香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

COMP 5212
Machine Learning
Lecture 13

Dimensionality Reduction

Junxian He
Oct 22, 2024

Midterm Exam

Thursday

Tomorrow (Oct 24), 1:20pm-2:40pm, one A4-size double-sided cheetsheet is allowe
(either printing or handwriting is fine)

We have two rooms for the exam for sparse seat plans:

1. For SIS ID ending with an even digit: Room 2303
2. For SIS ID ending with an odd digit: Room 2504

next Tuesday

High-Dimensional Data

- High-Dimensions = Lot of Features

High-Dimensional Data

- High-Dimensions = Lot of Features

Document classification

Features per document =

thousands of words/unigrams

millions of bigrams, contextual information

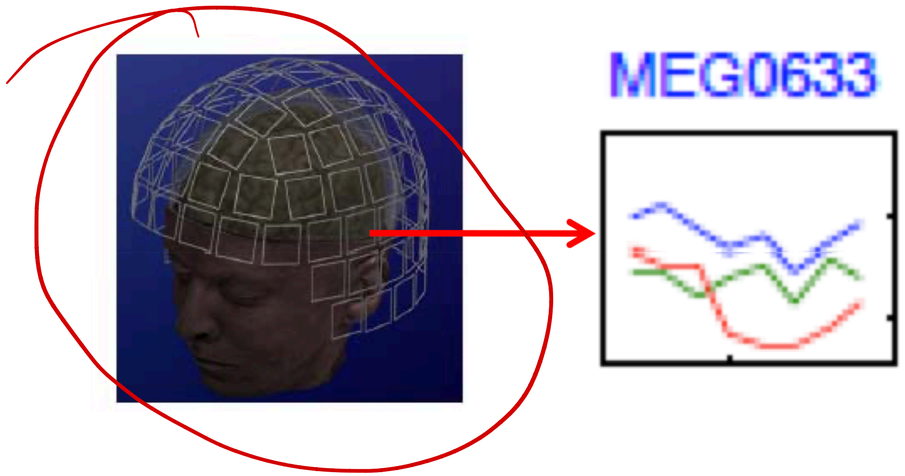
High-Dimensional Data

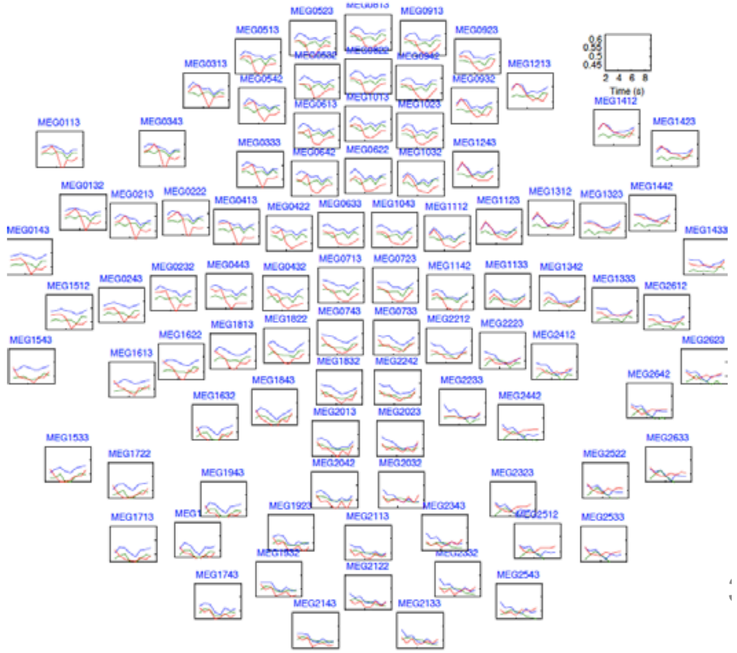
- High-Dimensions = Lot of Features

MEG Brain Imaging

120 locations x 500 time points

x 20 objects





Curse of Dimensionality

- Why are more features bad?

Curse of Dimensionality

- Why are more features bad?
  - Redundant features (not all words are useful to classify a document) more noise added than signal

Curse of Dimensionality

- Why are more features bad?
  - Redundant features (not all words are useful to classify a document) more noise added than signal
  - Hard to store and process data (computationally challenging)

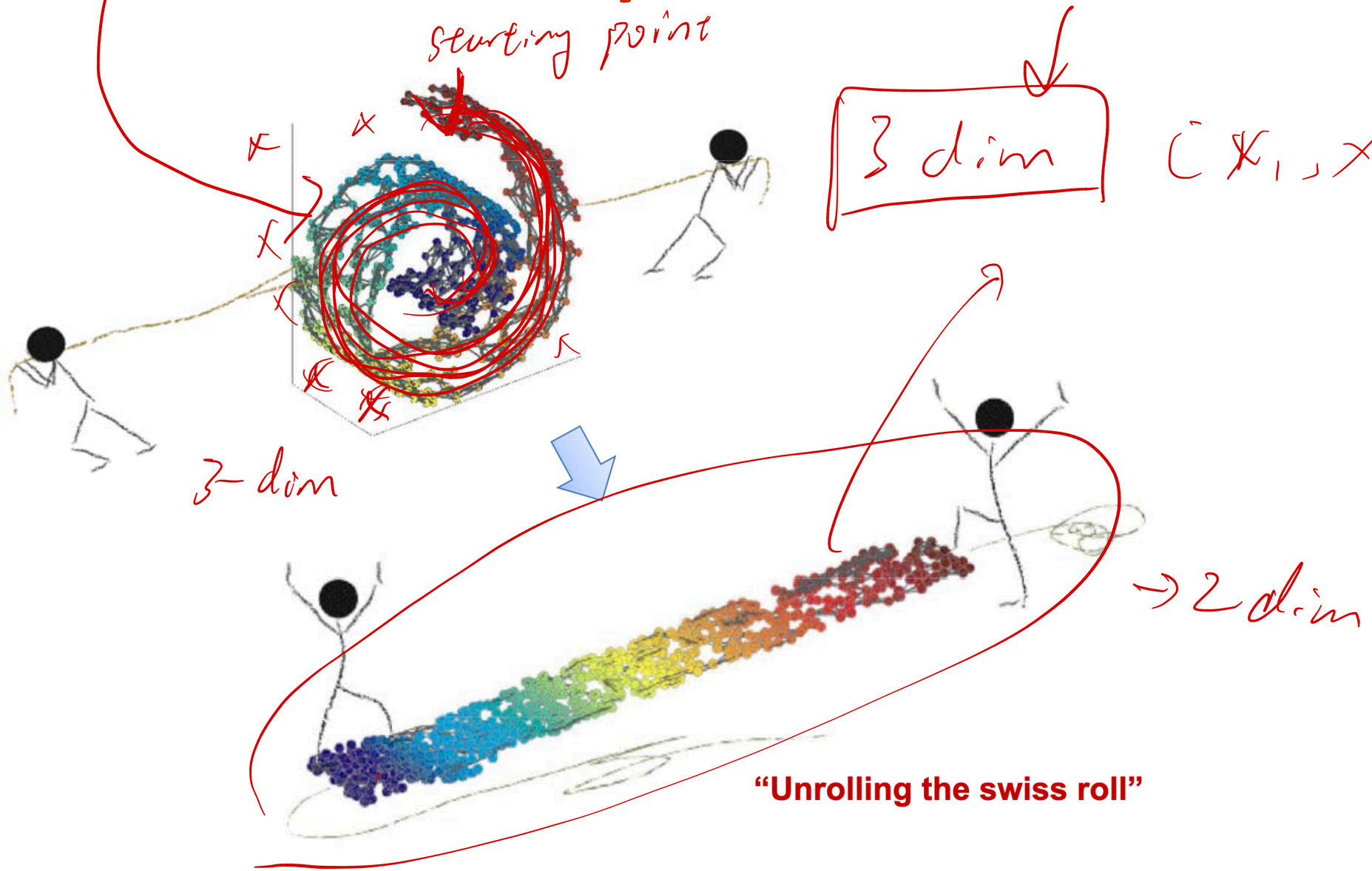
Curse of Dimensionality

- Why are more features bad?
  - Redundant features (not all words are useful to classify a document) more noise added than signal
  - Hard to store and process data (computationally challenging)
  - Hard to interpret and visualize

Curse of Dimensionality

- Why are more features bad?
  - Redundant features (not all words are useful to classify a document) more noise added than signal
  - Hard to store and process data (computationally challenging)
  - Hard to interpret and visualize
  - Complexity of decision rule tends to grow with # features

Dimensionality Reduction

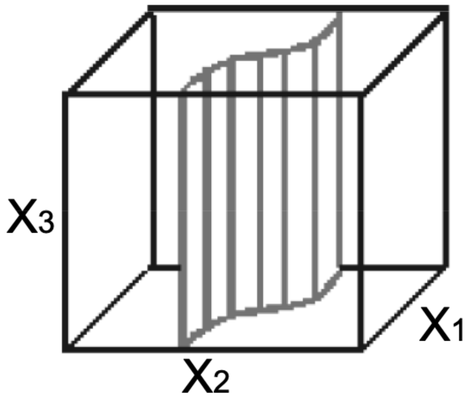


"Unrolling the swiss roll"

Dimensionality Reduction

Dimensionality Reduction

- Feature Selection – Only a few features are relevant to the learning task

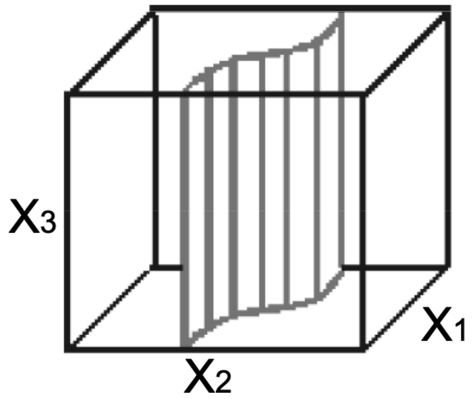


X3 - Irrelevant

supervised learning

Dimensionality Reduction

- Feature Selection – Only a few features are relevant to the learning task

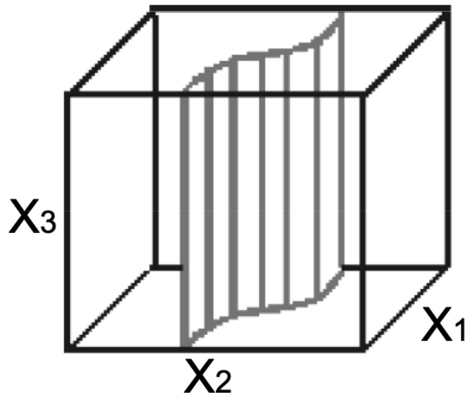


X_3 - Irrelevant

- Latent features – Some linear/nonlinear combination of features provides a more efficient representation than observed features

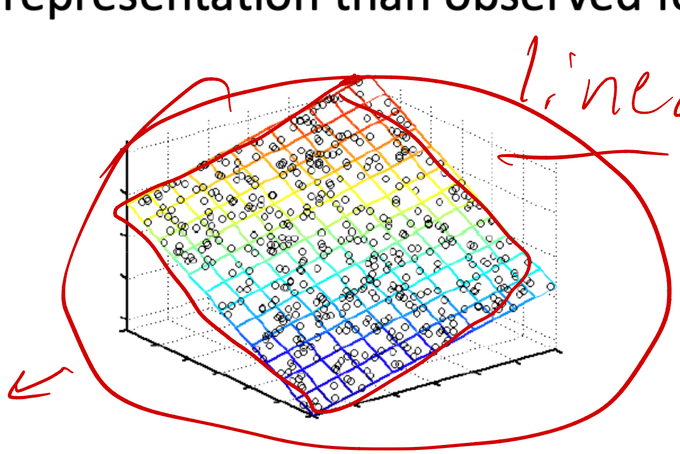
Dimensionality Reduction

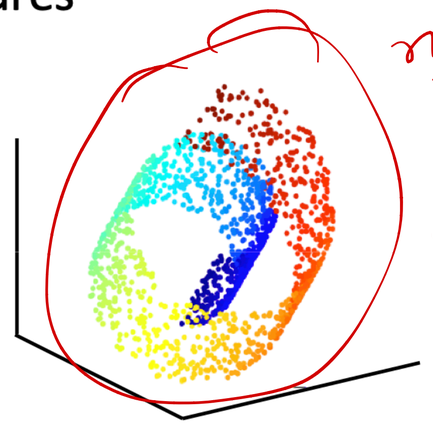
- Feature Selection – Only a few features are relevant to the learning task



X_3 - Irrelevant

- Latent features – Some linear/nonlinear combination of features provides a more efficient representation than observed features





unsupervised

Latent Feature Extraction

Latent Feature Extraction

Combinations of observed features provide more efficient representation, and capture underlying relations that govern the data

Latent Feature Extraction

Combinations of observed features provide more efficient representation, and capture underlying relations that govern the data

E.g. Ego, personality and intelligence are hidden attributes that characterize human behavior instead of survey questions

Topics (sports, science, news, etc.) instead of documents

Latent Feature Extraction

Combinations of observed features provide more efficient representation, and capture underlying relations that govern the data

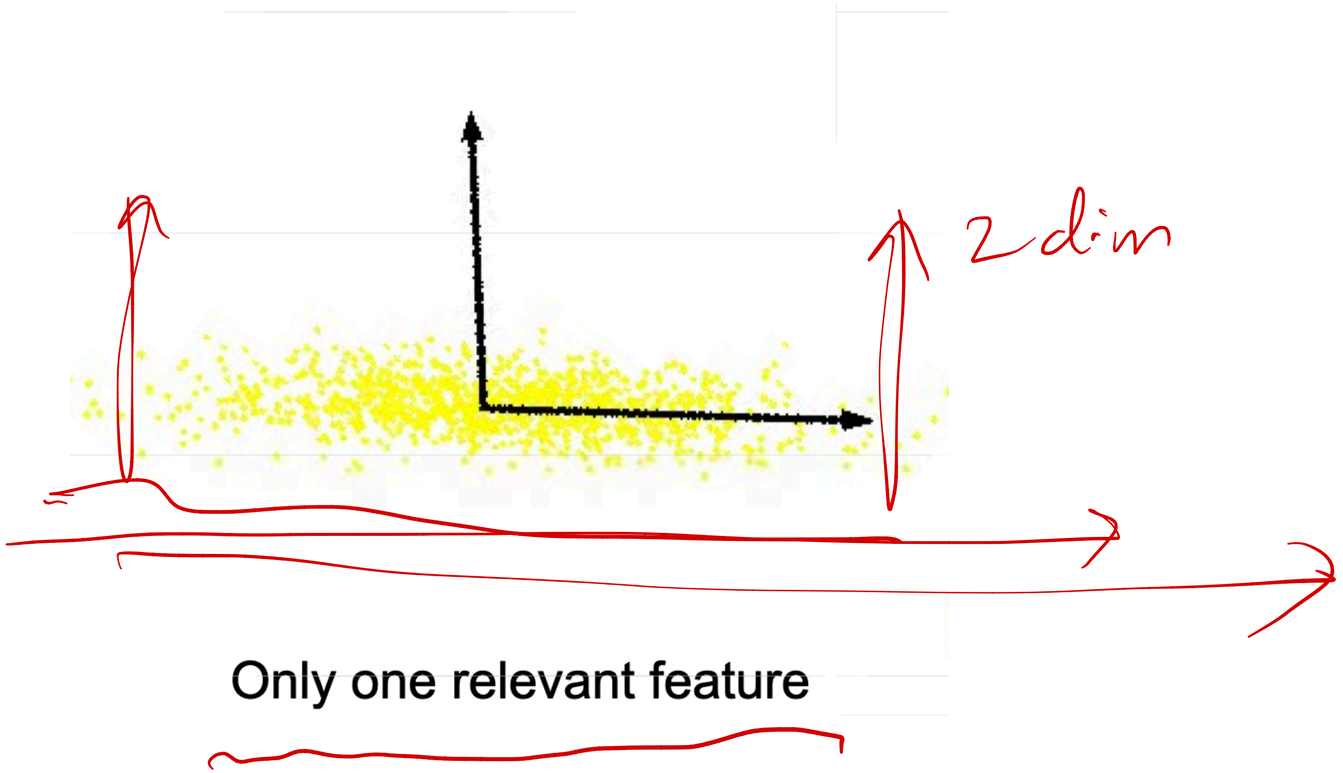
E.g. Ego, personality and intelligence are hidden attributes that characterize human behavior instead of survey questions

