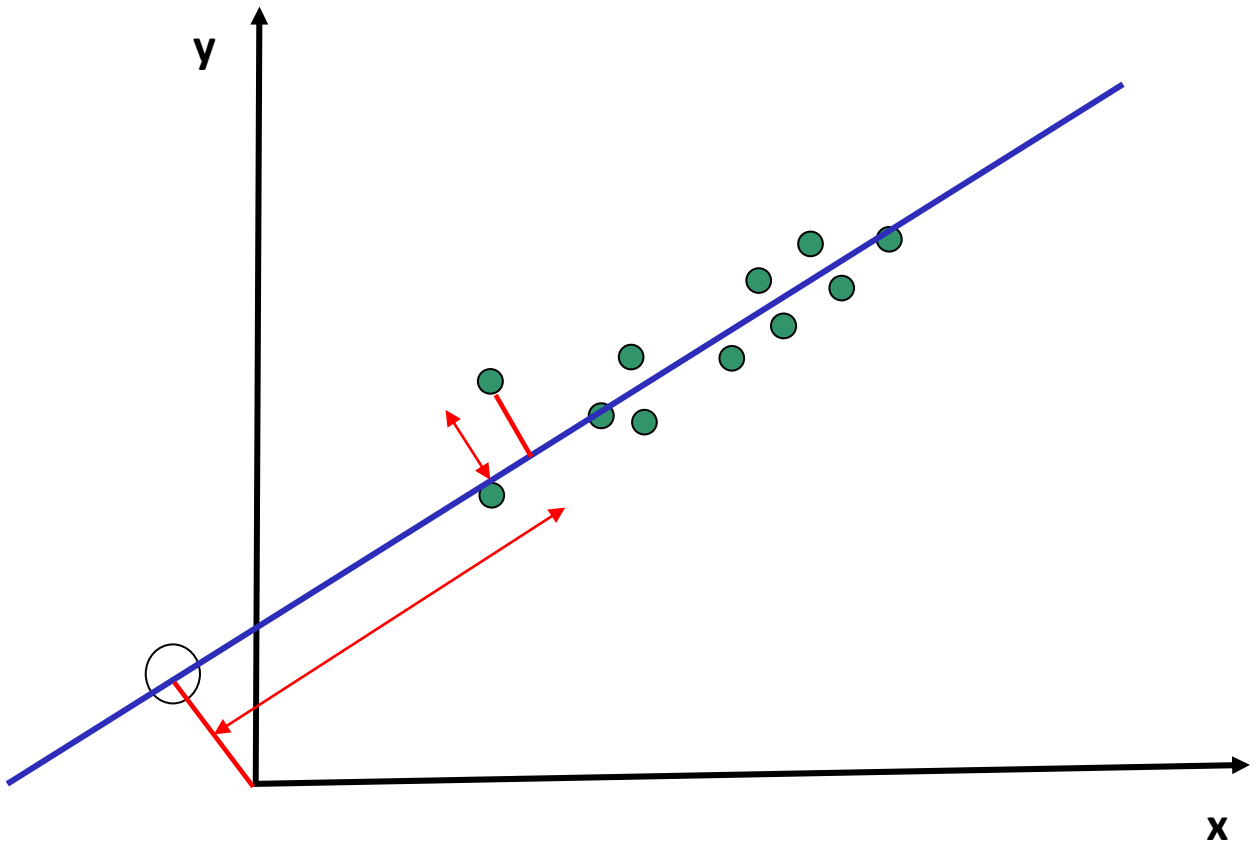
Application of changing basis



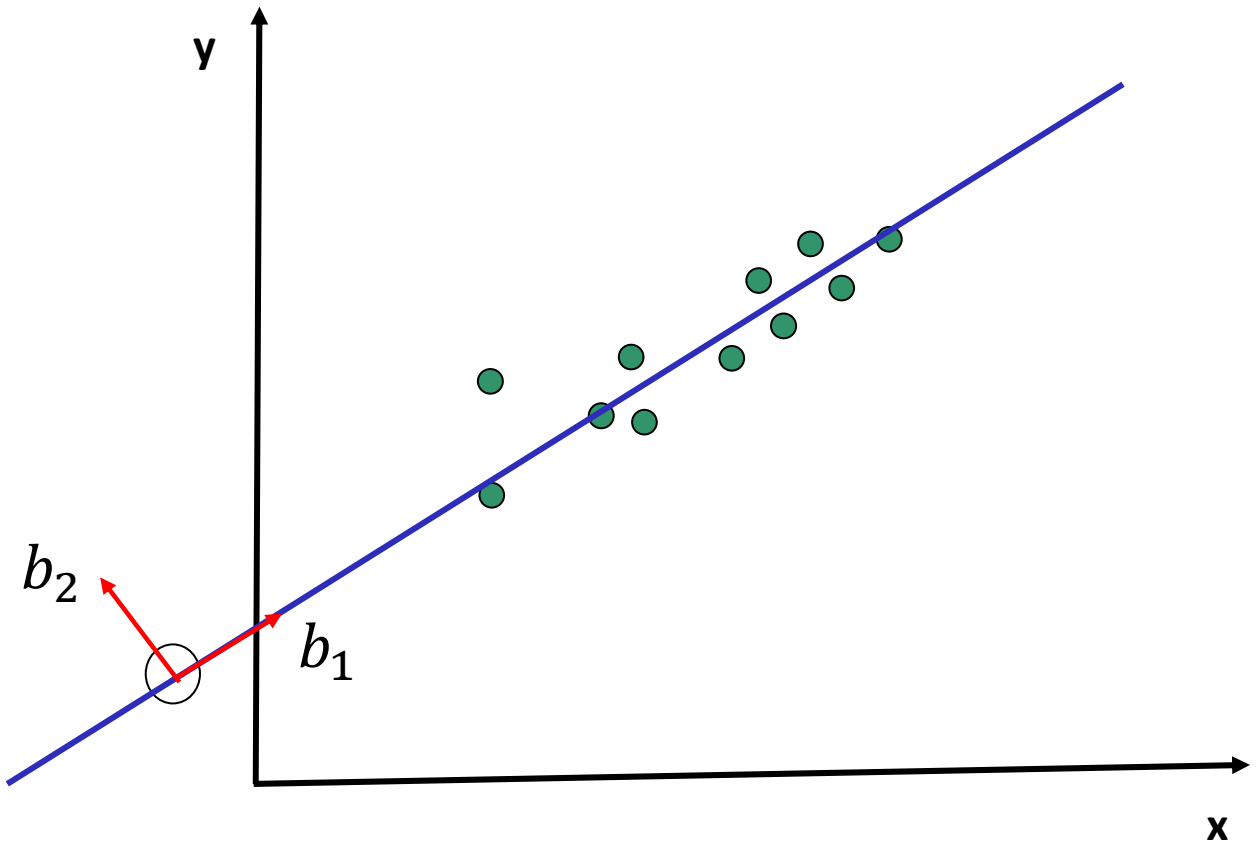
Application of changing basis



Application of changing basis



Principle Component Analysis



PCA Properties

- Finds directions that maximize variance
- These directions that are mutually orthogonal
- The first component has the highest variance
- The variation present in the PCs decrease as we move from the 1st PC to the last one
- The PCs are linear combinations of the original variables/basis.

$$\begin{bmatrix} 2 & 4 & 6 \\ -1 & -2 & -3 \end{bmatrix} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} [1 \quad 2 \quad 3]$$

