Clustering problems, mixture models and Bayesian nonparametrics

Nguyễn Xuân Long

Department of Statistics
Department of Electrical Engineering and Computer Science
University of Michigan

Vietnam Institute of Advanced Studies of Mathematics (VIASM),
30 Jul - 3 Aug 2012

Outline

What about . . .

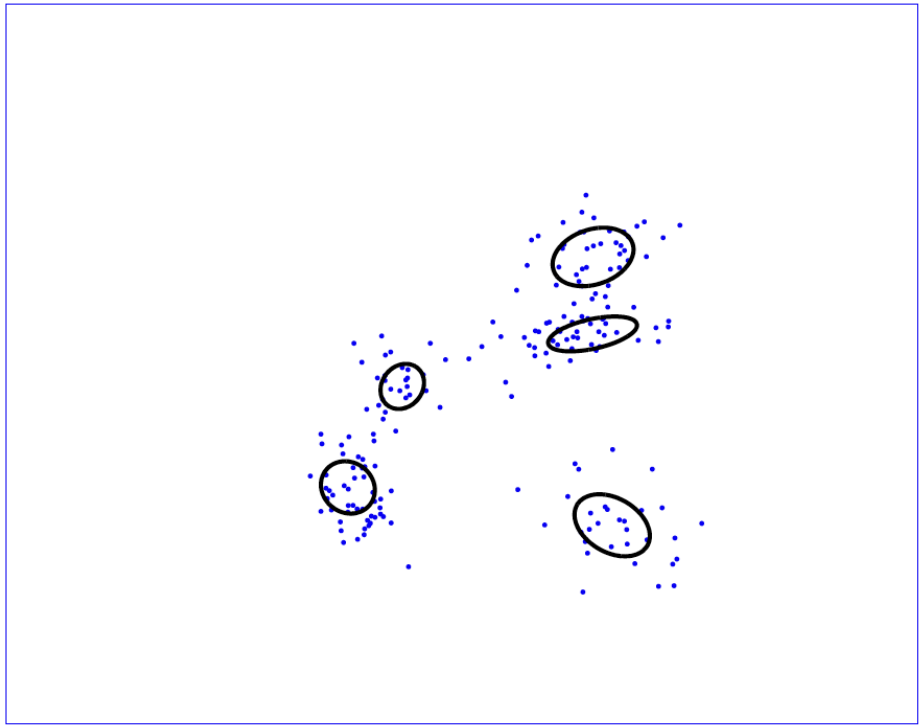
"Something old, something new, something borrowed, something blue . . ."

All these techniques center around clustering problems, but they illustrate a fairly large body of work in modern statistics and machine learning

- Part 1, 2, 3 focus on aspects of algorithms, optimization and stochastic simulations
- Part 4 is an in-depth excursion into the world of statistical modeling
- Part 5 has a good dose of probability theory and stochastic processes
- Part 6 delves deeper into the statistical theory

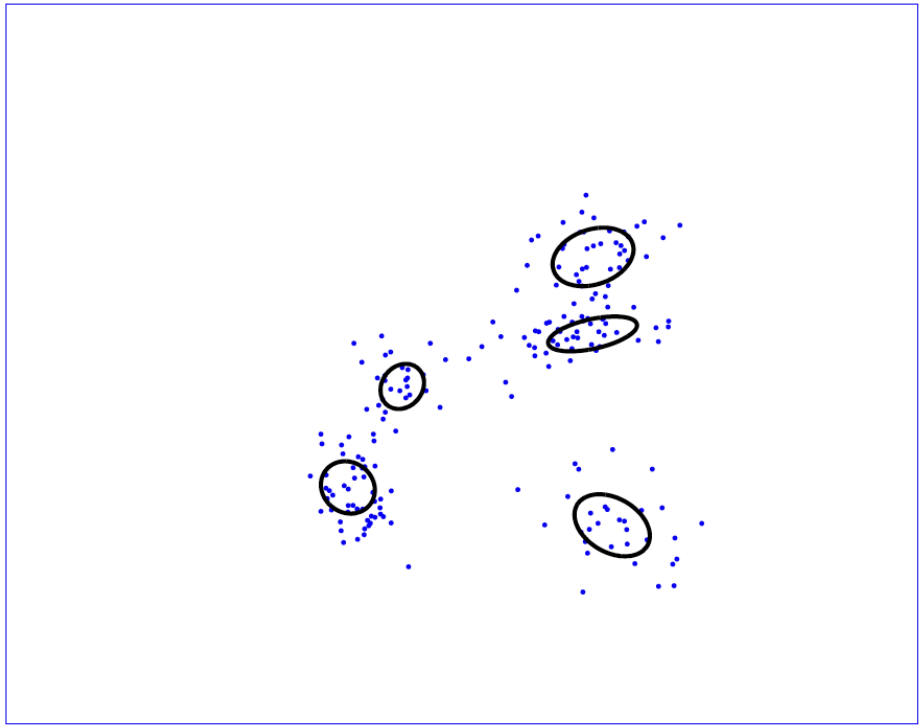
A basic clustering problem

Suppose we have data set $D = \{X_1, \ldots, X_N\}$ in some space.
How do we subdivide these data points to clusters?



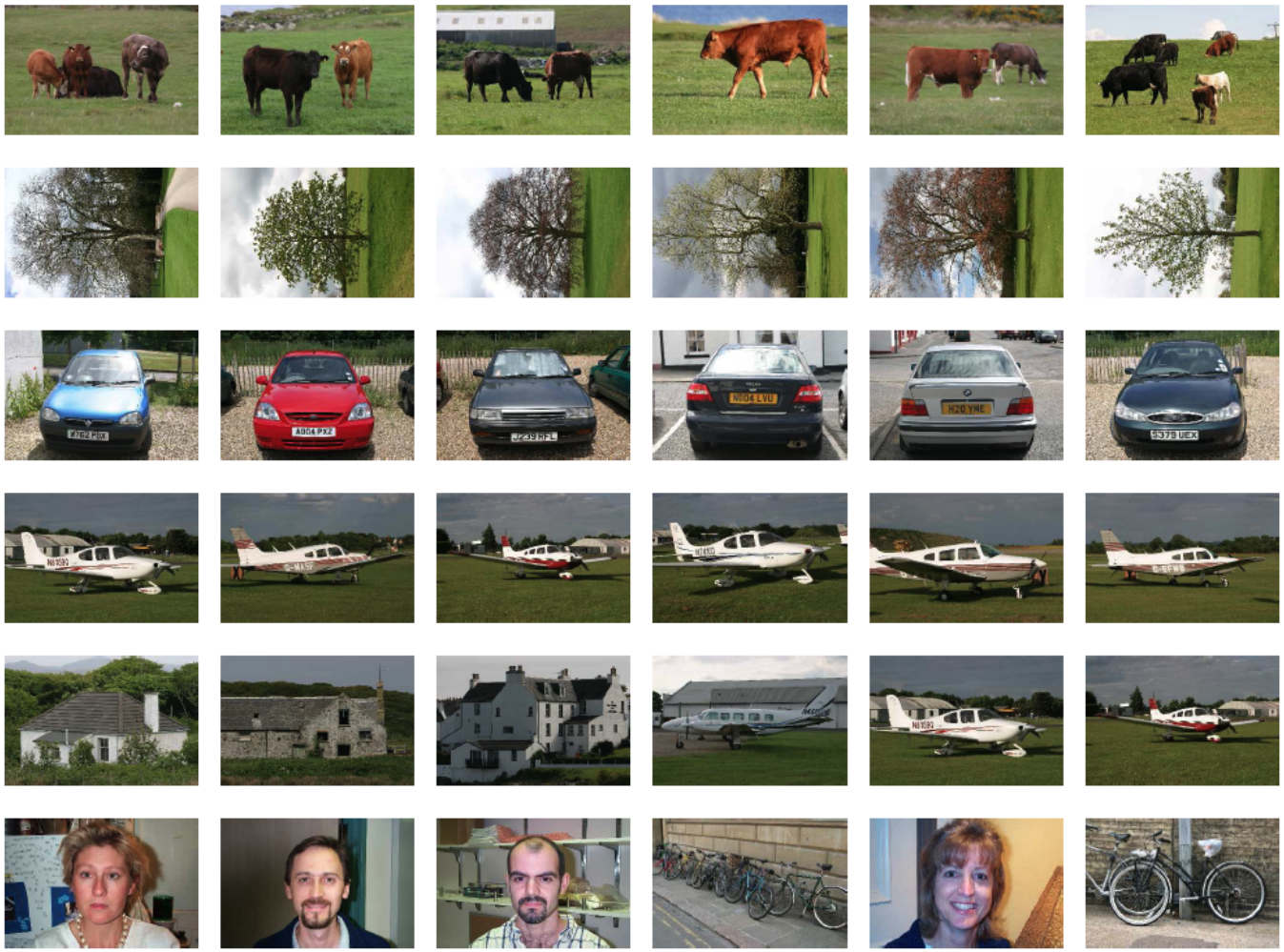
A basic clustering problem

Suppose we have data set $D = \{X_1, \ldots, X_N\}$ in some space.
How do we subdivide these data points to clusters?



Data points may represent scientific measurements, business transactions, text documents, images

Example: Clustering of images



Example: A data point is an article in *Psychology Today*

The Evolution of Moral Models

This is the 13th in the series "Religion and Science: A Beautiful Friendship".

Published on July 18, 2012 by Robert W. Fuller, Ph.D. in Somebodies and Nobodies

When religion has committed itself to a particular science model, it has often been left behind as the public embraced a new model. That's the position in which the Catholic Church found itself in defending Ptolemy's geocentric model of the solar system against the simpler heliocentric model of Copernicus. It's the situation in which supporters of "creationism"—and its offspring, "intelligent design"—find themselves today.

Example: A data point is an article in *Psychology Today*

The Evolution of Moral Models

This is the 13th in the series "Religion and Science: A Beautiful Friendship".

Published on July 18, 2012 by Robert W. Fuller, Ph.D. in Somebodies and Nobodies

When religion has committed itself to a particular science model, it has often been left behind as the public embraced a new model. That's the position in which the Catholic Church found itself in defending Ptolemy's geocentric model of the solar system against the simpler heliocentric model of Copernicus. It's the situation in which supporters of "creationism"—and its offspring, "intelligent design"—find themselves today.

Are Men Shallow?

A wealthy man's high dating standards may have an evolutionary basis.

Published on July 19, 2012 by Vinita Mehta, Ph.D., Ed.M. in Head Games

Does a man with money think he's a more worthy catch? Stereotypes have long depicted rich men as coveted romantic partners. Now, a new study further investigates how much truth there is to this supposed bias — and its evolutionary underpinnings.

Example: A data point is an article in *Psychology Today*

The Evolution of Moral Models

This is the 13th in the series "Religion and Science: A Beautiful Friendship".

Published on July 18, 2012 by Robert W. Fuller, Ph.D. in Somebodies and Nobodies

When religion has committed itself to a particular science model, it has often been left behind as the public embraced a new model. That's the position in which the Catholic Church found itself in defending Ptolemy's geocentric model of the solar system against the simpler heliocentric model of Copernicus. It's the situation in which supporters of "creationism"—and its offspring, "intelligent design"—find themselves today.

Are Men Shallow?

A wealthy man's high dating standards may have an evolutionary basis.

Published on July 19, 2012 by Vinita Mehta, Ph.D., Ed.M. in Head Games

Does a man with money think he's a more worthy catch? Stereotypes have long depicted rich men as coveted romantic partners. Now, a new study further investigates how much truth there is to this supposed bias — and its evolutionary underpinnings.

Learning Disabilities in Adulthood

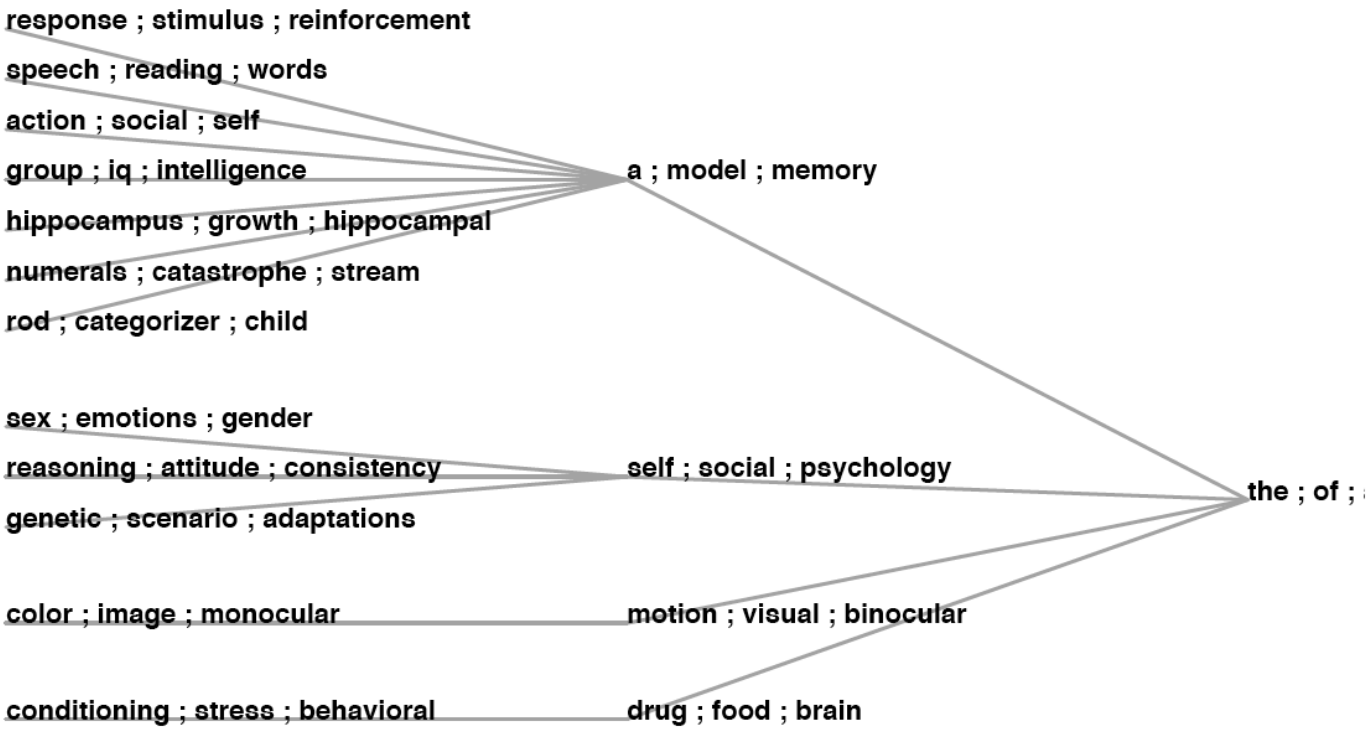
Barriers to proper accomodations

Published on July 20, 2012 by Becky Ready, Ph.D. in Your Quality of Life

Children with learning disabilities grow up to adults with learning disabilities. Learning disabilities are life-long disorders that have tremendous impact on one's educational and occupational achievement. Persons with learning disabilities are bright and multi-talented and capable of great accomplishment with the proper supports and accommodations.

Access to accommodations often relies on a neuropsychological assessment that includes, but is not

Obtain "clusters" organized by certain topics: (Blei et al, 2010)



K-means algorithm

maintain two kinds of variables:

$$\begin{cases} \text{cluster means: } \mu_k, & k = 1, \ldots, K; \\ \text{cluster assigment: } Z_n^k \in \{0, 1\}, & n = 1, \ldots, N. \end{cases}$$

number of clusters K assumed known.

K-means algorithm

maintain two kinds of variables:

$$\begin{cases} \text{cluster means: } \mu_k, & k = 1, \ldots, K; \\ \text{cluster assigment: } Z_n^k \in \{0, 1\}, & n = 1, \ldots, N. \end{cases}$$

number of clusters K assumed known.

Algorithm

1. Initialize $\{\mu_k\}_{k=1}^K$ arbitrarily.
2. Repeat (a) and (b) until convergence:

(a) update for all $n = 1, \ldots, N$:

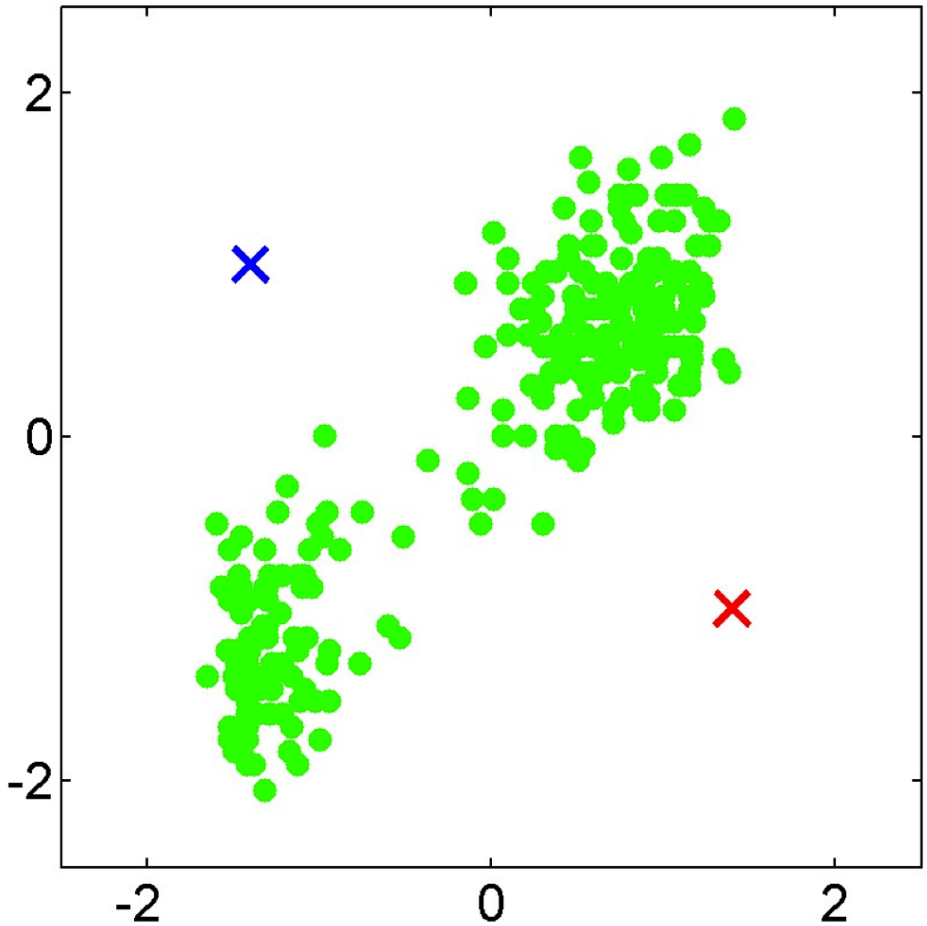
$$Z_n^k := \begin{cases} 1, & \text{if} \quad k = \arg\min_{i \leq K} \|X_n - \mu_i\|, \\ 0, & \text{otherwise.} \end{cases}$$