Topics (sports, science, news, etc.) instead of documents

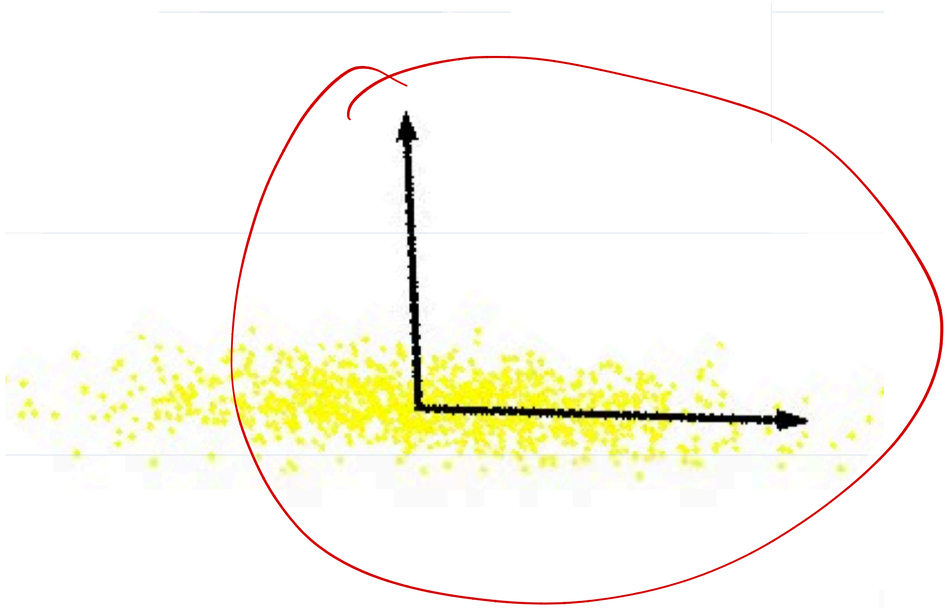
- Linear
  - Principal Component Analysis (PCA)
  - Factor Analysis
  - Independent Component Analysis (ICA)
- Nonlinear
  - ISOMAP
  - Local Linear Embedding (LLE)
  - Laplacian Eigenmaps

Principal Component Analysis (PCA)

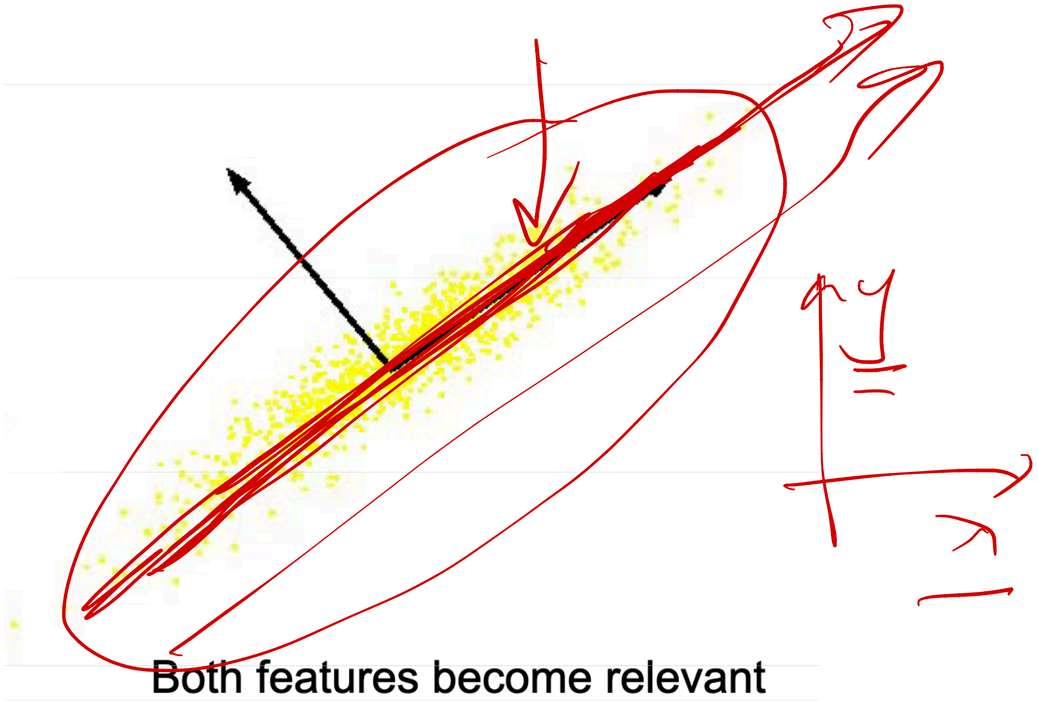
Principal Component Analysis (PCA)



Principal Component Analysis (PCA)

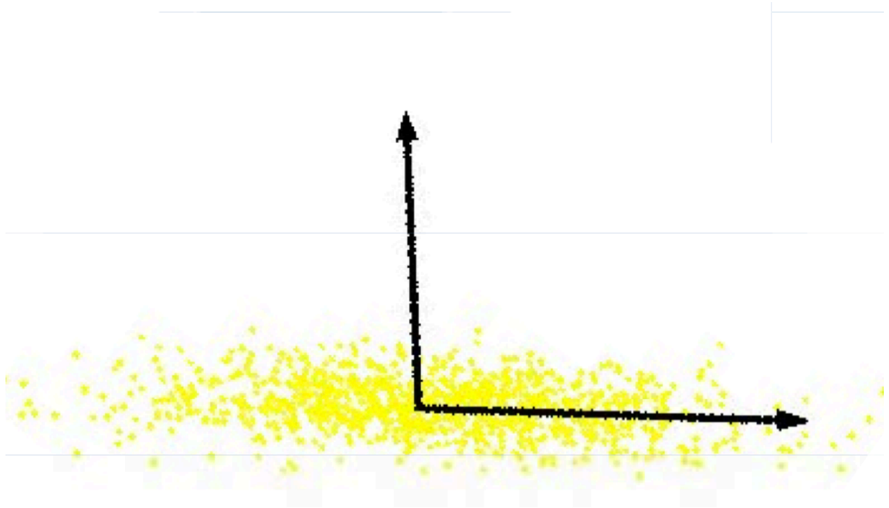


Only one relevant feature

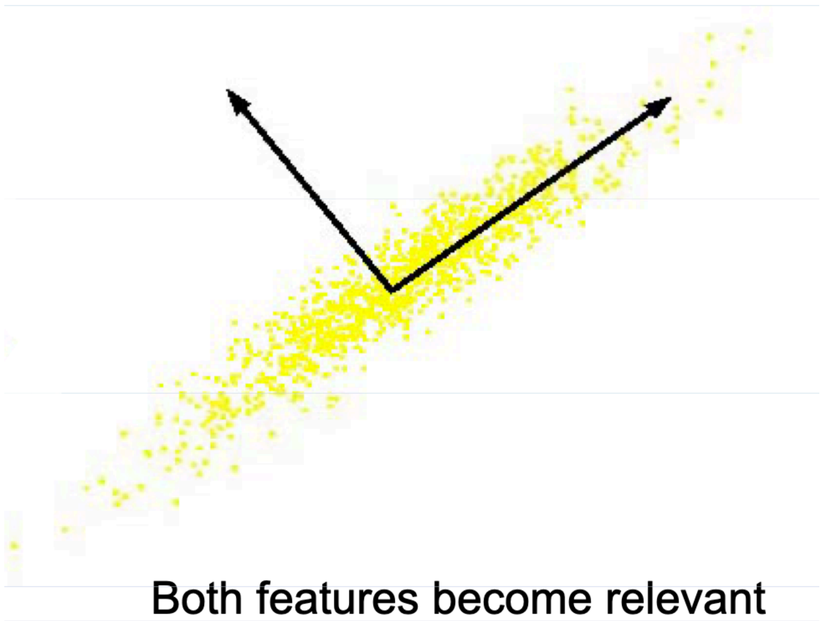


Both features become relevant

Principal Component Analysis (PCA)



Only one relevant feature

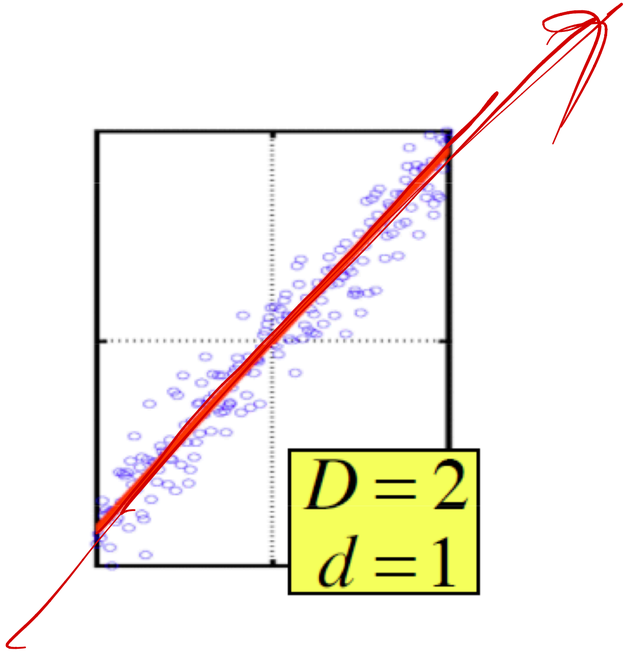


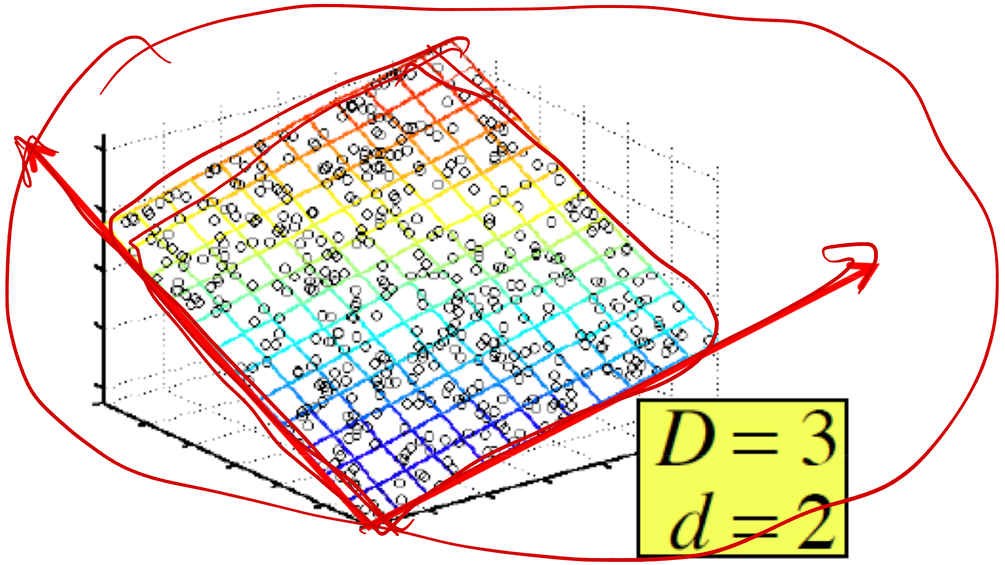
Both features become relevant

Can we transform the features so that we only need to preserve one latent feature? Find linear projection so that projected data is uncorrelated.

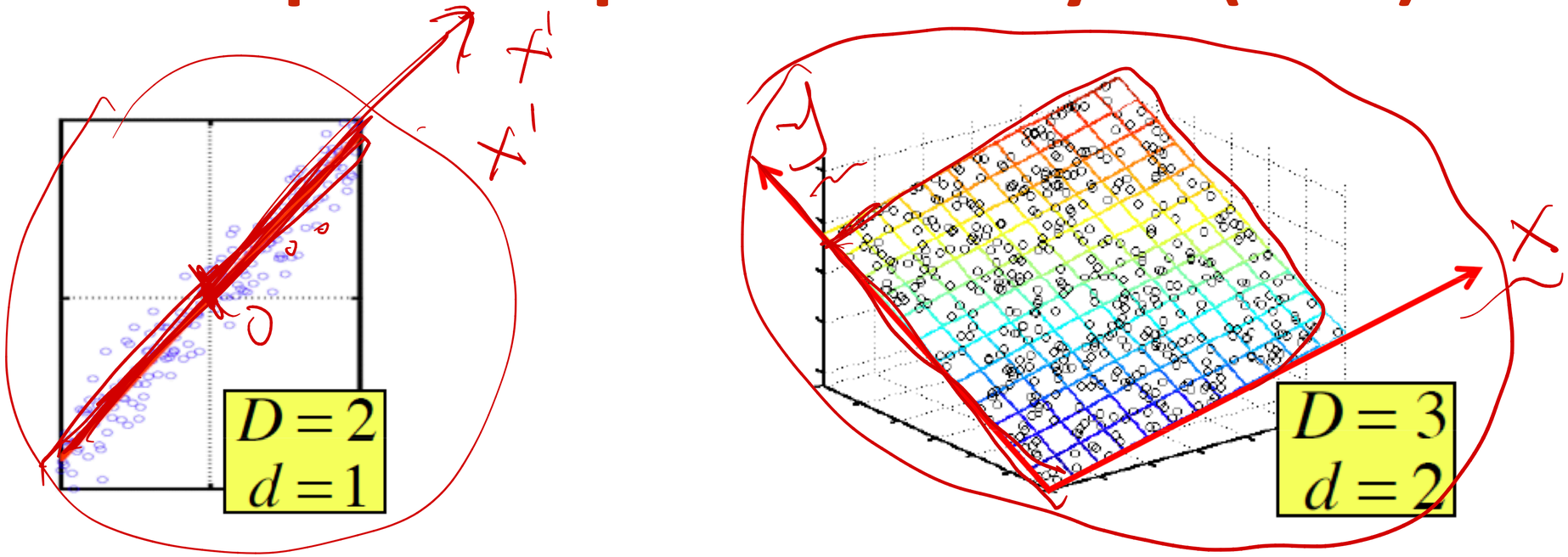
Principal Component Analysis (PCA)

