Other topics

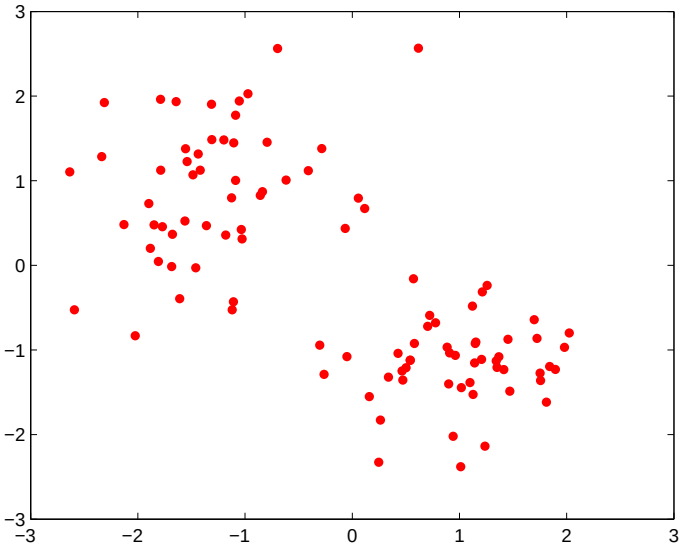
Mauricio A. Álvarez, PhD

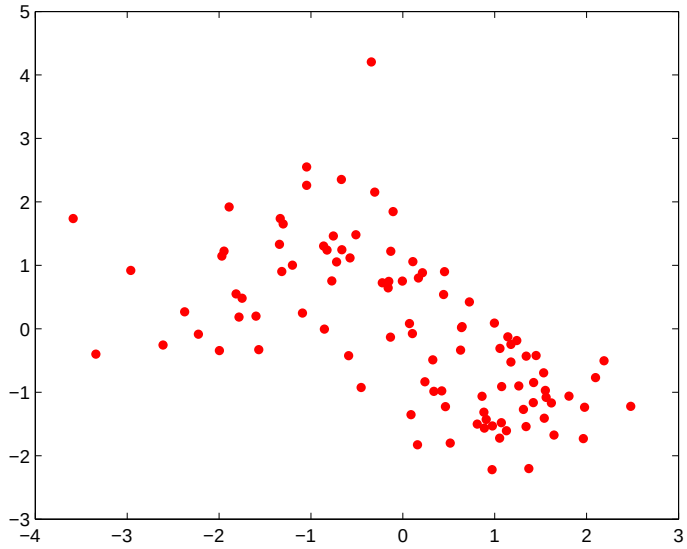
Curso de entrenamiento ArcelorMittal

Contents

Contents

Clustering





Mixture of probability functions

- One way to approximate multimodal probability functions is through a mixture of probability functions.

- From the mixtures of probability functions, the Gaussian mixture model (GMM) is one of the best known,

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

where K is the number of components in the mixture and the parameters π_k are probabilities that satisfy

$$0 \leq \pi_k \leq 1, \quad \sum_{k=1}^{K} \pi_k = 1.$$

Contents

Introduction

- The goal of the EM (Expectation - Maximization) algorithm is to find maximum likelihood solutions for models that use latent variables.
- Let $\mathbf{X} = \{\mathbf{x}_n\}_{n=1}^N$ be the set of observations and $\mathbf{Z} = \{\mathbf{z}_n\}_{n=1}^N$, the sequence of latent variables.
- The log-likelihood function is given as

$$\ln p(\mathbf{X}|\theta) = \ln \left\{ \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta) \right\}.$$

- The presence of the sum prevents the logarithm from acting directly on the joint distribution, which results in complex expressions for the maximum likelihood solution.

Incomplete data

- The set $\{\mathbf{X}, \mathbf{Z}\}$ is known as the set of complete data.
- The observed data $\mathbf{X}$ is known as incomplete data.
- From the set $\{\mathbf{X}, \mathbf{Z}\}$ we only know $\mathbf{X}$. The only information we have about $\mathbf{Z}$ is in the probability density function $p(\mathbf{Z}|\mathbf{X}, \theta)$.

What is the logic of the EM algorithm? (I)

- Since the complete data likelihood function cannot be used, consider using its expected value under the posterior distribution of the latent variables, which corresponds to step E of the EM algorithm.
- In step M, this expectation is maximized with respect to the parameters of interest.
- If the current estimate of the parameters is θ^{old}, a successive pair of steps E and M, give rise to a new estimate θ^{new}.

What is the logic of the EM algorithm? (II)

- In step E, the current parameters are used $\boldsymbol{\theta}^{\text{old}}$, to find the posterior distribution of the latent variables given by $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$.

- This posterior distribution is then used to find the expected logarithmic likelihood of the complete data evaluated for a general parameter vector $\boldsymbol{\theta}$.

- This expectation, denoted by $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$, is given as

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}).$$

- In step M, a new vector of parameters is determined by maximizing $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}).$$

Pseudo-code for the EM algorithm

Given a joint distribution $p(\mathbf{X}, \mathbf{Z}|\theta)$, the goal is to maximize the likelihood function $p(\mathbf{X}|\theta)$ wrt the parameters θ.