(b) update for all $k = 1, \ldots, K$:

$$\mu_k = \frac{\sum_{n=1}^N Z_n^k X_n}{\sum_{n=1}^N Z_n^k}.$$

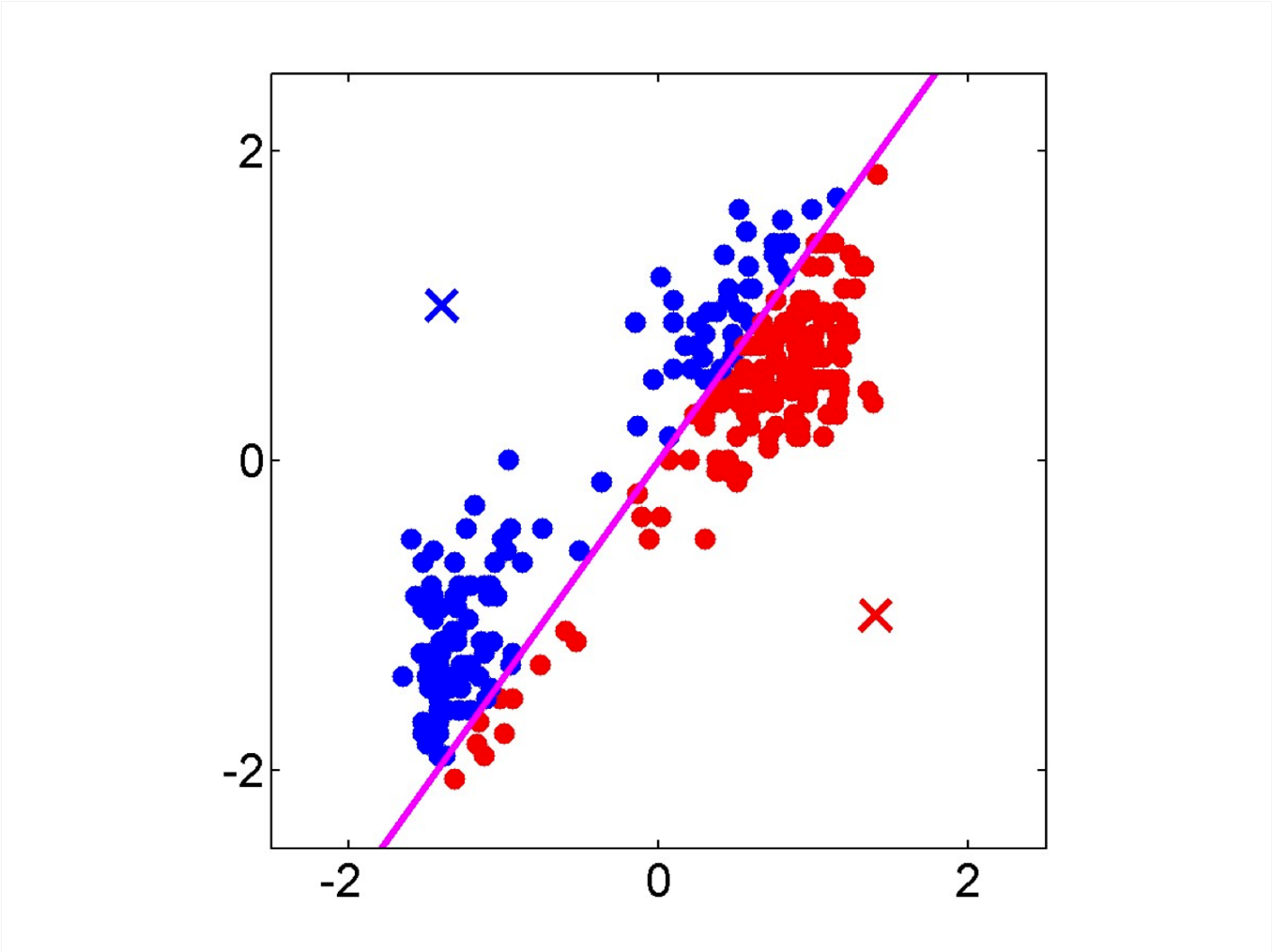
Illustration

$K = 2$



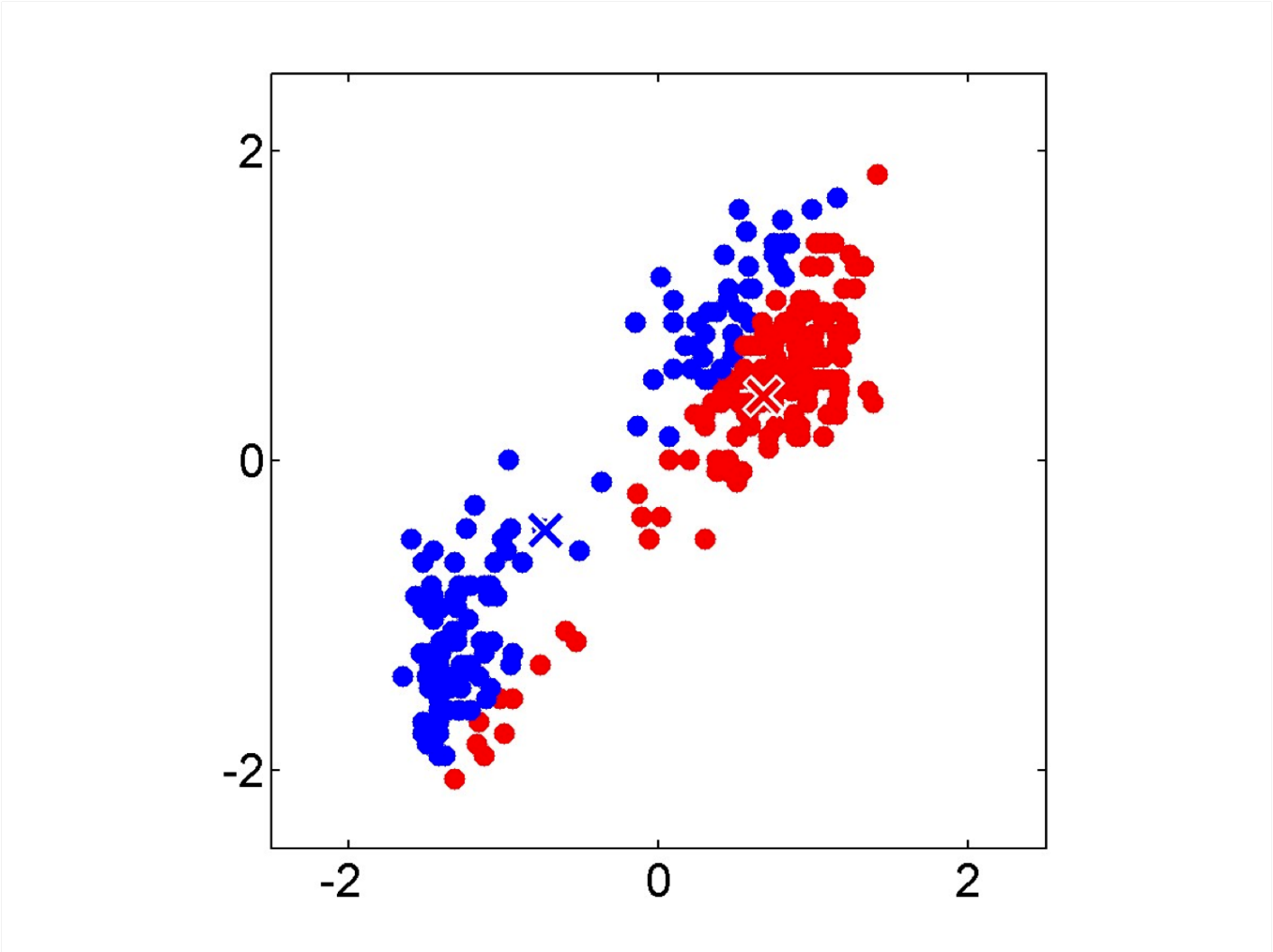
Illustration

$K = 2$



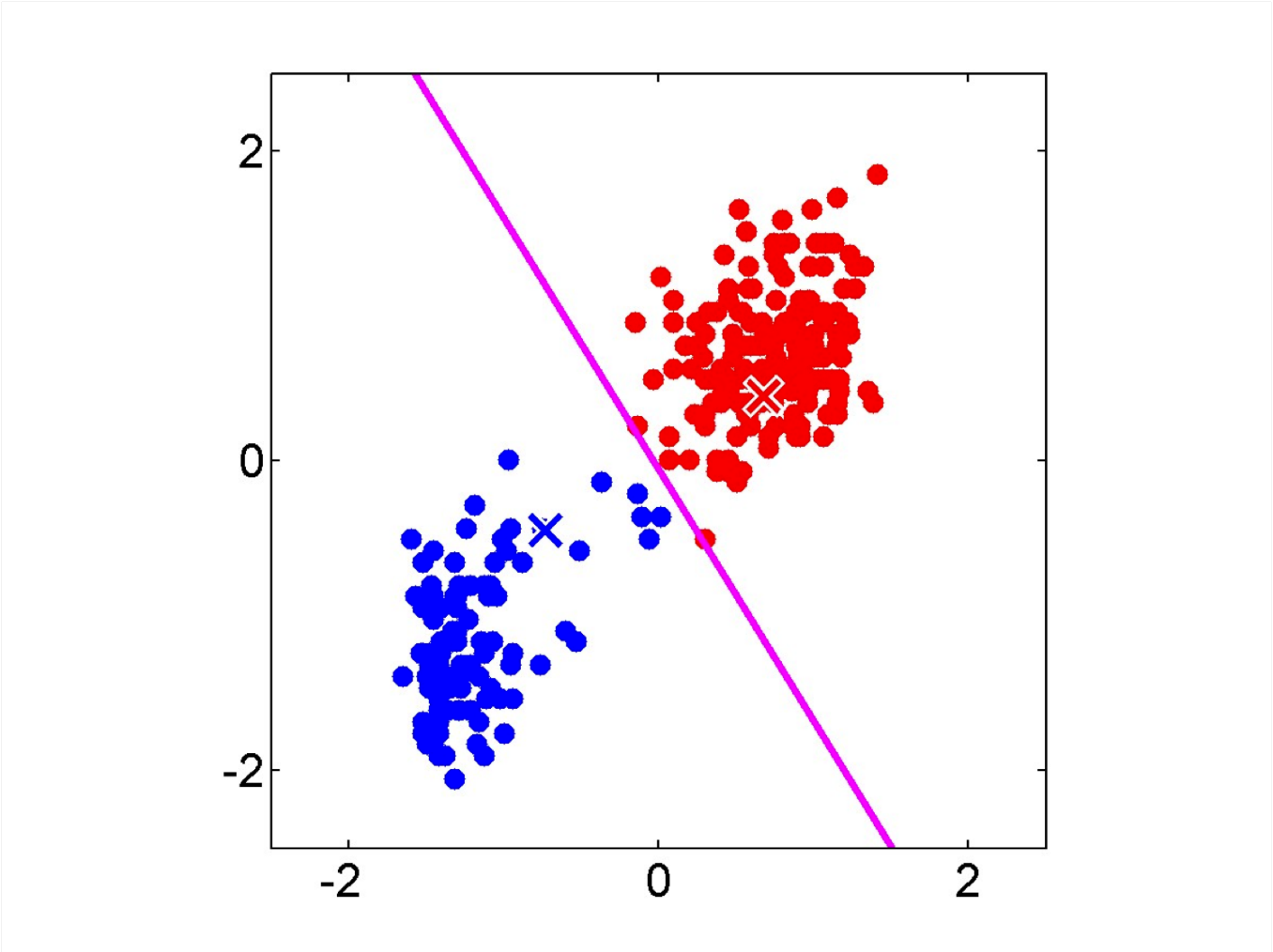
Illustration

$K = 2$



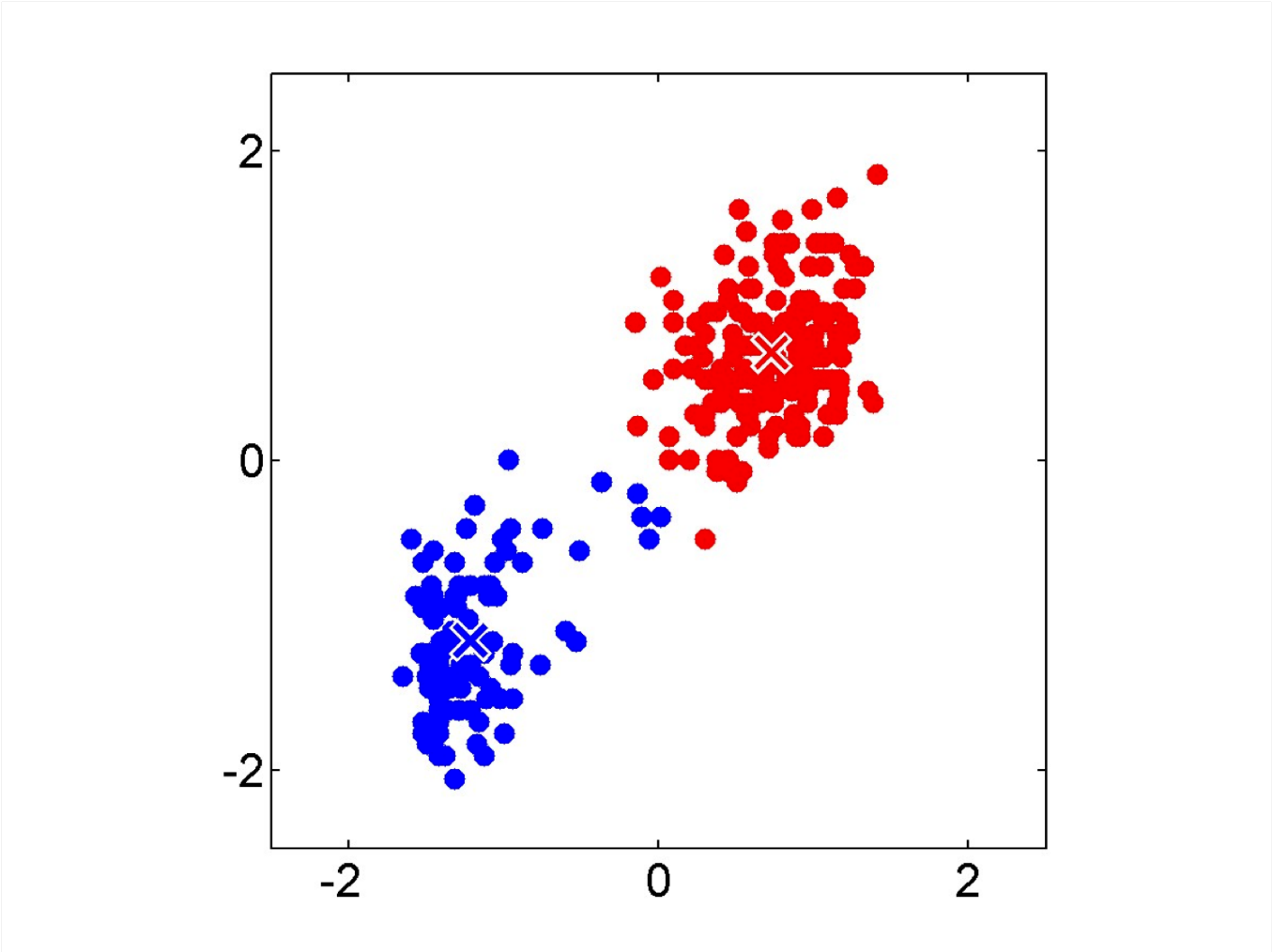
Illustration

$K = 2$



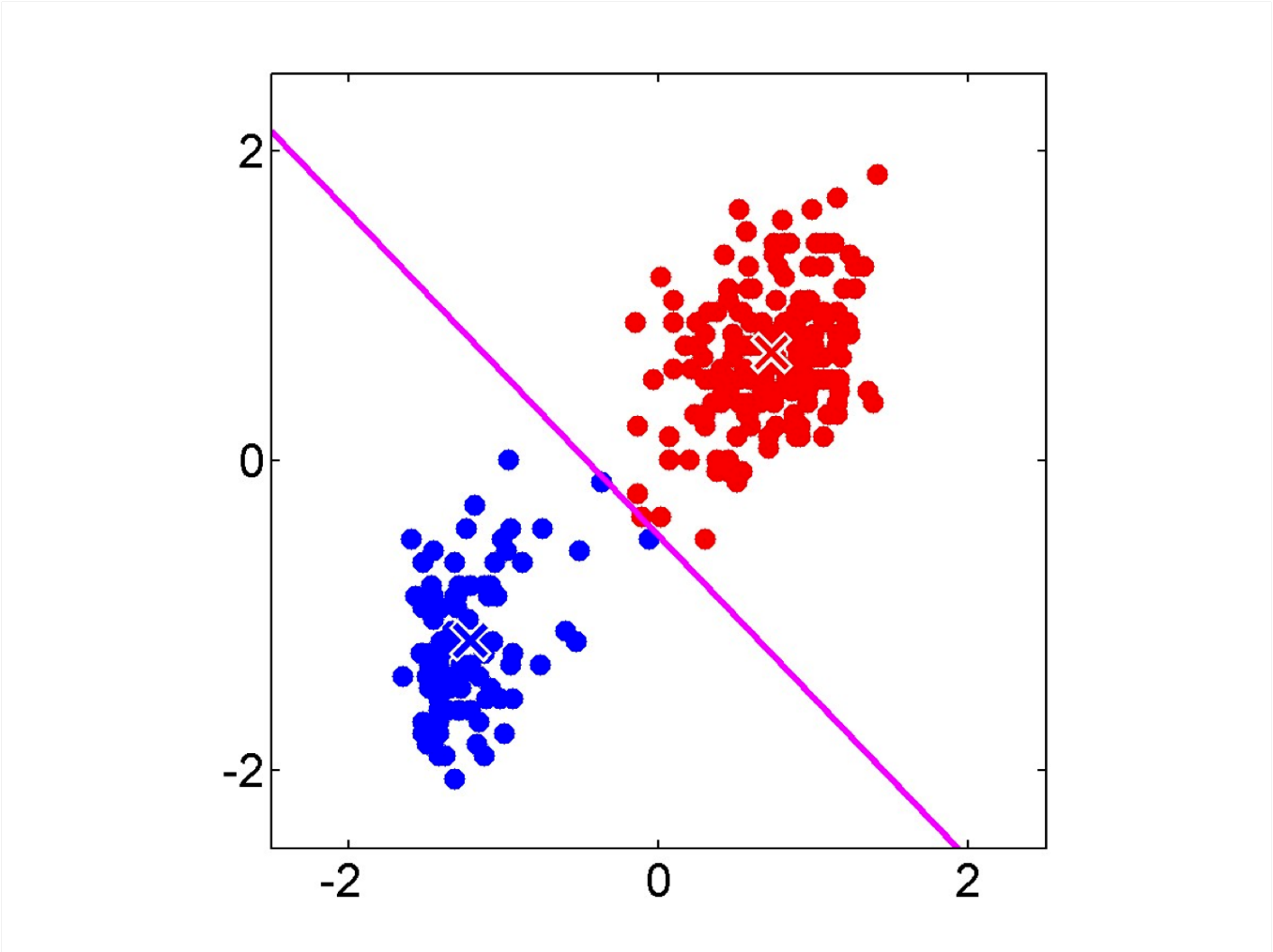
Illustration

$K = 2$



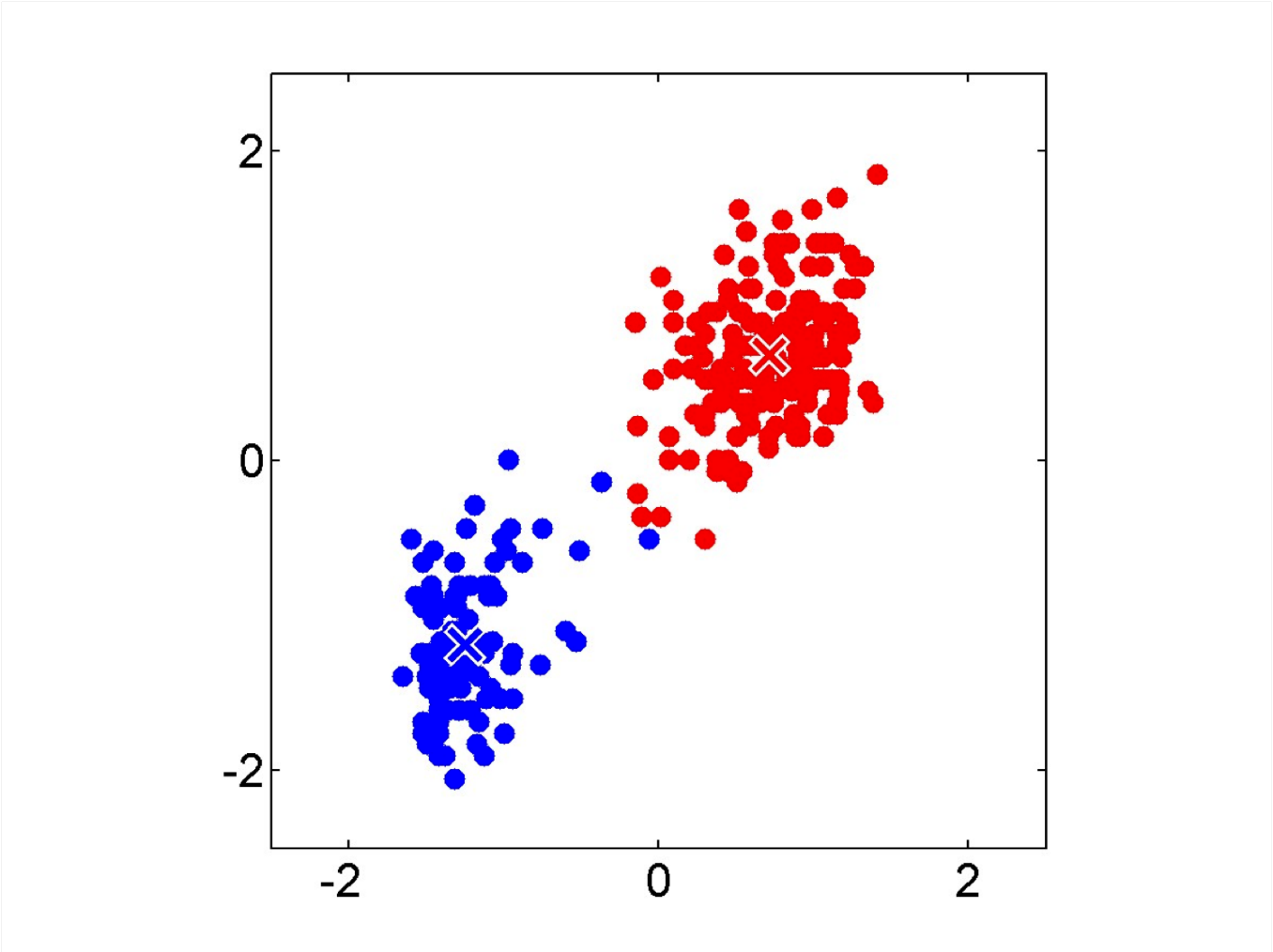
Illustration

$K = 2$



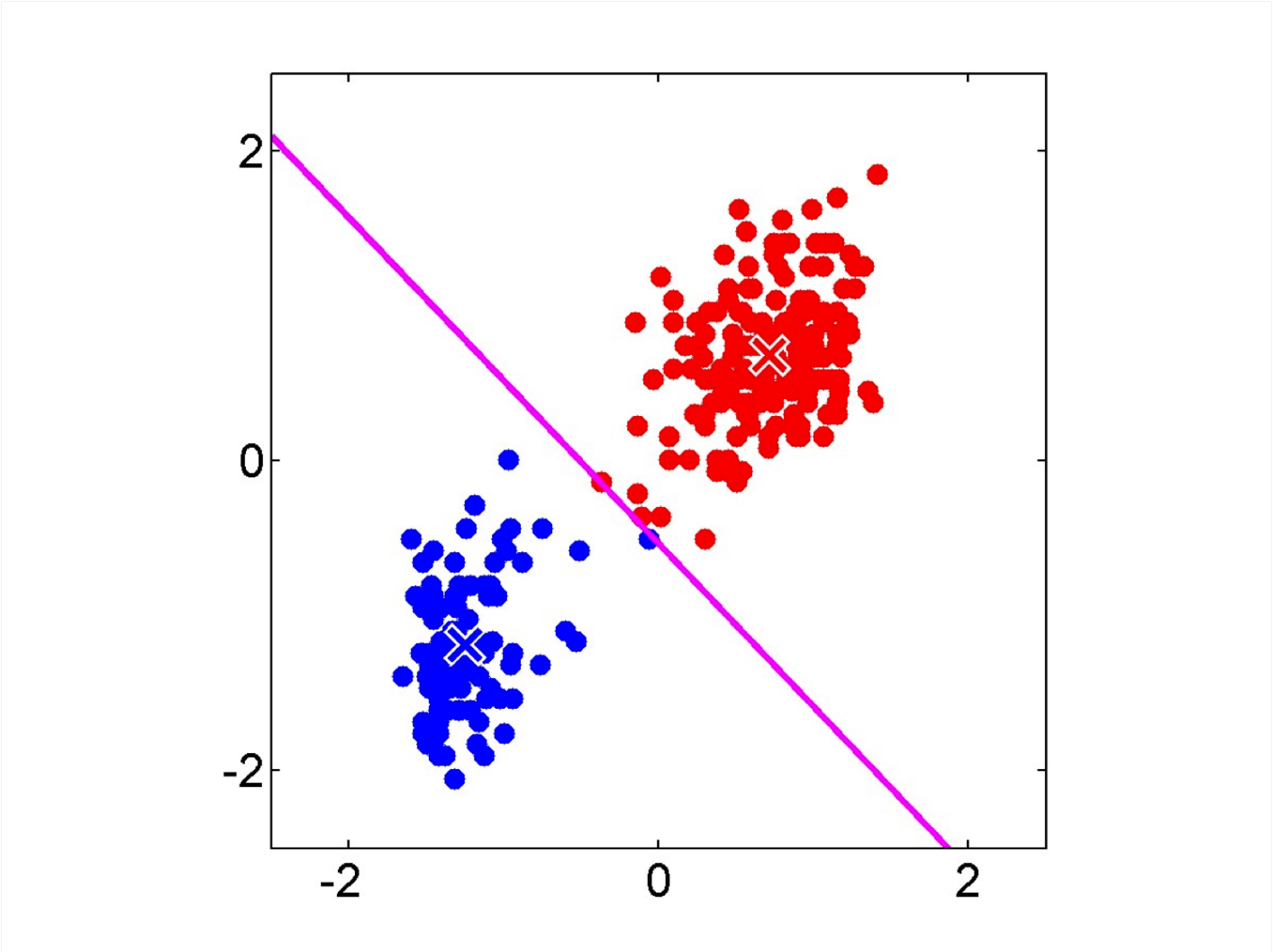
Illustration

$K = 2$



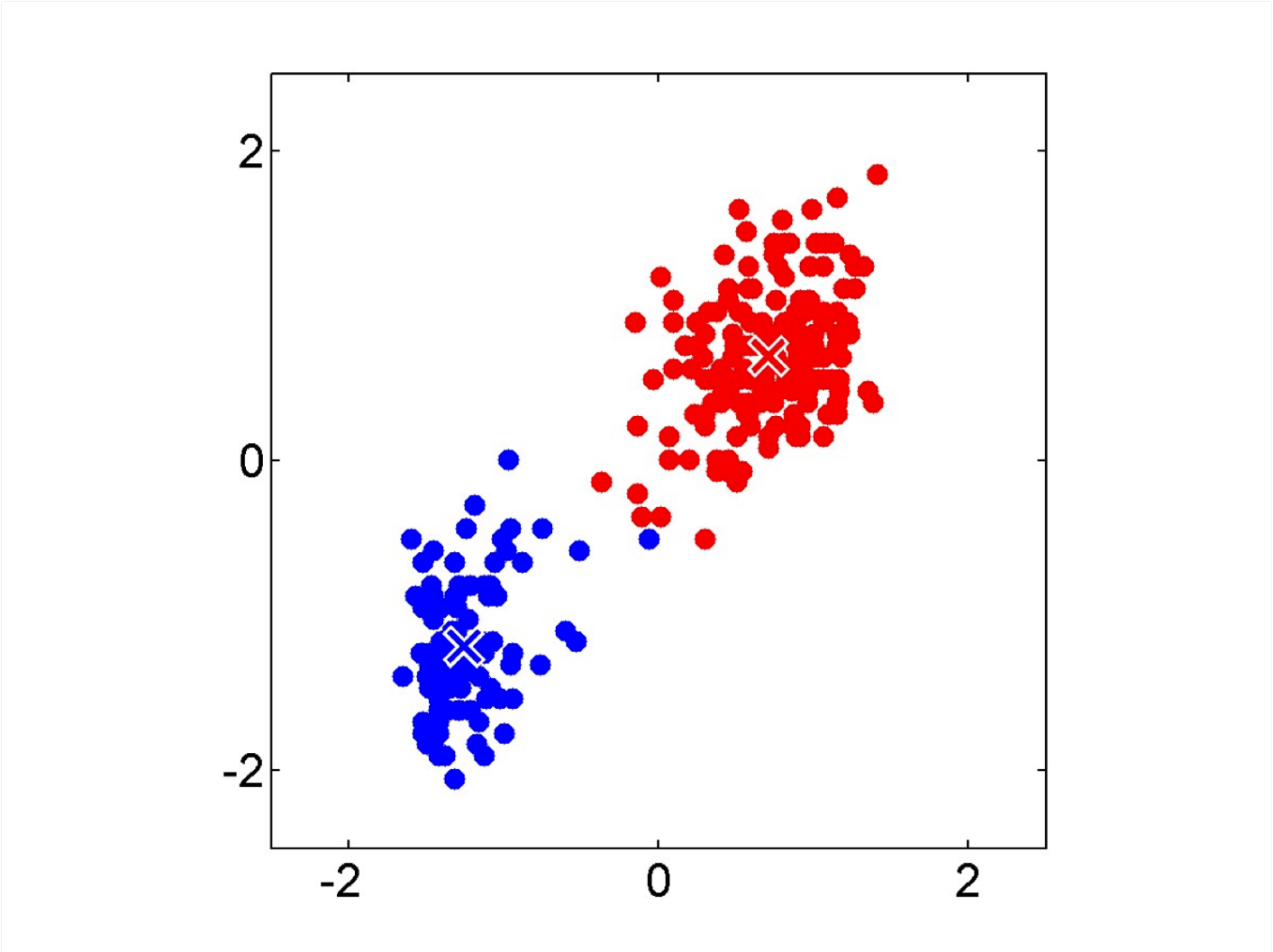
Illustration

$K = 2$



Illustration

$K = 2$



What does all this mean?

What does all this mean?

Operational/mechanical/algebraic meaning: It is easy to show that

K-means algorithm obtains a (locally) optimal solution to optimization problem:

$$\min_{\{\mathbf{Z},\boldsymbol{\mu}\}} \sum_{n=1}^{N} \sum_{k=1}^{K} Z_n^k \left\| X_n - \mu_k \right\|^2 .$$

What does all this mean?

Operational/mechanical/algebraic meaning: It is easy to show that

K-means algorithm obtains a (locally) optimal solution to optimization problem:

$$\min_{\{\mathbf{Z},\boldsymbol{\mu}\}} \sum_{n=1}^{N} \sum_{k=1}^{K} Z_n^k \left\| X_n - \mu_k \right\|^2 .$$

(Much) harder questions:

Why this optimization?
Does this give us the "true" clusters?
What if our assumptions are wrong?
What is the best possible algorithm for learning clusters?

What does all this mean?

Operational/mechanical/algebraic meaning: It is easy to show that

K-means algorithm obtains a (locally) optimal solution to optimization problem:

$$\min_{\{\mathbf{Z},\boldsymbol{\mu}\}} \sum_{n=1}^{N} \sum_{k=1}^{K} Z_n^k \left\| X_n - \mu_k \right\|^2 .$$

(Much) harder questions:

Why this optimization?
Does this give us the "true" clusters?
What if our assumptions are wrong?
What is the best possible algorithm for learning clusters?
How can we be certain of the "truth" from empirical data?

Statistical inference a.k.a. learning:

Learning "true" patterns from data

Statistical inference is concerned with procedures for extracting out pattern/law/value of certain phenomenon from empirical data

- the pattern is parameterized by $\theta \in \Theta$, while data are samples $X_1, \ldots, X_N$

Statistical inference a.k.a. learning:

Learning "true" patterns from data

Statistical inference is concerned with procedures for extracting out pattern/law/value of certain phenomenon from empirical data

- the pattern is parameterized by $\theta \in \Theta$, while data are samples $X_1, \ldots, X_N$

An inference procedure is called an *estimator* in mathematical statistics. It may be formalized as an algorithm, thus a *learning algorithm* in machine learning. Mathematically, it is a mapping from data to an estimate for θ:

$$X_1, \ldots, X_N \mapsto T(X_1, \ldots, X_N) \in \Theta$$

Statistical inference a.k.a. learning:

Learning "true" patterns from data

Statistical inference is concerned with procedures for extracting out pattern/law/value of certain phenomenon from empirical data

- the pattern is parameterized by $\theta \in \Theta$, while data are samples $X_1, \ldots, X_N$

An inference procedure is called an *estimator* in mathematical statistics. It may be formalized as an algorithm, thus a *learning algorithm* in machine learning. Mathematically, it is a mapping from data to an estimate for θ:

$$X_1, \ldots, X_N \mapsto T(X_1, \ldots, X_N) \in \Theta$$

The output of the learning algorithm, $T(X)$, is an estimate of the unknown "truth" θ.

What, How, Why

In clustering problem θ represents the variables used to describe cluster means and cluster assigments $\theta = \{\theta_k; Z_n^k\}$, as well as number of clusters K

What is the "right" inference procedure?

- traditionally studied by statisticians