Principal Component Analysis (PCA)





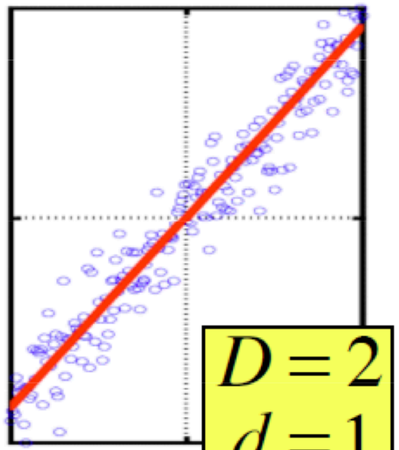
Principal Component Analysis (PCA)

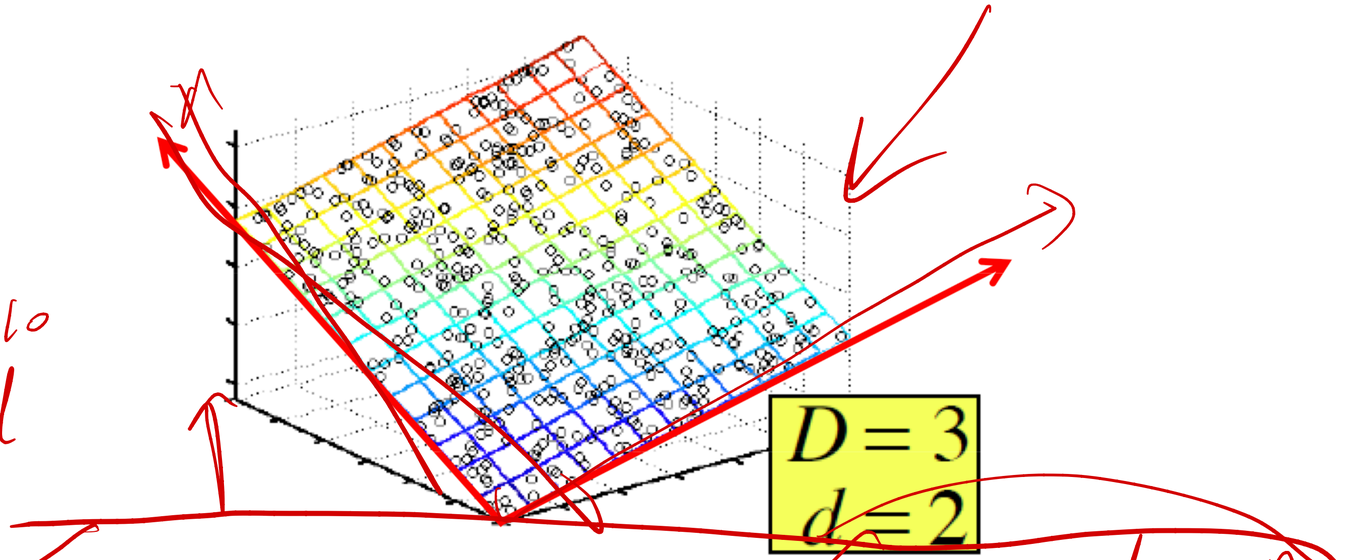


Assumption: Data lies on or near a low d-dimensional linear subspace.

Axes of this subspace are an effective representation of the data

Principal Component Analysis (PCA)





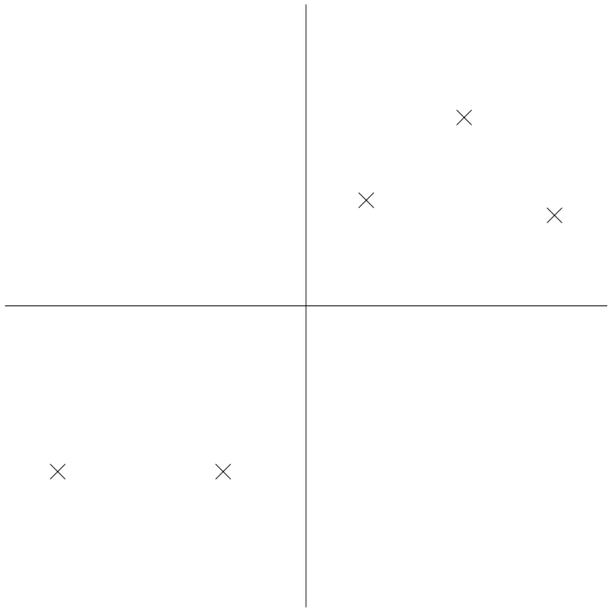
Assumption: Data lies on or near a low d-dimensional linear subspace

Axes of this subspace are an effective representation of the data

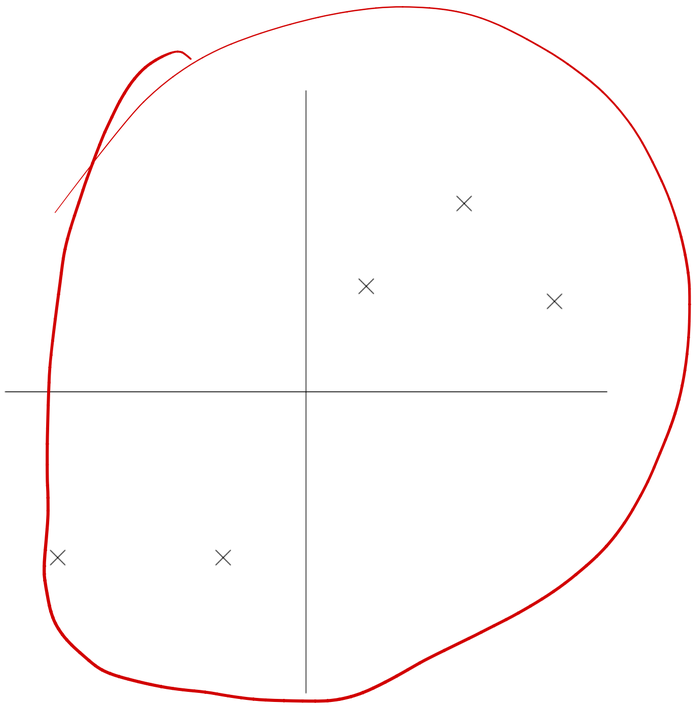
Identifying the axes is known as Principal Components Analysis, and can be obtained by Eigen or Singular value decomposition

Principal Component Analysis (PCA)

Principal Component Analysis (PCA)

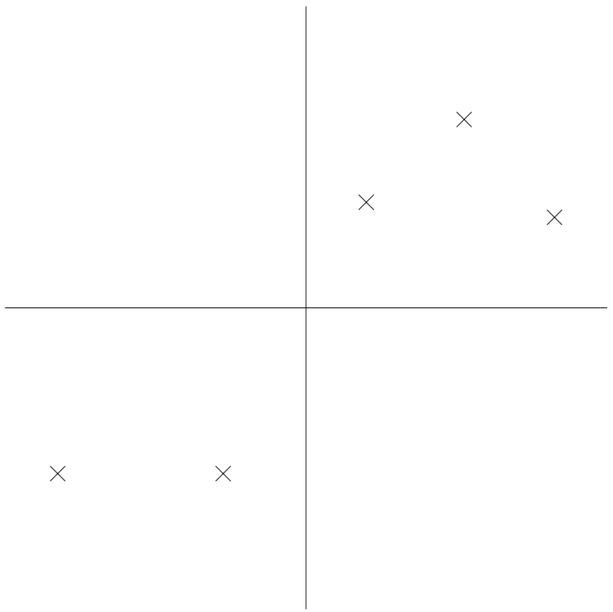


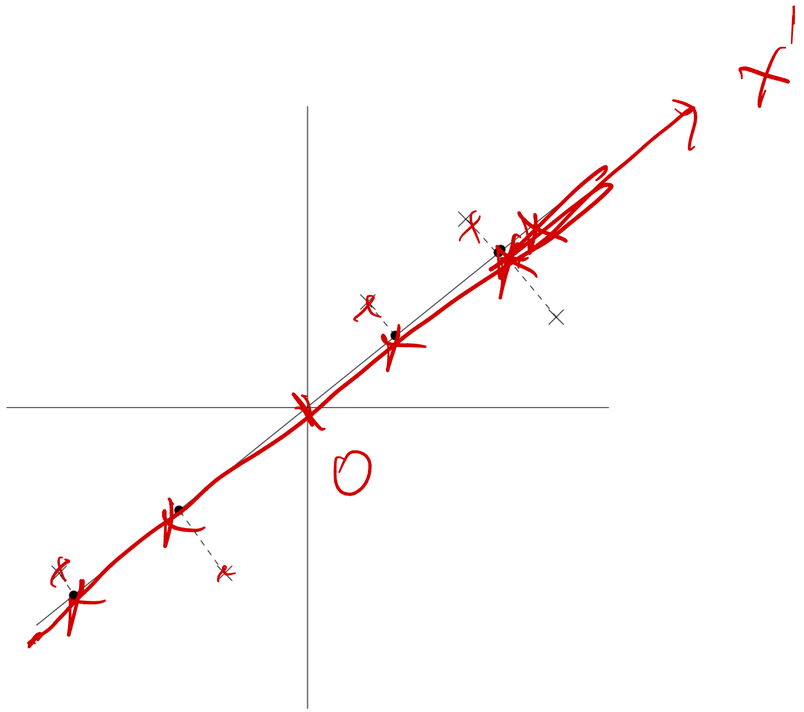
Principal Component Analysis (PCA)



Project the data onto different directions

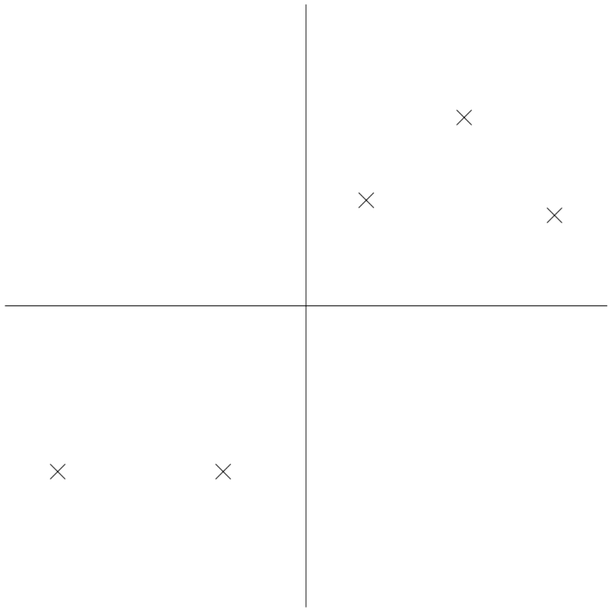
Principal Component Analysis (PCA)

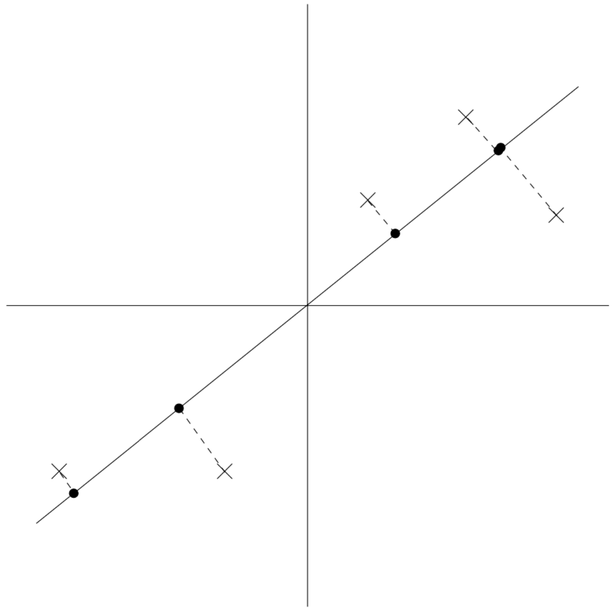


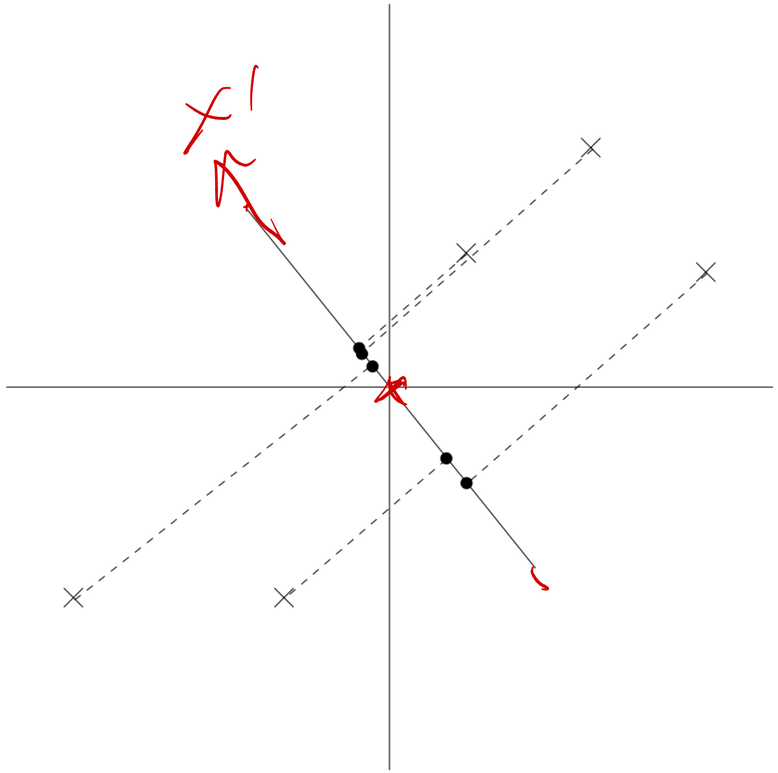


Project the data onto different directions

Principal Component Analysis (PCA)