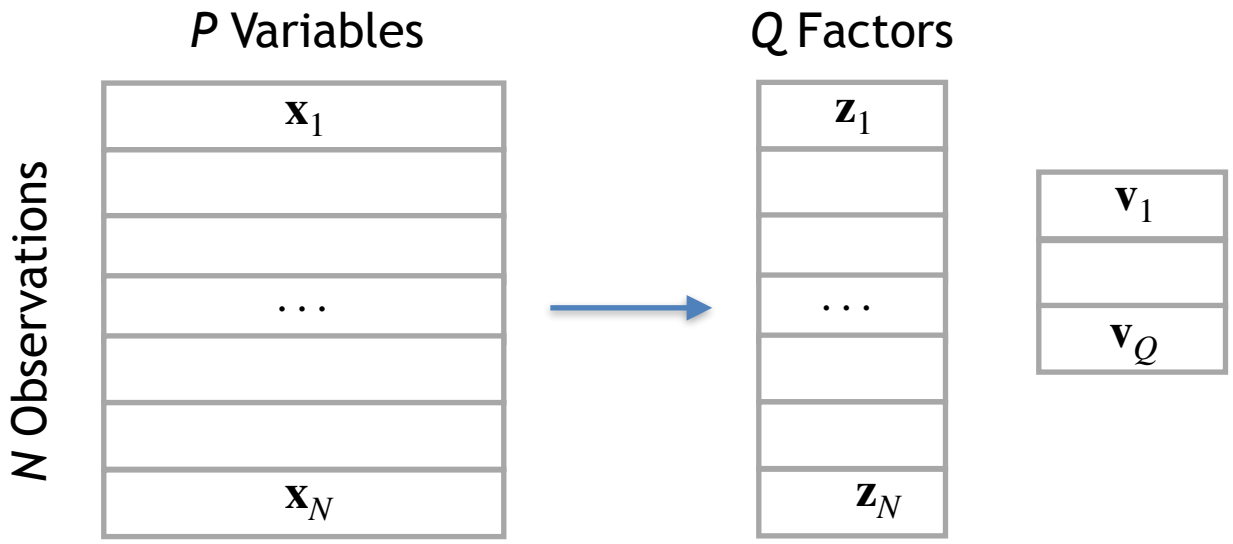
$$\begin{bmatrix} 2 & 4 & 6 \\ -1 & -2 & -3 \\ 3 & 7 & 10 \end{bmatrix} = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}$$

$$\mathbf{X} \quad = \quad \mathbf{z} \quad \mathbf{v}^T$$

$$\begin{bmatrix} 1.85 & 4.12 & 5.97 \\ -0.93 & -2.06 & -2.98 \\ 3.11 & 6.91 & 10.02 \\ 1.85 & 4.12 & 5.97 \\ -0.93 & -2.06 & -2.98 \\ 3.11 & 6.91 & 10.02 \\ 1.85 & 4.12 & 5.97 \\ -0.93 & -2.06 & -2.98 \\ 3.11 & 6.91 & 10.02 \end{bmatrix} = \begin{bmatrix} -1.93 \\ 0.96 \\ -3.23 \\ -1.93 \\ 0.96 \\ -3.23 \\ -1.93 \\ 0.96 \\ -3.23 \end{bmatrix} \begin{bmatrix} -0.96 & -2.13 & -3.10 \end{bmatrix}$$

General Setting

We have a large, multivariate data set,



and we want to condense every observation a smaller vector with less dimensions

Examples

- p Personality test -> q Personality factors
- p Neurons -> q neuronal dimensions
- p Customer choices -> q preference dimensions
- p Terms in documents -> q topic dimensions

Reasons why

- Data compression
- Noise reduction
- Feature detection for subsequent analysis - Clustering or similarity analysis (recommendation)
- Visualization
- Interpretation of the latent factors

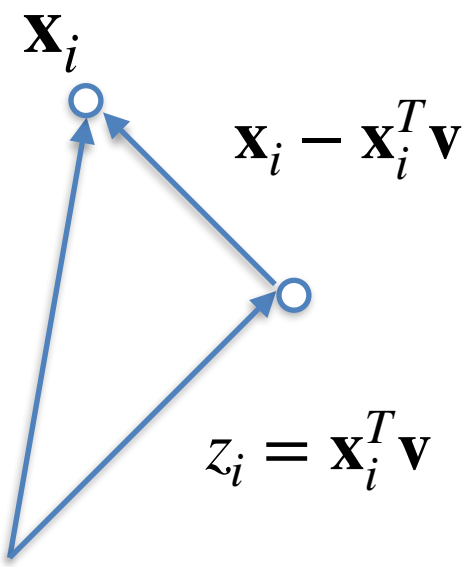
Some care and thought is necessary here!

Principle component analysis

$$J = \|\mathbf{X} - \mathbf{Z}\mathbf{v}\|^2$$
$$= \|\mathbf{X} - \mathbf{X}\mathbf{v}\mathbf{v}^\mathsf{T}\|^2$$

Subject to constraint

$$\mathbf{v}^\mathsf{T}\mathbf{v} = 1$$

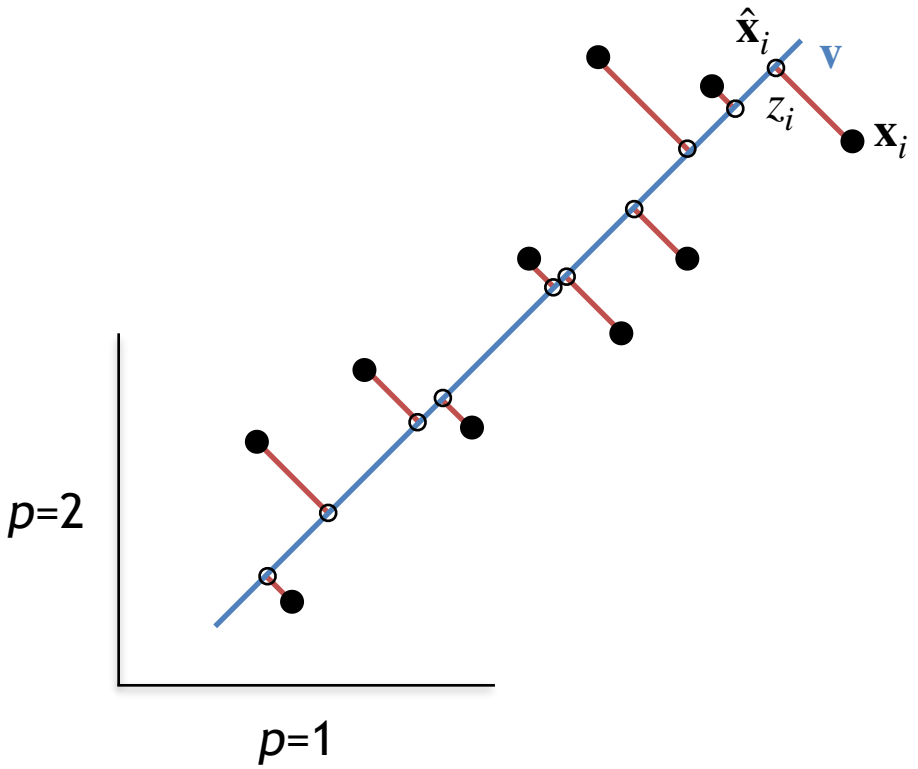


$$\|\mathbf{x}_i\|^2 = \|\mathbf{x}_i - z_i\mathbf{v}\|^2 + \|z_i\mathbf{v}\|^2$$

Minimizing error

Maximizing variance

Example for q=1



Minimize error

$$J = \|\mathbf{X} - \mathbf{X}\mathbf{v}\mathbf{v}^\mathsf{T}\|^2$$

Maximize variance of z_i

$$\lambda = \mathbf{z}^\mathsf{T}\mathbf{z} = \mathbf{v}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{v}$$