How to achieve this learning procedure in a computationally efficient manner?

- traditionally studied by computer scientists

Why is the procedure both "right" and "efficient"?

- how much data and how much computations do we need
- these questions drive asymptotic statistics and learning theory

We use clustering as a case study to illustrate these rather fundamental questions in statistics and machine learning, also because of

- vast range of modern applications
- fascinating recent research in algorithms and statistical theory motivated this type of problems
- interest links connecting optimization and numerical analysis to complex statistical modeling, probability theory and stochastic processes

Outline

Roles of probabilistic models

In order to infer about θ from data X, a probabilistic model is needed to provide the "glue" linking θ to X

Roles of probabilistic models

In order to infer about θ from data X, a probabilistic model is needed to provide the "glue" linking θ to X

A model is specified in the form of a probability distribution $P(X|\theta)$

Roles of probabilistic models

In order to infer about θ from data X, a probabilistic model is needed to provide the "glue" linking θ to X

A model is specified in the form of a probability distribution $P(X|\theta)$

Given the same probability model, statisticians may still disagree on how to proceed; there are two broadly categorized approaches to inference: frequentist and Bayes

Roles of probabilistic models

In order to infer about θ from data X, a probabilistic model is needed to provide the "glue" linking θ to X

A model is specified in the form of a probability distribution $P(X|\theta)$

Given the same probability model, statisticians may still disagree on how to proceed; there are two broadly categorized approaches to inference: frequentist and Bayes

- these two viewpoints are consistent mathematically, but can be wildly incompatible in terms of interpretation
- both are interesting and useful in different inferential situations
- roughly speaking, a frequentist method assumes that θ is a non-random unknown parameter, while a Bayesian method always treats θ as a random variable
- frequentists view data X as infinitely available as independent replicates, while a Bayesian does not worry about the data he hasn't seen (he cares more about θ)

Model-based clustering

Assume that data are generated according to a random process:

- pick one of K clusters from a distribution $\pi = (\pi_1, \ldots, \pi_K)$
- generate a data point from a cluster-specific probability distribution

Model-based clustering

Assume that data are generated according to a random process:

- pick one of K clusters from a distribution $\pi = (\pi_1, \ldots, \pi_K)$
- generate a data point from a cluster-specific probability distribution

This yields a mixture model:

$$P(X|\phi) = \sum_{k=1}^{K} \pi_k P(X|\phi_k),$$

where the collection of parameters is $\theta = (\boldsymbol{\pi}, \boldsymbol{\phi})$;

$\boldsymbol{\pi} = \pi_k$'s are mixing probabilities, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_K)$ are the parameters associated with the K clusters.

Model-based clustering

Assume that data are generated according to a random process:

- pick one of K clusters from a distribution $\pi = (\pi_1, \ldots, \pi_K)$
- generate a data point from a cluster-specific probability distribution

This yields a mixture model:

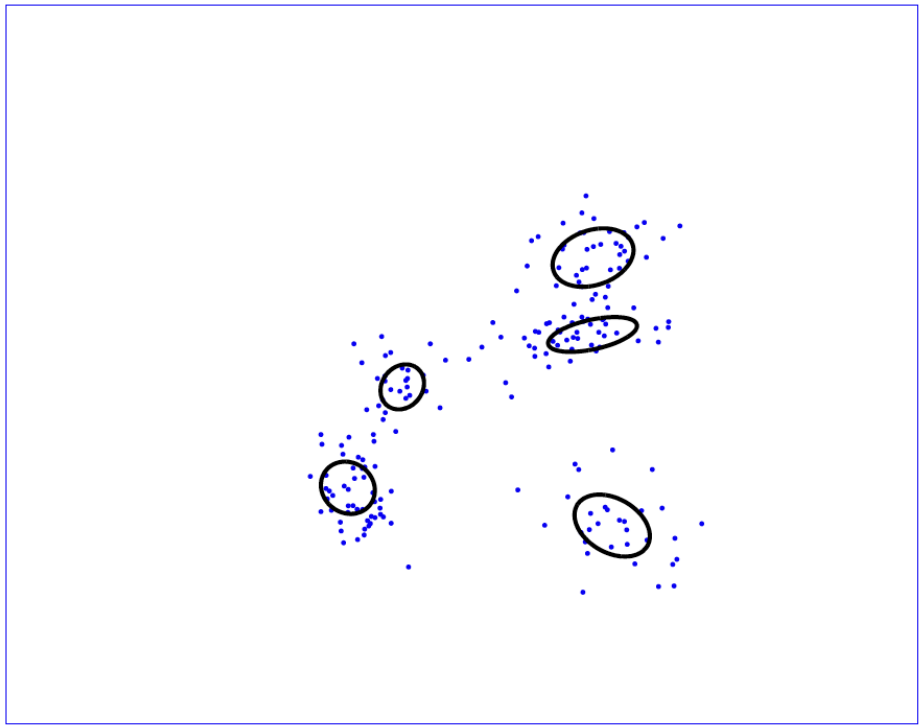
$$P(X|\phi) = \sum_{k=1}^{K} \pi_k P(X|\phi_k),$$

where the collection of parameters is $\theta = (\boldsymbol{\pi}, \boldsymbol{\phi})$;

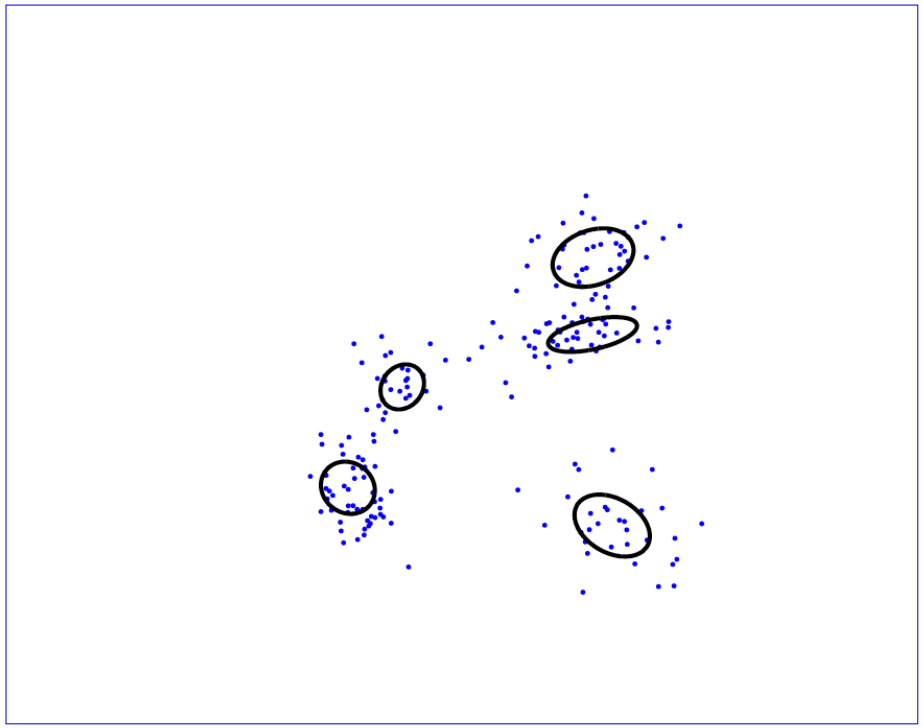
$\boldsymbol{\pi} = \pi_k$'s are mixing probabilities, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_K)$ are the parameters associated with the K clusters.

