McGill **Artificial Intelligence** Society

Lecture 4: Dimensionality Reduction

Announcements

INSERT ANNOUNCEMENTS HERE

Research Talks

- Monday, February 18 (17:30-19:00) at Trottier 1030

Assignment 4

- Message @Daoud on Slack if you have questions/clarifications.
- Learn How to PR on Github, if you don't know, please ask your buddy for help, no point in keeping it to yourself.

Final Project Demo Format

- Application demo (eg. webapp / mobile) or poster presentation
- Feedback by tomorrow!

Last Lecture By Me! Second Half of the Course Taught by Isaac Chan

Today's Lesson Plan

A Brief Note on Finding Good Features

Dimensionality Reduction Motivation

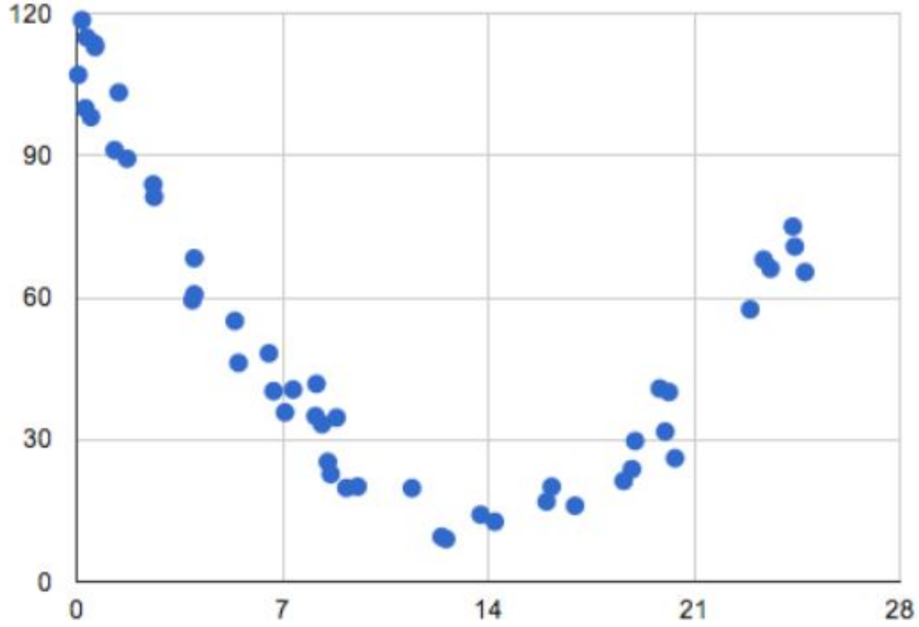
Deriving PCA

PCA Demo

Autoencoders

Autoencoder Demo

Motivation



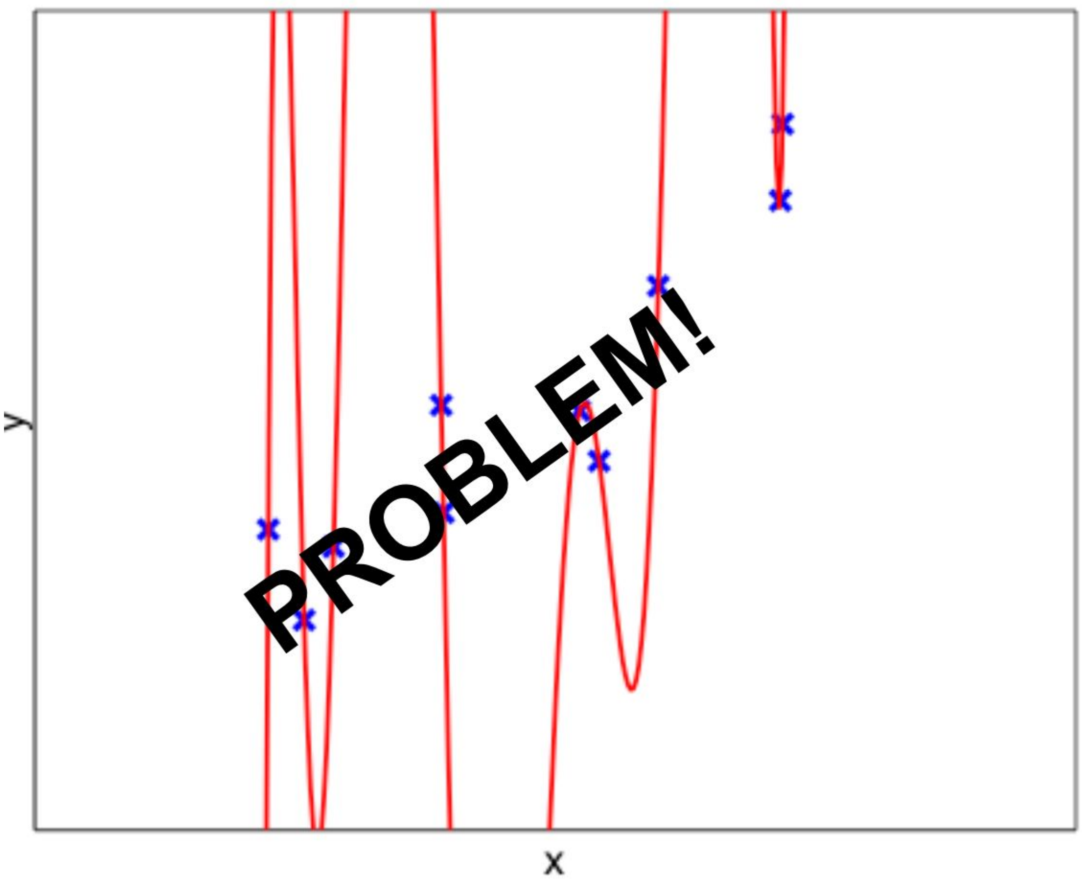
$$h(x) = b_0 + b_1 x + b_2 x^2$$

Motivation - A Note on Polynomial Features

$$\varnothing\begin{pmatrix}x_1\\x_2\end{pmatrix}=\begin{pmatrix}1\\x_1\\x_2\\{x_1}^2\\{x_2}^2\\x_1x_2\end{pmatrix}$$