λ: Variance of z_i

First eigenvalue of $\mathbf{X}^\mathsf{T}\mathbf{X}$

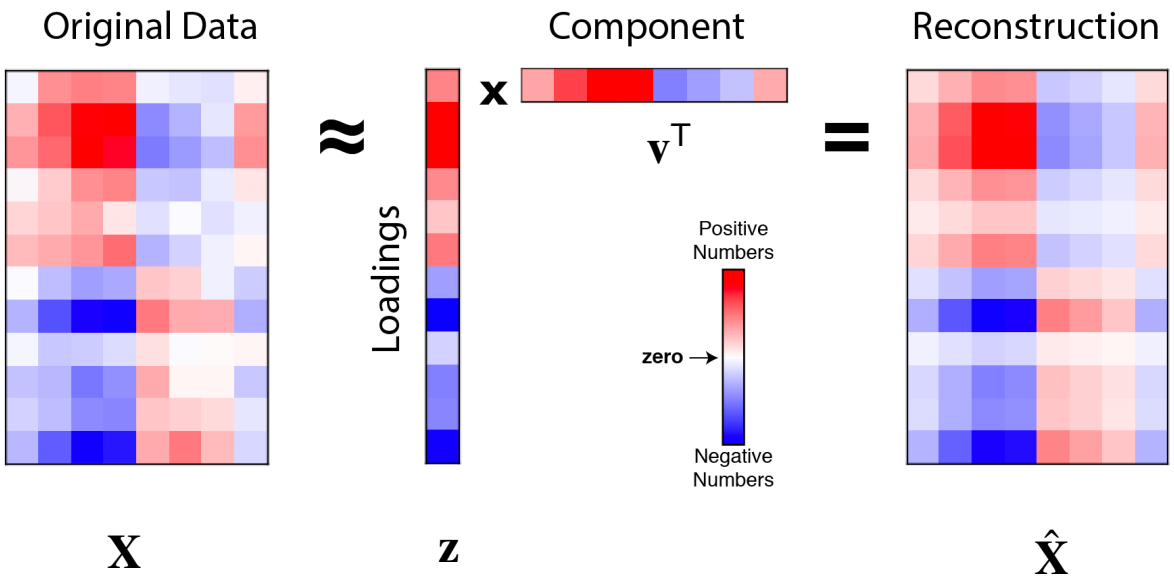
$\mathbf{v}$: First eigenvector

First eigenvector of $\mathbf{X}^\mathsf{T}\mathbf{X}$

Alex Williams: Everything you did and didn't know about PCA

HTF: Page 534-550

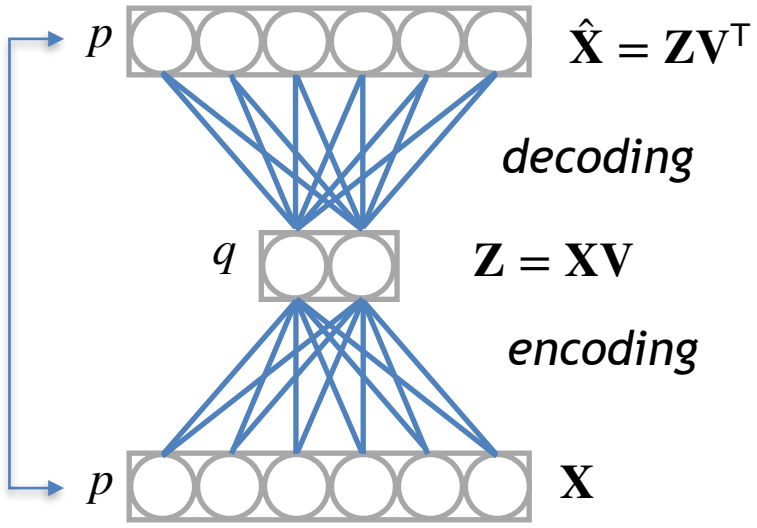
Principal component analysis



Alex Williams: [Everything you did and didn't know about PCA]

HTF: Page 534-550

Principal component analysis



- When considering multiple latent variables, we have multiple eigenvectors (each a column in $\mathbf{V}$)
- The eigenvectors have to have length 1 and are uncorrelated, $\mathbf{V}^\mathsf{T}\mathbf{V} = \mathbf{I}$
- The objective is to minimize the squared error

$$J = \|\mathbf{X} - \mathbf{X}\mathbf{V}\mathbf{V}^\mathsf{T}\|^2$$

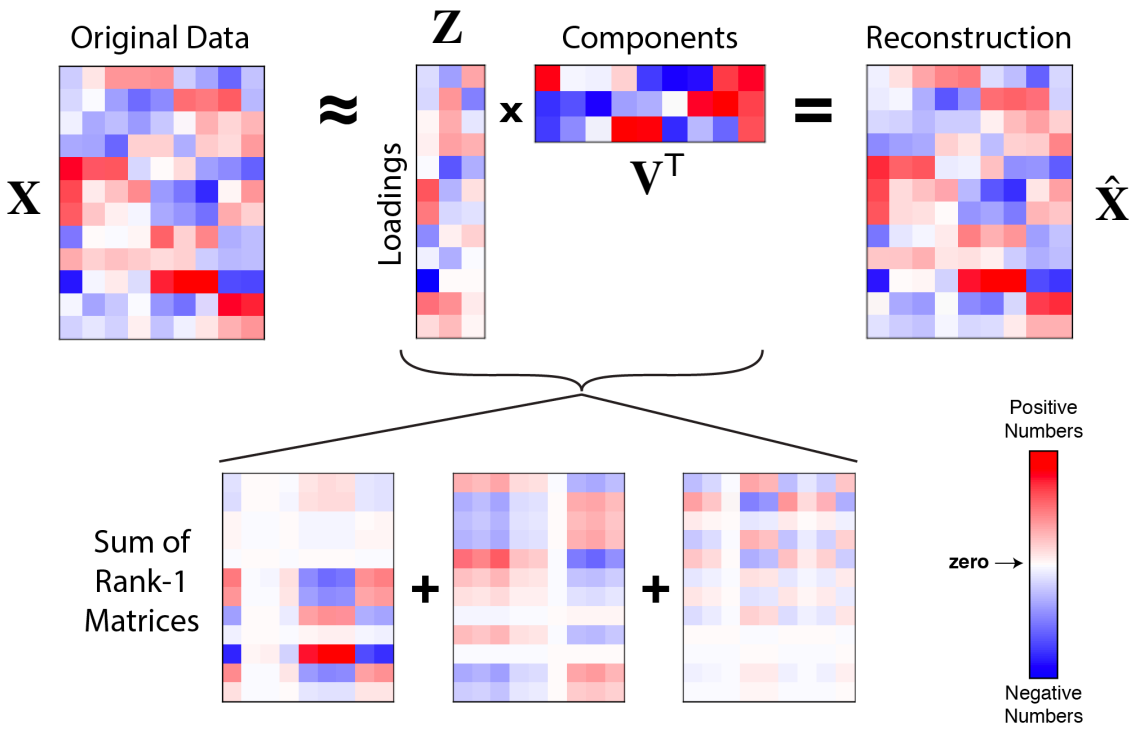
or equivalently, maximize the variance of z

$$\text{trace}(\mathbf{V}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{V})$$

Alex Williams: Everything you did and didn't know about PCA

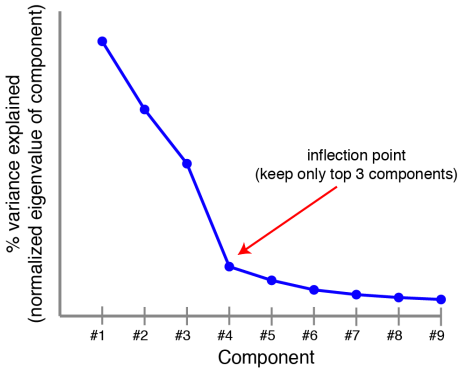
HTF: Page 534-550

Principal component analysis



- Each of the components has its own eigenvalue
- Eigenvalues = the sum-of-squares explained
- A scree-plot shows the ordered eigenvalues
- Often data is reduced to an inflection point

What are possible problems with this?