We still need to specify the cluster-specific distributions $P(X|\phi_k)$ for each k.

Example: for Gaussian mixtures, $\phi_k = (\mu_k, \Sigma_k)$ and $P(X|\phi_k)$ is a Gaussian distribution with mean μ_k and covariance matrix Σ_k.



Example: for Gaussian mixtures, $\phi_k = (\mu_k, \Sigma_k)$ and $P(X|\phi_k)$ is a Gaussian distribution with mean μ_k and covariance matrix Σ_k.



Why Gaussians? What is K, the number of Gaussians subpopulations?

Representation via latent variables

For each data point X, introduce a latent variable $Z \in \{1, \ldots, K\}$ that indicates which subpopulation X is associated with.

Generative model

$$Z \sim \text{Multinomial}(\pi),$$

$$X|Z = k \sim N(\mu_k, \Sigma_k).$$

Representation via latent variables

For each data point X, introduce a latent variable $Z \in \{1, \ldots, K\}$ that indicates which subpopulation X is associated with.

Generative model

$$Z \sim \text{Multinomial}(\pi),$$

$$X|Z = k \sim N(\mu_k, \Sigma_k).$$

Marginalizing out the latent Z, we obtain:

$$\begin{aligned} P(X|\theta) &= \sum_{k=1}^{K} P(X|Z = k, \theta) P(Z = k|\theta) \\ &= \sum_{k=1}^{K} \pi_k N(X|\mu_k, \Sigma_k). \end{aligned}$$

Representation via latent variables

For each data point X, introduce a latent variable $Z \in \{1, \ldots, K\}$ that indicates which subpopulation X is associated with.

Generative model

$$Z \sim \text{Multinomial}(\pi),$$

$$X|Z = k \sim N(\mu_k, \Sigma_k).$$

Marginalizing out the latent Z, we obtain:

$$\begin{aligned} P(X|\theta) &= \sum_{k=1}^{K} P(X|Z = k, \theta) P(Z = k|\theta) \\ &= \sum_{k=1}^{K} \pi_k N(X|\mu_k, \Sigma_k). \end{aligned}$$

Data set $D = (X_1, \ldots, X_N)$ are i.i.d. samples from this generating process.

Equivalent representation via mixing measure

Define the discrete probability measure

$$G = \sum_{k=1}^{K} \pi_k \delta_{\phi_k}$$

where δ_{ϕ_k} is an atom at ϕ_k

The mixture model is define as follows:

$$\begin{aligned} \theta_n := (\mu_n, \Sigma_n) &\sim G \\ X_n|\theta_n &\sim P(\cdot|\theta_n) \end{aligned}$$

Each θ_n is equal to the mean/variance of the cluster associated with data X_n. G is called a mixing measure.

Inference

Setup: Given data $D = \{X_1, ..., X_N\}$ assumed iid from a mixture model. $\{Z_1, \ldots, Z_N\}$ are associated (latent) cluster assignment variables.

Goal:

- Estimate parameters $\theta = (\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$.
- Estimate cluster assigment via calculation of conditional probability of cluster labels $P(Z_n|X_n)$

Maximum likelihood estimation

A frequentist method for estimation going back to Fisher (see, e.g., van der Vaart, 2000)

Maximum likelihood estimation

A frequentist method for estimation going back to Fisher (see, e.g., van der Vaart, 2000)

Likelihood function is a function of parameter:

$$L(\theta|\mathrm{Data}) \quad = \quad P(\mathrm{Data}|\theta) = \prod_{n=1}^{N} P(X_n|\theta).$$

Maximum likelihood estimation

A frequentist method for estimation going back to Fisher (see, e.g., van der Vaart, 2000)

Likelihood function is a function of parameter:

$$L(\theta|\mathrm{Data}) \quad = \quad P(\mathrm{Data}|\theta) = \prod_{n=1}^{N} P(X_n|\theta).$$

MLE gives the estimate:

$$\begin{aligned} \hat{\theta}_N \quad &:= \quad \arg\max_{\theta} L(\theta|\mathrm{Data}) \\ &= \quad \arg\max_{\theta} \sum_{n=1}^{N} \log P(X_n|\theta). \end{aligned}$$

Maximum likelihood estimation

A frequentist method for estimation going back to Fisher (see, e.g., van der Vaart, 2000)

Likelihood function is a function of parameter:

$$L(\theta|\mathrm{Data}) \quad = \quad P(\mathrm{Data}|\theta) = \prod_{n=1}^{N} P(X_n|\theta).$$

MLE gives the estimate:

$$\begin{aligned} \hat{\theta}_N \quad &:= \quad \arg\max_{\theta} L(\theta|\mathrm{Data}) \\ &= \quad \arg\max_{\theta} \sum_{n=1}^{N} \log P(X_n|\theta). \end{aligned}$$

A fundamental theorem in asymptotic statistics

Under regularity conditions, the maximum likelihood estimator is consistent and asymptotically efficient.

I.e., assuming that $X_1, \ldots, X_N \overset{i.i.d}{\sim} P(X|\theta^*)$, then $\theta_N \to \theta^*$ in probability (or almost surely), as $N \to \infty$.

MLE for Gaussian mixtures (cont)

Recall the mixture density:

$$P(X|\theta) = \sum_{k=1}^{K} \pi_k N(X|\mu_k, \Sigma_k)$$

$$\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma} := \arg\max \sum_{n=1}^{N} \log\{\sum_{k=1}^{K} \pi_k N(X_n|\mu_k, \Sigma_k)\}.$$

It is possible but cubersome to solve this optimization directly, a more practically convenient approach is via the EM (Expectation-Maximization) algorithm.

EM algorithm for Gaussian mixtures

Intuition

- For each data point X_n, if Z_n is known for all $n = 1, \ldots, N$, it would be easy to estimate the "cluster" means and covariances μ_k, Σ_k.
- But Z_n's are hidden — perhaps, we can "fill-in" the latent variable Z_n by an estimate, such as the conditional expectation $\mathbb{E}(Z_n|X_n)$. This can be done if all parameters are known.
- Classic "chicken-and-egg" situation!

EM algorithm for Gaussian mixture

1. Initialize $\{\mu_k, \Sigma_k, \pi_k\}_{k=1}^K$ arbitrarily.
2. Repeat (a) and (b) until convergence:

(a) For $k = 1, \ldots, K$, $n = 1, \ldots, N$, calculate conditional expectation of labels:

$$\begin{aligned} \tau_n^k &\longleftarrow P(Z = k | X_n) \\ &= \frac{P(X_n|Z=k)P(Z=k)}{\sum_{k=1}^K P(X_n|Z=k)P(Z=k)} \\ &= \frac{\pi_k N(X_n|\mu_k, \Sigma_k)}{\sum_{k=1}^K \pi_k N(X_n|\mu_k, \Sigma_k)}. \end{aligned}$$

(b) Update for $k = 1, \ldots, K$:

$$\begin{aligned} \mu_k &\longleftarrow \frac{\sum_{n=1}^N \tau_n^k x_n}{\sum_{n=1}^N \tau_n^k}, \\ \Sigma_k &\longleftarrow \frac{\sum_{n=1}^N \tau_n^k (x_n - \mu_k)(x_n - \mu_k)^T}{\sum_{n=1}^N \tau_n^k}, \\ \pi_k &\longleftarrow \frac{1}{N} \sum_{n=1}^N \tau_n^k. \end{aligned}$$

This algorithm is a "soft version" that generalizes the K-means algorithm!

Illustration of EM algorithm

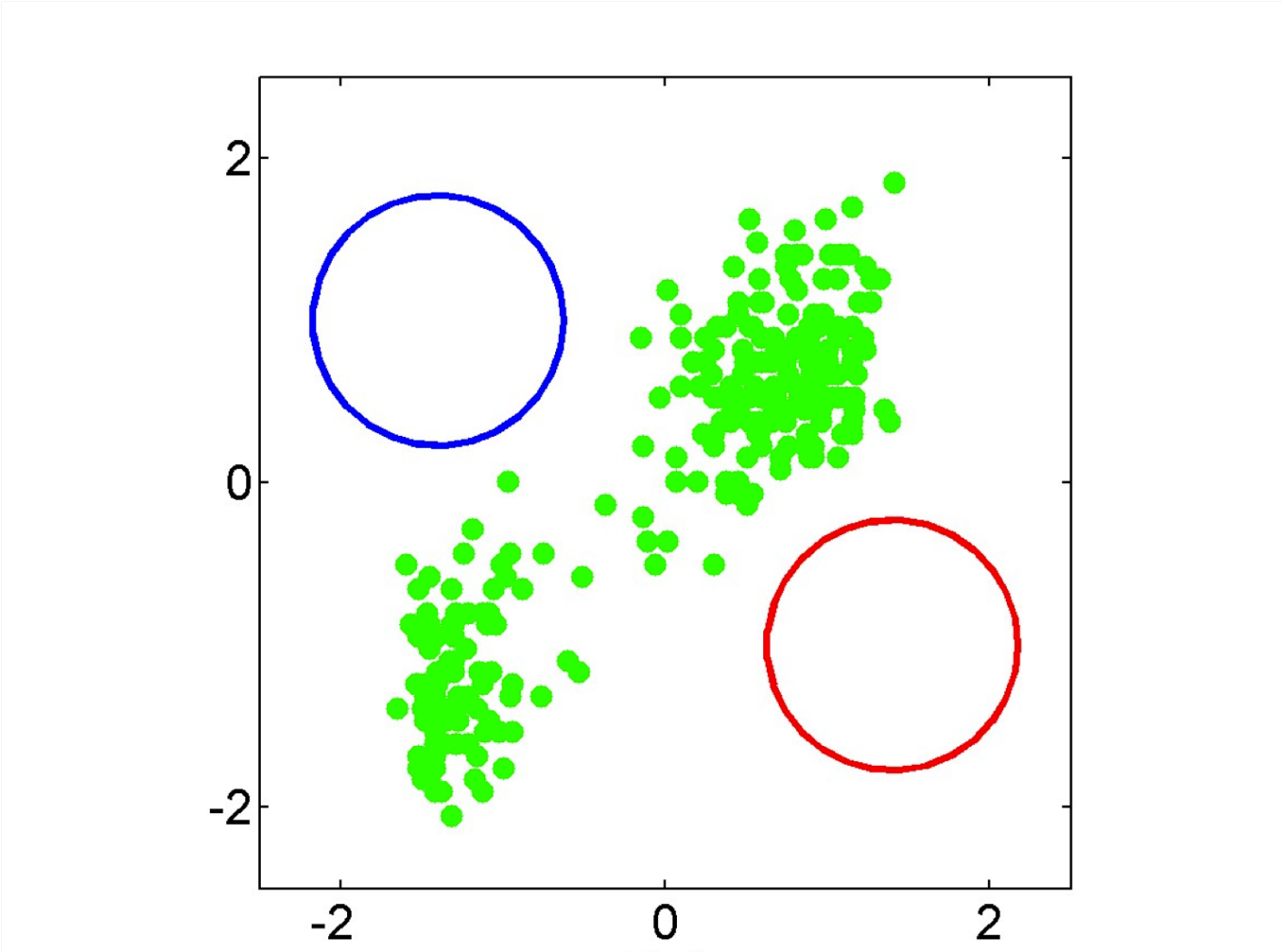


Illustration of EM algorithm

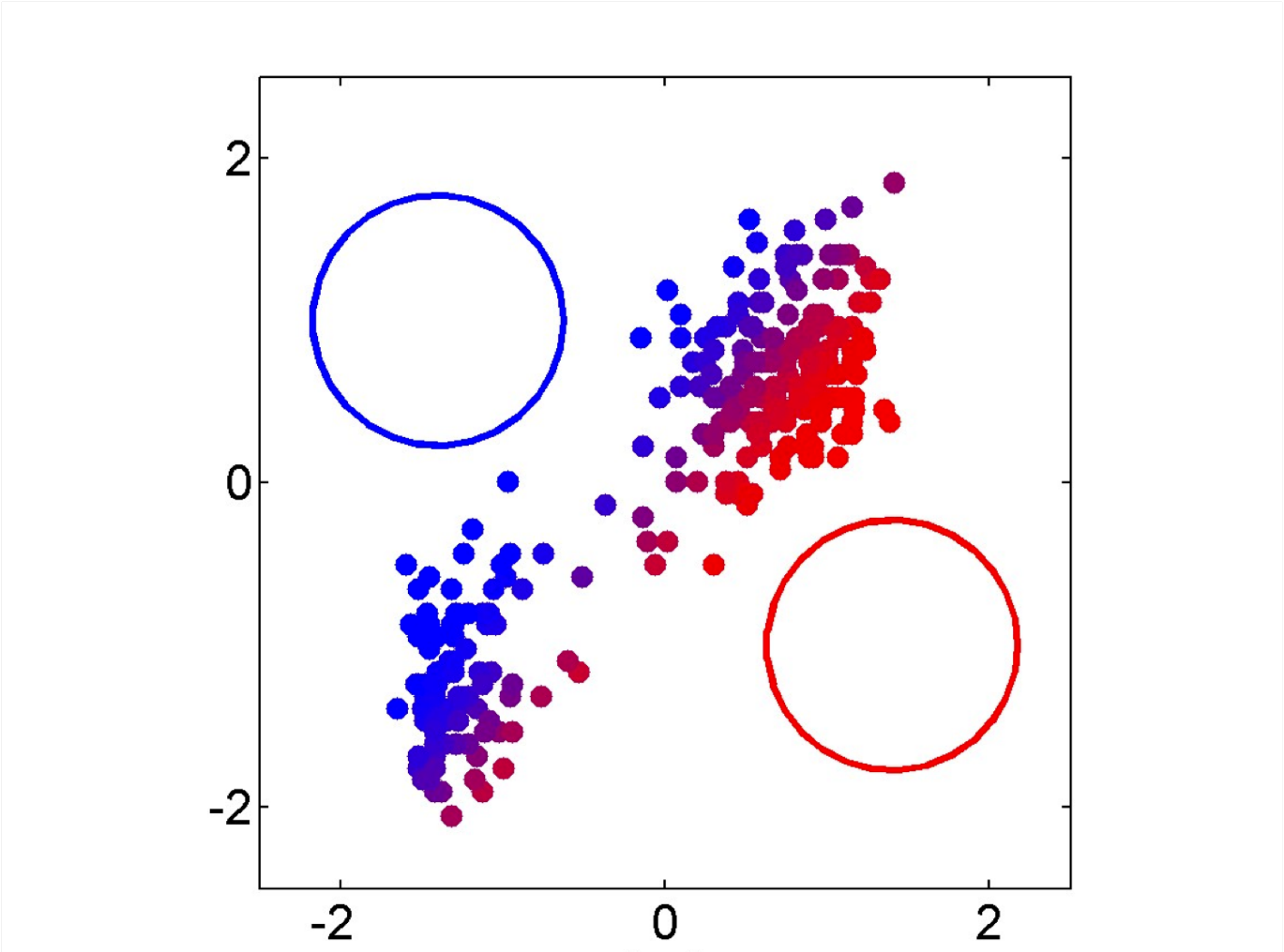


Illustration of EM algorithm

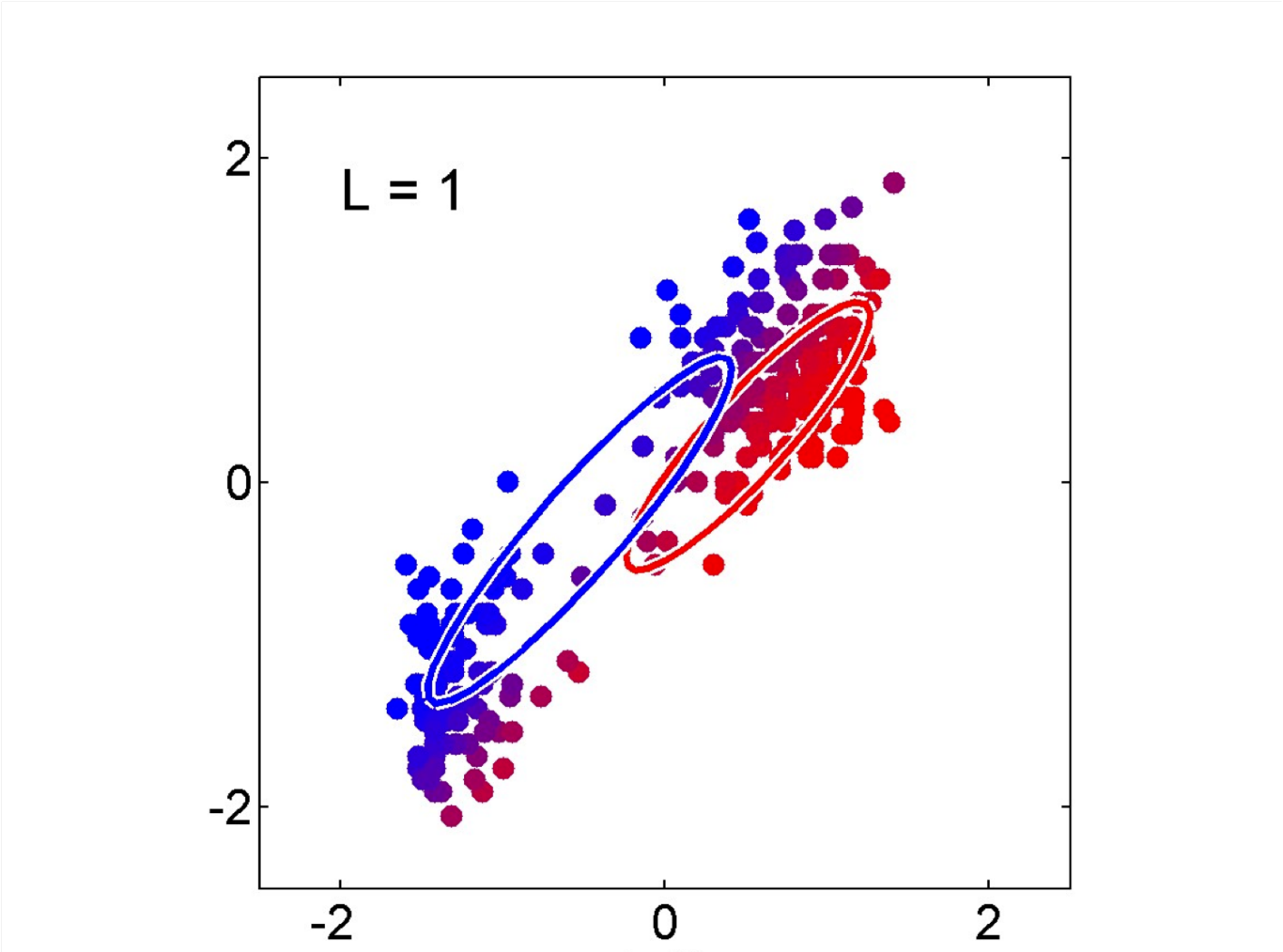
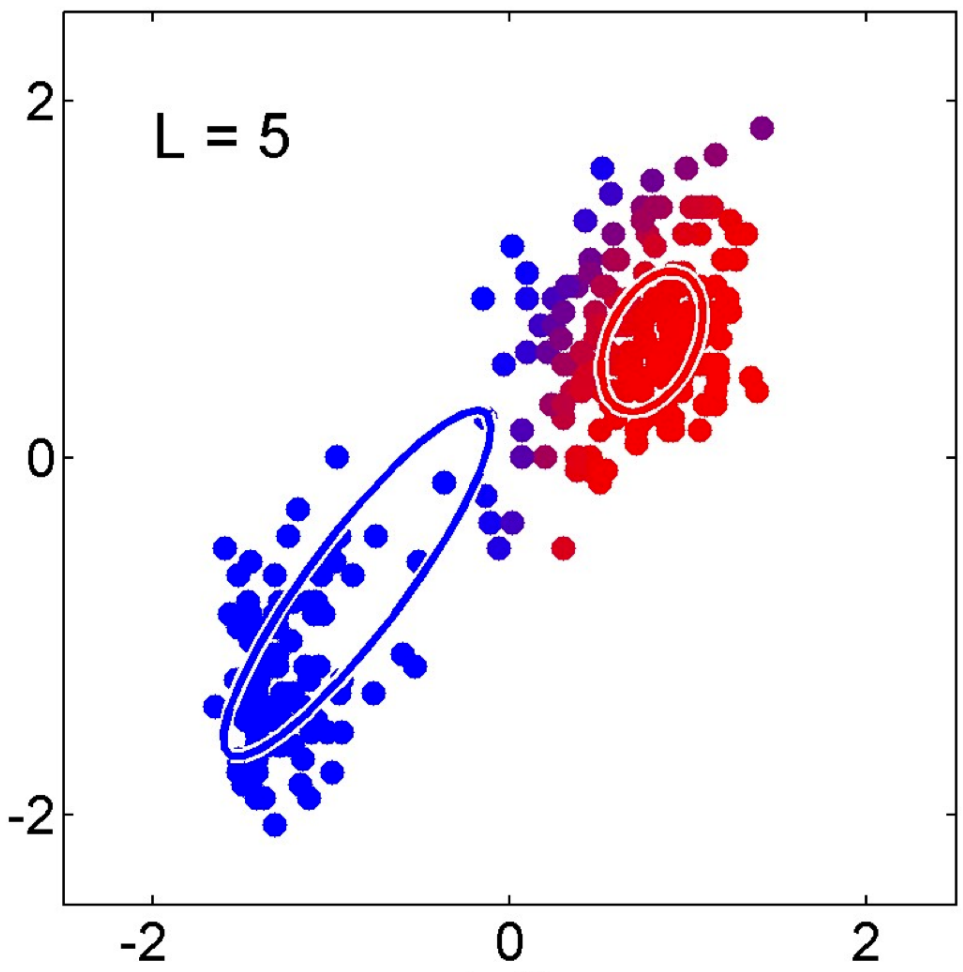


Illustration of EM algorithm



What does this algorithm really do?