Motivation - Overfitting Due To Number of Features

:

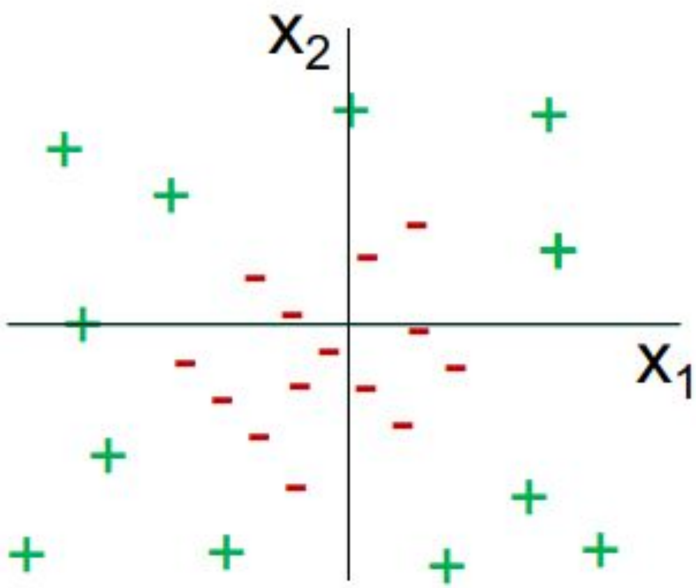


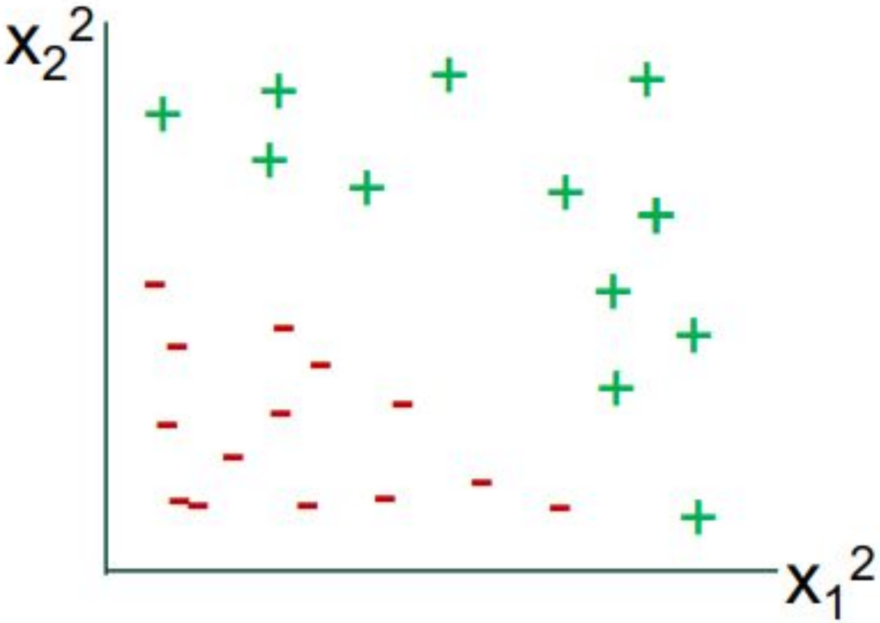
Finding Good Non-Linear Expansions

- Prior Knowledge
- Periodic Behaviour (sine, cosine)
- Knowledge of Independent Features
  - Avoid augmenting data with cross-terms of these independent features
- Etc.

Sometimes, we can find good features by analyzing our dataset!

Finding Good Features





Slide adapted from Herke Van Hoof's - COMP 551 Lecture 13 on Dimensionality Reduction

Features, so many of them...

- Images and Video Data
- Gene expression data
- NLP Tasks with very large vocabularies

Dimensionality Reduction

Dimensionality reduction is an unsupervised machine learning task!

Given dataset X with m features, we wish to learn X’ with m’ features, where m’ << m

PROS

- Reduces variance of the model
- Reduce model complexity and training time
- Find the most relevant features (or combinations of features)
- Data compression
- Better interpretability

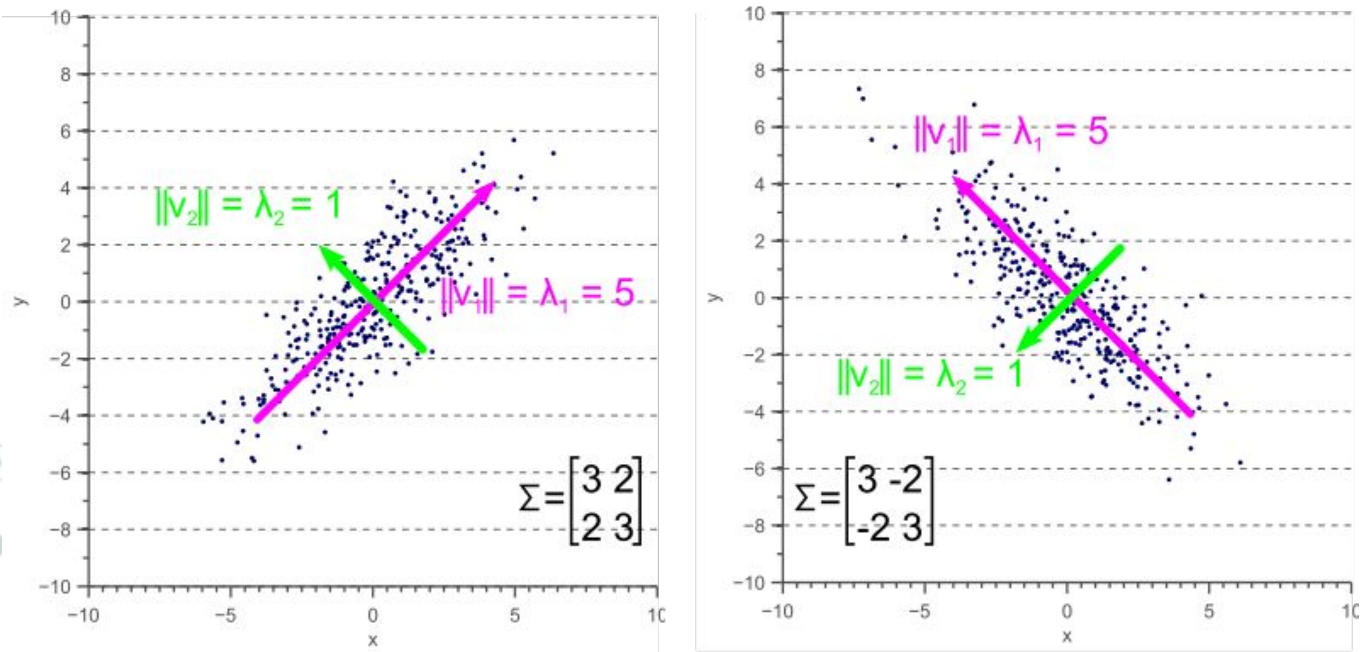
Principal Component Analysis (PCA)

