Machine Learning for Data Science (CS4786)
Lecture 26

Wrapping Up

Course Webpage :
http://www.cs.cornell.edu/Courses/cs4786/2016sp/

Competition I

- You guys all worked really really hard! Kudos!
- Task 1 was a harder than we intended!
- After report deadline we will post code/pseudo code of our solutions
- Task II you guys did better than us!

Real world data is hard!

Grades are for what you tried and how you tought about the problem, don't fret about accuracy!

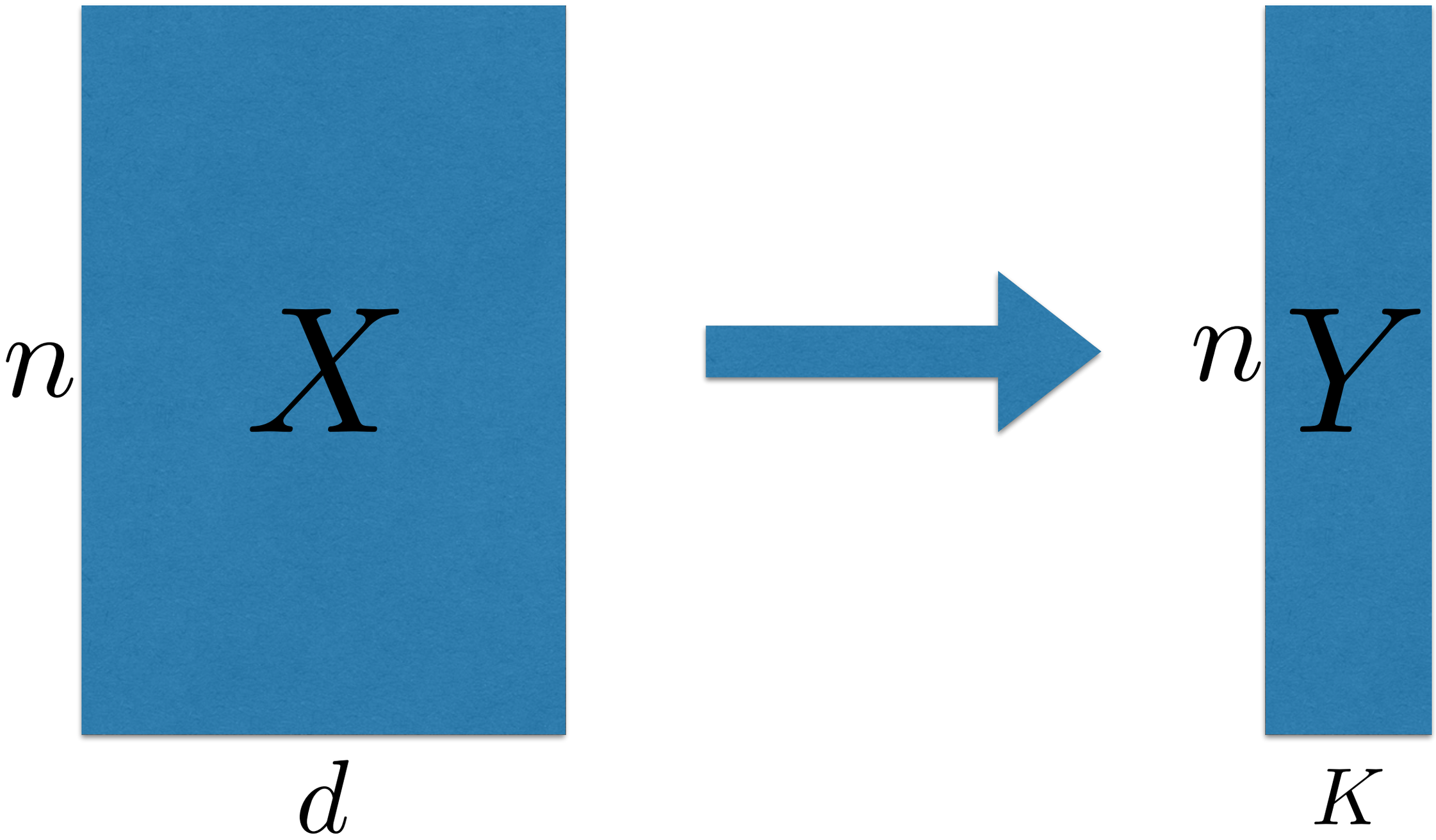
Competition II

- We are trying to set up the Kaggle part
- Data generated from HMM
- Labels are one of ‘1‘ to ‘5‘
- a small percentage (5%) of the sequence are generated in reverse order
- First 100 sequences are given completely
- 101-1000 sequences each have one missing value denoted by a ‘*’