We will show that this algorithm ultimately obtains a local optimum of the likelihood function. I.e., it is indeed an MLE method.

What does this algorithm really do?

We will show that this algorithm ultimately obtains a local optimum of the likelihood function. I.e., it is indeed an MLE method.
Suppose that we have "full data (complete data)" $D_c = \{(Z_n, X_n)_{n=1}^N\}$.
Then we can calculate the complete log-likelihood function:

$$\begin{aligned} l_c(\theta|D_c) &= \log P(D_c|\theta) \\ &= \sum_{n=1}^{N} \log P(X_n, Z_n|\theta) \\ &= \sum_{n=1}^{N} \log\left\{ \prod_{k=1}^{K} (\pi_k N(X_n|\mu_k, \Sigma_k))^{Z_n^k} \right\} \\ &= \sum_{n=1}^{N} \sum_{k=1}^{K} Z_n^k \log\{\pi_k N(X_n|\mu_k, \Sigma_k)\} \\ &= \sum_{n=1}^{N} \sum_{k=1}^{K} Z_n^k (\log \pi_k + \log N(X_n|\mu_k, \Sigma_k)). \end{aligned}$$

To estimate the parameters, we may wish to optimize the complete log-likelihood if we actually have full data $(Z_n, X_n)_{n=1}^N$.
Since Z_n's are actually latent, we settle for conditional expectation. In fact,

Easy exercise

The updating step (b) of the EM algorithm described earlier optimizes the conditional expectation of the complete log-likelihood:

$$\theta := \arg\max \mathbb{E}[l_c(\theta|D_c)|X_1, \ldots, X_N)],$$

where $\mathbb{E}[l_c(\theta|D_c)|X_1, \ldots, X_N] = \sum_{n=1}^N \sum_{k=1}^K \tau_n^k (\log \pi_k + \log N(X_n|\mu_k, \Sigma_k))$.

(Proof by taking gradient with respect to parameters and setting to 0).
Compare this to optimizing the original likelihood function:

$$l(\theta|D) = \sum_{n=1}^N \log \left\{ \sum_{i=1}^K \pi_k N(X_n|\mu_k, \Sigma_k) \right\}.$$

Summary

Rewrite the EM algorithm as maximization of expected complete log-likelihood:

Summary

Rewrite the EM algorithm as maximization of expected complete log-likelihood:

EM algorithm for Gaussian mixtures

1. Initialize randomly $\theta = \{\mu_k, \Sigma_k, \pi_k\}_{k=1}^K$.
2. Repeat (a) and (b) until convergence:

(a) "E-step": given current estimate of θ, compute $E[l_c(\theta|D_c)|D]$.
(b) "M-step": update θ by maximizing $E[l_c(\theta|D_c)|D]$;

Summary

Rewrite the EM algorithm as maximization of expected complete log-likelihood:

EM algorithm for Gaussian mixtures

1. Initialize randomly $\theta = \{\mu_k, \Sigma_k, \pi_k\}_{k=1}^K$.
2. Repeat (a) and (b) until convergence:

(a) "E-step": given current estimate of θ, compute $E[l_c(\theta|D_c)|D]$.
(b) "M-step": update θ by maximizing $E[l_c(\theta|D_c)|D]$;

It remains to show that maximizing the expected complete log-likelihood is equivalent to maximizing the log-likelihood function ...

EM algorithm for latent variable models

A model with latent variables abstractly defined as follows:

$$Z \sim P(\cdot|\theta)$$
$$X|Z \sim P(\cdot|Z,\theta).$$

EM algorithm for latent variable models

A model with latent variables abstractly defined as follows:

$$Z \sim P(\cdot|\theta)$$
$$X|Z \sim P(\cdot|Z,\theta).$$

This type of model includes

- mixture models, hierarchical models (will see later in this lecture)
- hidden Markov models, Kalman filters, etc

EM algorithm for latent variable models

A model with latent variables abstractly defined as follows:

$$Z \sim P(\cdot|\theta)$$
$$X|Z \sim P(\cdot|Z,\theta).$$

This type of model includes

- mixture models, hierarchical models (will see later in this lecture)
- hidden Markov models, Kalman filters, etc

Recall the log-likelihood function for observed data:

$$l(\theta|D) = \log p(D|\theta) = \sum_{n=1}^{N} \log p(X_n|\theta)$$

and the log-likelihood function for the complete data:

$$l_C(\theta|D_C) = \log p(D_C|\theta) = \sum_{n=1}^{N} \log p(X_n, Z_n|\theta).$$

EM algorithm maximizes the likelihood

EM algorithm for latent variable models

1. Initialize randomly θ.
2. Repeat (a) and (b) until convergence:

(a) "E-step": given current estimate of θ, compute $E[l_c(\theta|D_c)|D]$.
(b) "M-step": update θ by maximizing $E[l_c(\theta|D_c)|D]$;

EM algorithm maximizes the likelihood

EM algorithm for latent variable models

1. Initialize randomly θ.
2. Repeat (a) and (b) until convergence:

(a) "E-step": given current estimate of θ, compute $E[l_c(\theta|D_c)|D]$.
(b) "M-step": update θ by maximizing $E[l_c(\theta|D_c)|D]$;

Theorem

The EM algorithm is a coordinatewise hill-climbing algorithm with respect to the likelihood function.

For proof, see hand-written notes.

Outline

Bayesian estimation

In a Bayesian approach, parameters $\theta = (\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ are assumed to be random

Bayesian estimation

In a Bayesian approach, parameters $\theta = (\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ are assumed to be random

There needs to be a *prior* distribution for θ
Consequentially we obtain a Bayesian mixture model

Bayesian estimation

In a Bayesian approach, parameters $\theta = (\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ are assumed to be random

There needs to be a *prior* distribution for θ
Consequentially we obtain a Bayesian mixture model

Inference boils down to calculation of posterior probability:

Bayes' Rule

$$\begin{aligned} P(\theta|\text{Data}) &\equiv P(\theta|X) \\ &= \frac{P(\theta)P(X|\theta)}{\int P(\theta)P(X|\theta)d\theta} \\ \text{posterior} &\propto \text{prior} \times \text{likelihood} \end{aligned}$$

Prior distributions

$$\begin{aligned} \boldsymbol{\pi} = (\pi_1, \ldots, \pi_K) &\sim \mathrm{Dir}(\boldsymbol{\alpha}) \\ \mu_1, \ldots, \mu_K &\sim \mathrm{N}(\mathbf{0}, \mathbf{I}) \\ \Sigma_1, \ldots, \Sigma_K &\sim \mathrm{IW}(\Psi, m). \end{aligned}$$

Prior distributions

$$\begin{aligned} \boldsymbol{\pi} = (\pi_1, \ldots, \pi_K) &\sim \mathrm{Dir}(\boldsymbol{\alpha}) \\ \mu_1, \ldots, \mu_K &\sim \mathrm{N}(\mathbf{0}, \mathbf{I}) \\ \Sigma_1, \ldots, \Sigma_K &\sim \mathrm{IW}(\Psi, m). \end{aligned}$$

A million dollar question: how to choose prior distributions?
(such as Dirichlet, Normal, Inverse-Wishart, . . .)

Prior distributions

$$\begin{aligned} \boldsymbol{\pi} = (\pi_1, \ldots, \pi_K) &\sim \mathrm{Dir}(\boldsymbol{\alpha}) \\ \mu_1, \ldots, \mu_K &\sim \mathrm{N}(\mathbf{0}, \mathbf{I}) \\ \Sigma_1, \ldots, \Sigma_K &\sim \mathrm{IW}(\Psi, m). \end{aligned}$$

A million dollar question: how to choose prior distributions?
(such as Dirichlet, Normal, Inverse-Wishart, . . .)

. . . the pure Bayesian viewpoint
. . . the pragmatic viewpoint: computational convenience via conjugacy
. . . the theoretical viewpoint: posterior asymptotics

Dirichlet distribution

Let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ be a point in the $(K-1)$-simplex

- i.e., $0 \leq \pi_k \leq 1$, and $\sum_{k=1}^{K} = 1$

Dirichlet distribution

Let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ be a point in the $(K-1)$-simplex

- i.e., $0 \leq \pi_k \leq 1$, and $\sum_{k=1}^{K} = 1$

Let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ be a set of parameters, where $\alpha_k > 0$

Dirichlet distribution

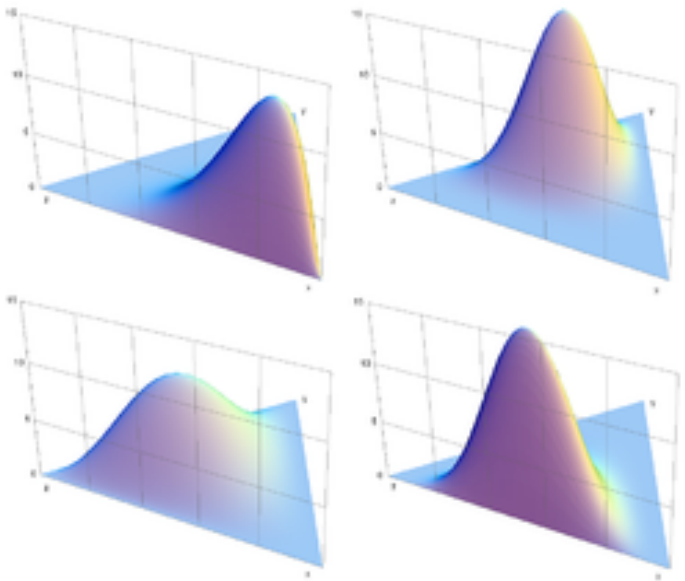
Let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ be a point in the $(K-1)$-simplex

- i.e., $0 \leq \pi_k \leq 1$, and $\sum_{k=1}^K = 1$

Let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ be a set of parameters, where $\alpha_k > 0$
The Dirichlet density is defined as

$$P(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \pi_1^{\alpha_1 - 1} \ldots \pi_K^{\alpha_K - 1}.$$

- $\mathbb{E}\pi_k = \alpha_k / (\sum_{k=1}^K \alpha_k)$



Multinomial-Dirichlet conjugacy

Let $\boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\alpha})$.
Let $Z \sim \mathrm{Multinomial}(\boldsymbol{\pi})$, i.e. $P(Z = k|\boldsymbol{\pi}) = \pi_k$ for $k = 1, \ldots, K$.
Write Z as indicator vector $Z = (Z^1 \ldots Z^k)$.

Multinomial-Dirichlet conjugacy

Let $\boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\alpha})$.
Let $Z \sim \mathrm{Multinomial}(\boldsymbol{\pi})$, i.e. $P(Z = k|\boldsymbol{\pi}) = \pi_k$ for $k = 1, \ldots, K$.
Write Z as indicator vector $Z = (Z^1 \ldots Z^k)$.

Then the posterior probability of π is:

$$\begin{aligned} P(\boldsymbol{\pi}|Z) &\propto P(\boldsymbol{\pi})P(Z|\boldsymbol{\pi}) \\ &\propto (\pi_1^{\alpha_1 - 1} \ldots \pi_K^{\alpha_K - 1}) \times (\pi_1^{Z^1} \ldots \pi_K^{Z^k}) \\ &\propto \pi_1^{\alpha_1 - 1} \ldots \pi_K^{\alpha_K - 1} \\ &= \pi_1^{\alpha_1 + Z^1 - 1} \ldots \pi_K^{\alpha_K + Z^k - 1}, \end{aligned}$$

which is again a Dirichlet density with modified parameter: $\mathrm{Dir}(\alpha + Z)$

Multinomial-Dirichlet conjugacy

Let $\boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\alpha})$.
Let $Z \sim \mathrm{Multinomial}(\boldsymbol{\pi})$, i.e. $P(Z = k|\boldsymbol{\pi}) = \pi_k$ for $k = 1, \ldots, K$.
Write Z as indicator vector $Z = (Z^1 \ldots Z^k)$.

Then the posterior probability of π is:

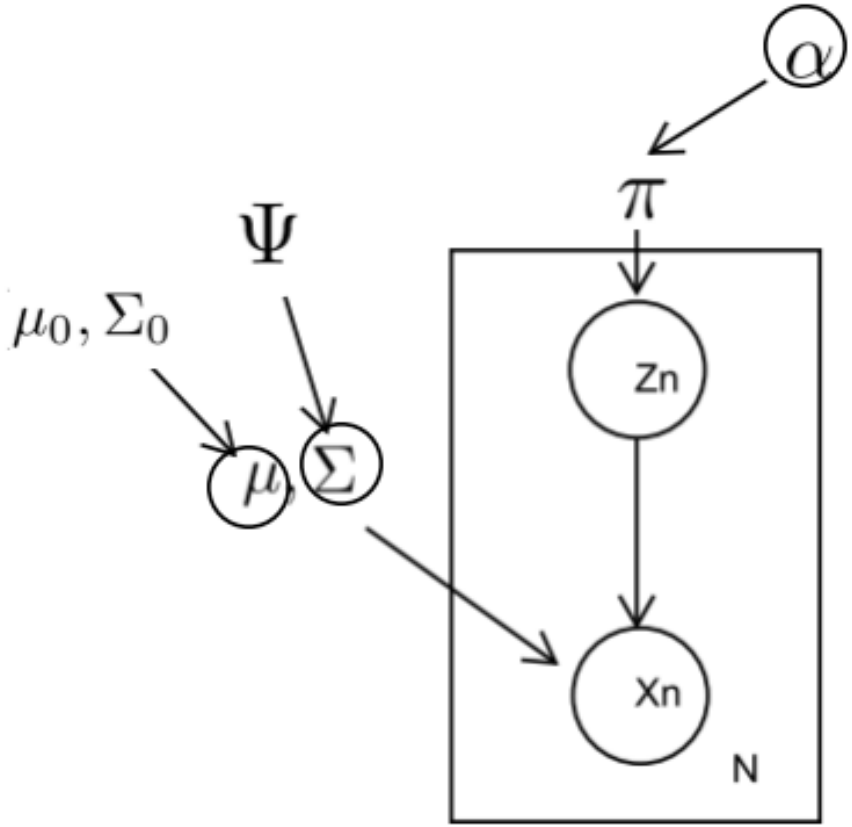
$$\begin{aligned} P(\boldsymbol{\pi}|Z) &\propto P(\boldsymbol{\pi})P(Z|\boldsymbol{\pi}) \\ &\propto (\pi_1^{\alpha_1 - 1} \ldots \pi_K^{\alpha_K - 1}) \times (\pi_1^{Z^1} \ldots \pi_K^{Z^k}) \\ &\propto \pi_1^{\alpha_1 - 1} \ldots \pi_K^{\alpha_K - 1} \\ &= \pi_1^{\alpha_1 + Z^1 - 1} \ldots \pi_K^{\alpha_K + Z^k - 1}, \end{aligned}$$

which is again a Dirichlet density with modified parameter: $\mathrm{Dir}(\alpha + Z)$

We say Multinomial-Dirichlet is a *conjugate pair*
Other conjugate pairs: Normal-Normal (for mean variable μ), Normal-Inverse Wishart (for covariance matrix Σ), etc

Bayesian mixture model

$$
\begin{aligned}
X_n | Z_n = k &\sim \mathrm{N}(\mu_k, \Sigma_k) \\
Z_n | \pi &\sim \mathrm{Multinomial}(\pi) \\
\pi &\sim \mathrm{Dir}(\alpha) \\
\mu_i | \mu_0, \Sigma_0 &\sim \mathrm{N}(\mu_0, \Sigma_0) \\
\Sigma_i | \Psi &\sim \mathrm{IW}(\Psi, m).
\end{aligned}
$$



$(\alpha, \mu_0, \Sigma_0, \Psi)$ are non-random parameters
(or they may be random and assigned with prior distributions as well)

Posterior inference

Posterior inference is about calculating conditional probability of latent variables and model parameters
i.e., $P((Z_n)_{n=1}^N, (\pi_k, \mu_k, \Sigma_k)_{k=1}^K | X_1, \ldots, X_N)$

Posterior inference

Posterior inference is about calculating conditional probability of latent variables and model parameters
i.e., $P((Z_n)_{n=1}^N, (\pi_k, \mu_k, \Sigma_k)_{k=1}^K | X_1, \ldots, X_N)$

This is usually difficult computationally

Posterior inference

Posterior inference is about calculating conditional probability of latent variables and model parameters
i.e., $P((Z_n)_{n=1}^N, (\pi_k, \mu_k, \Sigma_k)_{k=1}^K | X_1, \ldots, X_N)$

This is usually difficult computationally

An approach is via sampling, exploiting conditional independence

Posterior inference

Posterior inference is about calculating conditional probability of latent variables and model parameters
i.e., $P((Z_n)_{n=1}^N, (\pi_k, \mu_k, \Sigma_k)_{k=1}^K | X_1, \ldots, X_N)$