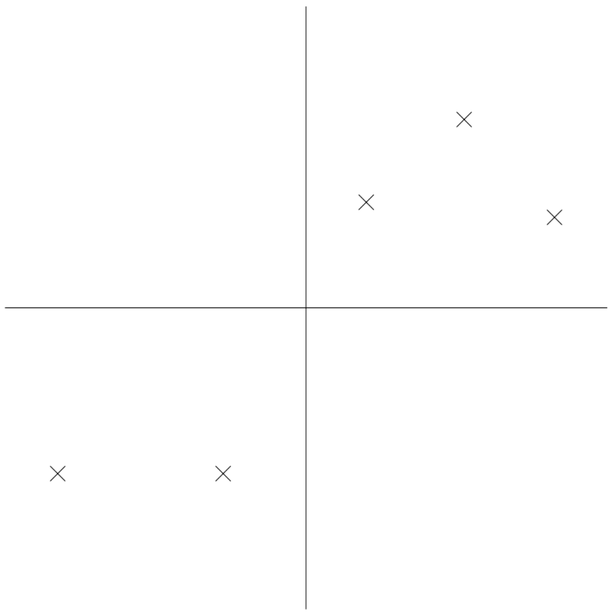


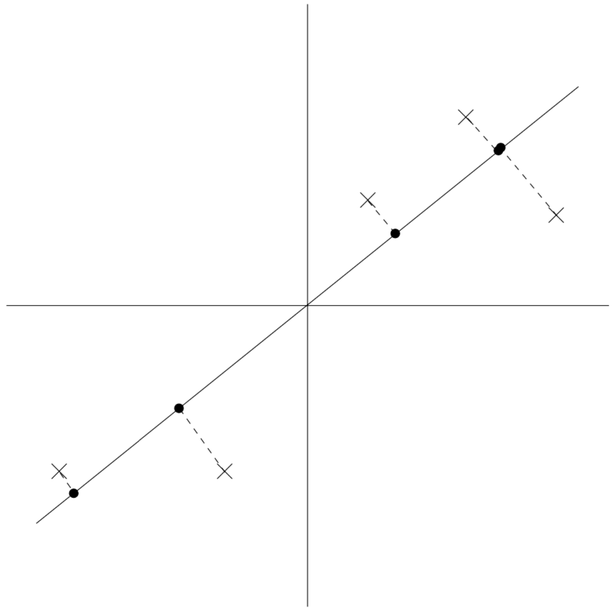


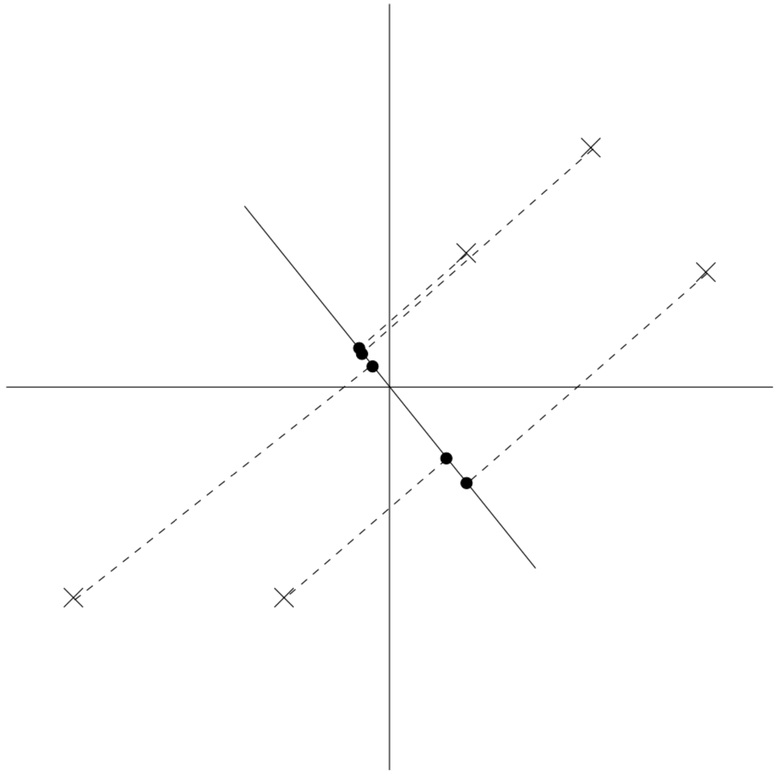


Project the data onto different directions

Principal Component Analysis (PCA)



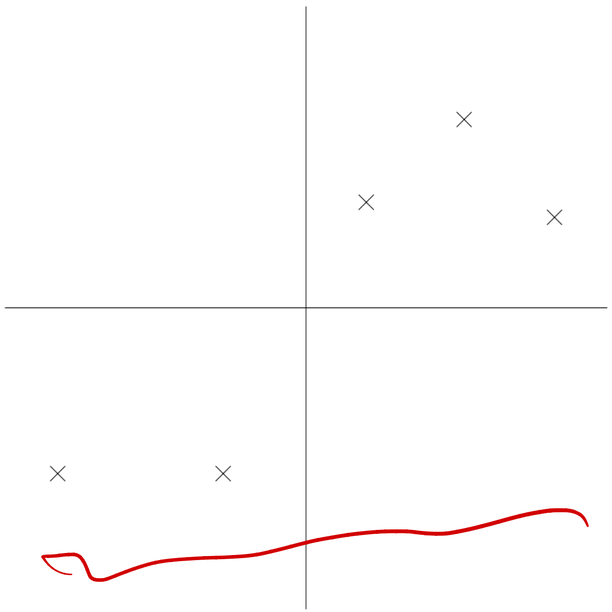


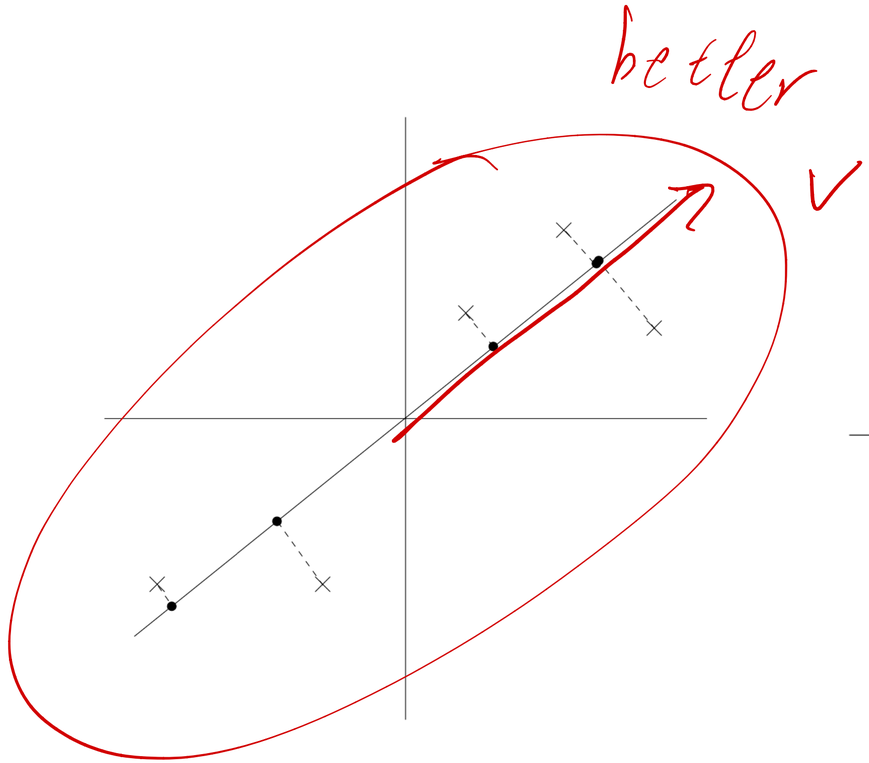


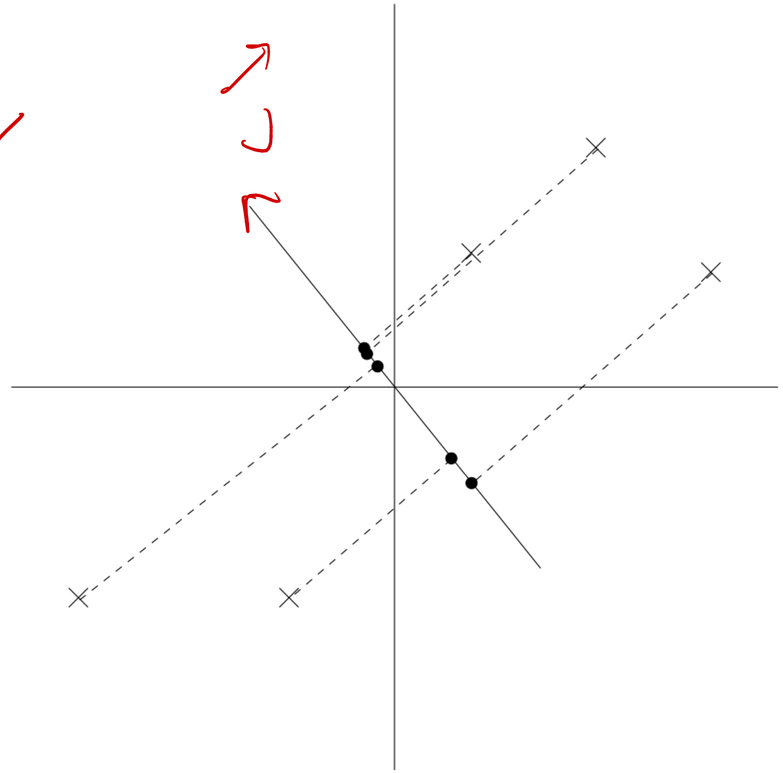
Project the data onto different directions

Which projection is better?

Principal Component Analysis (PCA)







Project the data onto different directions

Which projection is better?

We want the low-dim features that can discriminate the data the most

Normalizing Data

$$x_j^{(i)} \leftarrow \frac{x_j^{(i)} - \mu_j}{\sigma_j}$$

$$\mu_j = \frac{1}{n}\sum_{i=1}^{n} x_j^{(i)} \qquad \sigma_j^2 = \frac{1}{n}\sum_{i=1}^{n}(x_j^{(i)} - \mu_j)^2$$

Different features may have different scales

After normalization, each feature has 0 mean and variance 1

Principal Component Analysis (PCA)

Let v be the principal component

unit vector

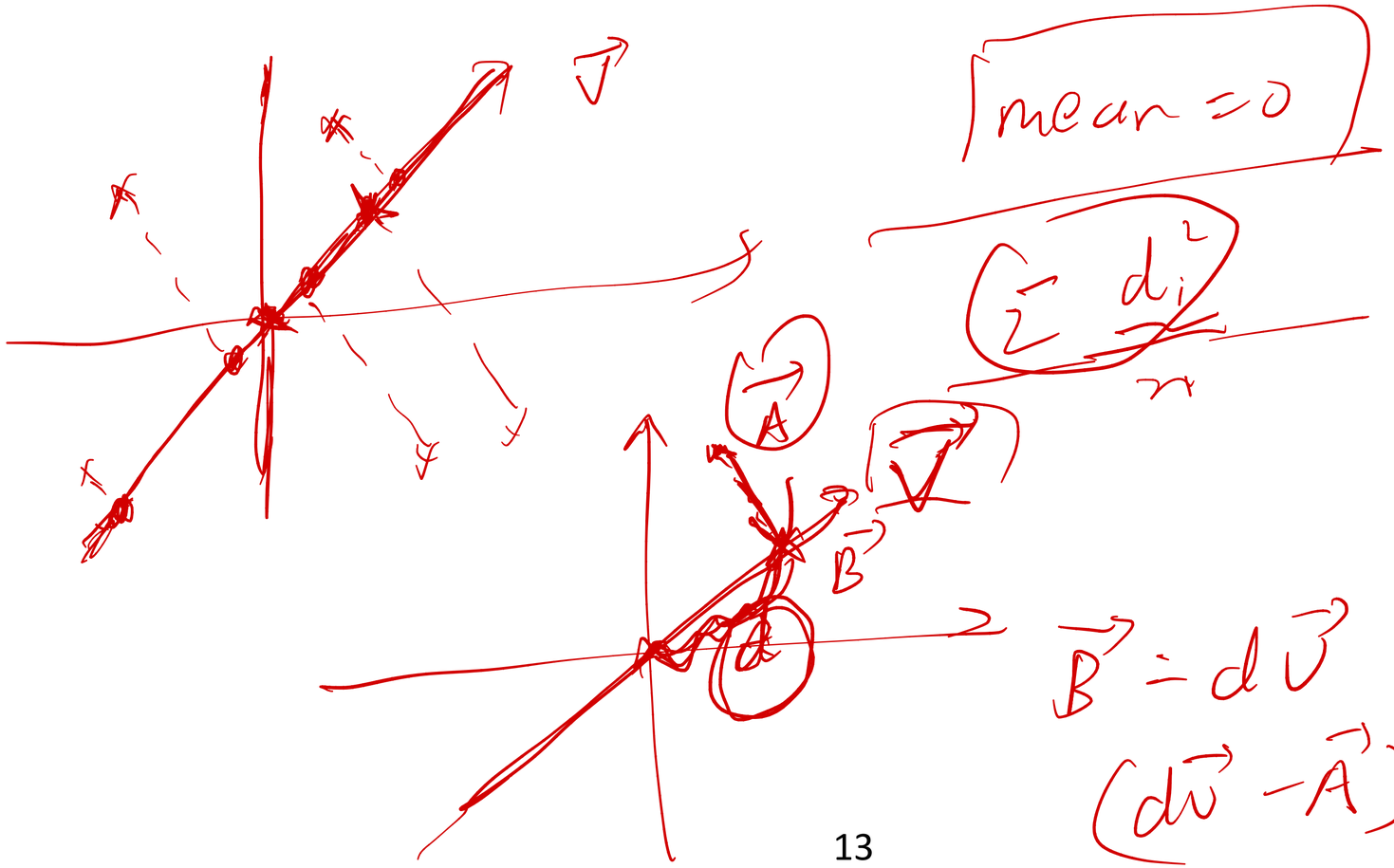
$\|v\| = 1$

x_j mean $(x_j) = 0$ Var $(x_j) = 1$

Principal Component Analysis (PCA)

Let v be the principal component

Find vector that maximizes sample variance of projection



Principal Component Analysis (PCA)

Let v be the principal component

Find vector that maximizes sample variance of projection

$$\frac{1}{n}\sum_{i=1}^{n}(\mathrm{v}^T\mathrm{x}_i)^2 = \frac{1}{n}\mathrm{v}^T X X^T \mathrm{v}$$

max

d_i^2

V?

$[V^T x_i]$?