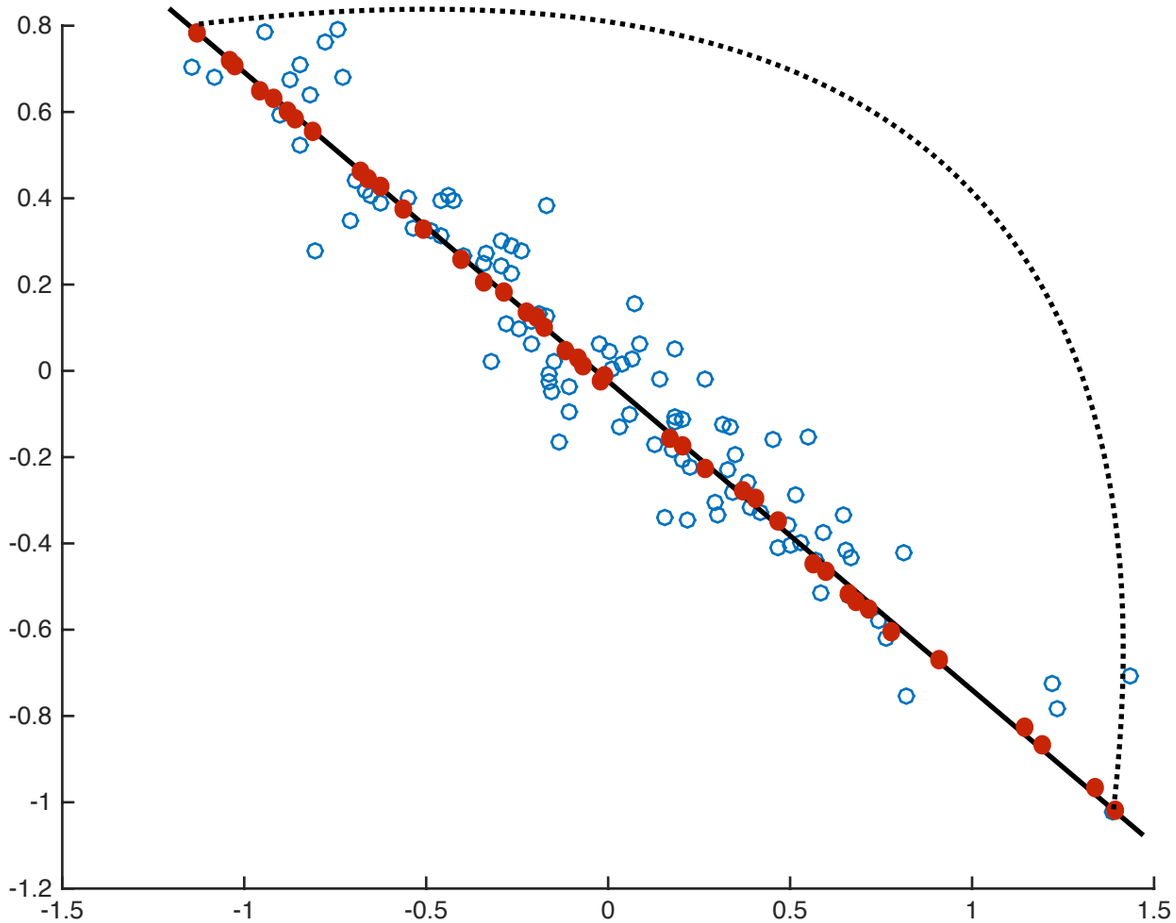
Task: Predict the missing values
Baseline: Accuracy of predicting majority label (=1) $\approx 25\%$

What have we covered so far?