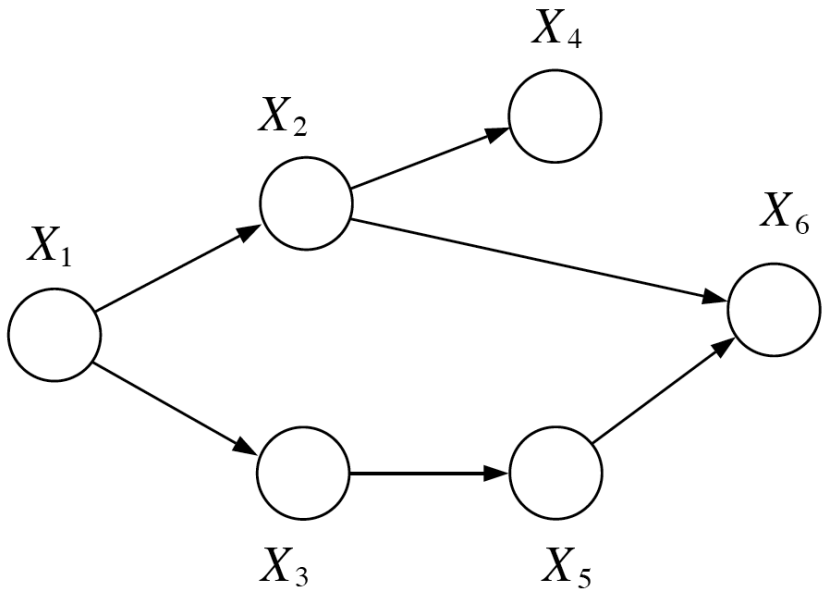
This is usually difficult computationally

An approach is via sampling, exploiting conditional independence

At this point we take a detour, discussing a general modeling and inference formalism known as graphical models

Directed graphical models

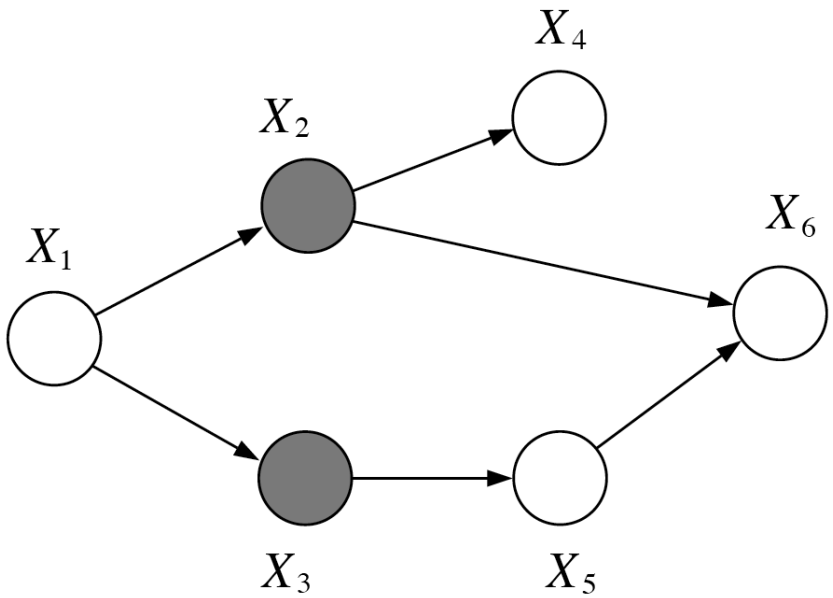
Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each node $v \in \mathcal{V}$ is associated with a random variable X_v:



The joint distribution on collection of variables $X_{\mathcal{V}} = \{X_v : v \in \mathcal{V}\}$ factorizes accoding to the "parent-of" relation defined by directed edges $\mathcal{E}$:

$$P(X_{\mathcal{V}}) = \prod_{v \in \mathcal{V}} P(X_v | X_{\text{parents}(v)})$$

Conditional independence



Observed variables are shaded

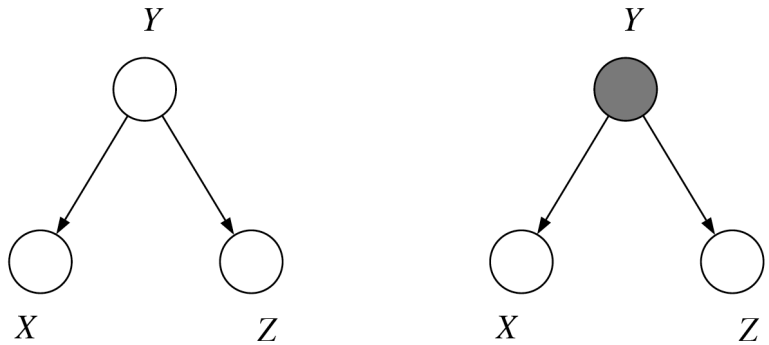
It can be shown that $X_1 \perp \{X_4, X_5, X_6 | X_2, X_3\}$.

Moreover we read off all such conditional independence from the graph structure.

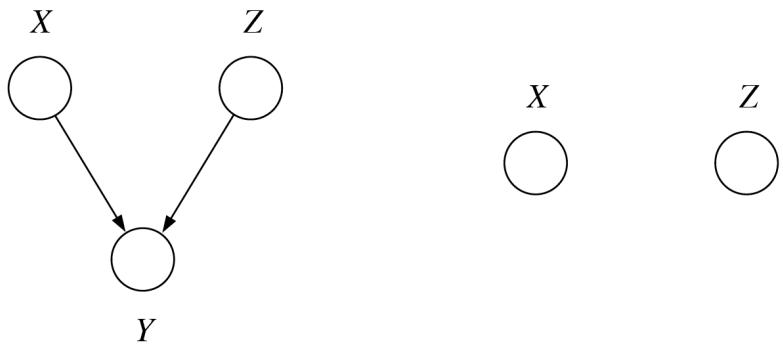
Basic conditional independence structure

"chain structure": $X \perp Z|Y$

"causal structure": $X \perp Z|Y$

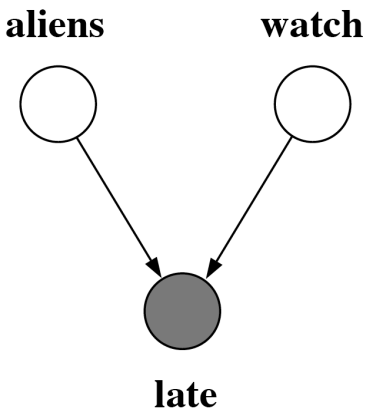


"explanation-away": $X \perp Z$ (marginally) but $X \not\perp Z|Y$



"Explanation-away"

aliens = "Alice was abducted by aliens"
watch = "forgot to set watch alarm before bed"
late = "Alice is late for class"

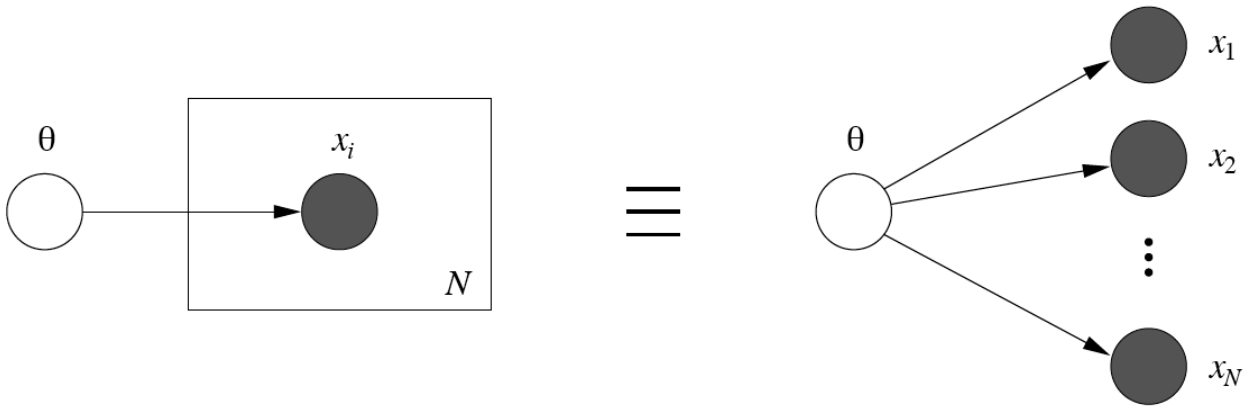


$$aliens \perp watch$$

$$aliens \not\perp watch \mid late$$

Condionally i.i.d.

"Conditional iid (identically and independently distributed)": this is represented by a plate notation that allows subgraphs to be replicated:



Note that this graph represents a mixture distribution for observed variables $(X_1, \ldots, X_N)$:

$$\begin{aligned} P(X_1, \ldots, X_N) &= \int P(X_1, \ldots, X_N|\theta) dP(\theta) \\ &= \int \prod_{i=1}^{N} P(X_i|\theta) dP(\theta) \end{aligned}$$

Gibbs sampling

A Markov chain Monte Carlo (MCMC) sampling method

Consider a collection of variables, say $X_1, \ldots, X_N$ with a joint distribution $P(X_1, \ldots, X_N)$ (which may be a conditional joint distribution in our specific problem)

A stationary Markov chain is a sequence of $\mathbf{X}^t = (X_1^t, \ldots, X_N^t)$ for $t = 1, 2, \ldots$ such that given $\mathbf{X}^t$, random variable $\mathbf{X}^{t+1}$ is conditionally independent of all variables before t, and

$$P(\mathbf{X}^{t+1}|\mathbf{X}^t) \text{ is invariant with respect to } t$$

Gibbs sampling (cont)

Gibbs sampling method sets up the Markov chain as follows

- at step $t = 1$, initialize $\mathbf{X}^1$ to arbitrary values
- at step t, choose n randomly among $1, \ldots, N$
- draw a sample for X_n^t from $P(X_n | X_1, \ldots, X_{n-1}, X_{n+1}, \ldots, X_N)$
- iterate

A fundamental theorem of Markov chain theory

Under mild conditions (ensuring ergodicity), $\mathbf{X}^t$ converges in the limit to the joint distribution of $\mathbf{X}$, namely $P(X_1, \ldots, X_N)$

Back to posterior inference

The goal is the sample from the (conditional) joint distribution
$P((Z_n)_{n=1}^N, (\pi_k, \mu_k, \Sigma_k)_{k=1}^K | X_1, \ldots, X_N)$

By Gibbs sampling, it is sufficient to be able to sample from conditional distributions of each of the latent variables and parameters given everything else (and conditionally on the data)

We will see that conditional independence helps in a big way

Sample π:

$$\begin{aligned} P(\pi|\mu, \Sigma, Z_1 ... Z_N, \text{Data}) &= P(\pi|Z_1 ... Z_N) \\ &\uparrow \quad \text{(conditional independence)} \\ &\propto P(Z_1 ... Z_n|\pi) P(\pi|\alpha) \\ &= \text{Dir}(\alpha_1 + n_1, \alpha_2 + n_2, ..., \alpha_K + n_K), \end{aligned}$$

where $n_j = \sum_{n=1}^{N} \mathbb{I}(Z_n = j)$.

Sample π:

$$\begin{aligned} P(\pi|\mu, \Sigma, Z_1 ... Z_N, \text{Data}) &= P(\pi|Z_1 ... Z_N) \\ &\uparrow \quad \text{(conditional independence)} \\ &\propto P(Z_1 ... Z_n|\pi) P(\pi|\alpha) \\ &= \text{Dir}(\alpha_1 + n_1, \alpha_2 + n_2, ..., \alpha_K + n_K), \end{aligned}$$

where $n_j = \sum_{n=1}^{N} \mathbb{I}(Z_n = j)$.

Sample Z_n:

$$\begin{aligned} P(Z_n = k|\text{everything else, including data}) &= P(Z_n = k|X_n, \pi, \mu, \Sigma) \\ &\uparrow \quad \text{(conditional independence)} \\ &= \frac{\pi_k N(X_n|\mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k N(X_n|\mu_k, \Sigma_k)}. \end{aligned}$$

Sample μ_k:

$$
\begin{aligned}
& P(\mu_k | \mu_0, \Sigma_0, \mathbf{Z}, \mathbf{X}, \mathbf{\Sigma}) \\
= \quad & P(\mu_k | \mu_0, \Sigma_0, \Sigma_k, \{Z_n, X_n \text{ such that } Z_n = k\}) \\
\propto \quad & P(\{X_n : Z_n = k\} | \mu_k, \Sigma_k) P(\mu_k | \mu_0, \Sigma_0) \\
\uparrow \quad & \text{Bayes' Rule} \\
= \quad & \prod_{n: Z_n = k} \exp\left\{ -\frac{1}{2} (X_n - \mu_k)^T \Sigma_k^{-1} (X_n - \mu_k) \right\} \\
& \times \exp -\frac{1}{2} (\mu_k - \mu_0)^T \Sigma_0^{-1} (\mu_k - \mu_0) \\
\propto \quad & \exp -\frac{1}{2} (\mu_k - \tilde{\mu_k})^T \tilde{\Sigma_k}^{-1} (\mu_k - \tilde{\mu_k}) \\
\equiv \quad & \mathrm{N}(\tilde{\mu_k}, \tilde{\Sigma_k})
\end{aligned}
$$

Here,

$$\tilde{\Sigma_k}^{-1} = \Sigma_0^{-1} + n_k \Sigma_k^{-1},$$

$$\text{where } n_k = \sum_{n=1}^{N} \mathbf{1}(Z_n = k)$$

$$\tilde{\Sigma_k}^{-1} \tilde{\mu_k} = \Sigma_0^{-1} \mu_0 + \Sigma_k^{-1} \sum_{X_n : Z_n = k} X_n$$

Hence, $\tilde{\mu_k} = \tilde{\Sigma_k}(\Sigma_0^{-1} \mu_0 + \Sigma_k^{-1} \sum_{X_n : Z_n = k} X_n)$.

Here,

$$\tilde{\Sigma_k}^{-1} = \Sigma_0^{-1} + n_k \Sigma_k^{-1},$$

$$\text{where } n_k = \sum_{n=1}^{N} \mathbf{1}(Z_n = k)$$

$$\tilde{\Sigma_k}^{-1} \tilde{\mu_k} = \Sigma_0^{-1} \mu_0 + \Sigma_k^{-1} \sum_{X_n : Z_n = k} X_n$$

Hence, $\tilde{\mu_k} = \tilde{\Sigma_k}(\Sigma_0^{-1}\mu_0 + \Sigma_k^{-1} \sum_{X_n : Z_n = k} X_n)$.

Notice that if $n_k \to \infty$, then $\tilde{\Sigma_k} \to 0$.
So, $\tilde{\mu_k} - \frac{1}{n_k} \sum_{n : Z_n = k} X_n \to 0$. (That is, the prior is taken over by data!)

To sample Σ_k, we use inverse Wishart distribution (a generalization of the chi-square distribution to multivariate cases) as prior:

$$B \sim W^{-1}(\Psi, m) \Leftrightarrow B^{-1} \sim W(\Psi, m)$$

$B, \Psi : p \times p$ PSD matrices, m : degree of freedom

Inverse-Wishart density:

$$P(B|\Psi, m) \propto \|\Psi\|^{\frac{m}{2}} \|B\|^{\frac{-(n+p+1)}{2}} \exp -\mathbf{tr}(\Psi B^{-1}/2)$$

To sample Σ_k, we use inverse Wishart distribution (a generalization of the chi-square distribution to multivariate cases) as prior:

$$B \sim W^{-1}(\Psi, m) \Leftrightarrow B^{-1} \sim W(\Psi, m)$$

$B, \Psi : p \times p$ PSD matrices, m : degree of freedom

Inverse-Wishart density:

$$P(B|\Psi, m) \propto \|\Psi\|^{\frac{m}{2}} \|B\|^{\frac{-(n+p+1)}{2}} \exp -\mathbf{tr}(\Psi B^{-1}/2)$$

Assume that, as a prior for Σ_k, $k = 1, ..., K$,

$$\Sigma_k|\Psi, m \sim \text{IW}(\Psi, m).$$

So the posterior distribution for Σ_k takes the form:

$$
\begin{aligned}
& P(\Sigma_k|\Psi, m, \boldsymbol{\mu}, \mathbf{pi}, \mathbf{Z}, \text{Data}) \\
= \quad & P(\Sigma_k|\Psi, m, \mu_k, \{X_n : Z_n = k\}) \\
\propto \quad & \prod_{n:Z_n=k} P(X_n|\Sigma_k, \mu_k) \times P(\Sigma_k|\Psi, m) \\
\propto \quad & \frac{1}{\|\Sigma_k\|^{\frac{n_k}{2}}} \exp\left\{ \sum_{n:Z_n=k} -\frac{1}{2}\mathbf{tr}[\Sigma_k^{-1}(X_k - \mu_k)(X_k - \mu_k)^T] \right\} \\
& \times \|\Psi\|^{\frac{m}{2}} \|\Sigma_k\|^{-\frac{m+p+1}{2}} \exp -\frac{1}{2}\mathbf{tr}[\Psi\Sigma_k^{-1}] \\
= \quad & \text{IW}(A + \Psi, n_k + m)
\end{aligned}
$$

where

$$
A = \sum_{n:Z_n=k} (X_n - \mu_k)(X_n - \mu_k)^T, \;\; n_k = \sum_{n=1}^{N} \mathbb{I}(Z_n = k)
$$

Summary of Gibbs sampling

Use Gibbs sampling to obtain samples from the posterior distribution $P(\boldsymbol{\pi}, \mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Sigma} | X_1, \ldots, X_N)$

Sampling algorithm

- Randomly generate $(\pi^{(1)}, Z^{(1)}, \mu^{,(1)}, \Sigma^{(1)})$
- For $t = 1, ..., T$, do the following:

  (1) draw $\pi^{(t+1)} \sim P(\pi | Z^{(t)}, \mu^{(t)}, \Sigma^{(t)}, \text{Data})$

  (2) draw $Z_n^{(t+1)} \sim P(Z_n | Z_{-n}^{(t)}, \mu^{(t)}, \Sigma^{(t)}, \pi^{(t+1)}, \text{Data})$

  (3) draw $\mu_k^{(t+1)} \sim P(\mu_k | \mu_k^{(t)}, Z_n^{(t+1)}, \Sigma^{(t)}, \pi^{(t+1)}, \text{Data})$

By the fundamental theorem of Markov chain, for t sufficiently large, $\pi^{(t)}, Z^{(t)}, \mu,^{(t)}, \Sigma^{(t)})$ can be viewed as a random sample of the posterior $P(\cdot|\text{Data})$

By the fundamental theorem of Markov chain, for t sufficiently large, $\pi^{(t)}, Z^{(t)}, \mu,^{(t)}, \Sigma^{(t)})$ can be viewed as a random sample of the posterior $P(\cdot|\text{Data})$

Suppose that we have drawn samples from the posterior distribution, posterior probabilities can be obtained via Monte Carlo approximation

By the fundamental theorem of Markov chain, for t sufficiently large, $\pi^{(t)}, Z^{(t)}, \mu,^{(t)}, \Sigma^{(t)})$ can be viewed as a random sample of the posterior $P(\cdot|\text{Data})$

Suppose that we have drawn samples from the posterior distribution, posterior probabilities can be obtained via Monte Carlo approximation

E.g., posterior probability of the cluster label for data point X_n:

$$P(Z_n|\text{Data}) \approx \frac{1}{T-s+1}\Sigma_{t=s}^{T}\delta_{Z_n^s}$$

Comparing Gibbs sampling and EM algorithm

Gibbs can be viewed as a stochastic version of the EM algorithm

EM algorithm

1. E step: given current value of parameter θ, calculate conditional expectation of latent variables $\mathbf{Z}$
2. M step: given the conditional expectations of $\mathbf{Z}$, update θ by maximizing the expected complete log-likelihood function

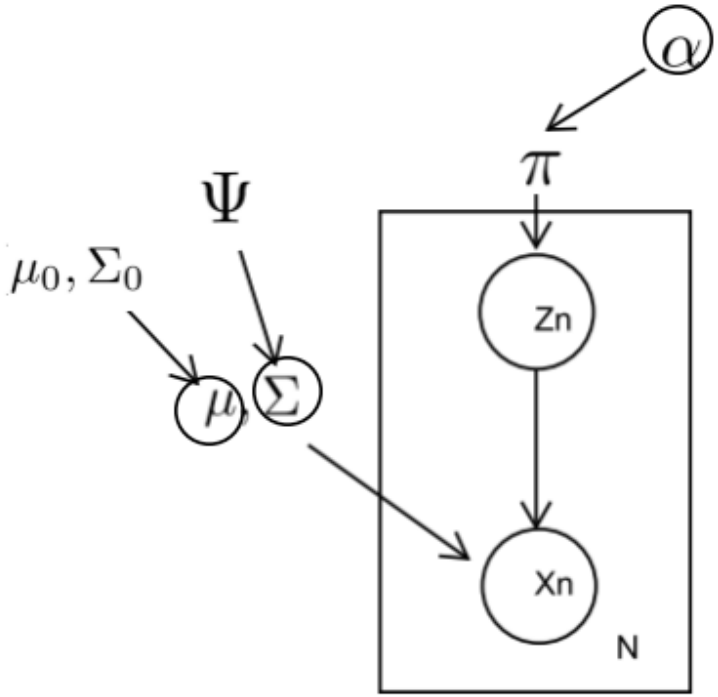
Gibbs sampling

1. given current values of θ, sample $\mathbf{Z}$
2. given current values of $\mathbf{Z}$, sample (random) θ