1. Initialize the parameter vector θ^{old}.
2. **E step**. Compute $p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}})$.
3. **M step**. Compute θ^{new} given as

$$\theta^{\text{new}} = \arg\max_{\theta} \mathcal{Q}(\theta, \theta^{\text{old}}),$$

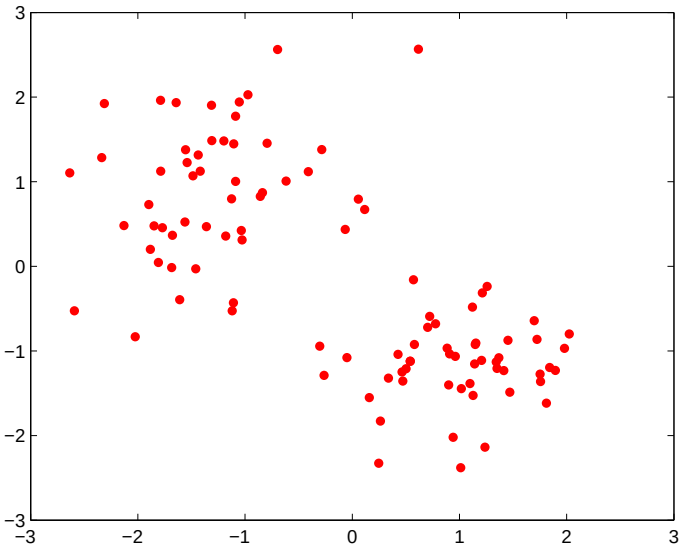
   where

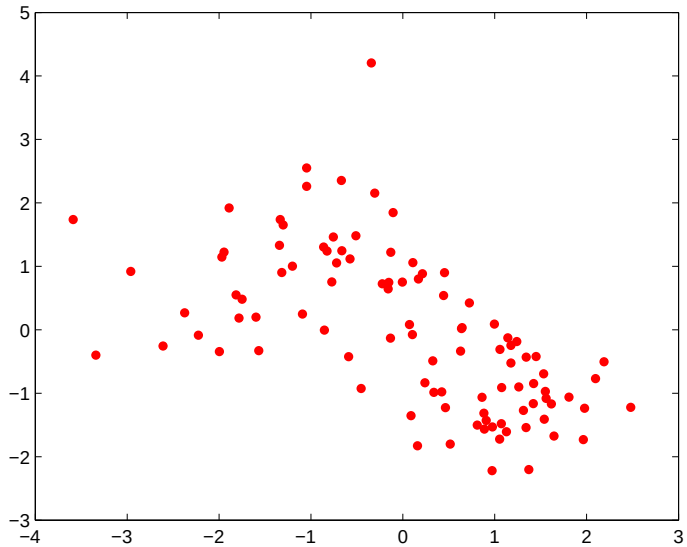
$$\mathcal{Q}(\theta, \theta^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\theta).$$

4. Check the convergence of the likelihood function or the parameters. If the convergence criterion is not satisfied, then $\theta^{\text{old}} \leftarrow \theta^{\text{new}}$ and go back to Step 2.

Contents

Clustering





Mixture of probability functions

- One way to approximate multimodal probability functions is through a mixture of probability functions.
- From the mixtures of probability functions, the Gaussian mixture model (GMM) is one of the best known,

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

where K is the number of components in the mixture and the parameters π_k are probabilities that satisfy

$$0 \leq \pi_k \leq 1, \quad \sum_{k=1}^{K} \pi_k = 1.$$

Latent variable $\mathbf{z}$

- Let $\mathbf{z}$ be a random binary vector of K dimensions with 1-of-K representation.
- The vector $\mathbf{z}$ can only take K states, according to which of its inputs is different from zero.
- The marginal distribution over $\mathbf{z}$ is given as

$$p(z_k = 1) = \pi_k.$$

- This distribution can be written in short as

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}.$$

Conditional distribution of **x** given **z**

- The conditional distribution of **x** given **z**, is Gaussian

$$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- It can also be expressed as,

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}$$

Marginal distribution for $\mathbf{x}$

- The joint distribution is given as $p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$.
- The marginal distribution for $\mathbf{x}$ follows as

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

- For each observed value $\mathbf{x}_n$ there is a corresponding latent variable $\mathbf{z}_n$.
- This is a new formulation of the mixture of distributions using latent variables, which allows working with the joint distribution $p(\mathbf{x}, \mathbf{z})$.