Singular Value Decomposition

$$\mathbf{X} \approx \mathbf{Z} \quad \mathbf{V}^T$$

Normalize latent variables

$$s_i = \sqrt{\lambda_i}$$

$$\mathbf{X} \approx \quad \mathbf{U} \quad \mathbf{S} \quad \mathbf{V}^T$$

left Eigenvectors

singular values

right Eigenvectors

Eigenvectors of

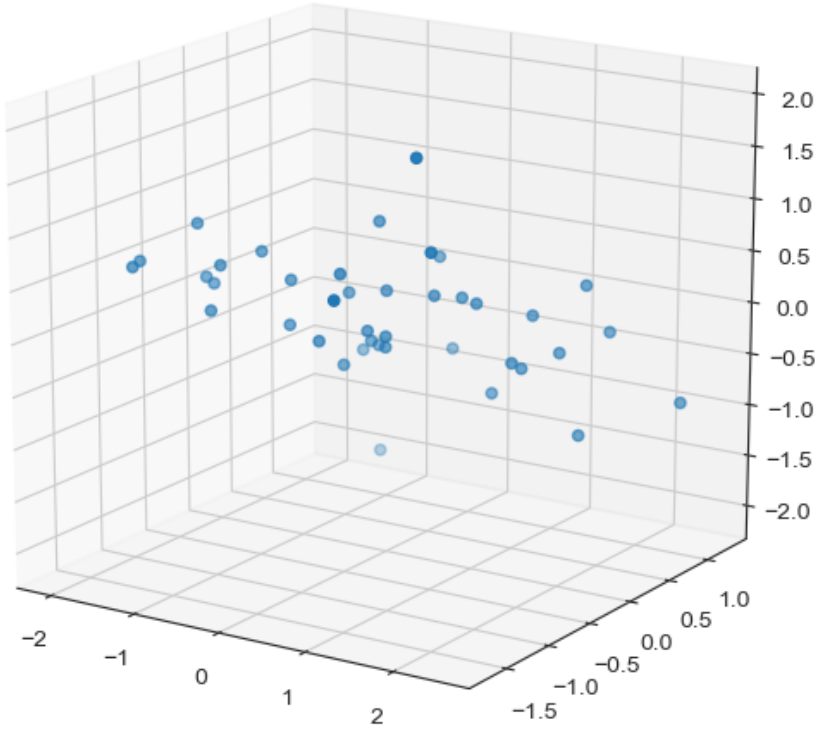
$$\mathbf{X}\mathbf{X}^T$$

Eigenvectors of

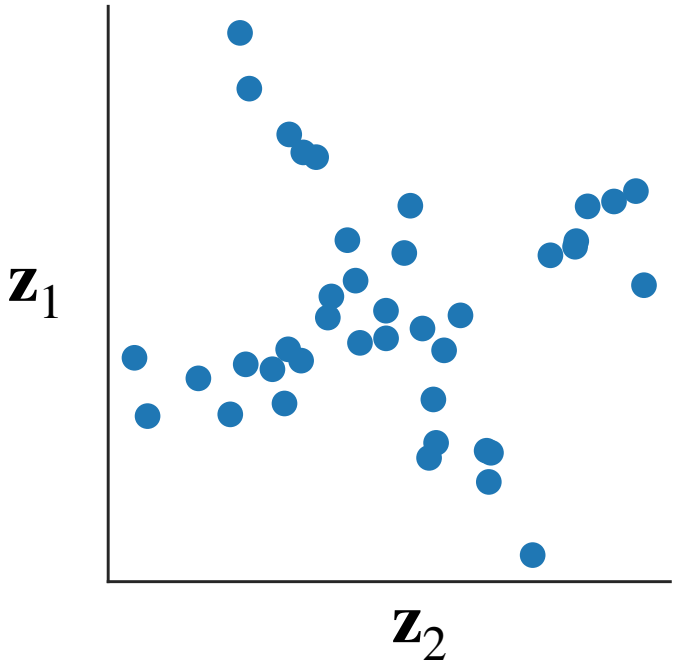
$$\mathbf{X}^T\mathbf{X}$$

PCA

Original Data Space (Y - 10D)



Latent Space (Z - 2D)

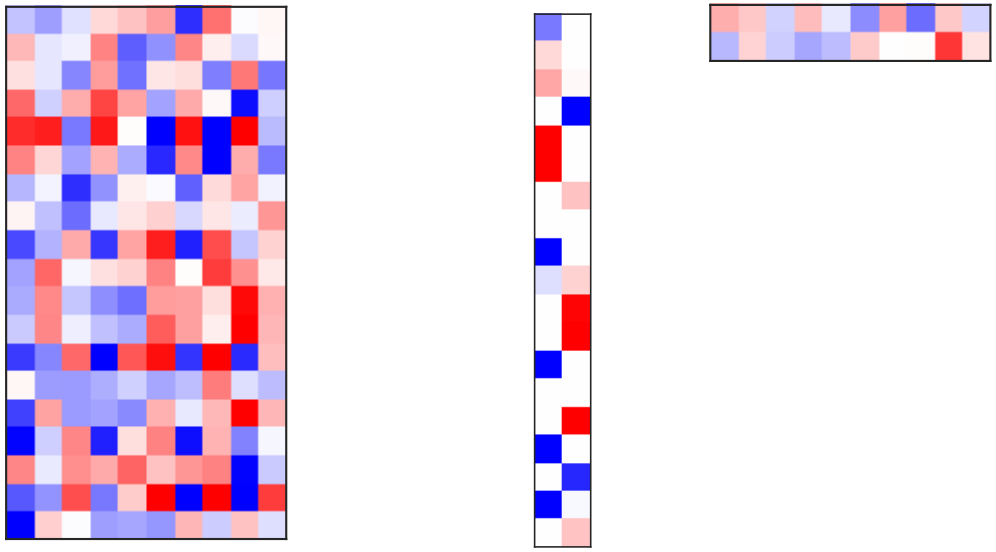


Although structure is become very clear, the principle components are not aligning with the main modes of the data -> Not easy to interpret PCs

Sparse PCA

Let's do a decomposition, where we look for a sparse solution in Z

$\mathbf{X} \approx \mathbf{Z} \quad \mathbf{V}^\mathsf{T}$



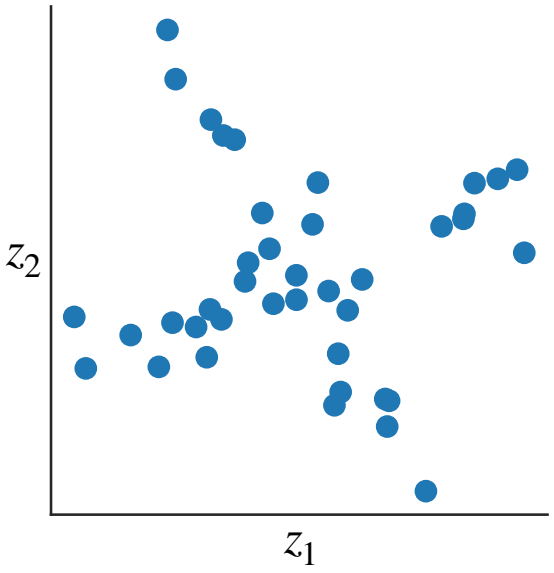
We can do this by imposing a L1-cost on the latent variables

$$J = \|\mathbf{X} - \mathbf{Z}\mathbf{V}^\mathsf{T}\|_2 + \alpha\|\mathbf{Z}\|_1$$