Goal: Project data into lower dimension subspace, given matrix X with m features, we wish to learn X' with m' features, where m' << m

$$Ax = \lambda x$$

How do we ensure the preservation of important information?

Criteria - Projection onto linear subspace must **maximize** variance!



Principal Component Analysis (PCA)

Goal: Project data into lower dimension subspace, given matrix X with m features, we wish to learn X' with m' features, where $\boxed{m' << m}$

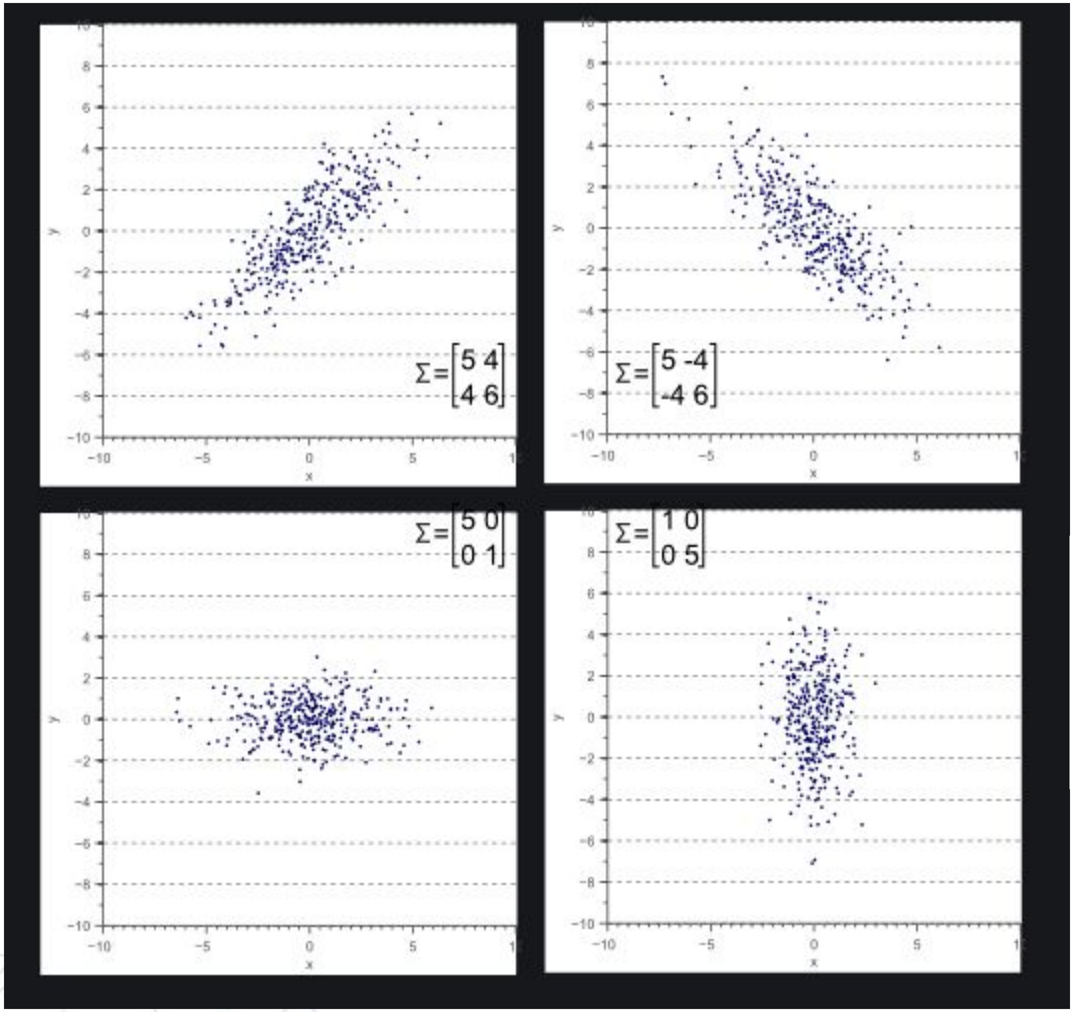
$$R^m \rightarrow R^{m'}$$

$$Given\ sample\ x_i \in R^m, Wx_i \rightarrow R^{m'}, where\ W \in R^{m' \times m}\ \ (compression\ matrix)$$

$$Assume\ \exists\ decompression\ matrix\ U \in R^{m \times m'}$$

$$argmin_{W,U} \sum_{i=1}^{n} ||x_i - UWx_i||^2$$

PCA - The Covariance Matrix



$$\Sigma = \begin{bmatrix} \sigma(x,x) & \sigma(x,y) \\ \sigma(y,x) & \sigma(y,y) \end{bmatrix}$$