Conditional distribution of **z** given **x**

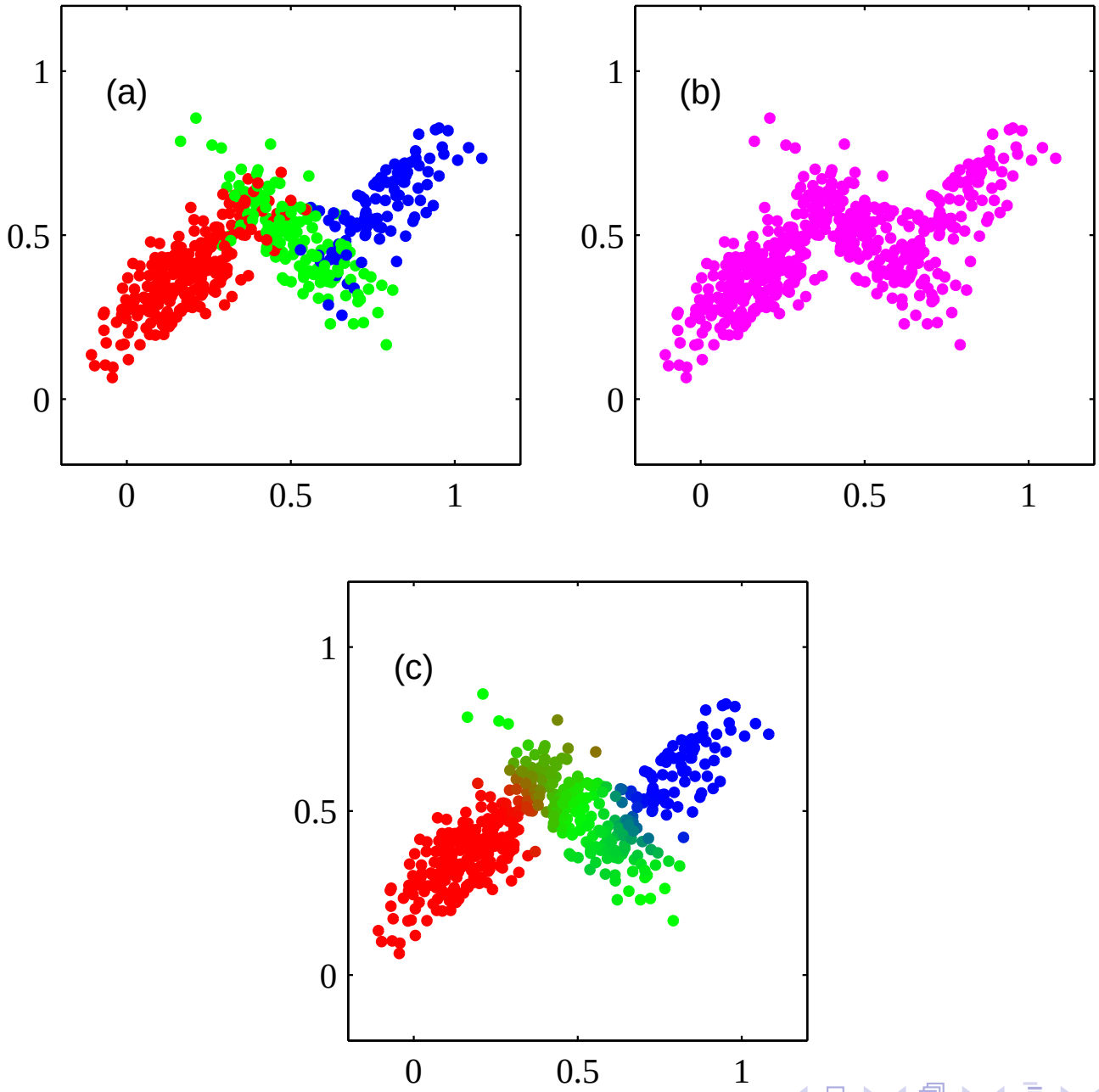
- Another quantity that plays an important role is the conditional probability of **z** given **x**.
- The probability is given as

$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)}$$

$$= \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

- The probability π_k can be seen as the prior probability $p(z_k = 1)$ and $\gamma(z_k)$ as the corresponding posterior probability once **x** has been observed.
- This quantity can be seen as the *responsibility* that the component k assumes to explain the observation **x**.

Incomplete and complete data



Log-likelihood function

- We start with a set of observations $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ that we want to model using a mixture of Gaussians.
- These observations are represented with a matrix $\mathbf{X}$ of dimensions $N \times D$ and row vectors $\mathbf{x}_n^\top$.
- Similarly, the corresponding latent variables are denoted by a matrix $\mathbf{Z}$ with row vectors $\mathbf{z}_n^\top$ and dimensions $N \times K$.
- The log-likelihood function is given as

$$\ln p(\mathbf{X}|\pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k) \right\}$$

- We want to find the parameters $\theta = \{\{\pi_k\}_{k=1}^K, \{\mu_k\}_{k=1}^K, \{\Sigma_k\}_{k=1}^K\}$, that maximize the likelihood of the incomplete data.

E step

- Starting with a value of θ^{old}, the posterior probability of the latent variables $\mathbf{Z}$ is calculated given the data $\mathbf{X}$ and the parameters θ^{old}.
- The probability function $p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}})$ has elements $\gamma(z_{n,k})$.
- The probabilities $\gamma(z_{n,k})$ are given as

$$\gamma(z_{n,k}) \equiv p(z_{n,k} = 1|\mathbf{x}_n) = \frac{p(z_{n,k} = 1)p(\mathbf{x}_n|z_{n,k} = 1)}{\sum_{j=1}^{K} p(z_{n,j} = 1)p(\mathbf{x}_n|z_{n,j} = 1)}$$

$$= \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

- $p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}})$ is a table of dimensions $N \times K$.