$X = \begin{bmatrix} x_1^T \\ x_2^T \\ x_3^T \\ \vdots \end{bmatrix}$

Principal Component Analysis (PCA)

Let v be the principal component

Find vector that maximizes sample variance of projection

$$\frac{1}{n}\sum_{i=1}^{n}(\mathrm{v}^T\mathrm{x}_i)^2 = \frac{1}{n}\mathrm{v}^T X X^T \mathrm{v}$$

$$\max_{\mathbf{v}} \ \mathbf{v}^T\mathbf{X}\mathbf{X}^T\mathbf{v} \quad \text{s.t.} \quad \mathbf{v}^T\mathbf{v} = 1$$

Lagrangian: $\max_{\mathrm{v}} \mathrm{v}^T X X^T \mathrm{v} - \lambda(\mathrm{v}^T\mathrm{v} - 1)$

$$\partial/\partial\mathbf{v} = 0 \qquad (\mathbf{X}\mathbf{X}^T - \lambda\mathbf{I})\mathbf{v} = 0$$

Principal Component Analysis (PCA)

Let v be the principal component

Find vector that maximizes sample variance of projection

$$\frac{1}{n}\sum_{i=1}^{n}(\mathrm{v}^T\mathrm{x}_i)^2 = \frac{1}{n}\mathrm{v}^T XX^T\mathrm{v}$$

$$\max_{\mathbf{v}} \ \mathbf{v}^T\mathbf{X}\mathbf{X}^T\mathbf{v} \quad \text{s.t.} \quad \mathbf{v}^T\mathbf{v} = 1$$

Lagrangian: $\max_{\mathrm{v}} \mathrm{v}^T XX^T\mathrm{v} - \lambda(\mathrm{v}^T\mathrm{v} - 1)$

$$\partial/\partial\mathbf{v} = 0 \qquad (\mathbf{X}\mathbf{X}^T - \lambda\mathbf{I})\mathbf{v} = 0 \qquad \Rightarrow \boxed{(\mathbf{X}\mathbf{X}^T)\mathbf{v} = \lambda\mathbf{v}}$$

Definition of eigenvectors

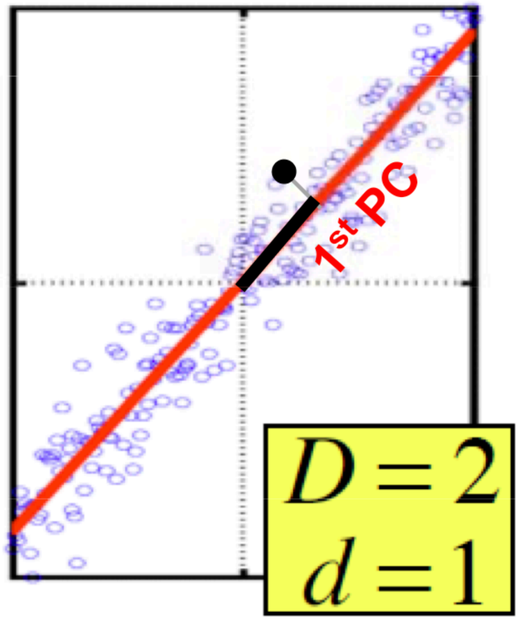
K-dimensional Cases

If we project our data into a k-dimensional subspace (k<d), we should choose $v_1, v_2, \ldots, v_k$ to be the top k eigenvectors of XX^T

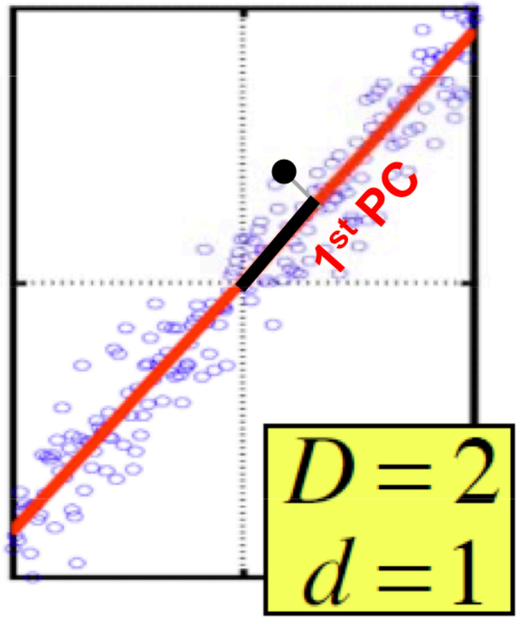
why? largest λ

For symmetric matrices, eigenvectors for distinct eigenvalues are orthogonal

Principal Component Analysis (PCA)

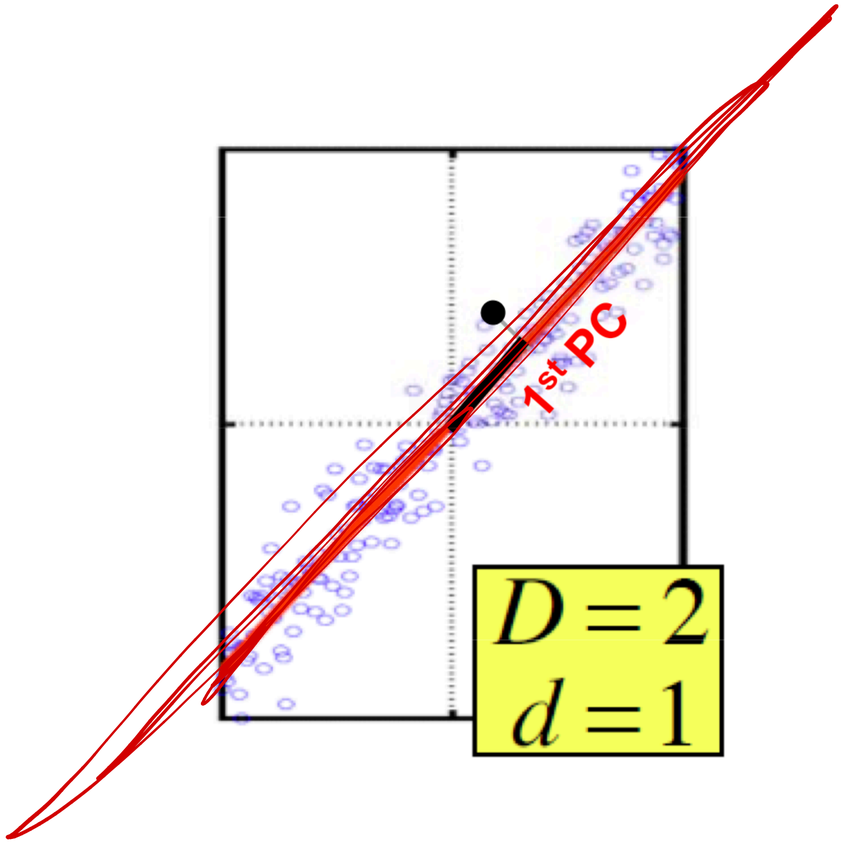


Principal Component Analysis (PCA)



Principal Components (PC) are orthogonal directions that capture most of the variance in the data

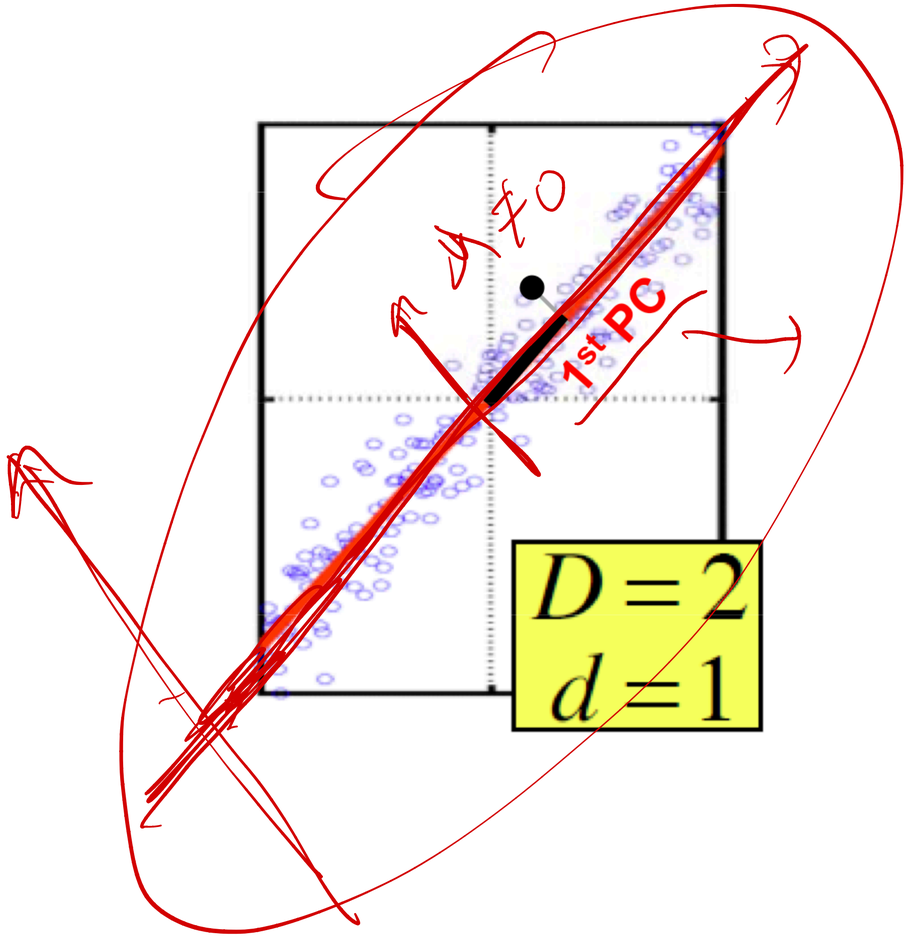
Principal Component Analysis (PCA)



Principal Components (PC) are orthogonal directions that capture most of the variance in the data

1st PC – direction of greatest variability in data

Principal Component Analysis (PCA)



Principal Components (PC) are orthogonal directions that capture most of the variance in the data

1st PC – direction of greatest variability in data

Projection of data points along 1st PC discriminate the data most along any one direction

Principal Component Analysis (PCA)

Principal Component Analysis (PCA)

Sample variance of projection $= \mathbf{v}^T \mathbf{X}\mathbf{X}^T \mathbf{v} = \lambda \mathbf{v}^T \mathbf{v} = \lambda$

Principal Component Analysis (PCA)

Sample variance of projection = $\mathbf{v}^T \mathbf{X}\mathbf{X}^T \mathbf{v} = \lambda \mathbf{v}^T \mathbf{v} = \lambda$

Thus, the eigenvalue λ denotes the amount of variability captured along that dimension.

Principal Component Analysis (PCA)

Sample variance of projection = $\mathbf{v}^T \mathbf{X}\mathbf{X}^T \mathbf{v} = \lambda \mathbf{v}^T \mathbf{v} = \lambda$

Thus, the eigenvalue λ denotes the amount of variability captured along that dimension.

$x \in R^d$

The 1st Principal component v_1 is the eigenvector of the sample covariance matrix XX^T associated with the largest eigenvalue λ_1

The 2nd Principal component v_2 is the eigenvector of the sample covariance matrix XX^T associated with the second largest eigenvalue λ_2

And so on …

d

d^{th}

Computing the Principal Components (PCs)

Computing the Principal Components (PCs)

Eigenvectors are solutions of the following equation:

$$(\mathbf{X}\mathbf{X}^T)\mathbf{v} = \lambda\mathbf{v} \qquad (\mathbf{X}\mathbf{X}^T - \lambda\mathbf{I})\mathbf{v} = 0$$

Computing the Principal Components (PCs)

Eigenvectors are solutions of the following equation:

$$(\mathbf{X}\mathbf{X}^T)\mathbf{v} = \lambda\mathbf{v} \qquad (\mathbf{X}\mathbf{X}^T - \lambda\mathbf{I})\mathbf{v} = 0$$

Non-zero solution $\mathbf{v} \neq 0$ possible only if

$$\det(\mathbf{X}\mathbf{X}^T - \lambda\mathbf{I}) = 0$$

Computing the Principal Components (PCs)

Eigenvectors are solutions of the following equation:

$$(\mathbf{X}\mathbf{X}^T)\mathbf{v} = \lambda\mathbf{v} \qquad (\mathbf{X}\mathbf{X}^T - \lambda\mathbf{I})\mathbf{v} = 0$$

Non-zero solution v ≠ 0 possible only if

$$\det(\mathbf{X}\mathbf{X}^T - \lambda\mathbf{I}) = 0 \Rightarrow \lambda_1\ \lambda_2\ \lambda_3\ \cdots$$

We can compute the eigenvalues from this equation

Computing the Principal Components (PCs)

Eigenvectors are solutions of the following equation:

$$(\mathbf{X}\mathbf{X}^T)\mathbf{v} = \lambda\mathbf{v} \qquad (\mathbf{X}\mathbf{X}^T - \lambda\mathbf{I})\mathbf{v} = 0$$

Non-zero solution v ≠ 0 possible only if

$$\det(\mathbf{X}\mathbf{X}^T - \lambda\mathbf{I}) = 0$$

We can compute the eigenvalues from this equation

This is a D^{th} order equation in λ, can have at most D distinct solutions (roots of the characteristic equation)

Computing the Principal Components (PCs)

Eigenvectors are solutions of the following equation:

$$(\mathbf{X}\mathbf{X}^T)\mathbf{v} = \lambda\mathbf{v} \qquad (\mathbf{X}\mathbf{X}^T - \lambda\mathbf{I})\mathbf{v} = 0$$

Non-zero solution v ≠ 0 possible only if

$$\det(\mathbf{X}\mathbf{X}^T - \lambda\mathbf{I}) = 0$$

We can compute the eigenvalues from this equation

This is a D^{th} order equation in λ, can have at most D distinct solutions (roots of the characteristic equation)

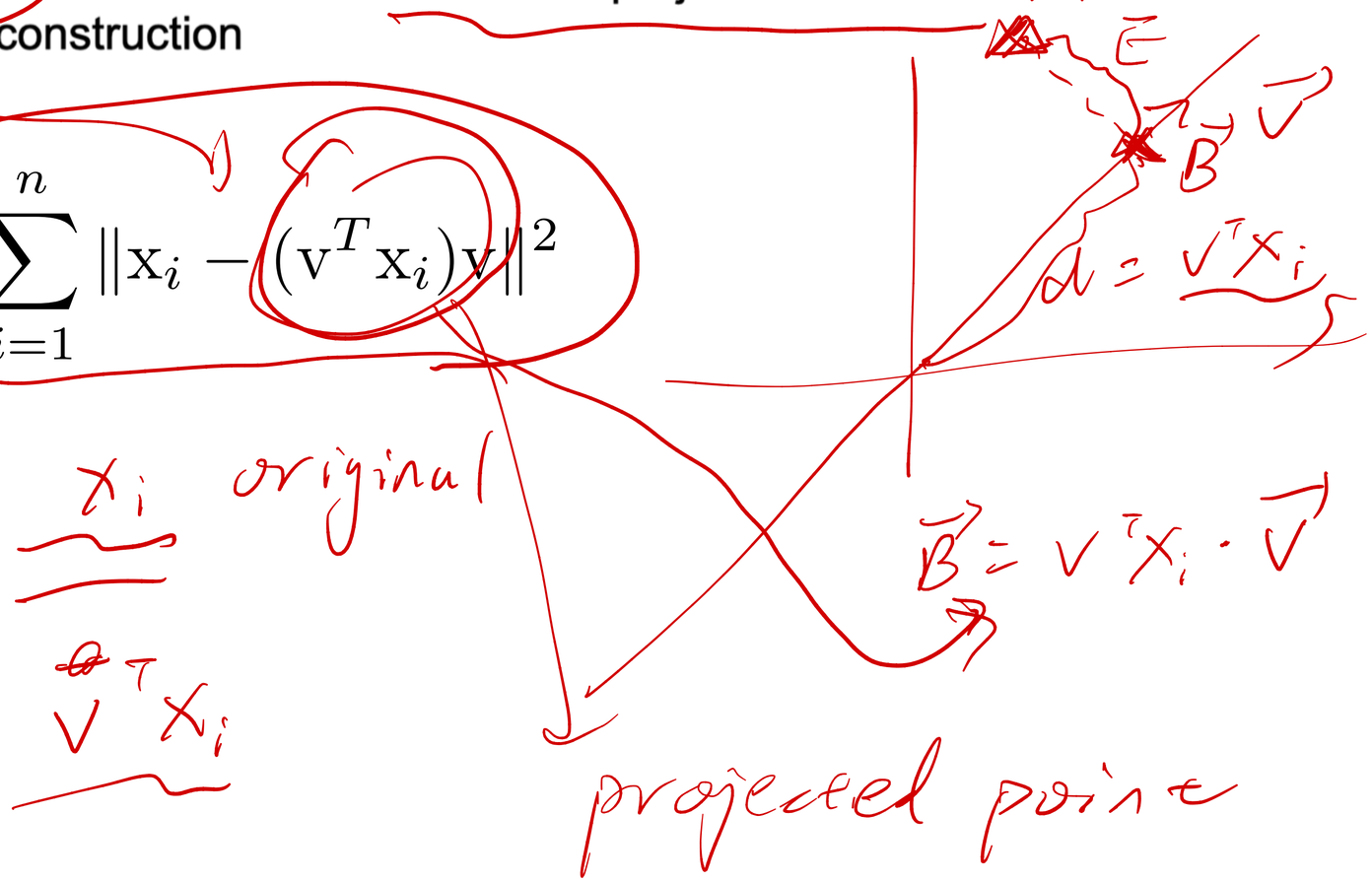
Once eigenvalues are computed, solve for eigenvectors (Principal Components) using

$$(\mathbf{X}\mathbf{X}^T - \lambda\mathbf{I})\mathbf{v} = 0$$

Another Interpretation

Minimum Reconstruction Error: PCA finds vectors v such that projection on to the vectors yields minimum MSE reconstruction

$$\frac{1}{n}\sum_{i=1}^{n}\|\mathrm{x}_i - (\mathrm{v}^T\mathrm{x}_i)\mathrm{v}\|^2$$