Outline

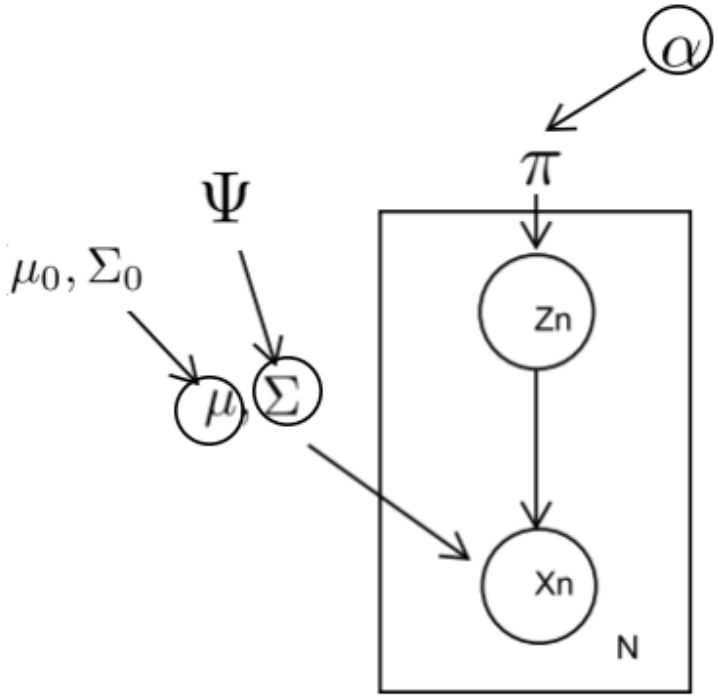
Recall our Bayesian mixture model: for each $n = 1, \ldots, N$, $k = 1, \ldots, K$,

$$
\begin{aligned}
X_n | Z_n = i &\sim \mathrm{N}(\mu_i, \Sigma_i), \\
Z_n | \pi &\sim \mathrm{Multinomial}(\pi) \\
\boldsymbol{\pi} &\sim \mathrm{Dir}(\boldsymbol{\alpha}) \\
\mu_| \mu_0, \Sigma_0 &\sim \mathrm{N}(\mu_0, \Sigma_0) \\
\Sigma_k | \Psi &\sim \mathrm{IW}(\Psi, m).
\end{aligned}
$$



Recall our Bayesian mixture model: for each $n = 1, \ldots, N$, $k = 1, \ldots, K$,

$$
\begin{aligned}
X_n | Z_n = i &\sim \mathrm{N}(\mu_i, \Sigma_i), \\
Z_n | \pi &\sim \mathrm{Multinomial}(\pi) \\
\boldsymbol{\pi} &\sim \mathrm{Dir}(\boldsymbol{\alpha}) \\
\mu_| \mu_0, \Sigma_0 &\sim \mathrm{N}(\mu_0, \Sigma_0) \\
\Sigma_k | \Psi &\sim \mathrm{IW}(\Psi, m).
\end{aligned}
$$



We may assume further that parameters $(\alpha, \mu_0, \Sigma_0, \Psi)$ are may be random and assigned by prior distributions

Thus we obtain a hierarchical model in which parameters appear as latent random variables

Exchangeability

The existence of latent variables can be motivated by De Finetti's theorem

Classical statistics often relies on assumption of i.i.d. data $X_1, \ldots, X_N$ with respect to some probability model parameterized by θ (non-random)

However, if $X_1, \ldots, X_N$ are exchangeable, then De Finetti's theorem establishes the existence of a latent *random* variable θ such that, $X_1, \ldots, X_N$ are conditionally i.i.d. given θ

This theorem is regarded by many as one of the results that provide the mathematical foundation for Bayesian statistics

Definition

Let I be a countable index set. A sequence $(X_i : i \in I)$ (finite or infinite) is **exchangeable** if for any permutation ρ of I

$$(X_{\rho(i)})_{i \in I} \stackrel{\text{law}}{=} (X_i)_{i \in I}$$

Definition

Let I be a countable index set. A sequence $(X_i : i \in I)$ (finite or infinite) is **exchangeable** if for any permutation ρ of I

$$(X_{\rho(i)})_{i \in I} \stackrel{\text{law}}{=} (X_i)_{i \in I}$$

De Finetti's theorem

If $(X_1, ..., X_n, ...)$ is an infinite exchangeable sequence of random variables on some probability space then there is a random variable $\theta \sim \pi$ such that $X_1, ..., X_n, ...$ are iid conditionally on θ. That is, for all n

$$P(X_1, ..., X_n, ...) = \int_\theta \prod_{i=1}^{n} P(X_i|\theta)\, d\pi(\theta)$$

Remarks

- Exchangeability is a weaker assumption than iid.
- θ may be (generally) an infinite dimensional variable

Latent Dirichlet allocation/ Finite admixture

Developed by Pritchard et al (2000), Blei et al (2001)

Widely applicable to data such as texts, images and biological data (Google Scholar has 12000 citations)

A canonical and simple example of hierarchical mixture model for discrete data

Latent Dirichlet allocation/ Finite admixture

Developed by Pritchard et al (2000), Blei et al (2001)

Widely applicable to data such as texts, images and biological data (Google Scholar has 12000 citations)

A canonical and simple example of hierarchical mixture model for discrete data

Some jargons:

Basic building blocks are *words* represented by random variables, say W, where $W \in \{1, ..., V\}$. V is the length of the vocabulary.

A *document* a sequence of words denoted by $\overline{W} = (W_1, ..., W_N)$.

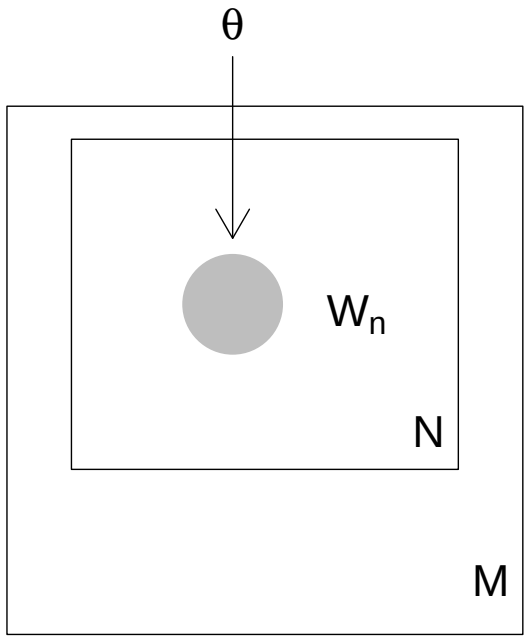
A *corpus* is a collection of documents $(\overline{W}^1, ..., \overline{W}^m)$.

General problems: Given a collection of documents, can we infer the topics that the documents may be clustered around?

Early modeling attempts

Unigram Model: All documents in corpus share the same "topic". I.e., For any document $\overline{W} = (W_1, \ldots, W_N)$,

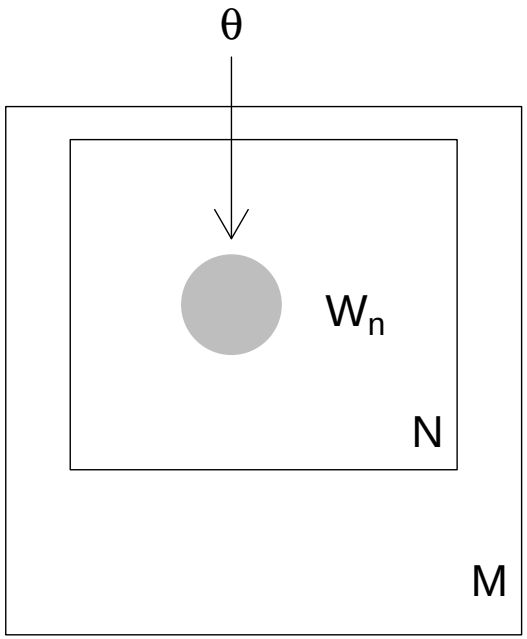
$$W_1, ..., W_N \sim \mathrm{Mult}(\theta)$$

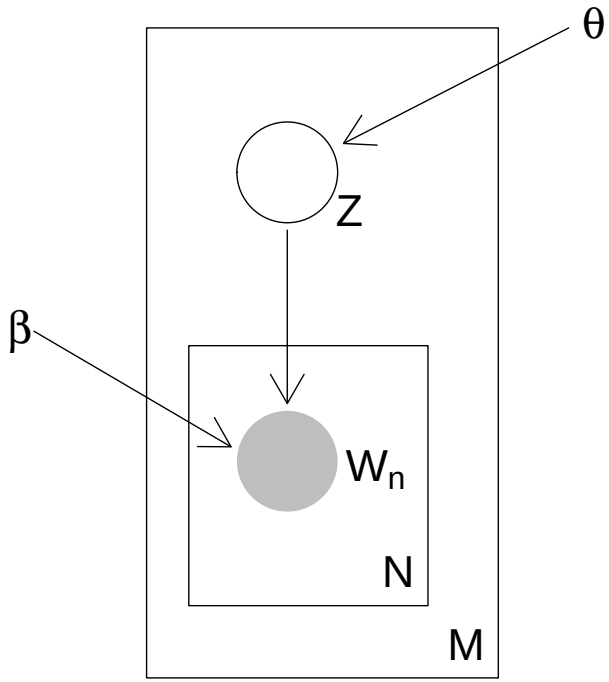


Mixture of Unigram Model: Each document $\overline{W}$ is associated with a latent "topic" variable Z. I.e.,

For each $d = 1, \ldots, m$, generate document $\overline{W} = (W_1, \ldots, W_N)$ as follows:

$$\begin{aligned} Z &\sim \text{Mult}(\theta) \\ W_1, ..., W_N | Z = k &\overset{iid}{\sim} \text{Mult}(\beta_k) \end{aligned}$$





Latent Dirichlet allocation model

Assume the following:

- The words within each document are exchangeable – this is the 'bag of words' assumption
- The documents within each corpus are exchangeable
- A document may be associated with K topics
- Each word within a document is associated with any topics

Latent Dirichlet allocation model

Assume the following:

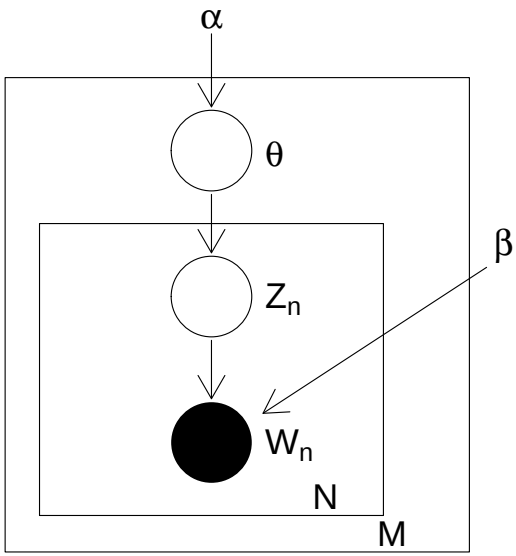
- The words within each document are exchangeable – this is the 'bag of words' assumption
- The documents within each corpus are exchangeable
- A document may be associated with K topics
- Each word within a document is associated with any topics

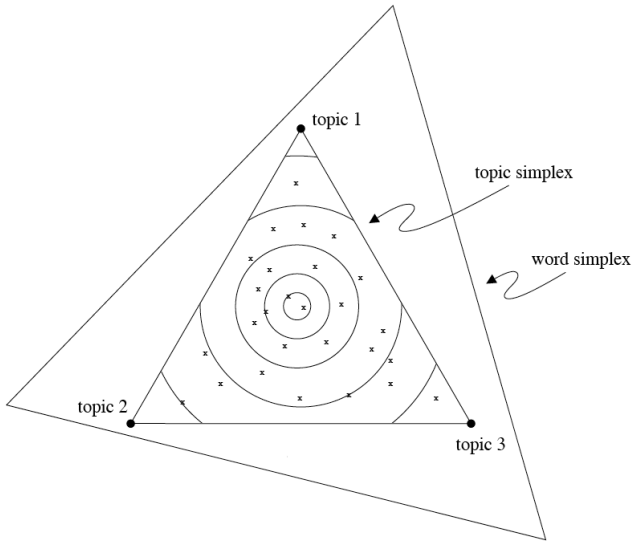
For $d = 1, \ldots, M$, generate document $\overline{W} = (W_1, \ldots, W_N)$ as follows:

- draw $\theta \sim \mathrm{Dir}(\alpha_1, ..., \alpha_K)$,
- for each $n = 1, ..., N$,

$$\begin{aligned} Z_n &\sim \mathrm{Mult}(\theta) \text{ i.e. } P(Z_n = k|\theta) = \theta_k \\ W_n|Z_n &\overset{iid}{\sim} \mathrm{Mult}(\boldsymbol{\beta}) \text{ i.e. } P(W_n = j|Z_n = k, \boldsymbol{\beta}) = \beta_{kj}, \beta \in \mathbb{R}^{K \times V} \end{aligned}$$

Geometric illustration





Each x dot in the topic polytope (topic simplex in illustration) corresponds to the word frequency vector for a random document

Extreme points of the topic polytope (e.g., topic 1, topic 2,...) in RHS are represented by vectors $\boldsymbol{\beta}_k$ for $k = 1, 2, \ldots$ in the hierarchical model in LHS

$$\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kV})$$

Posterior inference

The goal of inference includes:

1. Compute the posterior distribution, $P(\theta, Z|\overline{W}, \alpha, \beta)$ for each document $\overline{W} = (W_1, \ldots, W_N)$
2. Estimating α, β from the data, e.g., corpus of M documents $\overline{W}^1, \ldots, \overline{W}^M$

Both of the above are relatively easy to do using Gibbs Sampling, or Metropolis-Hastings, which will be left as an exercise.
Unfortunately a sampling algorithm may be extremely slow (to achieve convergence), this motivate a fast deterministic algorithm for posterior inference.

The posterior can be rewritten as

$$P(\theta, Z|\overline{W}, \alpha, \beta) = \frac{P(\theta, Z, \overline{W}, \alpha, \beta)}{P(\overline{W}|\alpha, \beta)}$$

The posterior can be rewritten as

$$P(\theta, Z|\overline{W}, \alpha, \beta) = \frac{P(\theta, Z, \overline{W}, \alpha, \beta)}{P(\overline{W}|\alpha, \beta)}$$

The numerator can be computed easily:

$$P(\theta, Z, \overline{W}, \alpha, \beta) = P(\theta|\alpha) \prod_{n=1}^{N} P(Z_n|\theta) P(W_n|Z_n, \beta)$$

The posterior can be rewritten as

$$P(\theta, Z|\overline{W}, \alpha, \beta) = \frac{P(\theta, Z, \overline{W}, \alpha, \beta)}{P(\overline{W}|\alpha, \beta)}$$

The numerator can be computed easily:

$$P(\theta, Z, \overline{W}, \alpha, \beta) = P(\theta|\alpha) \prod_{n=1}^{N} P(Z_n|\theta) P(W_n|Z_n, \beta)$$

Unfortunately, the denominator is difficult:

$$P(\overline{W}|\alpha, \beta) = \int_{Z,\theta} P(\theta, Z, \overline{W}|\alpha, \beta) =$$

$$\int \frac{\Gamma(\sum \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \prod_{n=1}^{N} \left[\sum_{k=1}^{K} \prod_{j=1}^{V} (\theta_k \beta_{kj})^{\mathbb{I}\{W_n = j\}} \right] d\theta$$

Variational inference

This is an alternative to sampling-based inference.
The main spirit is to turn a difficult computation problem into an optimization problem, one which can be modified/simplified.
We consider the simplest form of variational inference, known as "mean-field" approximation:

- Consider a family of tractable distributions $Q = \{q(\boldsymbol{\theta}, \mathbf{Z}|\overline{W}, \alpha, \beta)\}$
- Choose the one in Q, that is closest to the true posterior:

$$q^* = \arg\min_{q \in Q} \mathrm{KL}(q||p(\boldsymbol{\theta}, \mathbf{Z}|\overline{W}, \alpha, \beta))$$

- Use q^* instead of the true parameter $q \in Q$

In mean-field approximation, Q taken to be the family of "factorized" distributions, i.e.:

$$q(\theta, Z|\overline{W}, \gamma, \phi) = q(\theta|\overline{W}, \gamma)\Pi_{n=1}^{N} q(Z_n|\overline{W}, \phi)$$

In mean-field approximation, Q taken to be the family of "factorized" distributions, i.e.:

$$q(\theta, Z|\overline{W}, \gamma, \phi) = q(\theta|\overline{W}, \gamma)\Pi_{n=1}^{N} q(Z_n|\overline{W}, \phi)$$

Optimization is performed with rspect to *variational parameters* (γ, ϕ)

Under $q \in Q$, for $n = 1, \ldots, N; k = 1, \ldots, K$,

$$P(Z_n = k|\overline{W}, \phi_n) = \phi_{nk}$$

$$\boldsymbol{\theta}|\boldsymbol{\gamma} \sim \text{Dir}(\boldsymbol{\gamma}), \boldsymbol{\gamma} \in \mathbb{R}_+^K$$

The optimization of variational parameters can be achieved by implementing a simple system of updating equations.