M step (I)

- We first compute the function $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$.
- $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$ is given as

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})]$$

- The complete likelihood function for the Gaussian mixture is given as

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{nk}}$$

- The log-likelihood is given as

$$\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \{\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}.$$

M step (II)

- $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$ is given as

$$\begin{aligned}\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) &= \mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})] \\ &= \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(z_{nk}) \left\{\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right\}.\end{aligned}$$

- Notice from the equation above that $\mathbb{E}_{\mathbf{Z}}[z_{nk}]$ corresponds to $\gamma(z_{nk})$,

$$\mathbb{E}_{\mathbf{Z}}[z_{nk}] = \sum_{z_{nk}} z_{nk} p(z_{nk}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) = p(z_{nk} = 1|\mathbf{x}_n, \boldsymbol{\theta}^{\text{old}}) = \gamma(z_{nk}).$$

- Given $\gamma(z_{nk})$, we maximize $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$ with respect to the parameters $\boldsymbol{\theta} = \{\{\pi_k\}_{k=1}^K, \{\boldsymbol{\mu}_k\}_{k=1}^K, \{\boldsymbol{\Sigma}_k\}_{k=1}^K\}$.

M step (III)

- Maximizing $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$ wrt π_k leads to

$$\pi_k^{\text{new}} = \frac{1}{N} \sum_{n=1}^{N} \gamma(z_{nk}) = \frac{N_k}{N},$$

  where $N_k = \sum_{n=1}^{N} \gamma(z_{nk})$.

- Maximizing $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$ wrt $\boldsymbol{\mu}_k$ leads to

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n$$

- Maximizing $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$ wrt $\boldsymbol{\Sigma}_k$ leads to

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}) (\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}})^{\top}.$$

EM algorithm for the GMM

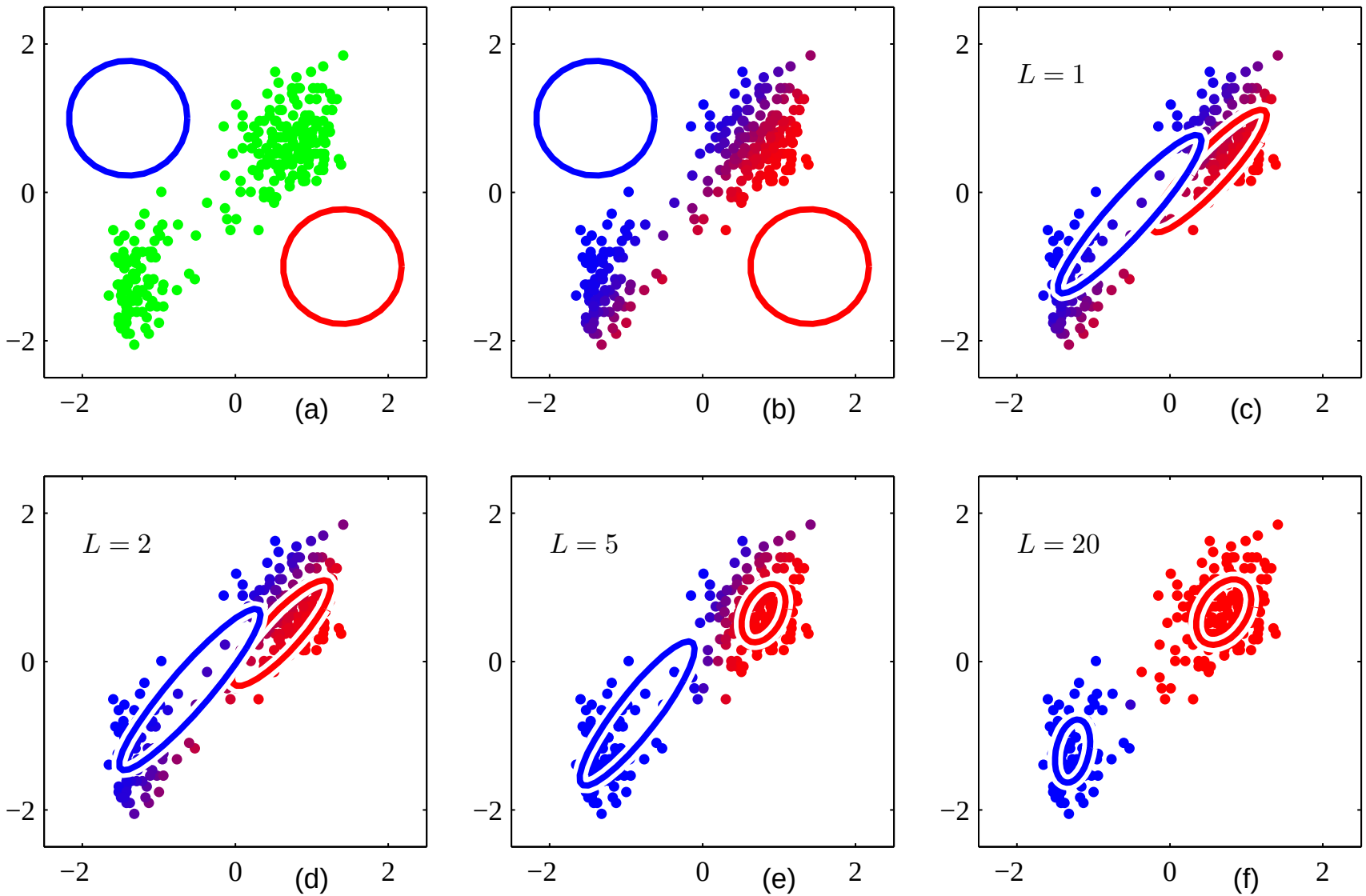
1. Initialize θ^{new}.