Dimensionality Reduction

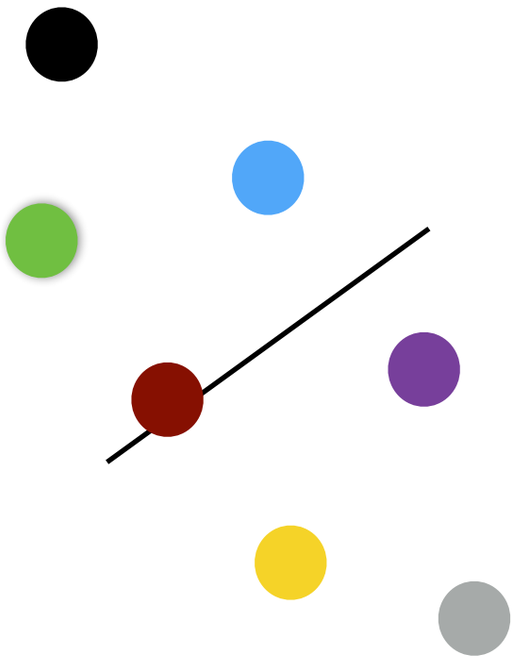


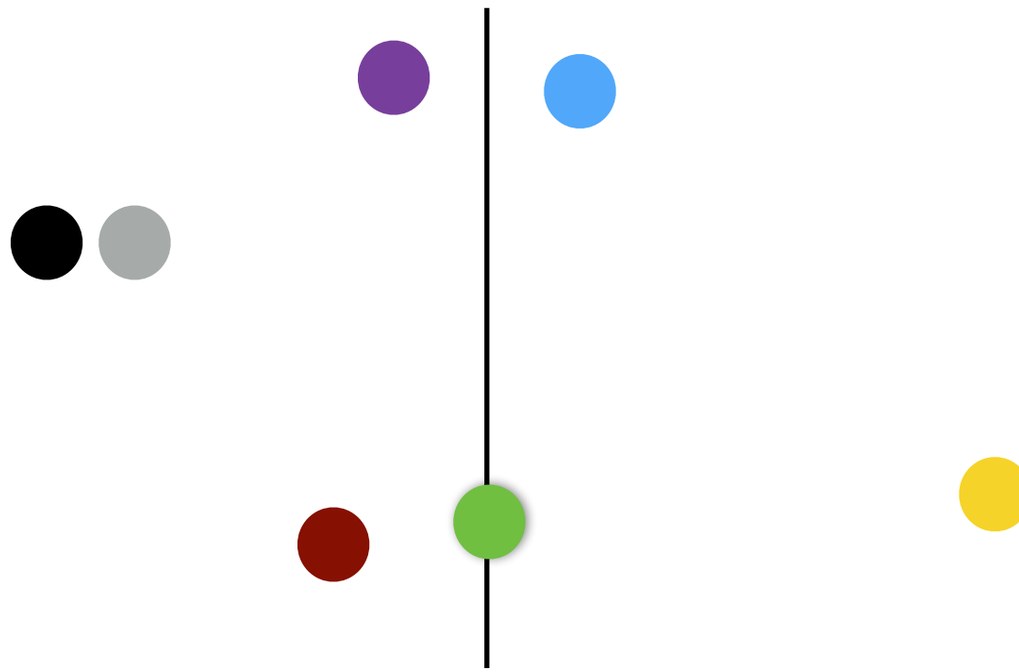
PCA: Variance Maximization



First principal direction = Top eigen vector

Which Direction to Pick?