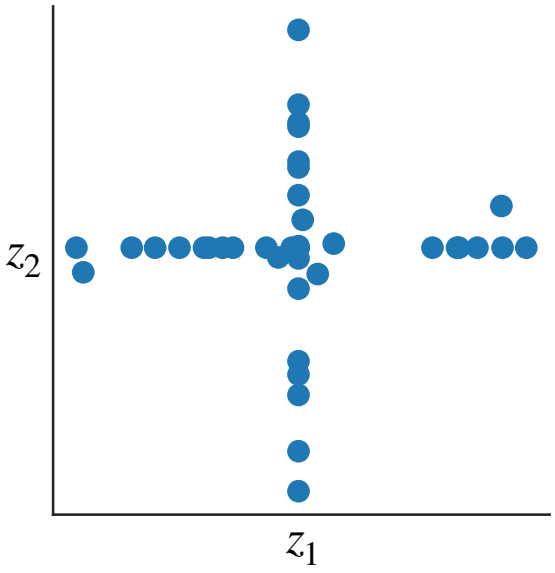
Subject to $\mathbf{V}\mathbf{V}^\mathsf{T} = \mathbf{I}$

Sparse PCA

PCA Solution



Sparse PCA Solution
(Z sparse)



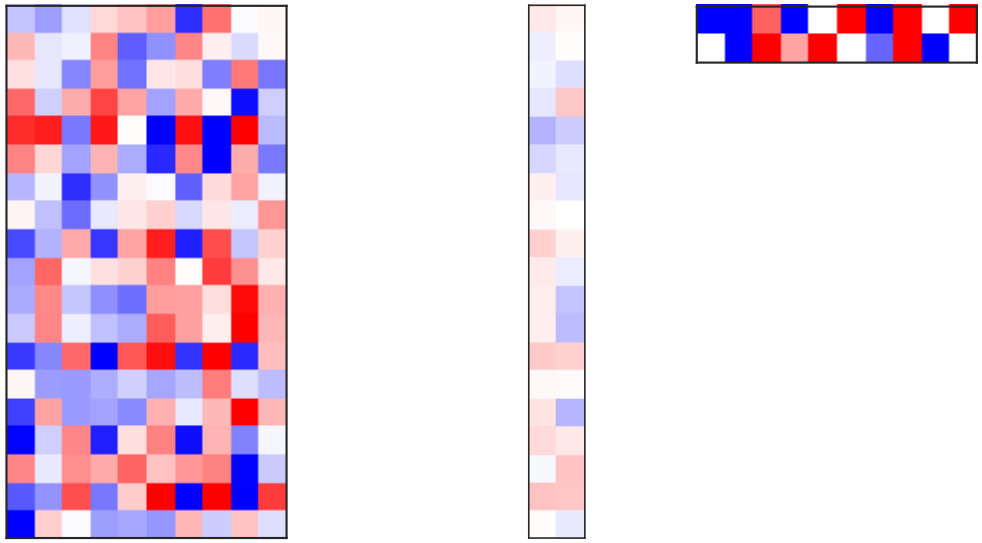
After imposing a spareness penalty:

- The principal axes are aligned with the major axis in the data
- The latent variable are shrunken towards the main axes

Sparse PCA

We can also impose the sparseness penalty of $\mathbf{V}$.

$\mathbf{X} \approx \mathbf{Z} \mathbf{V}^T$



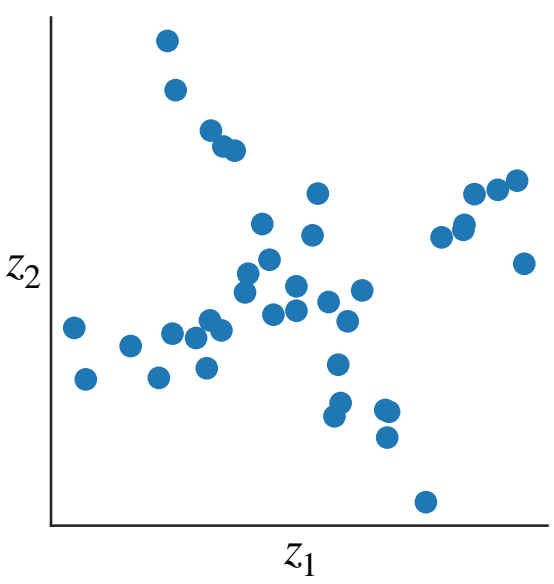
$$J = \|\mathbf{X} - \mathbf{Z}\mathbf{V}^T\|_2 + \alpha\|\mathbf{V}\|_1$$

Subject to $\mathbf{Z}^T\mathbf{Z} = \mathbf{I}$

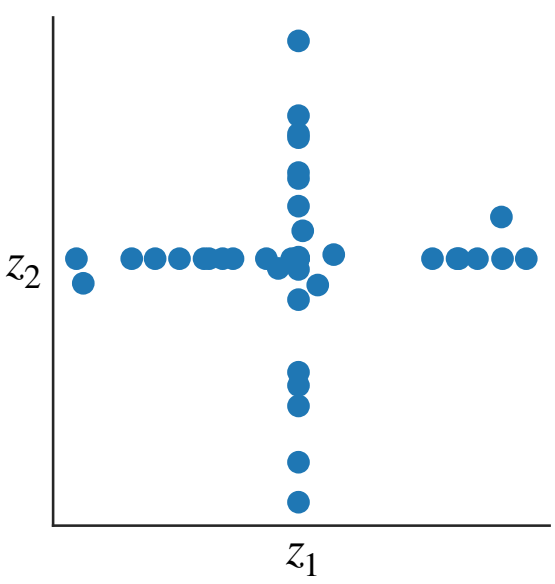
Indeed, this is what the sparsePCA function in sklearn does.

Sparse PCA

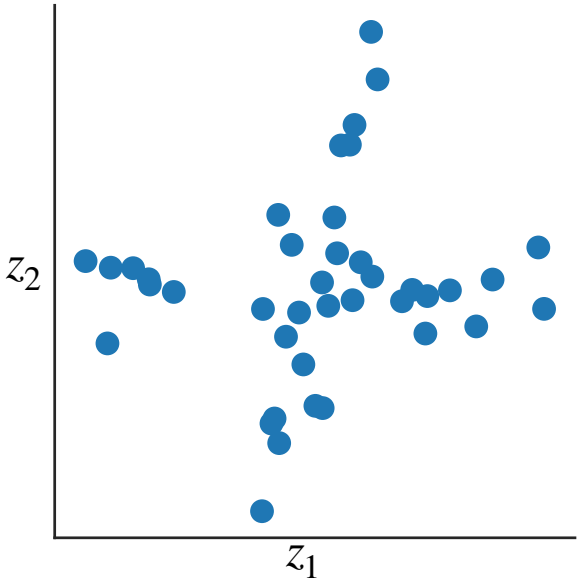
PCA Solution



Sparse PCA Solution
(Z sparse)



Sparse PCA Solution
(V sparse)



Imposing spareness on Z makes each observation loading on few principal vectors.

Imposing spareness on V makes each latent variable only involve a few of the output variables.