2. **E** step. Compute $\gamma(z_{nk})$, for $n = 1, \ldots, N$ y $k = 1, \ldots, K$.

3. **M** step. Use the update equations for π_k^{new}, μ_k^{new} and Σ_k^{new}, for $k = 1, \ldots, K$.

4. The convergence of the likelihood function or the parameters is verified. If the convergence criterion is not satisfied, then $\theta^{\text{old}} \leftarrow \theta^{\text{new}}$ and repeats from step 2.

Example



Contents

Infinite mixture models

- The simplest approach to (flat) clustering is to use a finite mixture model.
- The principle problem with finite mixture models is how to choose the number of components K.
- We discuss **infinite mixture models**, in which we do not impose any a priori bound on K.
- To do this, we will use a **non-parametric prior** based on the **Dirichlet process** (DP).
- This allows the number of clusters to grow as the amount of data increases.

From finite to infinite mixture models

- The representation for the finite mixture model that we saw before follows

$$p(\mathbf{x}_i|z_i = k, \theta) = p(\mathbf{x}_i|\theta_k)$$
$$p(z_i = k|\boldsymbol{\pi}) = \pi_k.$$

- We can additionally consider a distribution over the parameters π_k, e.g. the Dirichlet distribution

$$p(\boldsymbol{\pi}|\alpha) = \text{Dir}(\boldsymbol{\pi}|(\alpha/K)\mathbf{1}_K),$$

where $\alpha > 0$ and $\mathbf{1}_K$ is a vector of ones of dimension K.

Dirichlet distribution (I)

- A multivariate generalization of the beta distribution is the Dirichlet distribution.
- It has support over the **probability simplex**, defined by

$$S_K = \left\{ \mathbf{x} : 0 \leq x_k \leq 1, \sum_{k=1}^{K} x_k = 1 \right\}.$$

- The pdf is defined as follows:

$$\text{Dir}(\mathbf{x}|\boldsymbol{\alpha}) \triangleq \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} x_k^{\alpha_k - 1} \mathbb{I}(\mathbf{x} \in S_K),$$

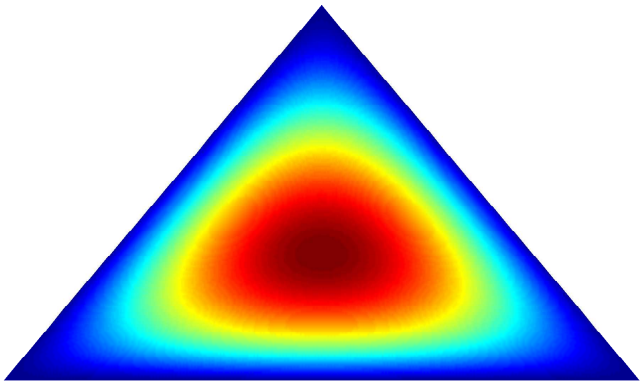
where $B(\alpha_1, \ldots, \alpha_K)$ is the natural generalization of the beta function to K variables

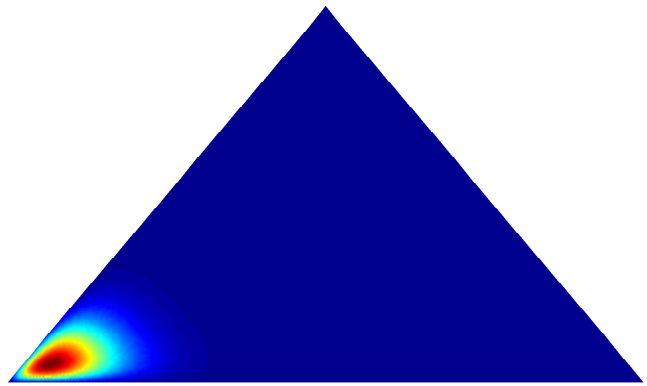
$$B(\boldsymbol{\alpha}) \triangleq \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\alpha_0)},$$

with $\alpha_0 \triangleq \sum_{k=1}^{K} \alpha_k$.