Image Credits

Principal Component Analysis (PCA)

Goal: Project data into lower dimension subspace, given matrix X with m features, we wish to learn X' with m' features, where $\boxed{m' << m}$

- Select the projection dimension m' using cross-validation
- Typically, we center the examples by subtracting the mean

$$\bar{X} = X - \mu(X)$$

Numerical Example

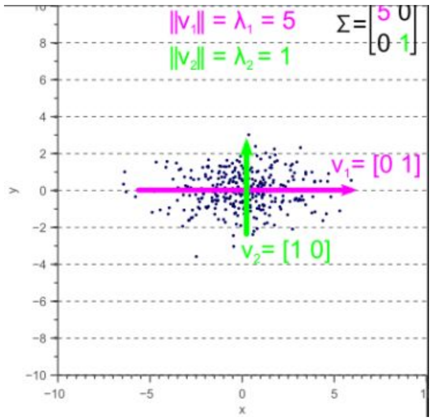
$$X = \begin{array}{c} \begin{array}{cc} x_1 & x_2 \end{array} \\ \begin{bmatrix} 3 & 7 \\ 3 & 3 \\ 3 & 7 \\ 4 & 5 \\ 2 & 3 \end{bmatrix} \end{array}, \quad \mu(X) = \begin{bmatrix} 3 & 5 \\ 3 & 5 \\ 3 & 5 \\ 3 & 5 \\ 3 & 5 \end{bmatrix} \quad \bar{X} = \begin{bmatrix} 0 & 2 \\ 0 & -2 \\ 0 & 2 \\ 1 & 0 \\ -1 & -2 \end{bmatrix}$$

$$\bar{X} = X - \mu(X)$$

Principal Component Analysis (PCA)

Typically, we center the examples by subtracting the mean

$$\bar{X} = X - \mu(X)$$



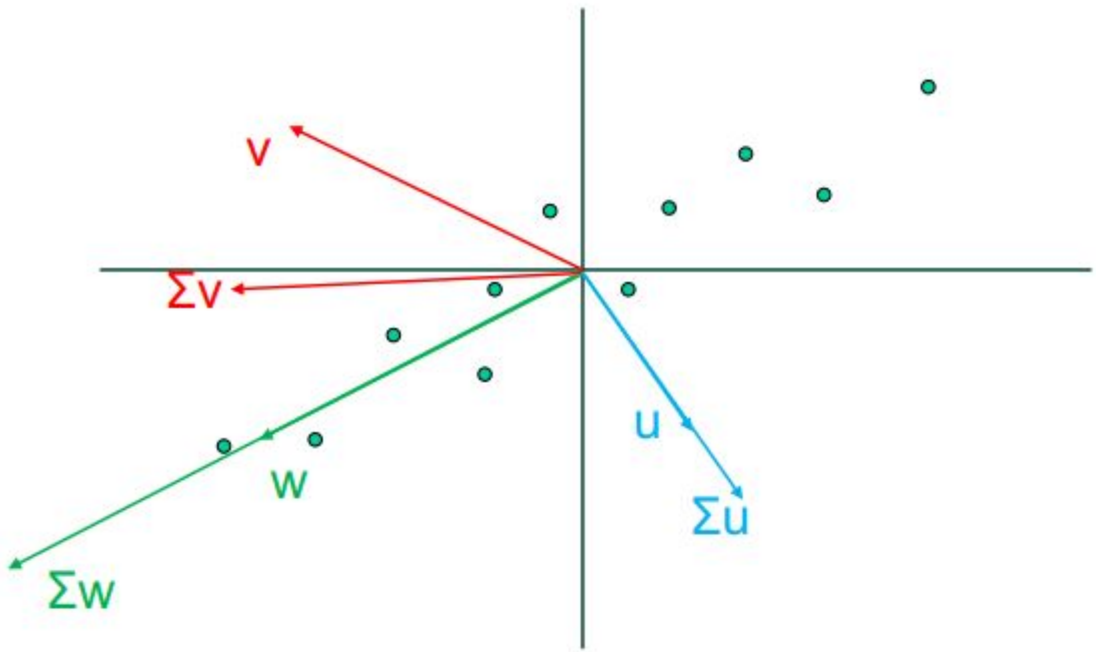
Closed form solution: W matrix composed of the m’ eigenvectors corresponding to the largest eigenvalues of the covariance matrix of X

Then, we compute the covariance matrix (this formula only holds because we subtracted the mean from X):