Topic analysis

A common problem is to analyze the topics of documents

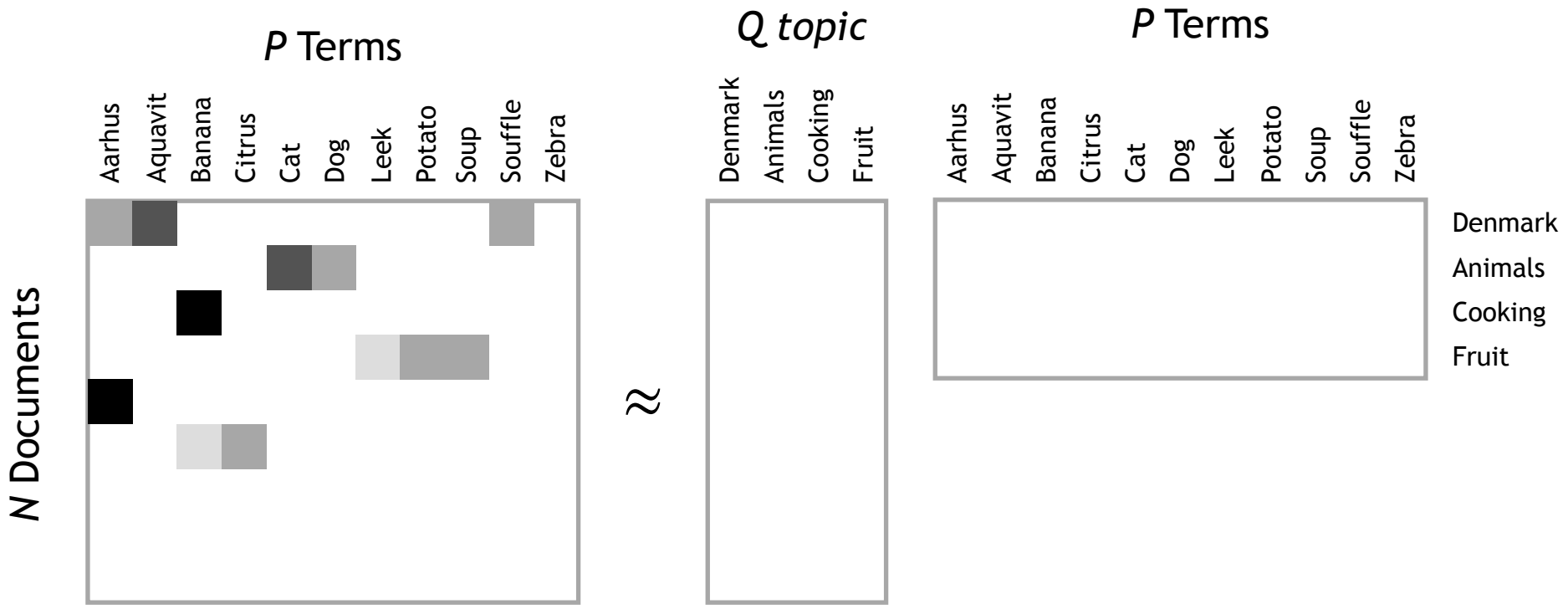
- Find documents for a specific topic
- Recommend similar documents to reader
- Cluster documents in groups

Data is often “tf-idf”

- **T**erm **F**requency: Relative frequency of word in document
- **I**nverse **D**ocument **F**requency: Divided by frequency of documents having this term

  “the” would have high term frequency, but low tf-idf

Topic analysis



What type of dimensionality reduction would be good here?

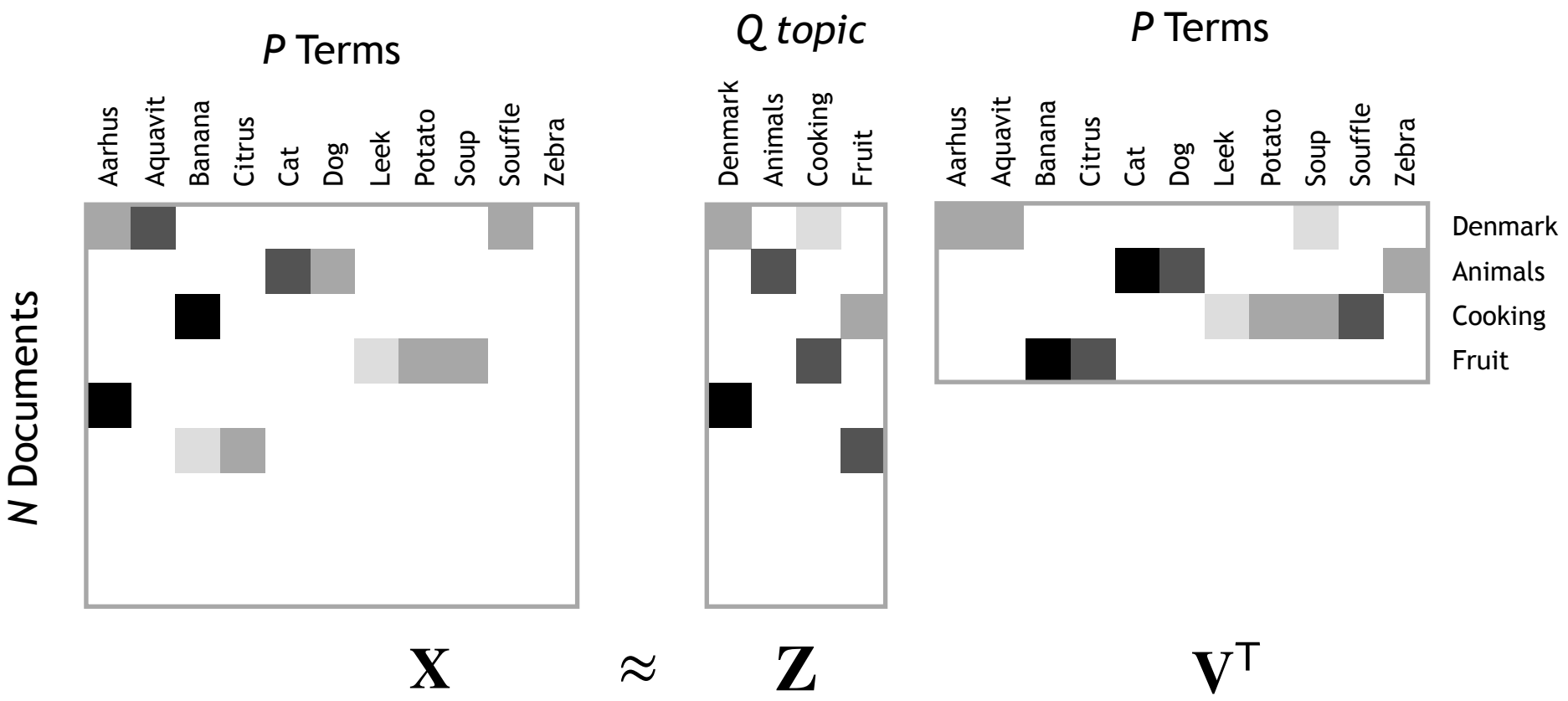
Sparseness would be good

How would we interpret negative weights in topic or term matrix?

i.e. how a document that has a -1.2 weight on Denmark?

-> negative weights do not make much sense here!