Dirichlet distribution (II)

Examples of the distribution





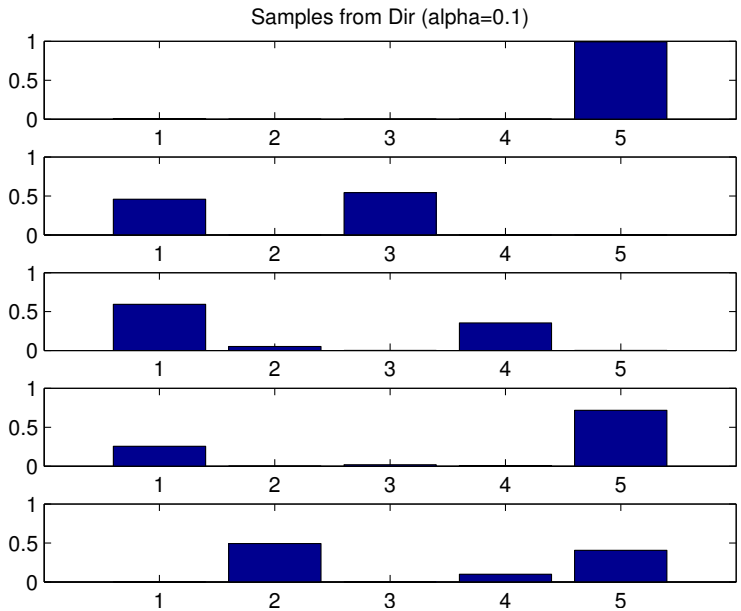
$$\boldsymbol{\alpha} = (2, 2, 2), \text{ y } \boldsymbol{\alpha} = (20, 2, 2).$$

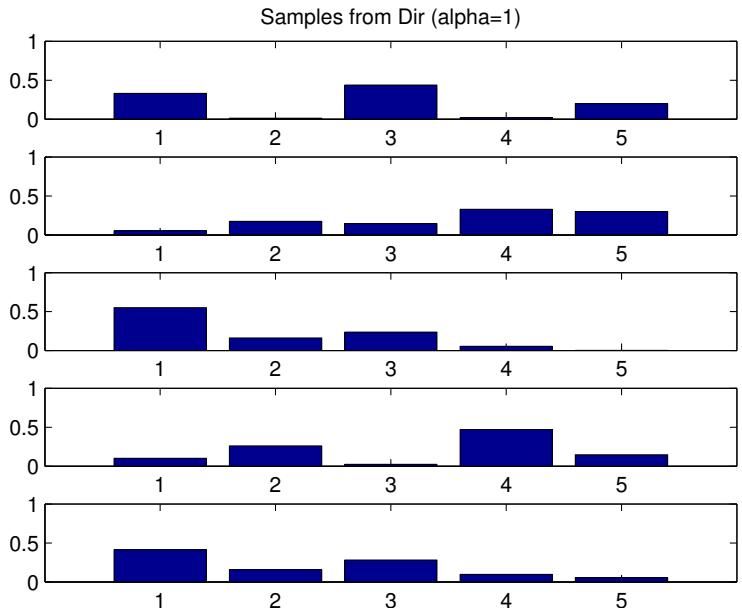
Properties

$$\mathbb{E}[x_k] = \frac{\alpha_k}{\alpha_0}, \text{mode}[x_k] = \frac{\alpha_k - 1}{\alpha_0 - K}, \text{var}[x_k] = \frac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0^2(\alpha_0 + 1)}$$

Often we use a symmetric Dirichlet prior of the form $\alpha_k = \alpha/K$. In this case, the mean becomes $1/K$ and the variance becomes $\text{var}[x_k] = \frac{K-1}{K^2(\alpha+1)}$.