$$\sum = \bar{X}^T \bar{X}$$

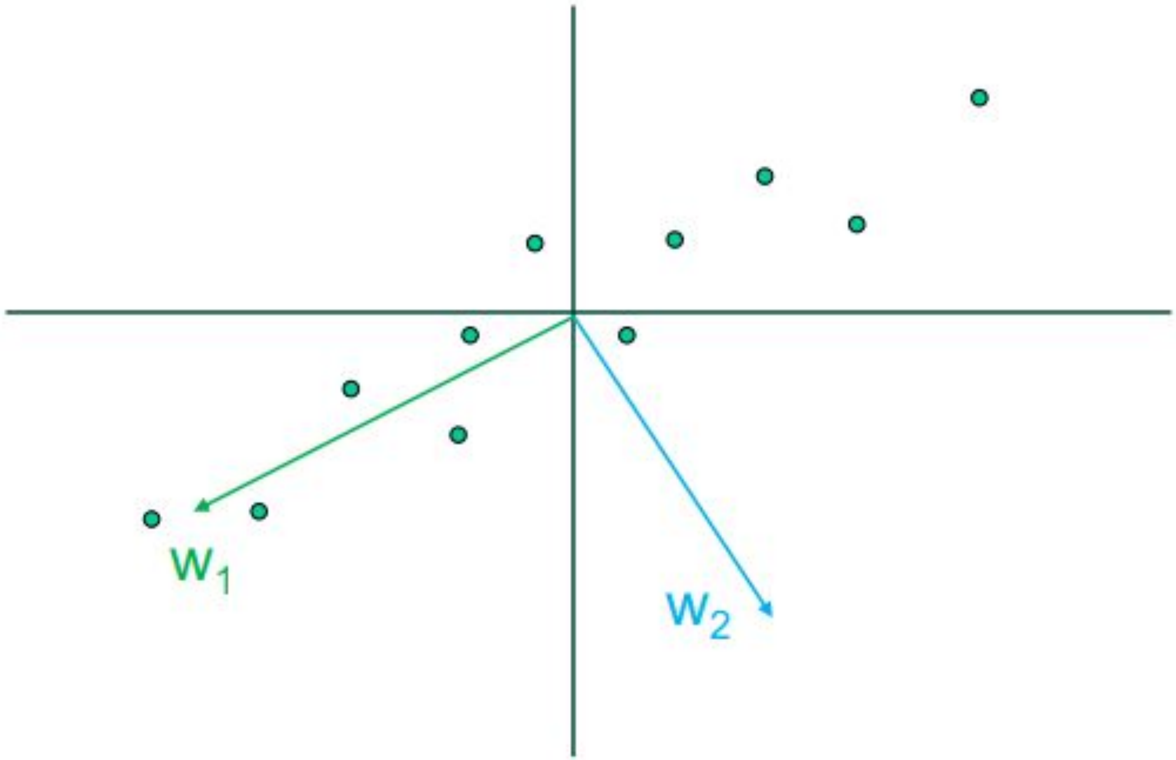
Slides 16-21 Taken from Herke Van Hoof’s COMP551 Slides Lecture 13 - Dimensionality Reduction

Principal Component Analysis (PCA)



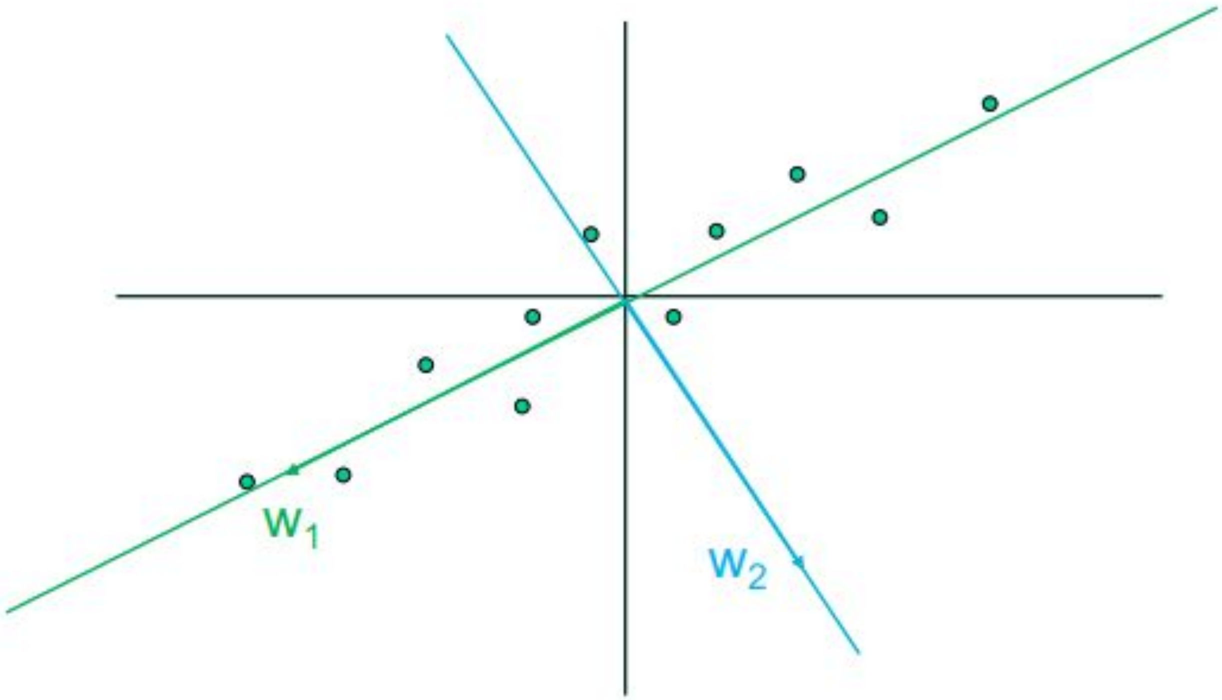
Eigenvector equation: $\lambda w = \Sigma w$

Principal Component Analysis (PCA)