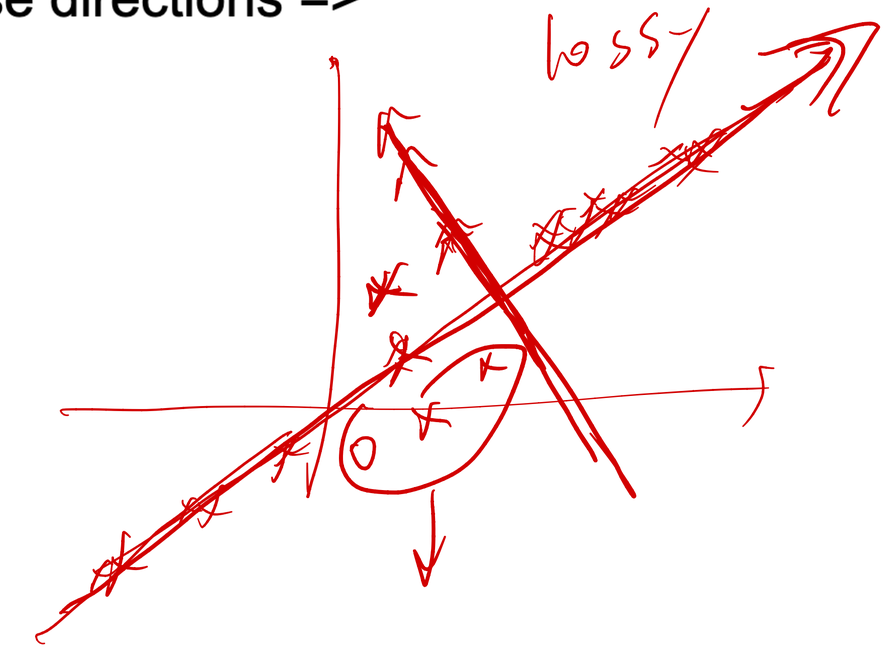


Dimensionality Reduction using PCA

Dimensionality Reduction using PCA

The eigenvalue λ denotes the amount of variability captured along that dimension.

Zero eigenvalues indicate no variability along those directions => data lies exactly on a linear subspace



Dimensionality Reduction using PCA

The eigenvalue λ denotes the amount of variability captured along that dimension.

Zero eigenvalues indicate no variability along those directions => data lies exactly on a linear subspace

Only keep data projections onto principal components with non-zero eigenvalues, say $v_1, \ldots, v_d$ where $d = \text{rank}(XX^T)$

lossless dim Reduction

$D \rightarrow d = \text{rank}(XX^T)$

Dimensionality Reduction using PCA

The eigenvalue λ denotes the amount of variability captured along that dimension.

Zero eigenvalues indicate no variability along those directions => data lies exactly on a linear subspace

Only keep data projections onto principal components with non-zero eigenvalues, say $v_1, \ldots, v_d$ where $d = \text{rank}(XX^T)$

Original Representation

data point

$x_i = [x_i^1, x_i^2, \ldots. x_i^D]$

(D-dimensional vector)

Dimensionality Reduction using PCA

The eigenvalue λ denotes the amount of variability captured along that dimension.

Zero eigenvalues indicate no variability along those directions => data lies exactly on a linear subspace

Only keep data projections onto principal components with non-zero eigenvalues, say $v_1, \ldots, v_d$ where $d = \text{rank}\ (XX^T)$

Original Representation

data point

$x_i = [x_i^1, x_i^2, \ldots. x_i^D]$

(D-dimensional vector)

Transformed representation

projections

$[v_1^T x_i, v_2^T x_i, \ldots v_d^T x_i]$

(d-dimensional vector)

Dimensionality Reduction using PCA

Dimensionality Reduction using PCA

Usually data lies near a linear subspace, as noise introduces small variability

Only keep data projections onto principal components with **large** eigenvalues

Can *ignore* the components of lesser significance.

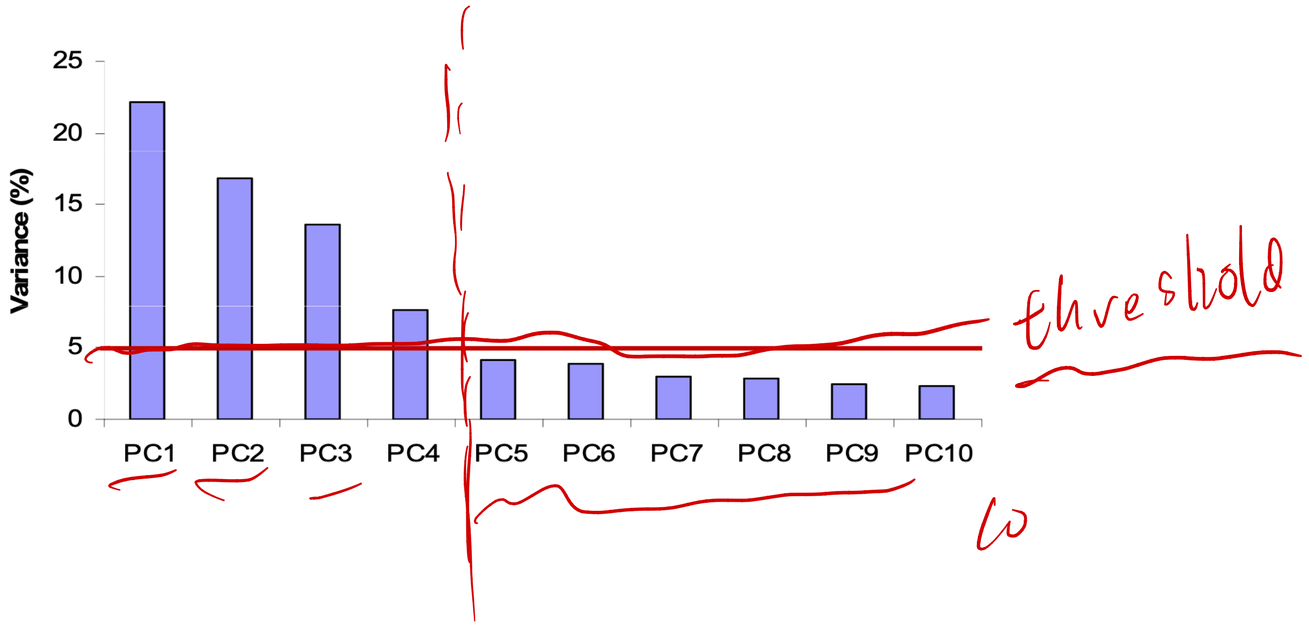
small eigenvalue

Dimensionality Reduction using PCA

Usually data lies near a linear subspace, as noise introduces small variability

Only keep data projections onto principal components with **large** eigenvalues

Can *ignore* the components of lesser significance.

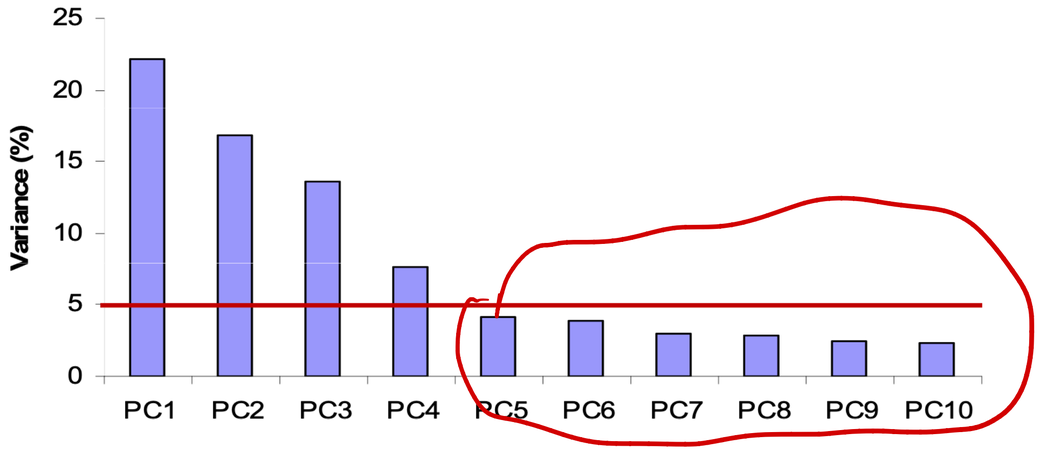


Dimensionality Reduction using PCA

Usually data lies near a linear subspace, as noise introduces small variability

Only keep data projections onto principal components with **large** eigenvalues

Can *ignore* the components of lesser significance.



You might lose some information, but if the eigenvalues are small, you don't lose much

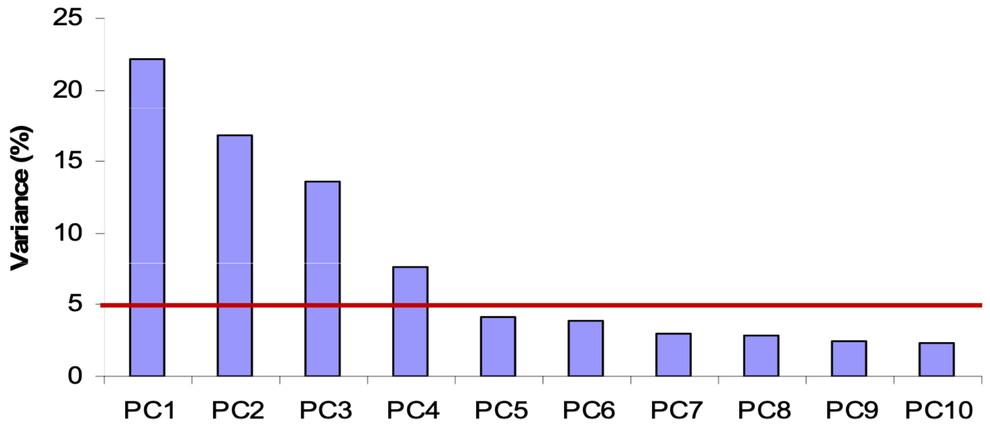
lossy

Dimensionality Reduction using PCA

Usually data lies near a linear subspace, as noise introduces small variability

Only keep data projections onto principal components with **large** eigenvalues

Can *ignore* the components of lesser significance.



You might lose some information, but if the eigenvalues are small, you don't lose much

It is not lossless compression

Example: faces

Example: faces

Eigenfaces from 7562 images:

top left image is linear combination of rest.

Sirovich & Kirby (1987)
Turk & Pentland (1991)

10000 images

A ⟶ B

Source coding theorem

[200 x 200]

10000 x [200 x 200]

Original

compression is intelligent

PCA

15 eigen images

A ⟶ B

10000 images ⟶

PCA

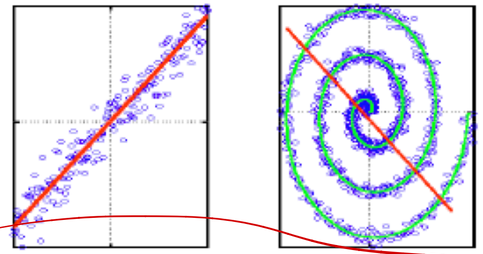
15 x [200 x 200]

+

10000 x 15

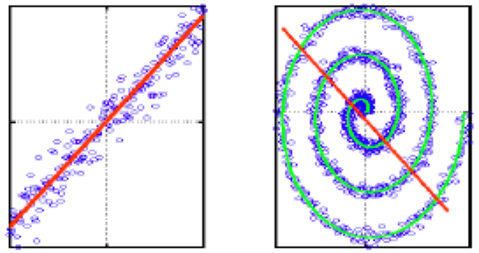
Properties of PCA

- **Strengths**
  - Eigenvector method
  - No tuning parameters
  - Non-iterative
  - No local optima
- **Weaknesses**
  - Limited to second order statistics
  - Limited to linear projections

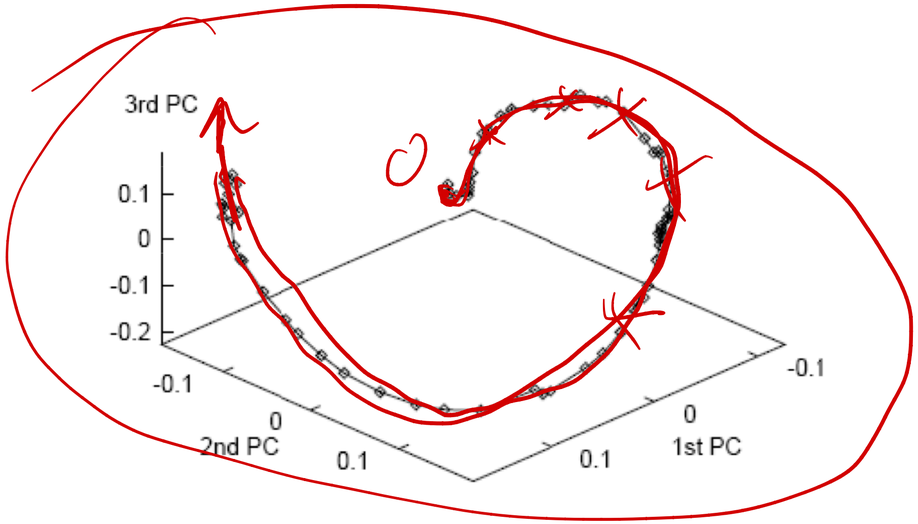


Properties of PCA

- **Strengths**
  - Eigenvector method
  - No tuning parameters
  - Non-iterative
  - No local optima
- **Weaknesses**
  - Limited to second order statistics
  - Limited to linear projections