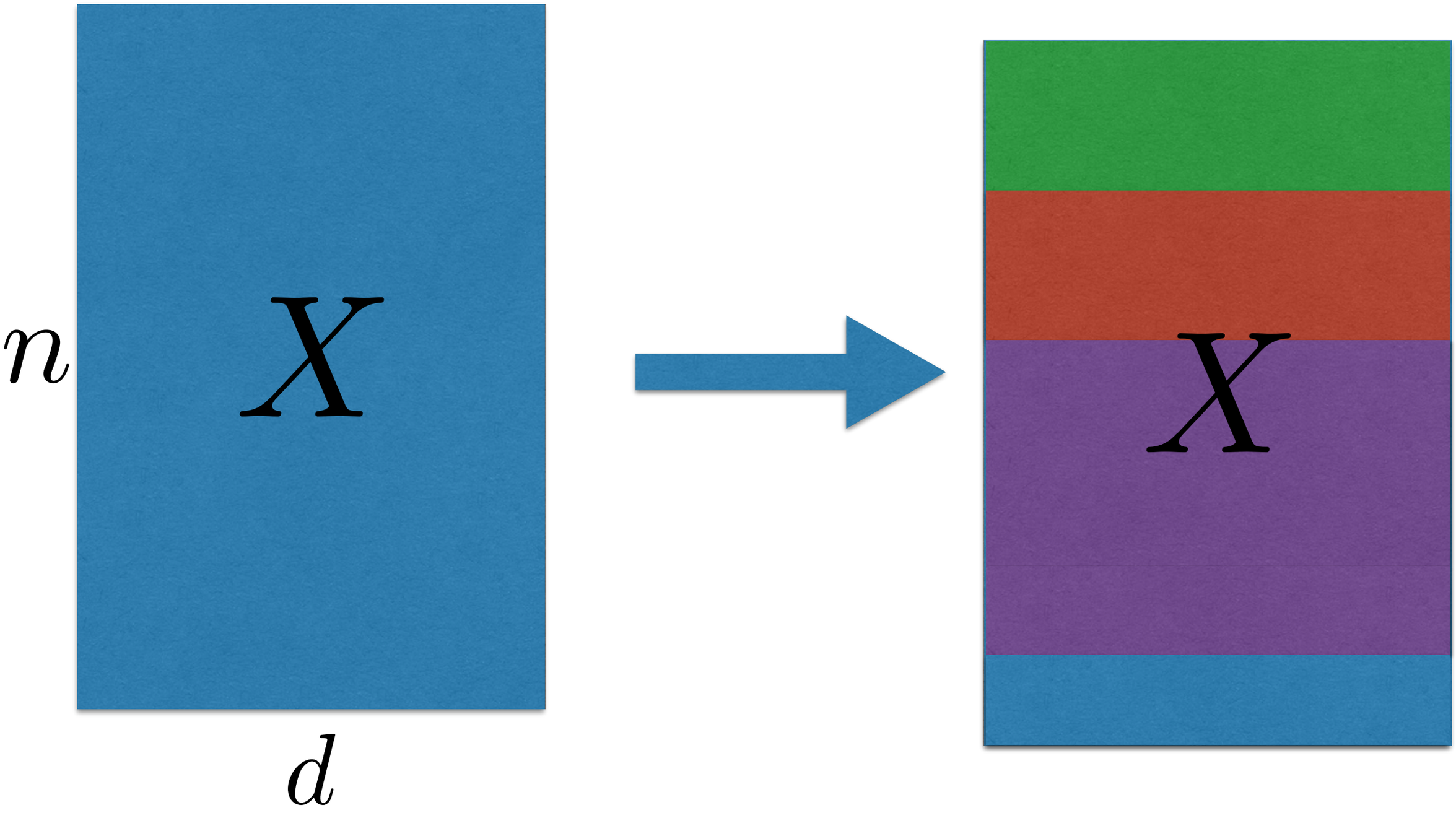


Direction has large correlation

Pick a Random W

$$Y = X \times \begin{bmatrix} +1 & \dots & -1 \\ -1 & \dots & +1 \\ +1 & \dots & -1 \\ & \vdots & \\ +1 & \dots & -1 \end{bmatrix}_K \Big/ \sqrt{d}$$

Clustering



K-means Clustering

- For all $j \in [K]$, initialize cluster centroids $\hat{\mathbf{r}}_j^1$ randomly and set $m = 1$
- Repeat until convergence (or until patience runs out)
  1. For each $t \in \{1, \ldots, n\}$, set cluster identity of the point

$$\hat{c}^m(\mathbf{x}_t) = \underset{j \in [K]}{\operatorname{argmin}} \left\| \mathbf{x}_t - \hat{\mathbf{r}}_j^m \right\|$$

  2. For each $j \in [K]$, set new representative as

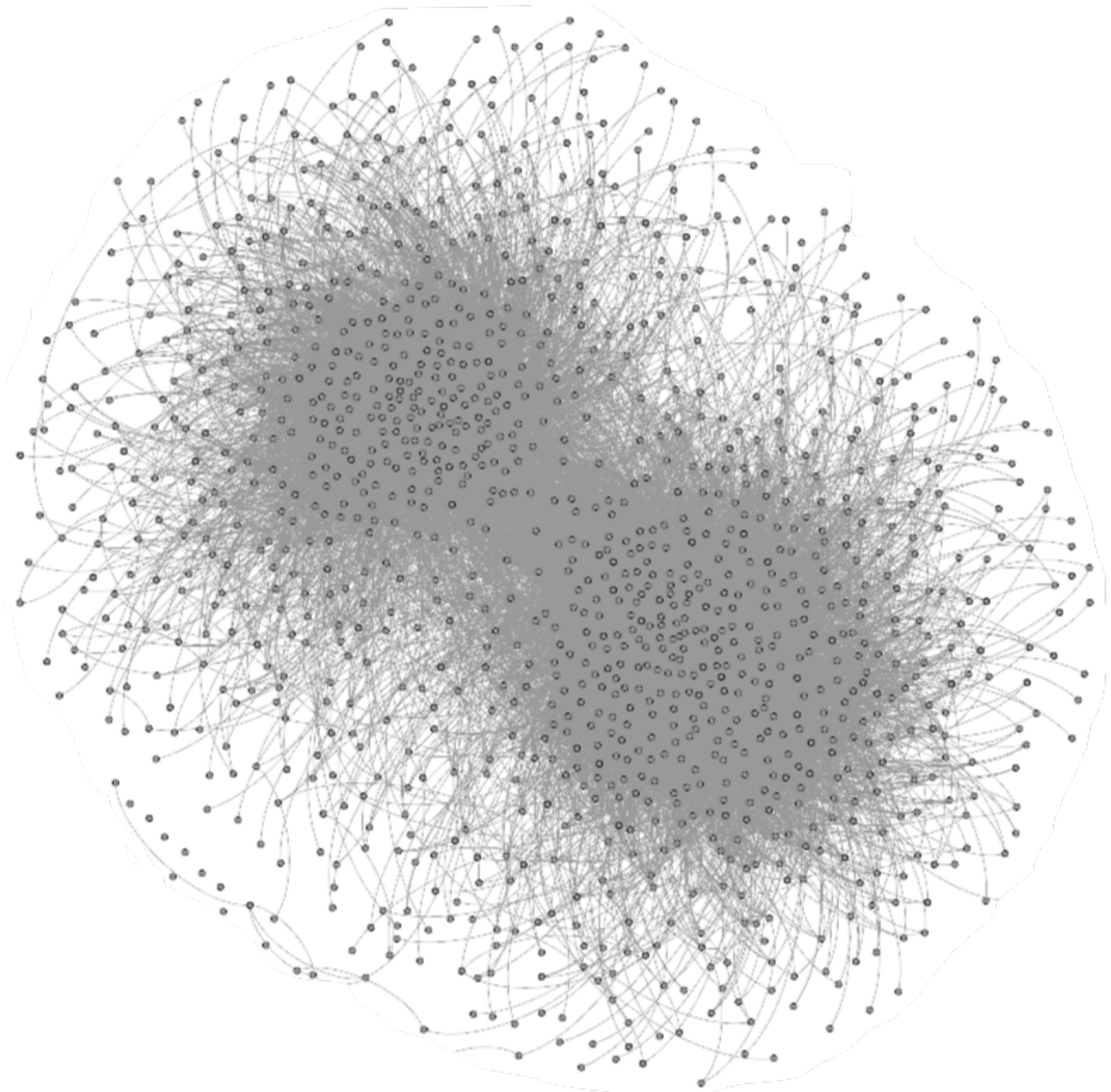
$$\hat{\mathbf{r}}_j^{m+1} = \frac{1}{|\hat{C}_j^m|} \sum_{\mathbf{x}_t \in \hat{C}_j^m} \mathbf{x}_t$$