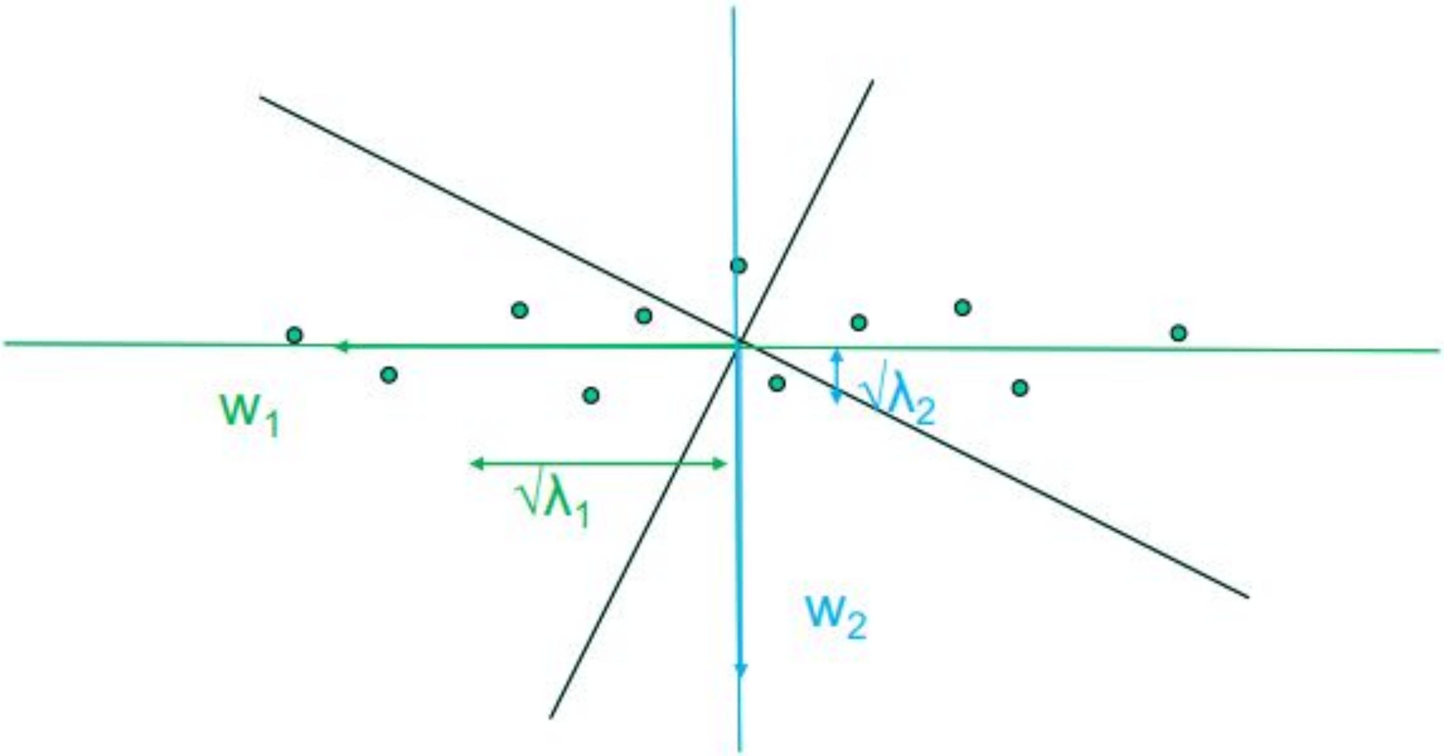
Eigenvector equation: $\lambda w = \Sigma w$

Principal Component Analysis (PCA)



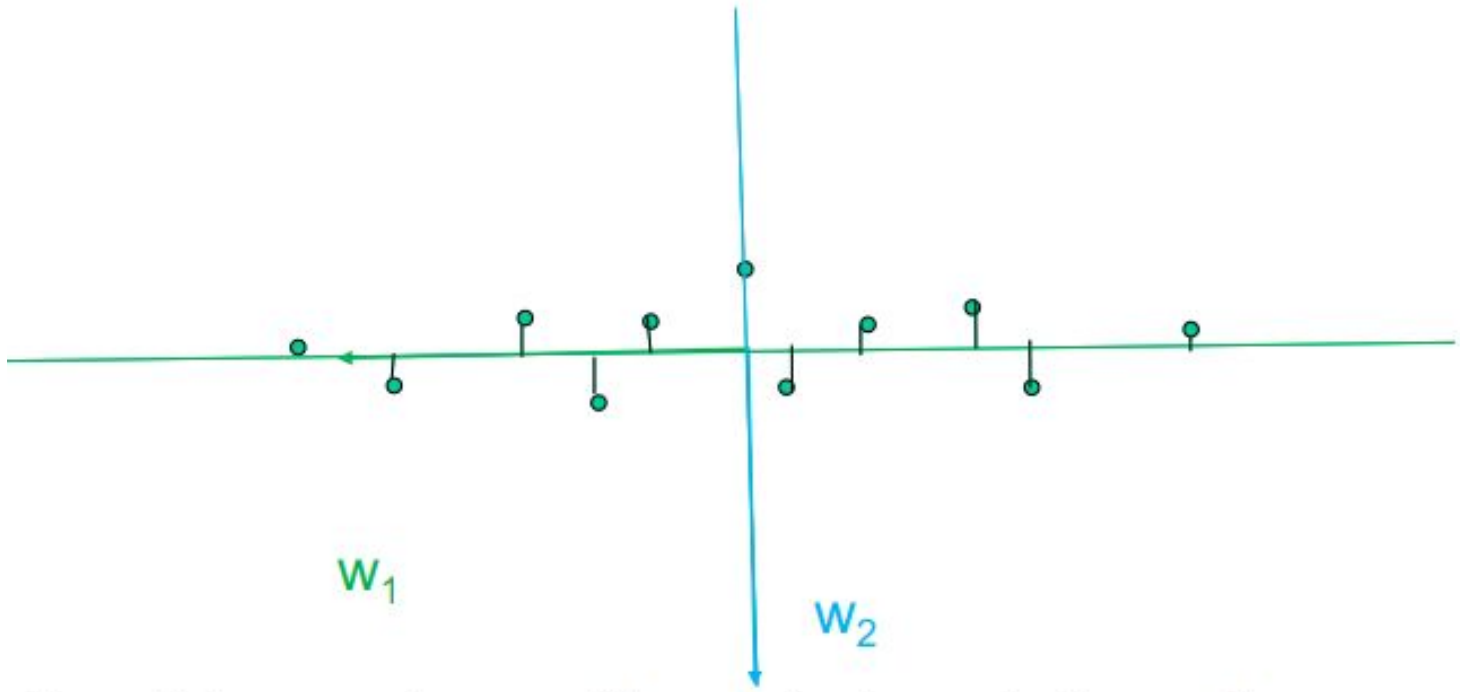
W specifies a new basis

Principal Component Analysis (PCA)



W specifies a new basis

Principal Component Analysis (PCA)



Now it is easy to specify most relevant dimension:
Reduce the dimensionality 1d
Found dimension is a combination of x1 & x2

Principal Component Analysis (PCA)

PCA Summary

In summary, to compute m’ principal components:

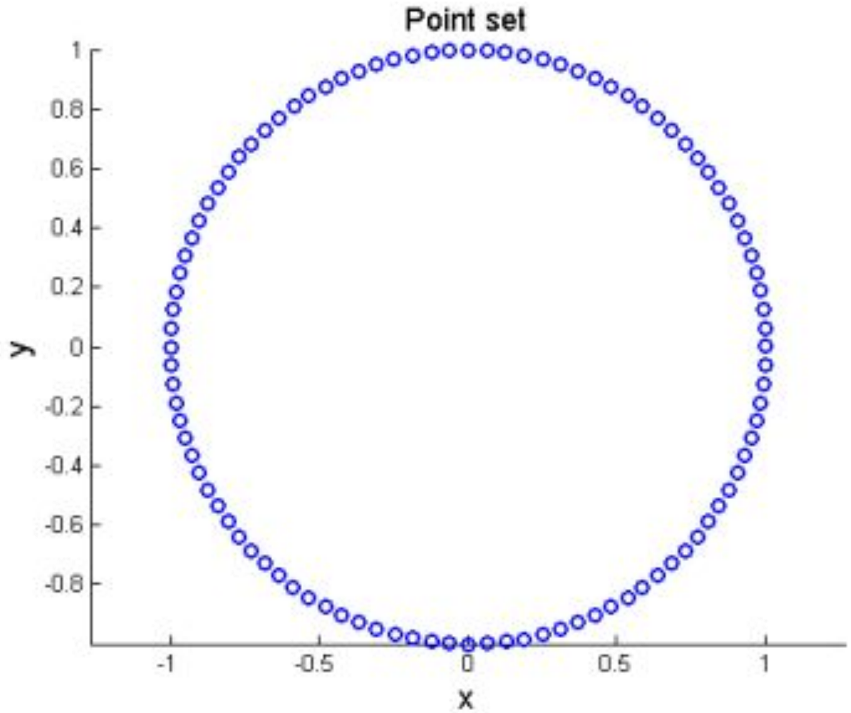
- Calculate matrix W, with matrix columns being the eigenvectors of covariance
- Take m’ most relevant components (determined via cross-validation)

PCA Demo

Autoencoders

Motivation

What's the problem with applying PCA to this data (What assumption does PCA make when reducing dimensions?)



Autoencoders

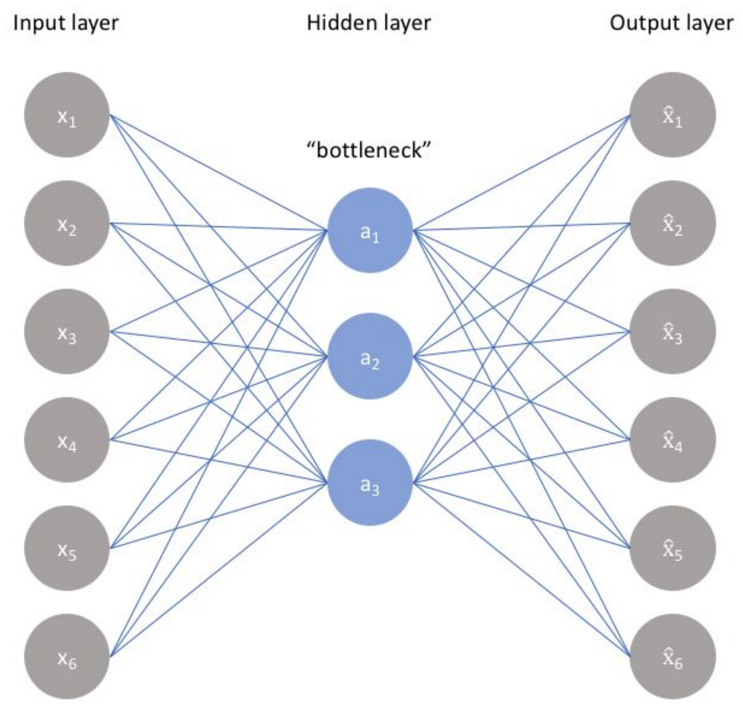


Image Credits

- Learning data features for future tasks
- Useful as extra data for supervised tasks
- Learned features can be used for clustering, visualization

Autoencoders

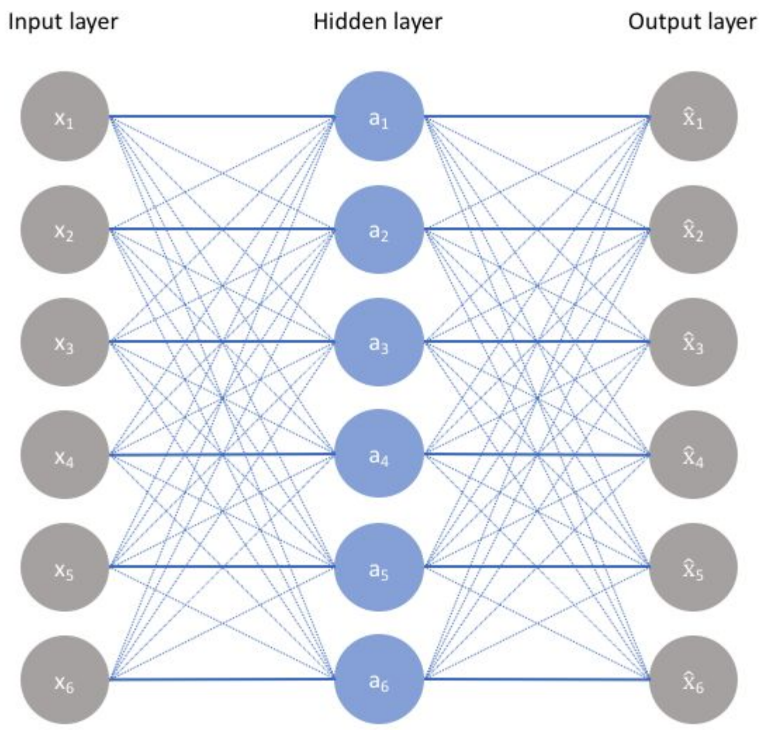


Image Credits

$L(x, \hat{x})$

Continuous inputs: Squared-error loss
Binary inputs: Cross-entropy loss (log-loss)

Autoencoders - Dealing with Sensitive Inputs

Goal: Accurately build a reconstruction, without simply memorizing or overfitting training data!

$$L(x, \hat{x}) + regularization\ term$$

Without regularization, we can only hope to prevent overfitting by limiting the number of nodes present in the hidden layer(s)!