Mean-field algorithm

1. Initialize ϕ, γ arbitrarily.
2. Keep updating until convergence:

$$\begin{aligned} \phi_{nk} &\propto \beta_{kW_n} \exp\{\mathbb{E}_q[\log \theta_k | \gamma]\} \\ \gamma_k &= \alpha_k + \sum_{n=1}^{N} \phi_{nk}. \end{aligned}$$

In the first updating equation, we use a fact of Dirichlet distribution: if $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K) \sim \mathrm{Dir}(\boldsymbol{\gamma})$, then

$$\mathbb{E}[\log \theta_k | \gamma] = \Psi(\gamma_k) - \Psi(\sum_{k=1}^{K} \gamma_k),$$

where Ψ is the digamma function:

$$\Psi(x) = \frac{d \log \Gamma}{dx} = \frac{\Gamma'(x)}{\Gamma(x)}.$$

Derivation of the mean-field approximation

Jensen's Inequality

$$\begin{aligned}
\log P(\overline{W}|\alpha,\beta) &= \log \int_\theta \sum_Z P(\boldsymbol{\theta}, \mathbf{Z}, \overline{W}|\alpha,\beta) d\theta \\
&= \log \int_\theta \sum_Z \frac{P(\boldsymbol{\theta}, \mathbf{Z}, \overline{W}|\alpha,\beta)}{q(\boldsymbol{\theta}, \mathbf{Z})} q(\boldsymbol{\theta}, \mathbf{Z})\, d\theta \\
&\geq \int_\theta \sum_Z q(\boldsymbol{\theta}, \mathbf{Z}) \log \frac{P(\boldsymbol{\theta}, \mathbf{Z}, \overline{W}|\alpha,\beta)}{q(\boldsymbol{\theta}, \mathbf{Z})} d\theta \\
&= \mathbb{E}_q \log P(\boldsymbol{\theta}, \mathbf{Z}, \overline{W}|\alpha,\beta) - E_q \log q(\boldsymbol{\theta}, \mathbf{Z}) \\
&=: L(\gamma, \phi; \alpha, \beta)
\end{aligned}$$

The gap of the bound:

$$\log P(\overline{W}|\alpha,\beta) - L(\gamma,\phi;\alpha,\beta) = \mathrm{KL}(q(\boldsymbol{\theta}, \mathbf{Z})||P(\boldsymbol{\theta}, \mathbf{Z}|\overline{W}, \alpha, \beta)).$$

So q^* solves the following maximization:

$$\max_{\gamma,\phi} L(\gamma, \phi; \alpha, \beta).$$

So q^* solves the following maximization:

$$\max_{\gamma,\phi} L(\gamma, \phi; \alpha, \beta).$$

Note that

$$\log P(\boldsymbol{\theta}, \mathbf{Z}, \overline{W}|\alpha, \beta) = \log P(\theta|\alpha) + \sum_{n=1}^{N} \Big(\log P(Z_n|\theta) + \log P(W_n|Z_n, \beta) \Big)$$

So q^* solves the following maximization:

$$\max_{\gamma,\phi} L(\gamma, \phi; \alpha, \beta).$$

Note that

$$\log P(\boldsymbol{\theta}, \mathbf{Z}, \overline{W}|\alpha, \beta) = \log P(\theta|\alpha) + \sum_{n=1}^{N} \Big(\log P(Z_n|\theta) + \log P(W_n|Z_n, \beta) \Big)$$

So,

$$\begin{aligned} L(\gamma, \phi; \alpha, \beta) &= \mathbb{E}_q \log P(\theta|\alpha) + \sum_{n=1}^{N} \{\mathbb{E}_q \log P(Z_n|\theta) + \mathbb{E}_q \log P(W_n|Z_n, \beta)\} \\ &\quad - \mathbb{E}_q \log q(\theta|\gamma) - \sum_{n=1}^{N} \mathbb{E}_q \log q(Z_n|\phi_n). \end{aligned}$$

Let's go over each term in the previous expression:

$$\begin{aligned}
\log P(\theta|\alpha) &= \frac{\Gamma\left(\sum \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}, \\
\log P(\theta|\alpha) &= \sum_{k=1}^{K} (\alpha_{k-1}) \log \theta_k + \log \Gamma\left(\sum \alpha_k\right) - \sum_{k=1}^{K} \log \Gamma(\alpha_k), \\
\mathbb{E}_q \log P(\theta|\alpha) &= \sum_{k=1}^{K} (\alpha_k - 1) \left(\Psi(\gamma_k) - \Psi\left(\sum_{k=1}^{K} \gamma_k\right)\right) \\
&\quad + \log \Gamma\left(\sum \alpha_k\right) - \sum_{k=1}^{K} \log \Gamma(\alpha_k).
\end{aligned}$$

Next term:

$$
\begin{aligned}
P(Z_n|\theta) &= \prod_{k=1}^{K} \theta_k^{\mathbb{I}(Z_n=k)}, \\
\log P(Z_n|\theta) &= \sum_{k=1}^{K} \mathbb{I}\left(Z_n = k\right) \log \theta_k, \\
\mathbb{E}_q \log P(Z_n|\theta) &= \sum_{k=1}^{K} \phi_{nk} \left(\Psi\left(\gamma_k\right) - \Psi\left(\sum_{k=1}^{K} \gamma_k \right) \right).
\end{aligned}
$$

And next:

$$
\begin{aligned}
\log P\left(W_n|Z_n, \beta\right) &= \log \prod_{k=1}^{K} \prod_{j=1}^{V} \left(\beta_{kj}\right)^{\mathbb{I}(W_n=j, Z_n=k)}, \\
\mathbb{E}_q \log P(Z_n|\theta) &= \sum_{k=1}^{K} \sum_{j=1}^{V} \mathbb{I}\left(Z_n = k\right) \log \beta_{kj}.
\end{aligned}
$$

And next:

$$q(\theta|\gamma) = \frac{\Gamma(\sum \gamma_k)}{\sum_{k=1}^{K} \Gamma(\gamma_k)} \prod_{k=1}^{K} \theta_k^{\gamma_k - 1},$$

so,

$$\begin{aligned} \mathbb{E}_q \log q(\theta|\gamma) &= \sum_{k=1}^{K} (\gamma_k - 1) \left(\Psi(\gamma_k) - \Psi\left(\sum_{k=1}^{K} \gamma_k \right) \right) \\ &\quad + \log \Gamma(\sum_{k=1}^{K} \gamma_k) - \sum_{k=1}^{K} \log \Gamma(\gamma_k). \end{aligned}$$

And next:

$$q(\theta|\gamma) = \frac{\Gamma(\sum \gamma_k)}{\sum_{k=1}^{K} \Gamma(\gamma_k)} \prod_{k=1}^{K} \theta_k^{\gamma_k - 1},$$

so,

$$\begin{aligned} \mathbb{E}_q \log q(\theta|\gamma) &= \sum_{k=1}^{K} (\gamma_k - 1) \left(\Psi(\gamma_k) - \Psi\left(\sum_{k=1}^{K} \gamma_k \right) \right) \\ &\quad + \log \Gamma(\sum_{k=1}^{K} \gamma_k) - \sum_{k=1}^{K} \log \Gamma(\gamma_k). \end{aligned}$$

And next:

$$q(Z_n|\phi_n) = \prod_{k=1}^{K} \phi_{nk}^{\mathbb{I}(Z_n = k)},$$

So $\mathbb{E}_q \log q(Z_n|\gamma_n) = \sum_{k=1}^{K} \phi_{nk} \log \phi_{nk}$.

To summarize, we maximize the lower bound of the likelihood function:

$$L(\gamma, \phi; \alpha, \beta) := \mathbb{E}_q[\log P(\boldsymbol{\theta}, \mathbf{Z}, \overline{W}|\alpha, \beta)] - \mathbb{E}_q[\log q(\boldsymbol{\theta}, \mathbf{Z})]$$

with respect to variational parameters satisfying constraints, for all $n = 1, \ldots, N; k = 1, \ldots, K$

$$\sum_{k=1}^{K} \phi_{nk} = 1,$$
$$\phi_{nk} \geq 0,$$
$$\gamma_k \geq 0.$$

To summarize, we maximize the lower bound of the likelihood function:

$$L(\gamma, \phi; \alpha, \beta) := \mathbb{E}_q[\log P(\boldsymbol{\theta}, \mathbf{Z}, \overline{W}|\alpha, \beta)] - \mathbb{E}_q[\log q(\boldsymbol{\theta}, \mathbf{Z})]$$

with respect to variational parameters satisfying constraints, for all $n = 1, \ldots, N; k = 1, \ldots, K$

$$\sum_{k=1}^{K} \phi_{nk} = 1,$$
$$\phi_{nk} \geq 0,$$
$$\gamma_k \geq 0.$$

L decomposes nicely into a sum, which admits simple gradient-based update equations:

$$\text{Fix } \gamma, \text{ maximize w.r.t. } \phi, \quad \Rightarrow \phi_{nk} \propto \beta_{kW_n} \exp(\Psi(\gamma_k) - \Psi(\sum_{k=1}^{k} \gamma_k)),$$

$$\text{Fix } \phi, \text{ maximize w.r.t. } \gamma, \quad \Rightarrow \gamma_k = \alpha_k + \sum_{n=1}^{N} \phi_{nk}.$$

Variational EM algorithm

It remains to estimate parameters α and β,

$$\text{Data} \quad D = \text{set of documents } \{\overline{W}^1, \overline{W}^2, \ldots, \overline{W}^M\}$$

$$\text{Log-likelihood} \quad L(D) = \sum_{d=1}^{M} \log p(\overline{W_d}|\alpha, \beta)$$

$$= \sum_{d=1}^{M} L(\gamma_d, \phi_d|\alpha, \beta) + KL(q \| P(\boldsymbol{\theta}, \mathbf{z}|D, \alpha, \beta))$$

EM algorithm involves alternating between E step and M step until convergence:

Variational E step

For each $d = 1, \ldots, M$, let $(\gamma_d, \phi_d) := \arg\max L(\gamma_d, \phi_d; \alpha, \beta)$.

M step

Solve $(\alpha, \beta) = \arg\max \sum_{d=1}^{M} L(\gamma_d, \phi_d|\alpha, \beta)$.

Example

An example article from the AP corpus (Blei et al, 2003)

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE

Example

An example article from *Science* corpus (1880–2002) (Blei & Lafferty, 2009)

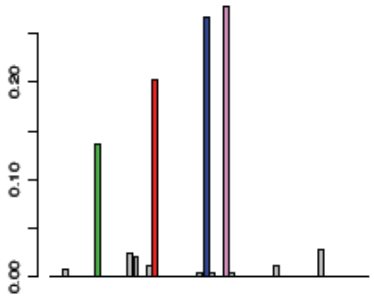
Chance and Statistical Significance in Protein and DNA Sequence Analysis

Samuel Karlin and Volker Brendel

Top words from the top topics (by term score)

sequence	measured	residues	computer
region	average	binding	methods
pcr	range	domains	number
identified	values	helix	two
fragments	different	cys	principle
two	size	regions	design
genes	three	structure	access
three	calculated	terminus	processing
cdna	two	terminal	advantage
analysis	low	site	important

Expected topic proportions



Abstract with the most likely topic assignments

Statistical approaches help in the determination of significant configurations in protein and nucleic acid sequence data. Three recent statistical methods are discussed: (i) score-based sequence analysis that provides a means for characterizing anomalies in local sequence text and for evaluating sequence comparisons; (ii) quantile distributions of amino acid usage that reveal general compositional biases in proteins and evolutionary relations; and (iii) *r*-scan statistics that can be applied to the analysis of spacings of sequence markers.

Top Ten Similar Documents

Exhaustive Matching of the Entire Protein Sequence Database
How Big Is the Universe of Exons?
Counting and Discounting the Universe of Exons
Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment
Ancient Conserved Regions in New Gene Sequences and the Protein Databases

Variations of the same theme

Graphical model for dynamic topic modeling (Blei & Lafferty, 2009)

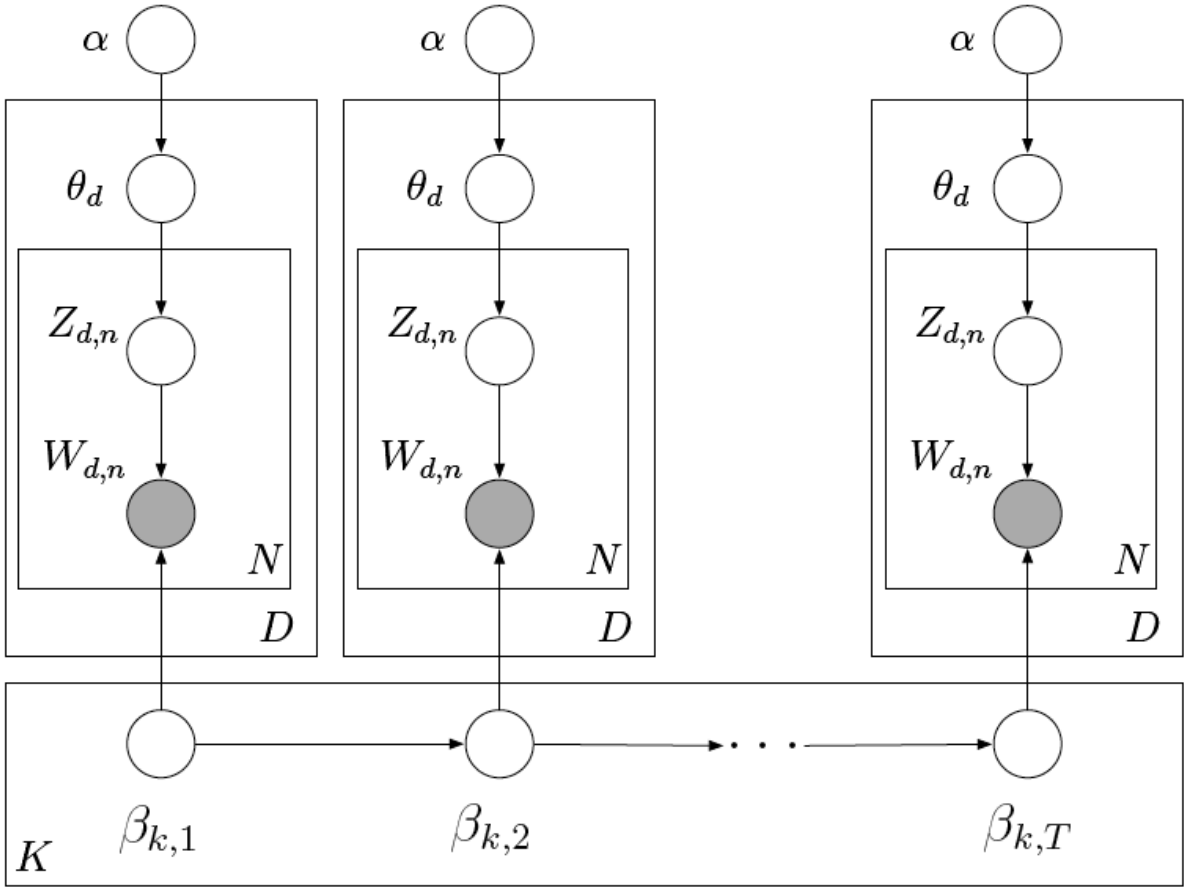
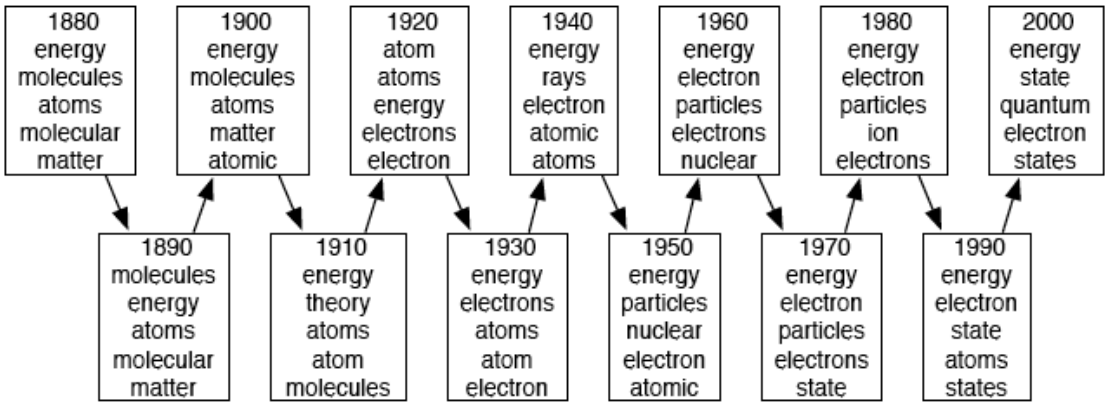
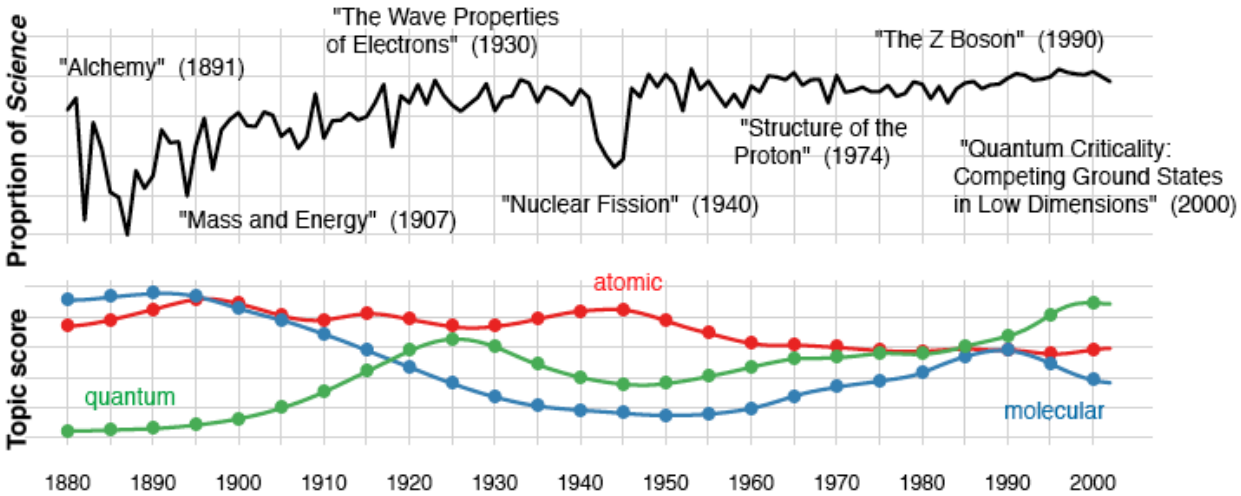
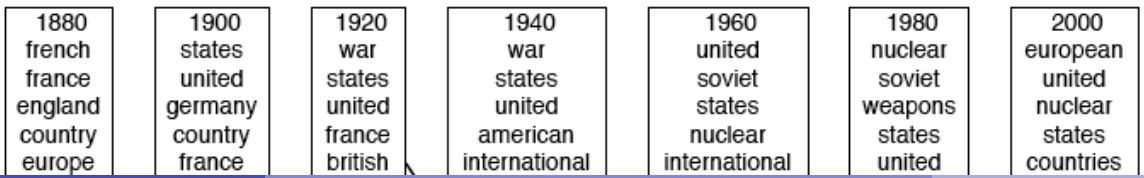


FIGURE 8. A graphical model representation of a dynamic topic model (for three time slices). Each topic's parameters $\beta_{t,k}$ evolve over time.

Evolution of two topics from the dynamic model fitted to the *Science* archive (1880–2002)







Outline

See hand-written notes.

Outline

See hand-written notes.

Outline

Incomplete References

For Part 1,2,3:
J. Berger. Statistical decision theory and Bayesian analysis, Springer 1985.
P. Bickel & K. Doksum. Mathematical statistics: basic ideas and selected topics, vol. 1, Prentice Hall, 2000.
C. Bishop. Pattern recognition and machine learning, 2007.
T. Hastie, R. Tibshirani & J. Friedman. Elements of statistical learning, Springer, 2009.
O. Kallenberg. Foundations of modern probability. Springer, 2010.
M. I. Jordan, Introduction to probablistic graphical models. Unpublished text book.
C. Robert. The Bayesian choice: From decision-theoretic foundation to computational implementation, Springer, 2007.
A. van der Vaart. Asymptotic Statistics, Cambridge University Press, 2000. y L. Wasserman. All of Statistics: a concise course in statistical inference, Springer, 2004.

For Part 4:

D. Blei, A. Ng, & M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 2003.

D. Blei & J. Lafferty, Topic models: A review, 2009.

J. Pritchard, M. Stephen & P. Donnelly, Inference of population structure using multilocus genotype data. *Genetics*, 2000.

M. J. Wainwright & M. I. Jordan, Graphical models, exponential families and variational inference, *Foundations and trends in machine learning*, 2008.

For Part 5, 6:

T. Ferguson. A Bayesian analysis of some nonpametric problems. *Annals of Statistics*, 1973.

J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 1994.

H. Ishwaran & L. James. Gibbs sampling methods of stick-breaking priors. *Joural of American Statistical Association*, 2001.

R. Neal. Markov Chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, 2000.

N. Hjort, C. Holmes, P. Mueller & S. Walker (Eds). Bayesian nonparametrics: principles and practices. Cambridge University Press, 2010.

M. I. Jordan, Dirichlet processes, Chinese restaurant processes and all that. *NIPS 2005 Tutorial*, 2005.

Y. W. Teh, M. I. Jordan, D. Blei & M. Beal. Hierarchical Dirichlet Processes, *Journal of American Statistical Association*, 2006.

J. Ghosh & R. Ramamoorthi. Bayesian nonparametrics, Springer 2003.

S. Ghosal, J. Ghosh & A. van der Vaart, Convergence of posterior distributions.*Annals of Statistics*, 2000.

X. Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. Technical Report 527, Department of Statistics, University of Michigan, 2011.

C. Villani. Optimal transport: Old and new topics, Springer 2008.