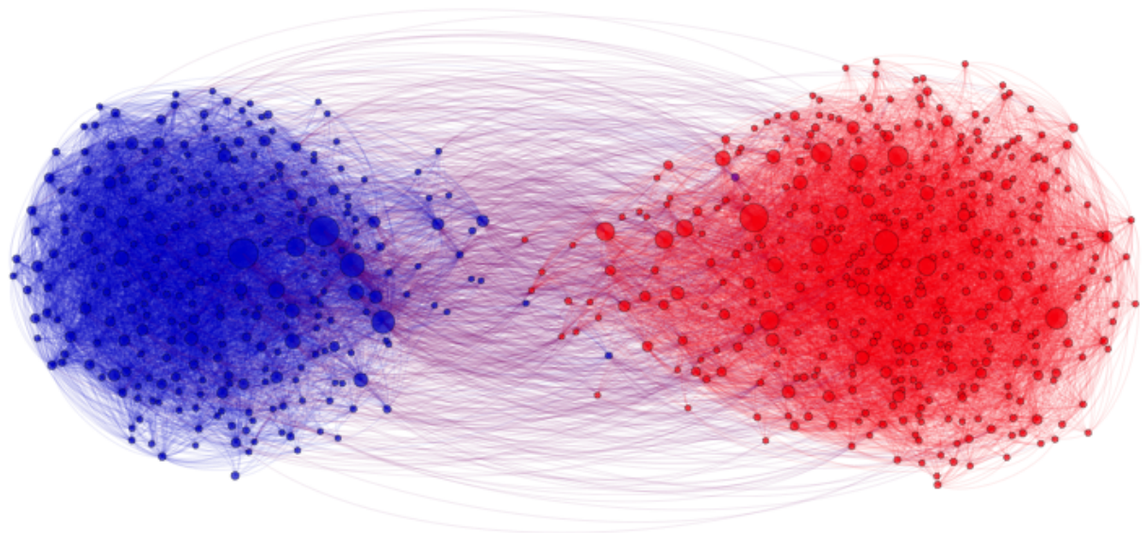
  3. $m \leftarrow m + 1$

Single Link Clustering

- Initialize n clusters with each point $\mathbf{x}_t$ to its own cluster
- Until there are only K clusters, do
  1. Find closest two clusters and merge them into one cluster
  2. Update between cluster distances (called proximity matrix)

Tell me who your friends are . . .





- Cluster nodes in a graph.
- Analysis of social network data.

Impossibility

Theorem

Any clustering algorithm that has scale invariance and consistency does not have richness.

No Free Lunch!