E.g. L1 regularization, sparse autoencoders, contractive autoencoders

Denoising Autoencoders

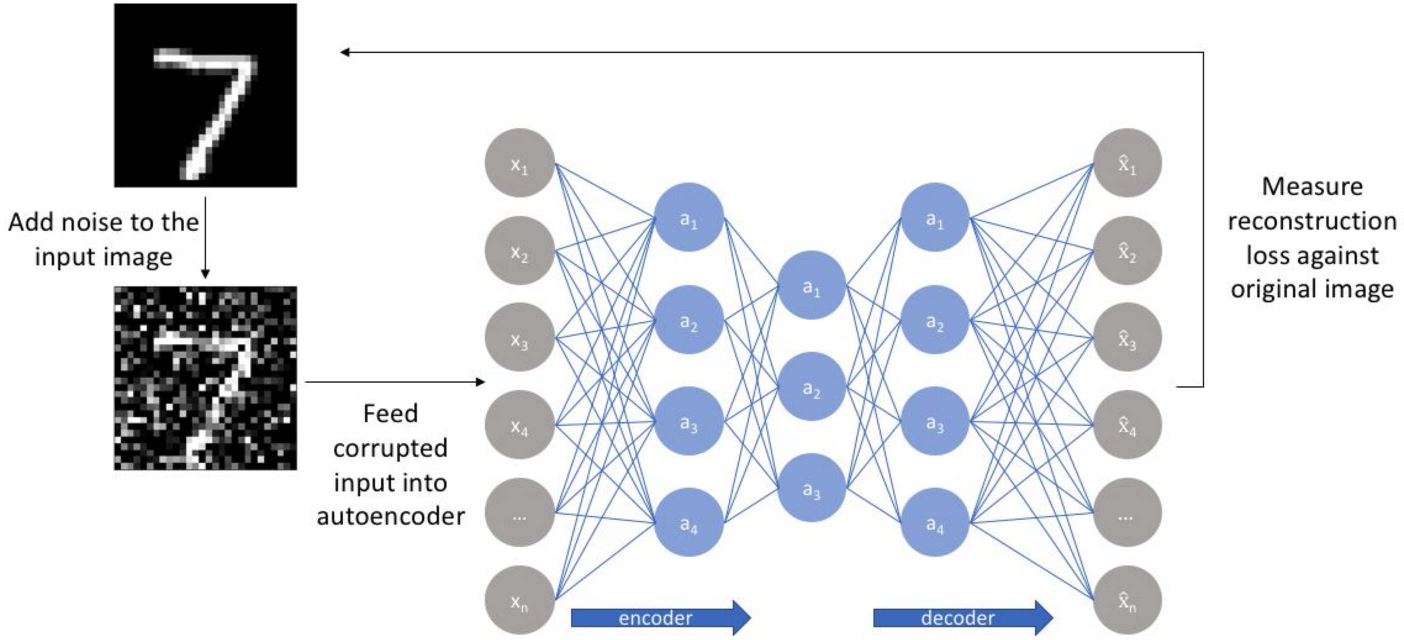


Image Credits

PCA vs Autoencoders

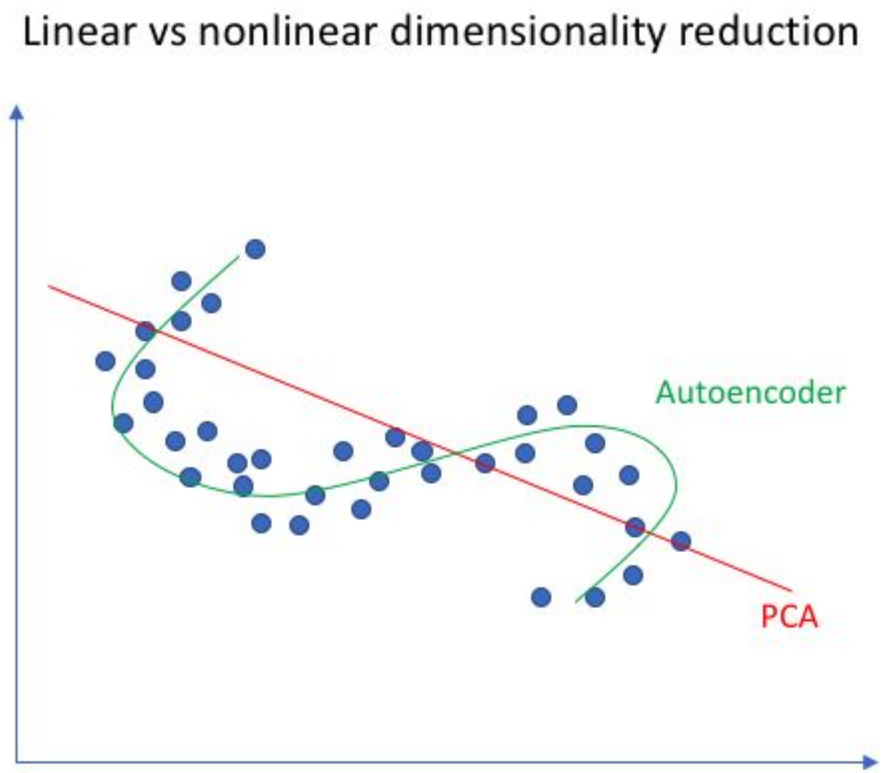


Image Credits

Autoencoder Demo

Thanks!

Any questions?

Reminders:

Homework 3 Due Tomorrow, Homework 4 Due Before Next Lecture