Non-Negative Matrix factorisation (NMF)



Minimize

$$\frac{1}{2}\|\mathbf{X} - \mathbf{Z}\mathbf{V}^{\mathsf{T}}\|_F^2$$

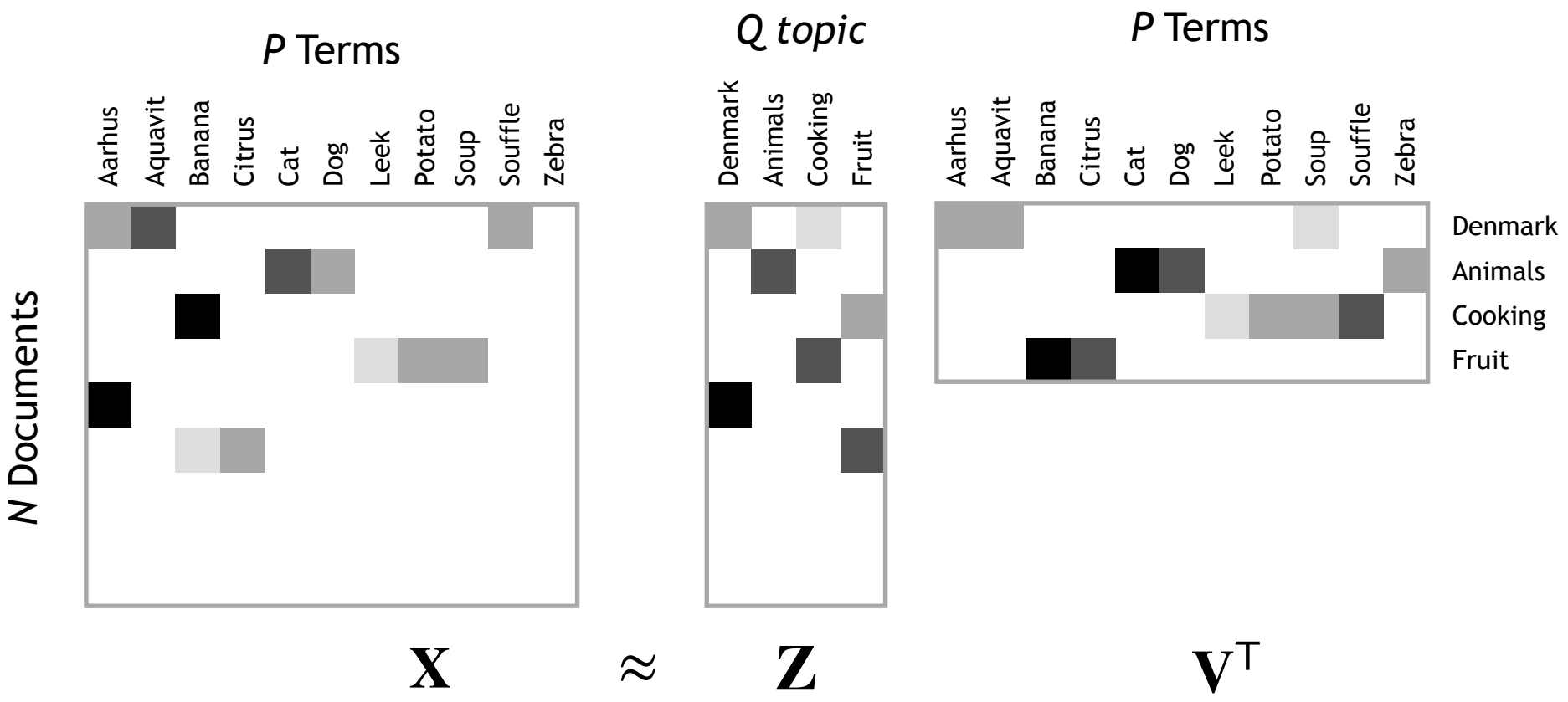
With constraints

$$z_{i,j} > 0 \qquad v_{i,j} > 0$$

Data can only be positive!

Alone does not lead to unique solutions

Non-Negative Matrix factorisation (NMF)



Minimize

$$\frac{1}{2}\|\mathbf{X} - \mathbf{Z}\mathbf{V}^{\mathsf{T}}\|_F^2$$

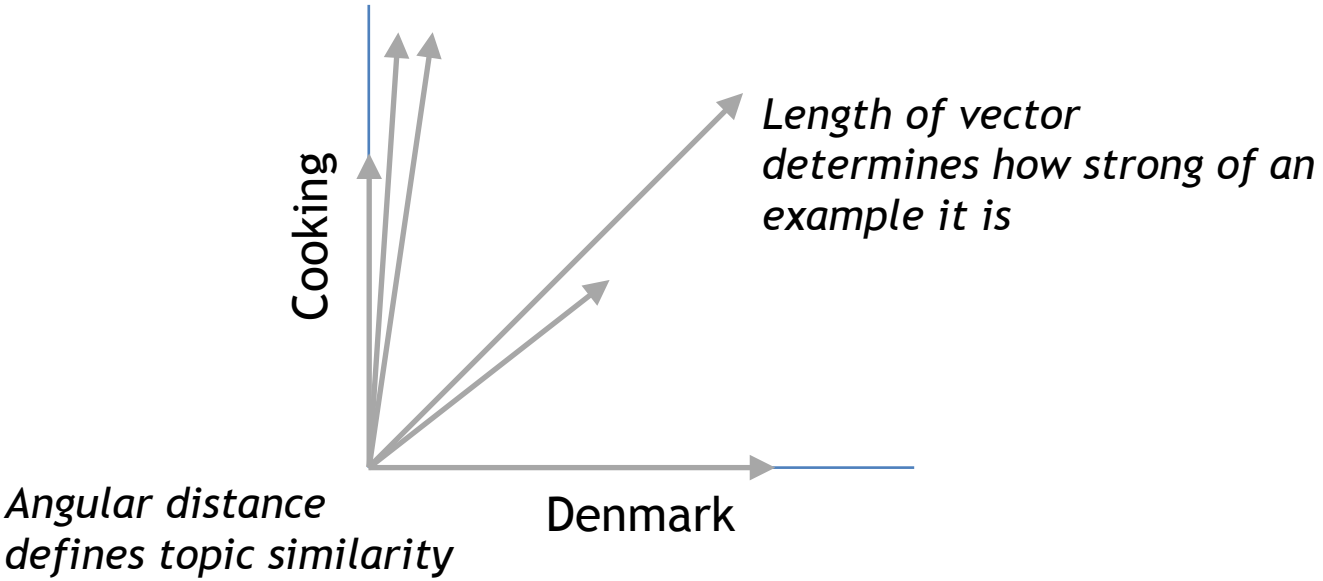
With constraints

$$z_{i,j} > 0 \qquad v_{i,j} > 0$$

So, often we add L1 and L2 penalties on *z* and *v*

Topic analysis

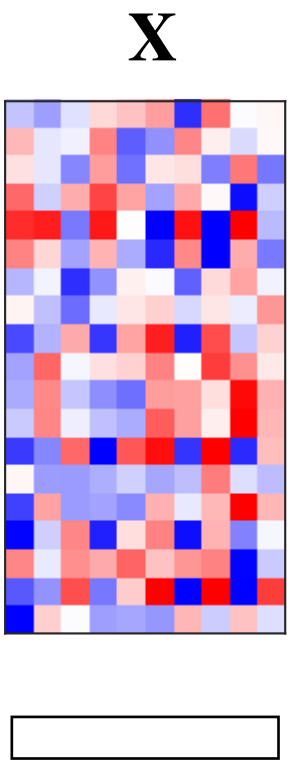
If somebody searches for “Danish cooking”



Two documents may be similar
even though they don’t share a single term
i.e. two documents about:

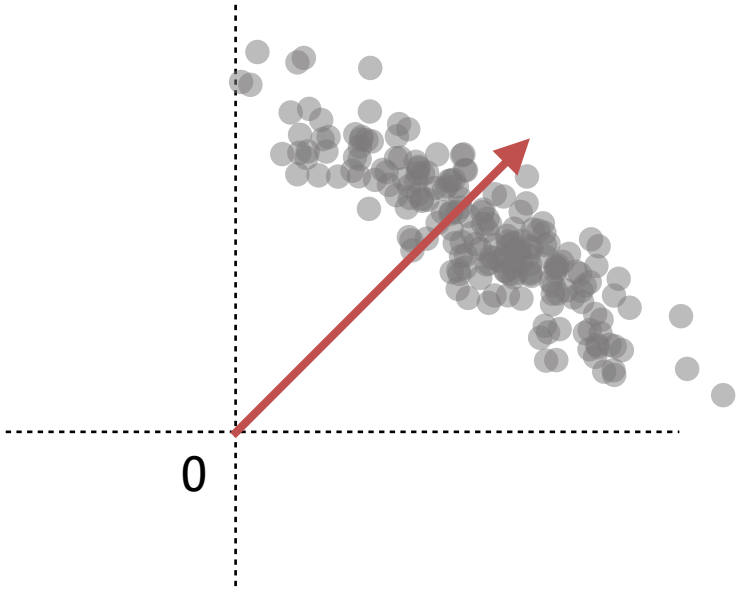
- *Aarhus meatballs*
- *Copenhagen recipes*

Centering

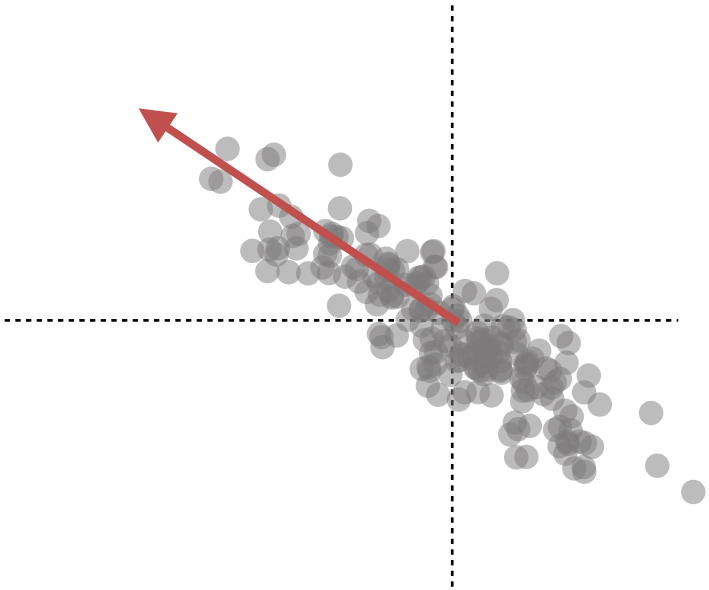


- Removing the mean from each data column before dimensionality reduction
- Pretty standard for most implementations (sometimes you can't control it)
- Not always the right thing

Centering

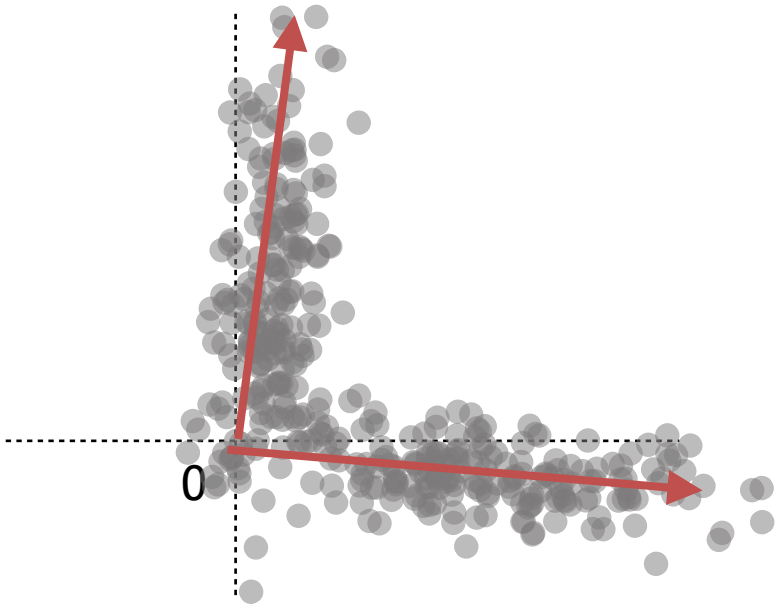


Before centering, the first PC reflects the direction of the biggest sums-of-squares (driven by the mean)

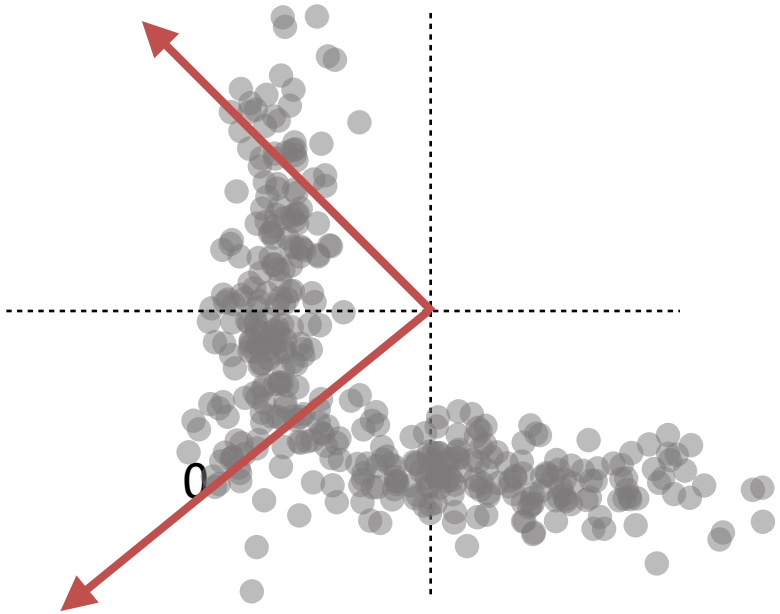


After centering, the first PC reflects the direction with the most variance

Centering continued

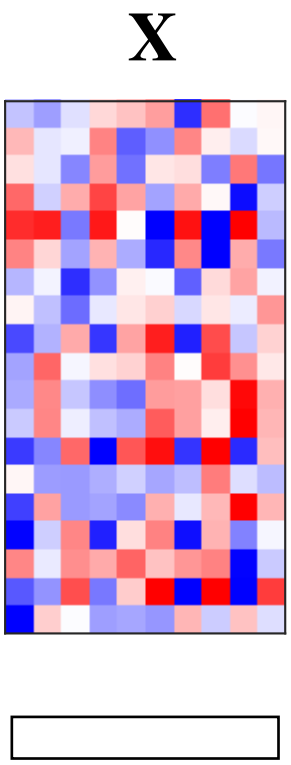


Before centering, the data can be well described by 2 sparse components



After centering, a sparse description of the data is no longer possible

Centering



- Always figure out when and how the algorithm you use centers the data
- Always think critically if centering makes sense

Flexible menu of dimensionality reduction techniques

Choosing a dimensionality reduction technique is a bit like ordering a Pizza. Lots of ways to combine toppings!

- Reconstruction cost
- Constraints on **Z**
- Constraints on **V**
- Regularisation on **Z**
- Regularisation on **V**
- Centering

- PCA
- Sparse PCA
- Dictionary learning
- Regularized PCA
- Nonnegative matrix factorization
- Semi-nonnegative matrix factorization
- Independent component analysis (ICA)

- Non-linear transforms of data

- Kernel PCA
- t-SNE
- ISOMAP
- Deep auto encoding

How to choose?

Depends a lot on what the goal is

Always try to find an independent evaluation criterion (maybe predict new data)?

Respect natural constraints in the data

Interpretability of latent factors is easier if the dimensionality reduction matches your generative model of the data