Probabilistic Models

- Θ consists of set of possible parameters
- We have a distribution P_θ over the data induced by each $\theta \in \Theta$
- Data is generated by one of the $\theta \in \Theta$
- Learning: Estimate value or distribution for $\theta^* \in \Theta$ given data

Expectation Maximization Algorithm

- For demonstration we shall consider the problem of finding MLE (MAP version is very similar)
- Initialize $\theta^{(0)}$ arbitrarily, repeat unit convergence:

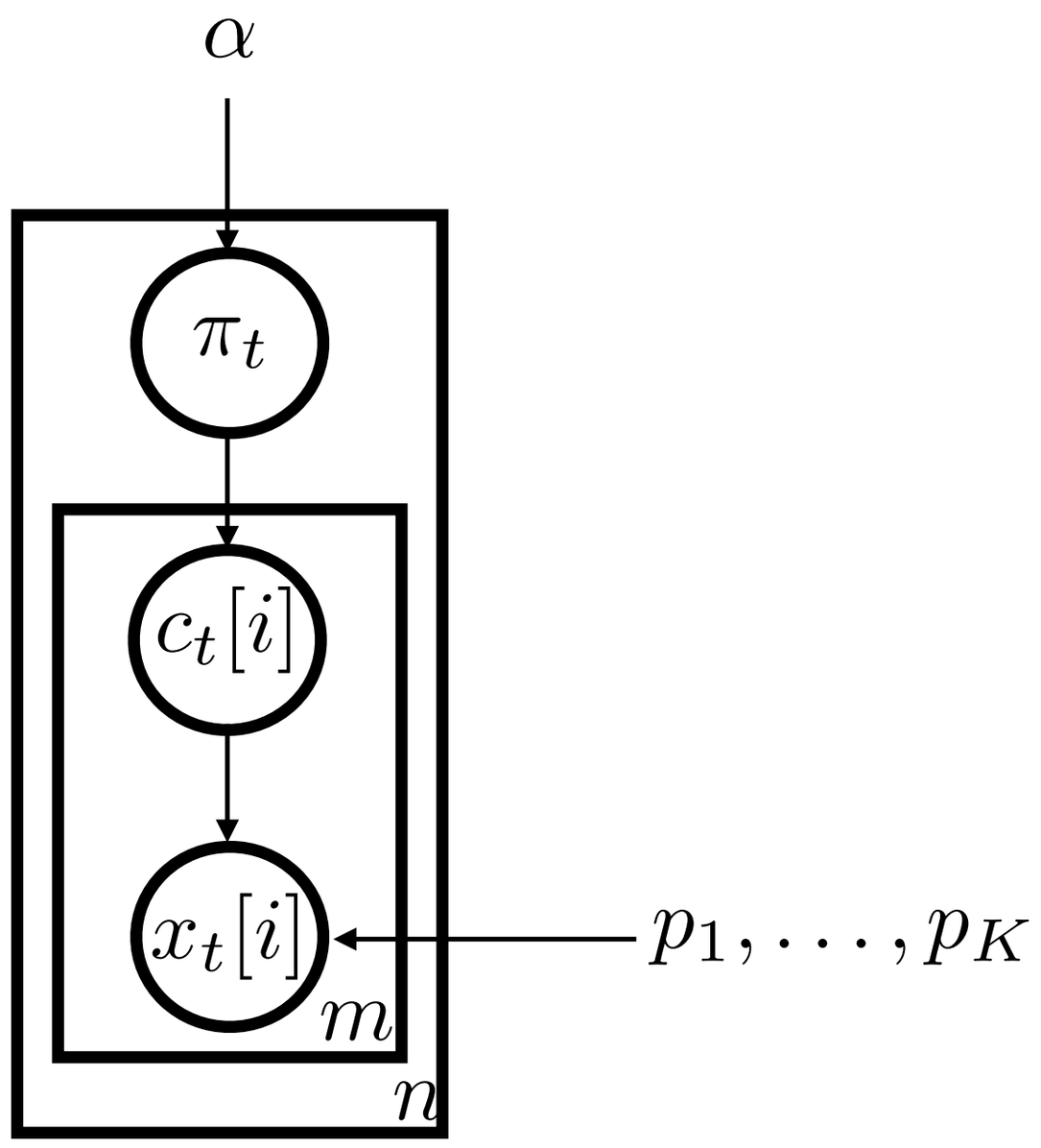
(E step) For every t, define distribution Q_t over the latent variable c_t as:

$$Q_t^{(i)}(c_t) = P(c_t|x_t, \theta^{(i-1)})$$

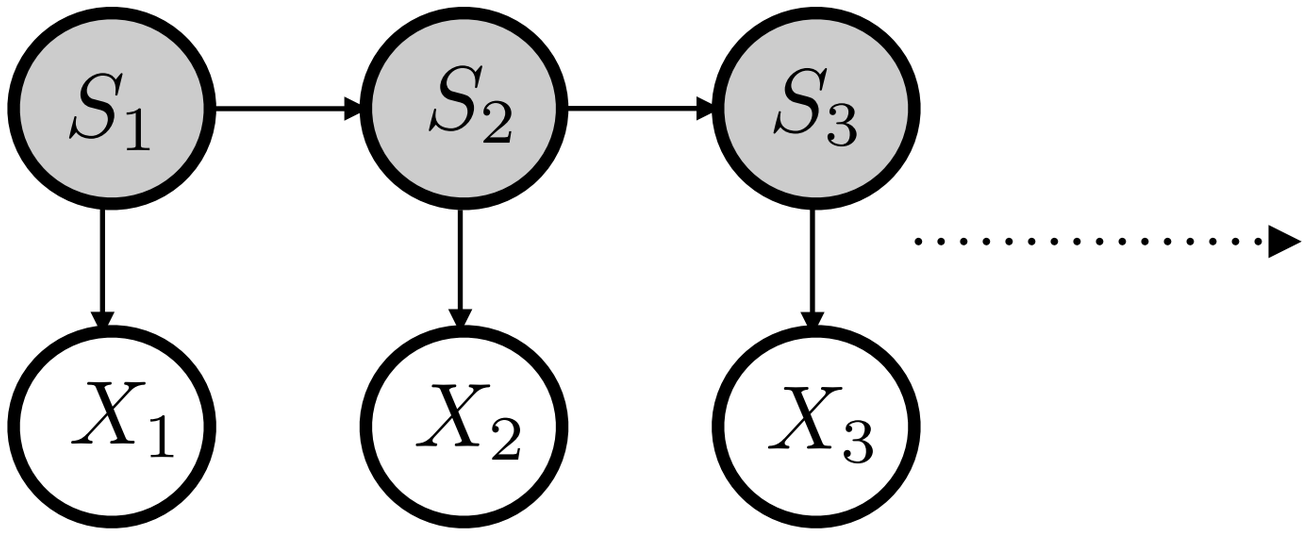
(M step)

$$\theta^{(i)} = \text{argmax}_{\theta \in \Theta} \sum_{t=1}^{n} \sum_{c_t} Q_t^{(i)}(c_t) \log P(x_t, c_t|\theta)$$

Latent Dirichlet Allocation



Example: Hidden Markov Model



Bayesian Networks

- Directed acyclic graph (DAG): $G = (V, E)$
- Joint distribution P_θ over $X_1, \ldots, X_n$ that factorizes over G:

$$P_\theta(X_1, \ldots, X_n) = \prod_{i=1}^{N} P_\theta(X_i | \text{Parent}(X_i))$$

- Hence Bayesian Networks are specified by G along with CPD's over the variables (given their parents)

Graphical Models

. Variable Elimination

. Message Passing

. Approximate Inference

. Parameter Estimation/learning using EM

Bigger Picture

- Dimensionality reduction, clustering and more generally learning

There are no free lunches :(

- Probabilistic modeling makes assumptions or guesses about way data is generated or how variables are related
- Caution:
  - In the real world no modeling assumption is really true . . . there are good fits and bad fits
  - Choosing a model: Bias Vs Variance, Approximation error Vs estimation error, Expressiveness Vs amount of data
  - Choose the right model for the right job, there are no universally good answers
  - Feature extraction is an art (not covered in class)

Some of what have we have not covered?

Supervised Learning

- Training data: $(x_1, y_1), \ldots, (x_n, y_n)$ provided (typically assumed to be drawn from a fixed unknown distribution)
- Goal: Find a mapping h from input instances to outcome that minimizes $\mathbb{E}\Big[\ell(\hat{h}(x), y)\Big]$
  (ℓ is a loss function that measures error in prediction)

Generative Vs Discriminative Approaches

Generative approach:

- Input instances x_t's are generated based on/by y_t's
- We try to model $P(y, x) = P(x|y)P(y)$
- Example: Naive Bayes

Discriminative approach:

- We model $P(Y|X)$ or the boundary of classification
- Rationale: we are only concerned with predicting output y's given input x
- Example: linear regression, logistic regression

Probabilistic Story Vs Optimization Story

- Maximizing likelihood is same as minimizing negative log likelihood.
- Think of - log likelihood as loss function

$$-\log(P_\theta(Y|X)) \rightarrow \text{loss}(h_\theta(X), Y)$$

ie. θ parameterizes hypothesis for prediction or boundary

- MLE = Find hypothesis minimizing empirical loss on data
- Log Prior can be viewed as "regularization" of hypothesis

$$-\log(P(Y|X,\theta)) - \log(P(\theta)) \rightarrow \text{loss}(h_\theta(X), Y) + R(\theta)$$

- MAP = Find hypothesis minimizing empirical loss + regularization term
- Not all losses can be viewed as negative log likelihood

Semi-supervised Learning

- Can we used unlabeled examples to learn better?
- For instance, if we had a generative graphical model for the data: do example
- If we had prior information about the marginal distribution of X's and its relation to $P(Y|X)$

Active Learning

- Humans label the examples, can we get the learning algorithm in the loop?
- Learning algorithm picks the examples it wants labeled
- Eg. Margin based active learning, query points where model that fits observed data well so far disagree most