Samples from a Dirichlet distribution

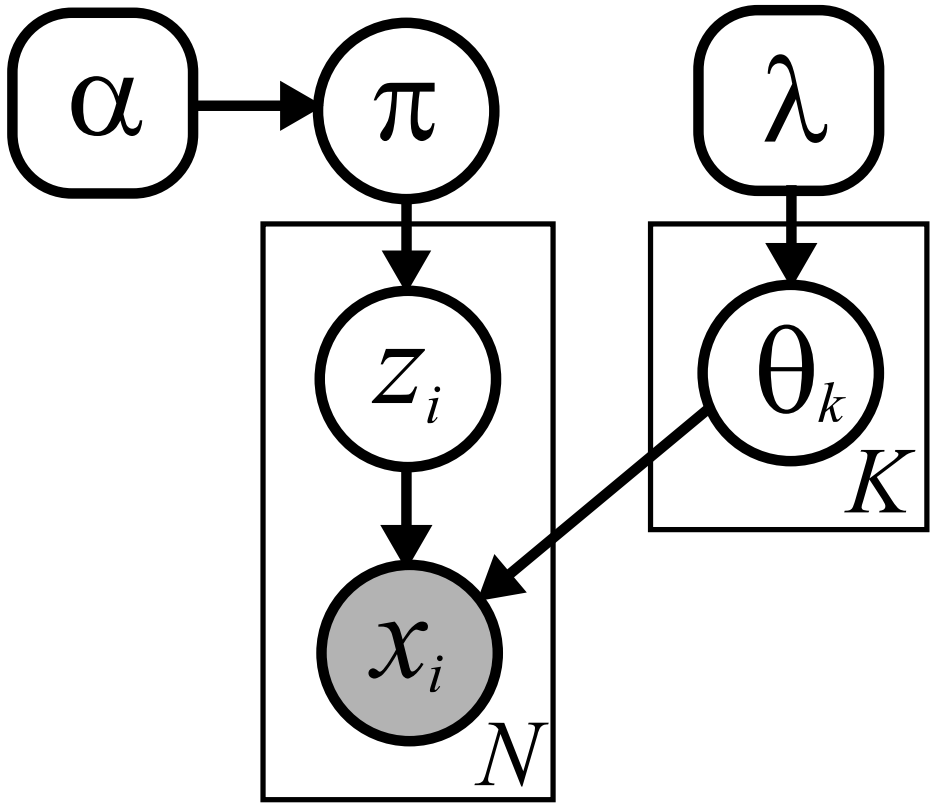




From finite to infinite mixture models

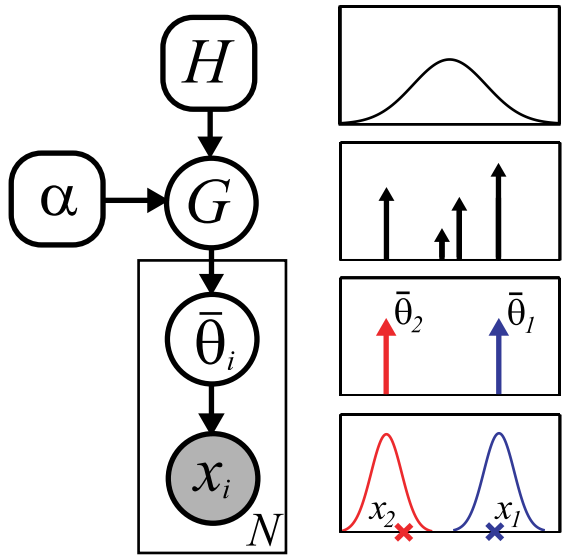
- We can also place priors over θ_k.
- The form of $p(\theta_k|\lambda)$ is chosen to be conjugate to $p(\mathbf{x}_i|z_i = k, \theta)$.
- We can write $p(\mathbf{x}_i|\theta_k)$ as $\mathbf{x}_i \sim F(\theta_{z_i})$, where F is the observation distribution.
- Similarly, we can write $\theta_k \sim H(\lambda)$ where H is the prior.

Graphical model



Alternative representation of the finite mixture model

- Another representation for the finite mixture model is shown here



- $\bar{\theta}_i$ is the parameter used to generate observation $\mathbf{x}_i$
- These parameters are sampled from distribution G,

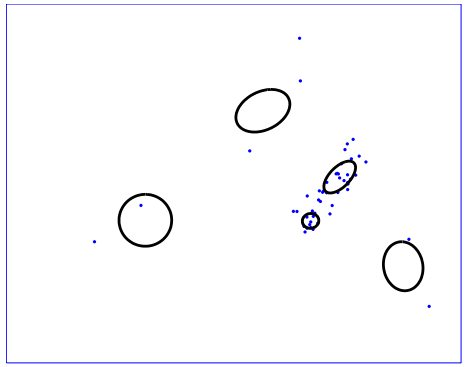
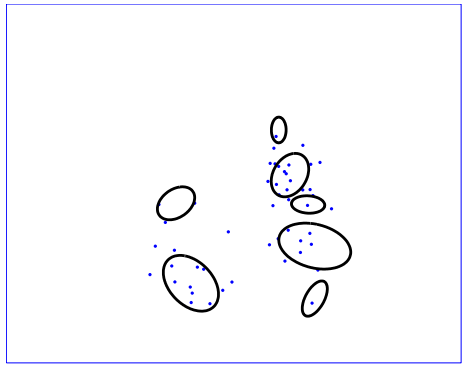
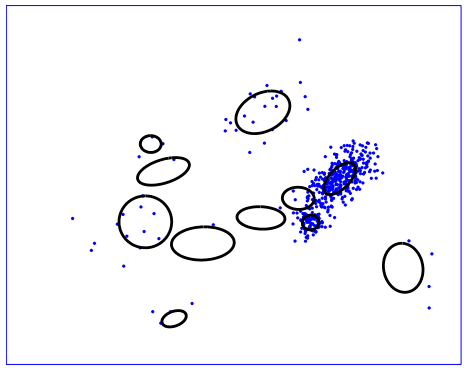
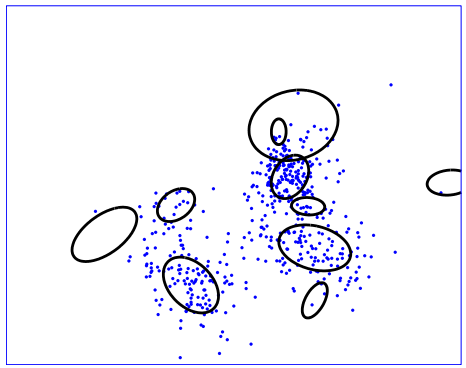
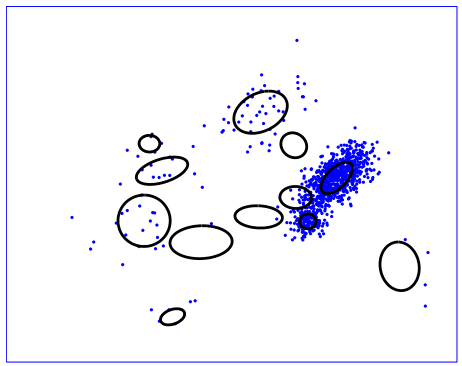
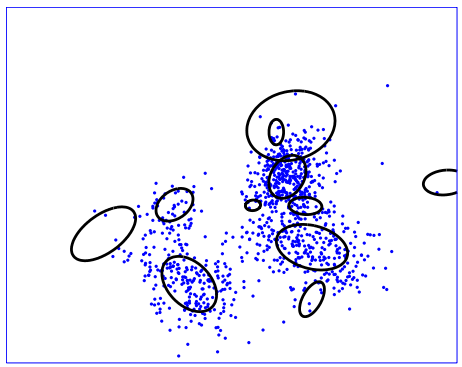
$$G(\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \delta_{\boldsymbol{\theta}_k}(\boldsymbol{\theta}),$$