Nonlinear example

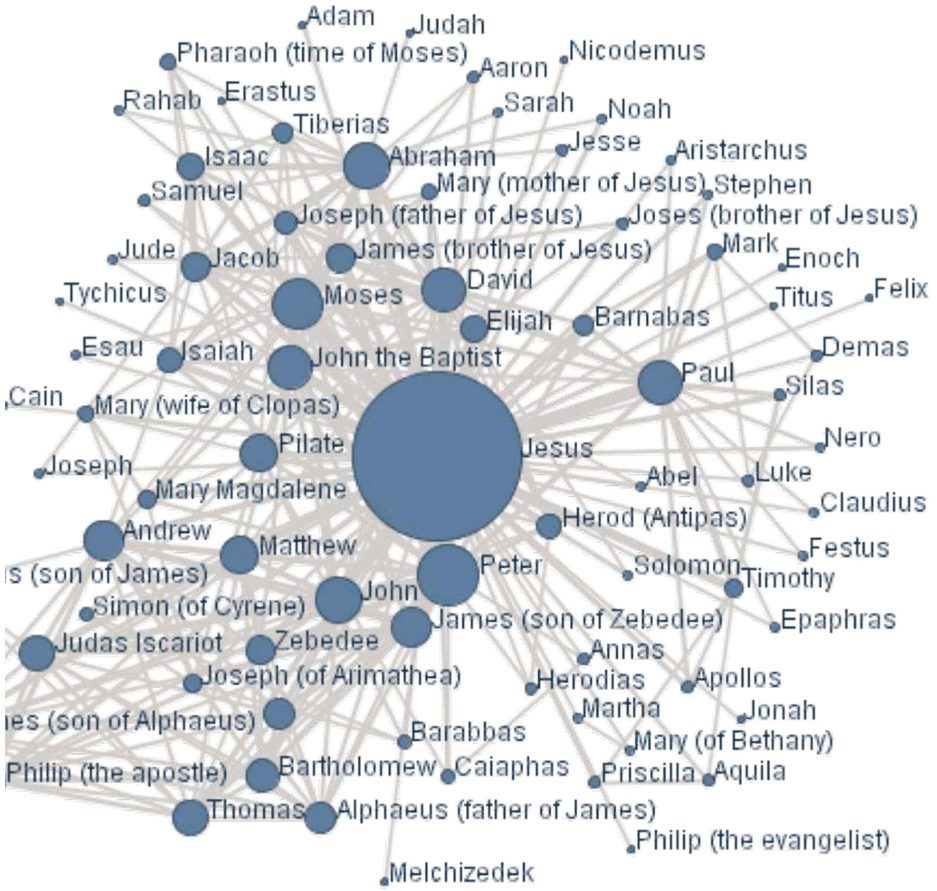


香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

COMP 5212
Machine Learning
Lecture 14

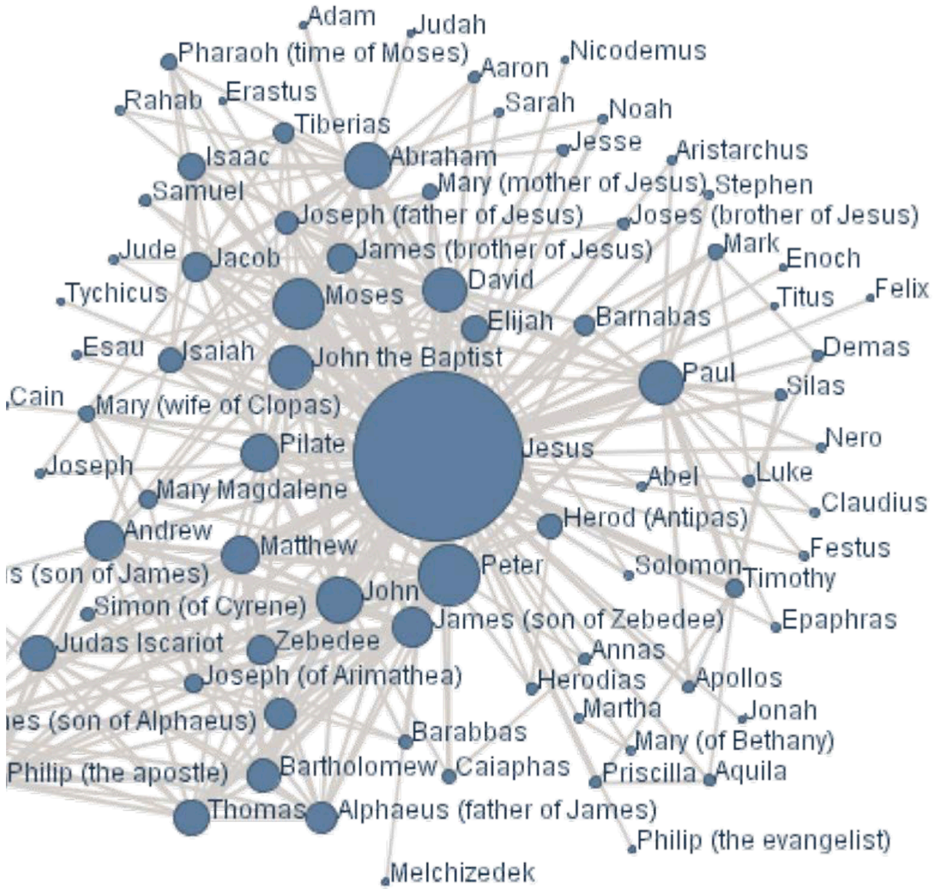
Probabilistic Graphical Models

What Are Graphical Models?



What Are Graphical Models?

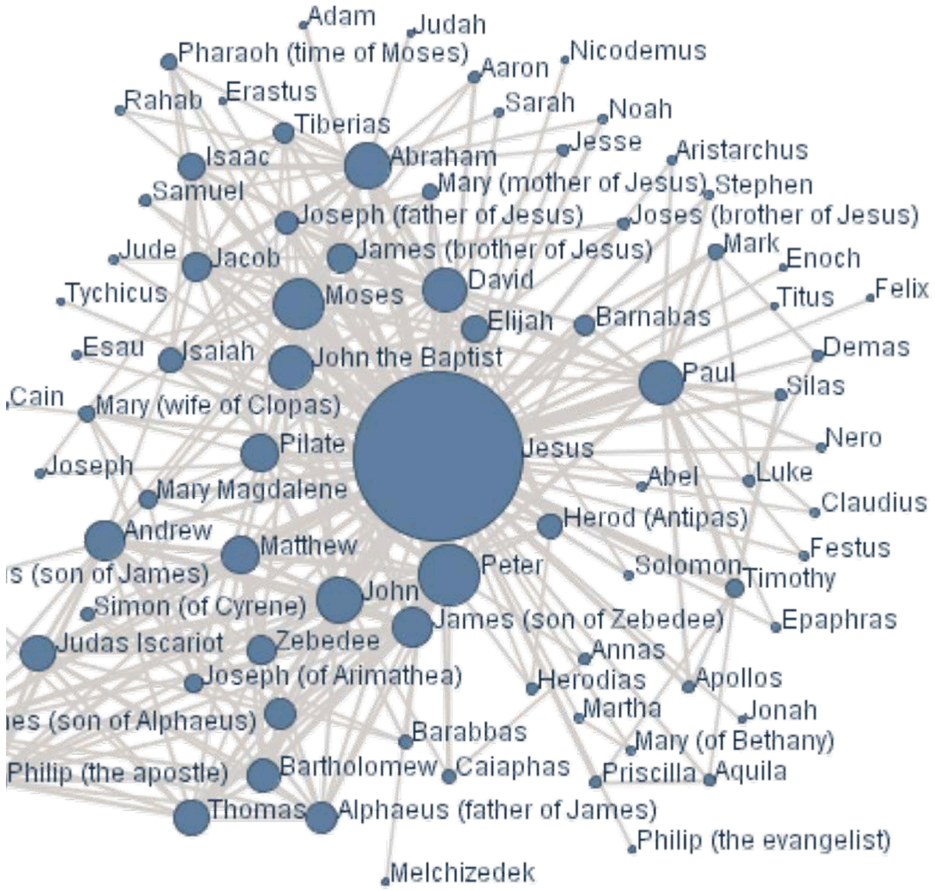
- Informally, a GM is just a graph representing relationship among random variables
  - Nodes: random variables (features, not examples)
  - Edges (or absence of edges): relationship



What Are Graphical Models?

- Informally, a GM is just a graph representing relationship among random variables
  - Nodes: random variables (features, not examples)
  - Edges (or absence of edges): relationship

- Looks simple!
  - But detail matters, as always.
  - What exactly do we mean by relationship?



Relationship between two random variables

- Many types of relationships exist:
  - X and Y are correlated
  - X and Y are dependent
  - X and Y are independent
  - X and Y are partially correlated given Z
  - X and Y are conditionally dependent given Z
  - X and Y are conditionally independent given Z
  - X causes Y
  - Y causes X
  - ...

Relationship between two random variables

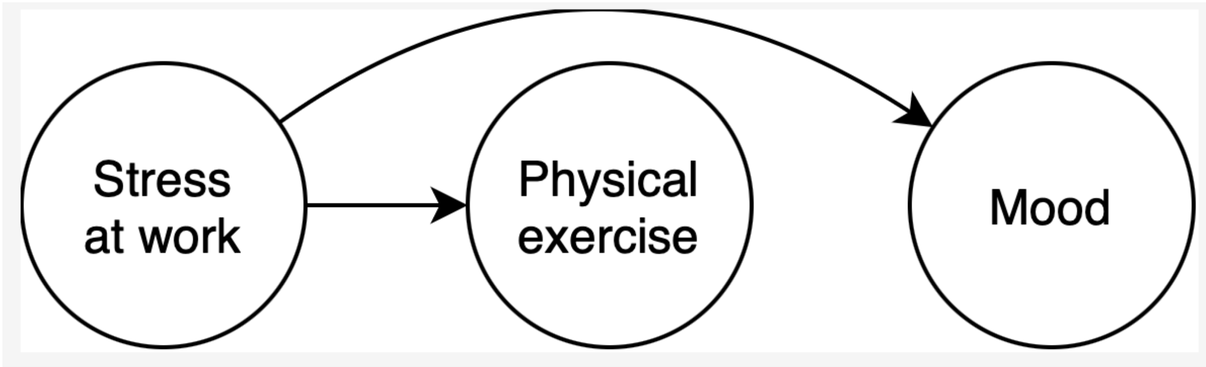
- Many types of relationships exist:
  - X and Y are correlated
  - X and Y are dependent
  - X and Y are independent
  - X and Y are partially correlated given Z
  - X and Y are conditionally dependent given Z
  - X and Y are conditionally independent given Z
  - X causes Y
  - Y causes X
  - ...

Correlation does not imply causation

Relationship between two random variables

- Many types of relationships exist:
  - X and Y are correlated
  - X and Y are dependent
  - X and Y are independent
  - X and Y are partially correlated given Z
  - X and Y are conditionally dependent given Z
  - X and Y are conditionally independent given Z
  - X causes Y
  - Y causes X
  - ...

Correlation does not imply causation



What is a Graphical Model?

What is a Graphical Model?

Graphical model represents a multivariate distribution in High-D space

What is a Graphical Model?

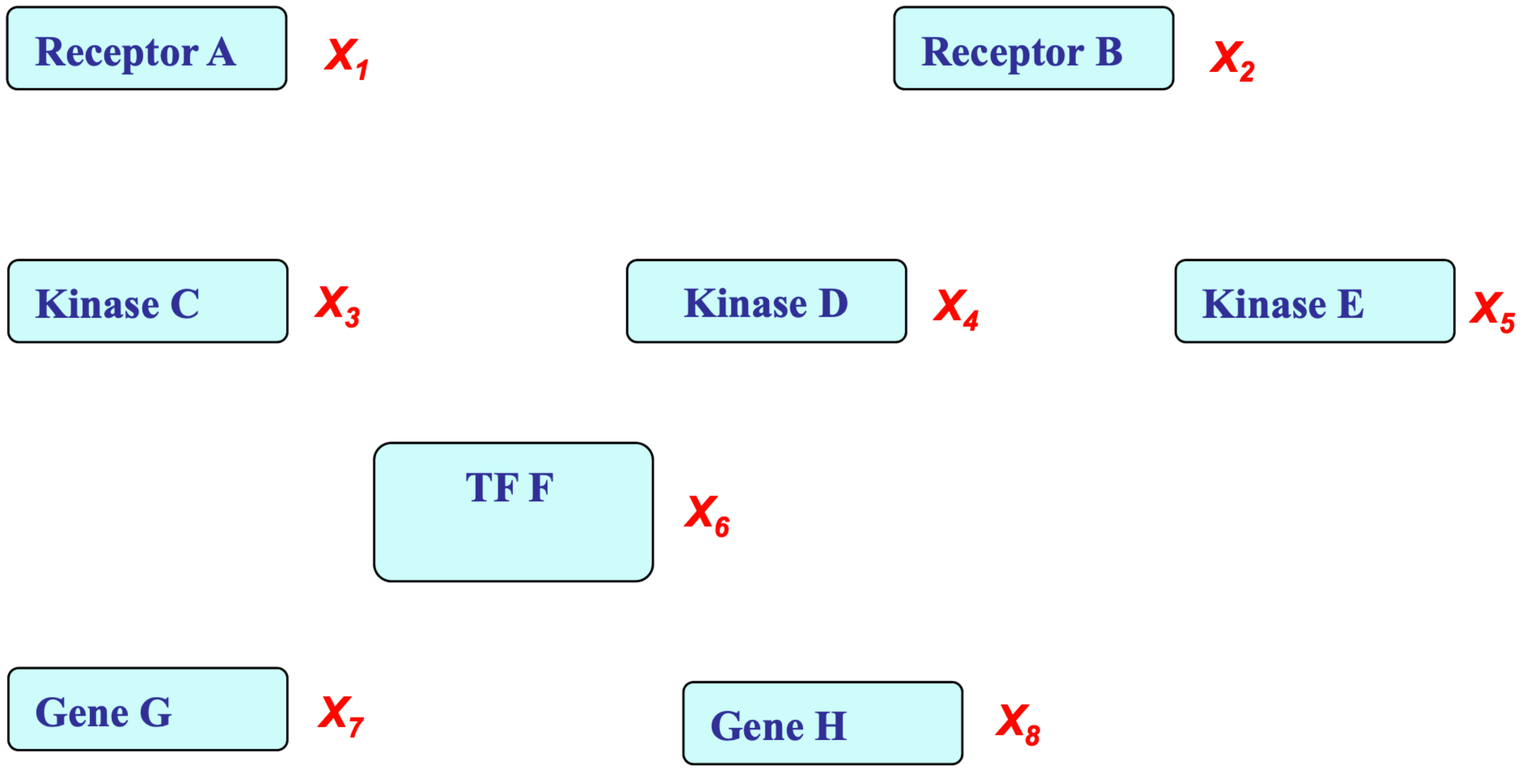
Graphical model represents a multivariate distribution in High-D space

A possible world for celluar signal transduction:

What is a Graphical Model?