Domain Adaptation

- We learn a particular task on one corpus but want to use this learnt model to adapt with much fewer examples on another corpus
- Typical assumption: $P(Y|X)$ in both corpus remain fixed
- Marginal distributions change across the corpuses

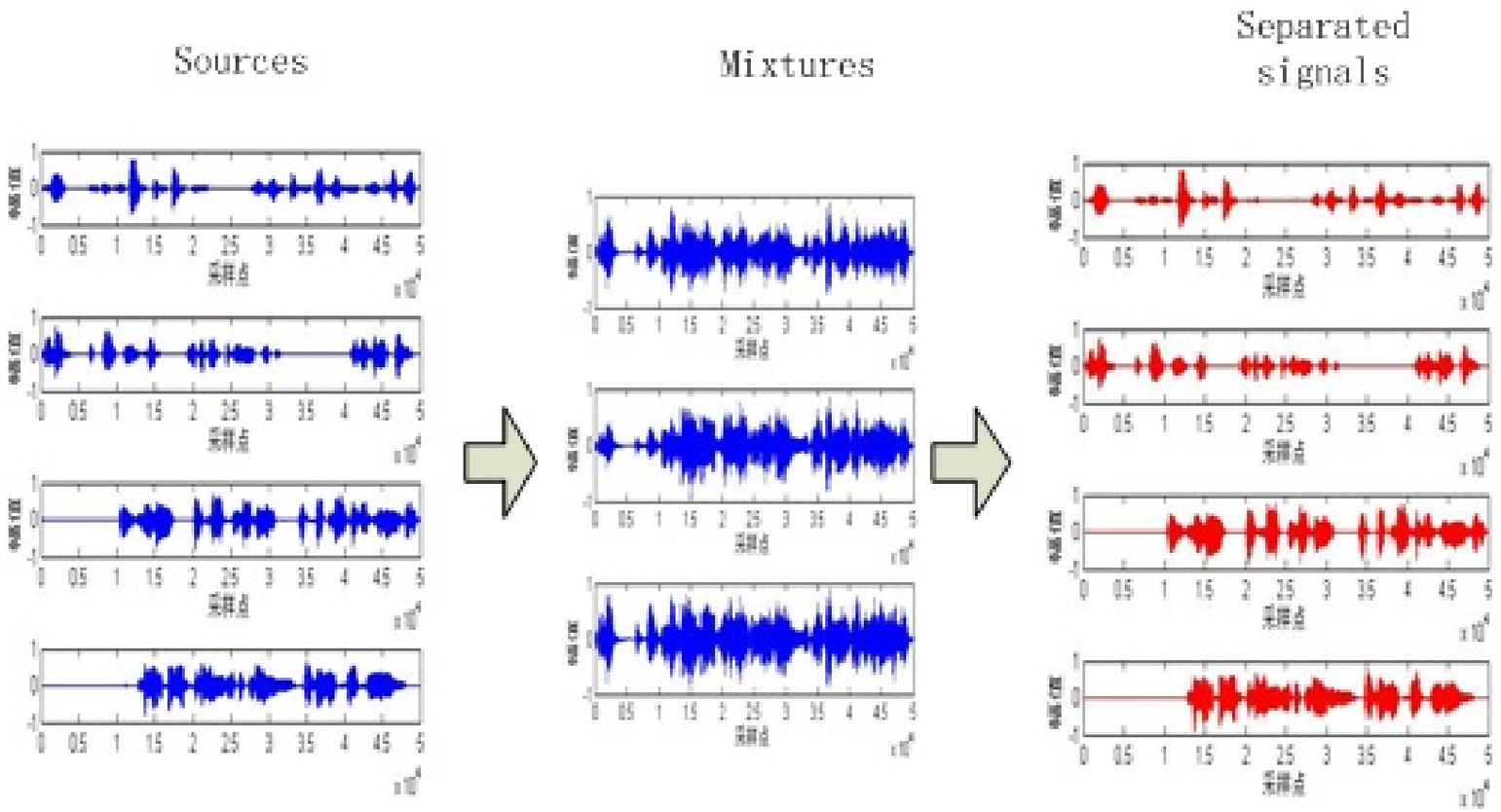
Independent Component Analysis (ICA)

Cocktail Party

- You are at a cocktail party, people are speaking all around you
- But you are still able to follow conversation with your group?
- Can a computer do this automatically?

Independent Component Analysis (ICA)

Blind Source Seperation



- Can do this as long as the sources are independent
- Represent data points as linear (or non-linear) combination of independent sources

Other Learning Frameworks

1. Transfer learning, multitask learning
2. Collaborative Filtering
3. Structured prediction
4. Online learning
5. ...