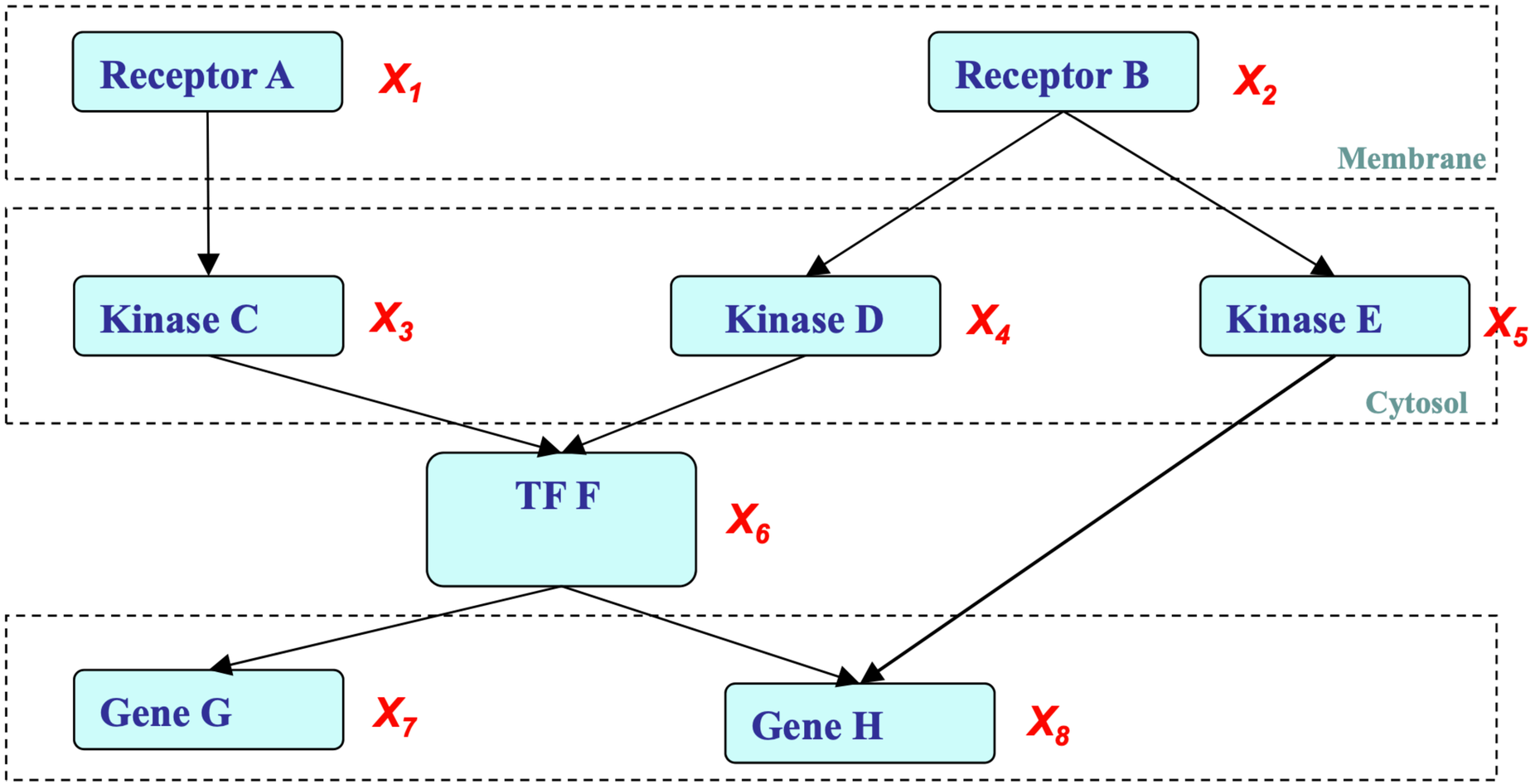
Graphical model represents a multivariate distribution in High-D space

A possible world for celluar signal transduction:



Structure Simplifies Representation

Dependencies among variables



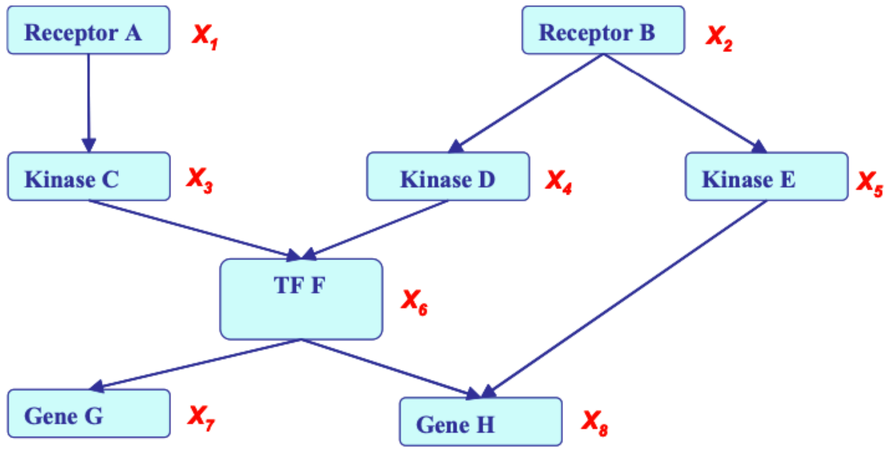
Probabilistic Graphical Models

Probabilistic Graphical Models

- If X_i's are **conditionally independent** (as described by a **PGM**), the joint can be factored to a product of simpler terms, e.g.,

Probabilistic Graphical Models

- If X_i's are **conditionally independent** (as described by a **PGM**), the joint can be factored to a product of simpler terms, e.g.,

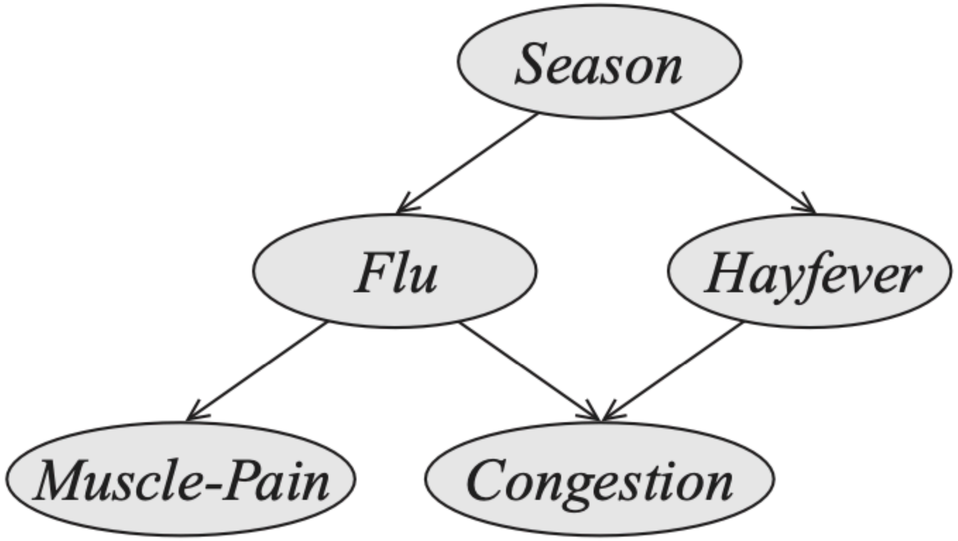


$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$

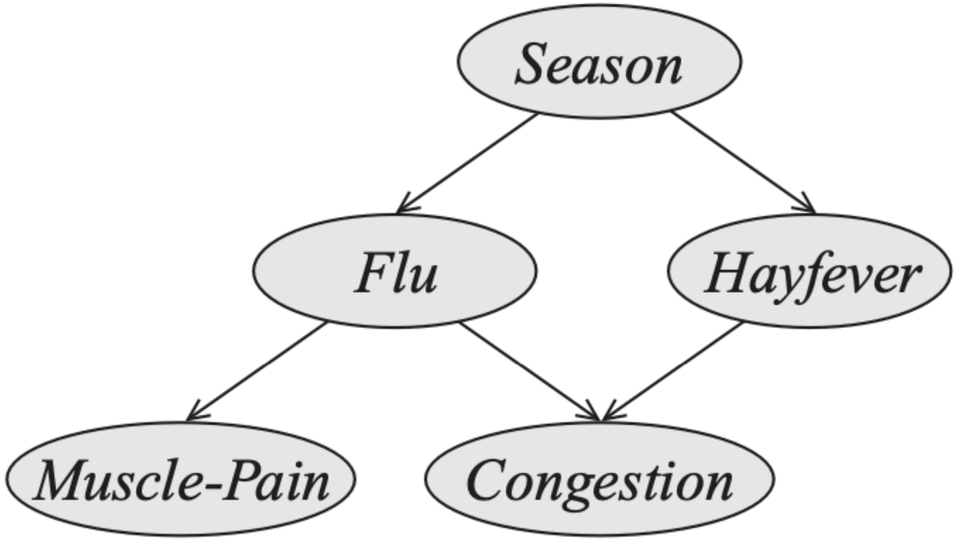
$$= P(X_1)\, P(X_2)\, P(X_3|\, X_1)\, P(X_4|\, X_2)\, P(X_5|\, X_2)$$
$$P(X_6|\, X_3, X_4)\, P(X_7|\, X_6)\, P(X_8|\, X_5, X_6)$$

Stay tune for what are these independencies!

Another Example



Another Example



$$P(Congestion \mid Flu, Hayfever, Season) = P(Congestion \mid Flu, Hayfever);$$

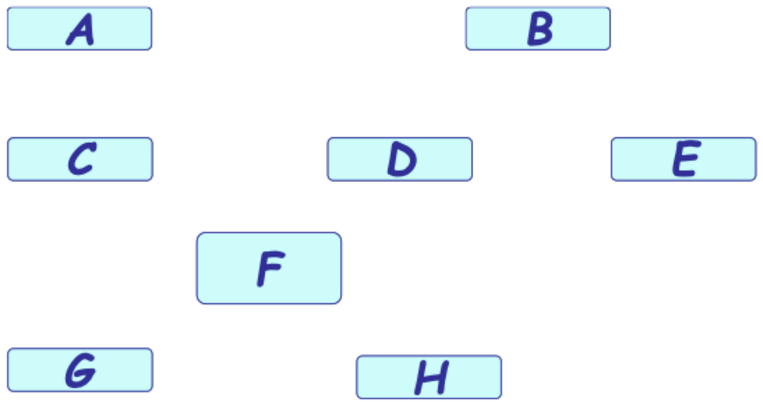
What is a PGM After All

What is a PGM After All

It is a smart way to write/specify/compose/design exponentially-large probability distributions without paying an exponential cost, and at the same time endow the distributions with ***structured semantics***

What is a PGM After All

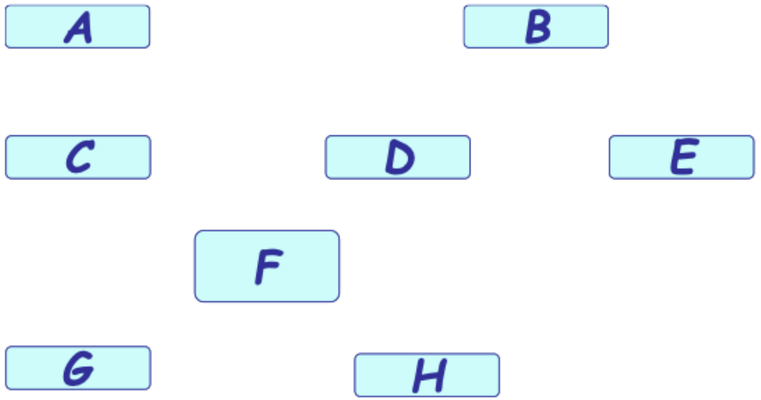
It is a smart way to write/specify/compose/design exponentially-large probability distributions without paying an exponential cost, and at the same time endow the distributions with ***structured semantics***

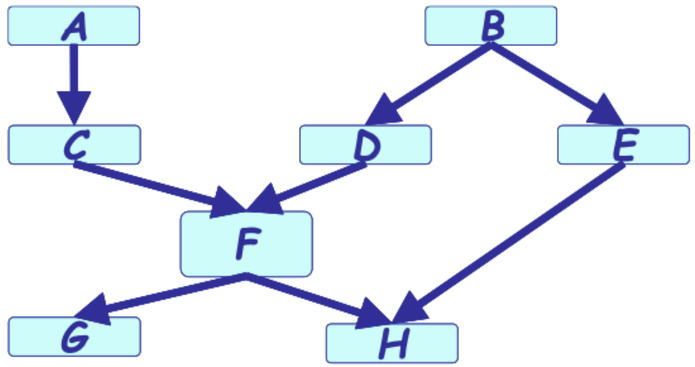


$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$

What is a PGM After All

It is a smart way to write/specify/compose/design exponentially-large probability distributions without paying an exponential cost, and at the same time endow the distributions with ***structured semantics***



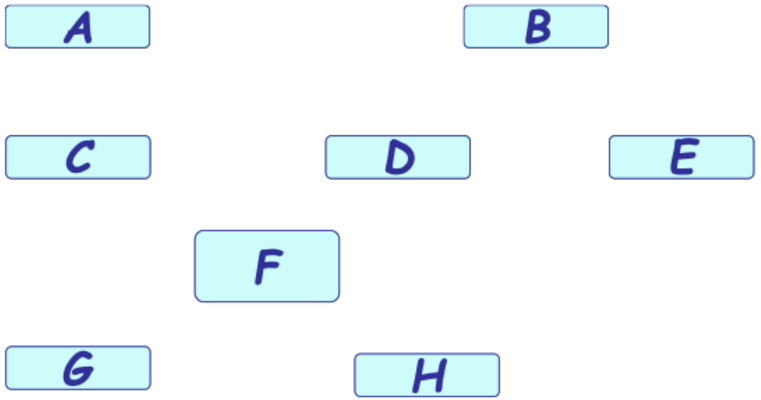


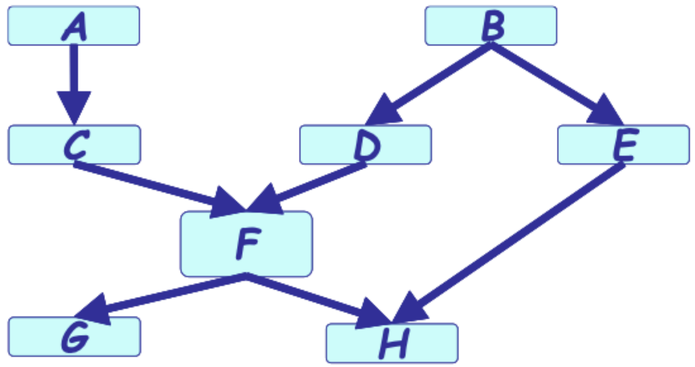
$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$

$$P(X_{1:8}) = P(X_1)P(X_2)P(X_3 \mid X_1 X_2)P(X_4 \mid X_2)P(X_5 \mid X_2)$$
$$P(X_6 \mid X_3, X_4)P(X_7 \mid X_6)P(X_8 \mid X_5, X_6)$$

What is a PGM After All

It is a smart way to write/specify/compose/design exponentially-large probability distributions without paying an exponential cost, and at the same time endow the distributions with ***structured semantics***





$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$

$$P(X_{1:8}) = P(X_1)P(X_2)P(X_3 \mid X_1 X_2)P(X_4 \mid X_2)P(X_5 \mid X_2)$$
$$P(X_6 \mid X_3, X_4)P(X_7 \mid X_6)P(X_8 \mid X_5, X_6)$$

More formal definition:

It refers to a family of distributions on a set of random variables that are compatible with all the probabilistic independence propositions encoded by a graph that connects these variables

Probabilistic Graphical Model is a graphical language to express conditional independence

Thank You!
Q & A