  where $\boldsymbol{\pi} \sim \text{Dir}\left(\frac{\alpha}{K}\mathbf{1}\right)$ and $\boldsymbol{\theta}_k \sim H$.
- G is a finite mixture of delta functions, centered on $\boldsymbol{\theta}_k$.
- The probability that $\bar{\theta}_i$ is equal to $\boldsymbol{\theta}_k$ is exactly π_k, the prior probability for that cluster.

Alternative representation of the finite mixture model

- If we sample from this model, we will always get exactly K clusters, with data points scattered around the cluster centers.
- We would like a more flexible model, that can generate a variable number of clusters
  - the more data we generate, the more likely we should be to see a new cluster.
- The way to do this is to replace the discrete distribution G with a random probability measure.
- The Dirichlet process, denoted $G \sim \text{DP}(\alpha, H)$, is one way to do this.

Infinite mixture model ($N = 50, 500, 1000$)

The Dirichlet process

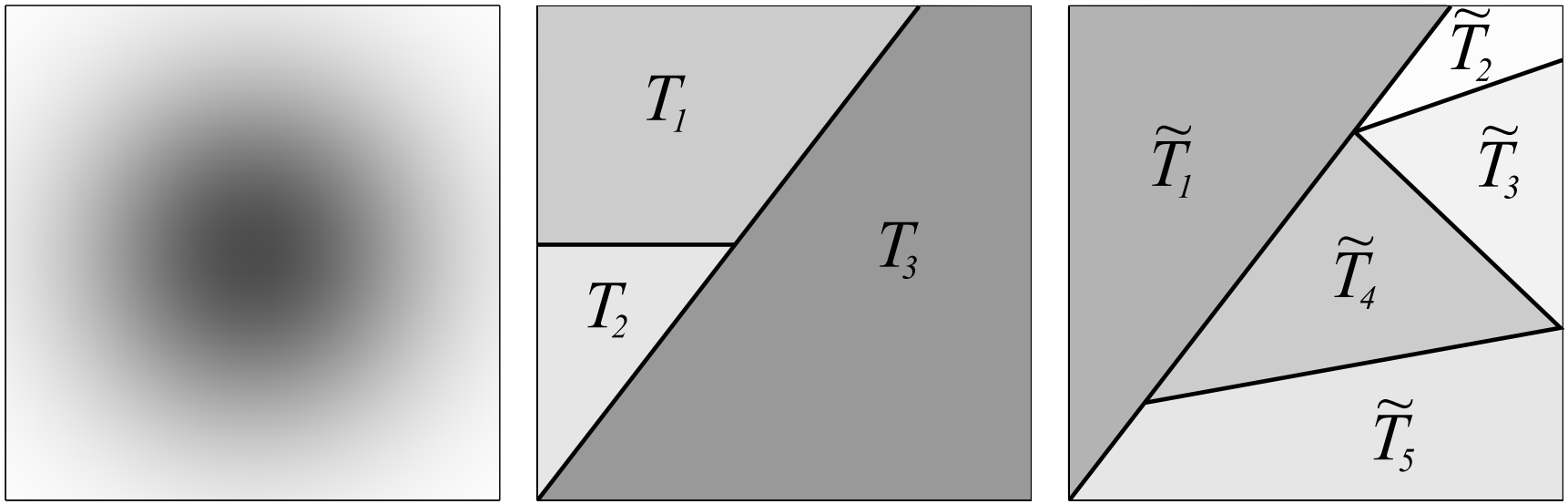
- A Dirichlet process is a distribution over probability measures $G : \Theta \to \mathbb{R}^+$.
- It is required that $G(\theta) \geq 0$ and $\int_\Theta G(\theta) d\theta = 1$.
- The DP is defined implicitly by the requirement that $(G(T_1), \ldots, G(T_K))$ has a joint Dirichlet distribution

$$\text{Dir}(\alpha H(T_1), \ldots, \alpha H(T_K)),$$

for any finite partition $(T_1, \ldots, T_K)$ of Θ.

- We write $G \sim \text{DP}(\alpha, H)$ where α is the **concentration parameter** and H is called the **base measure**.

Example



The base measure is a 2d Gaussian.

Dirichlet distribution as a prior distribution

- The Dirichlet distribution is commonly used as a prior over the parameters of a categorical or discrete distribution.

- The categorical distribution for a vector $\mathbf{z}$ with 1-of-K representation is given as

$$\mathrm{Cat}(\mathbf{z}|\boldsymbol{\pi}) = \prod_{j=1}^{K} \pi_j^{\mathbb{I}(z_j=1)},$$

  where $\pi_j = p(z_j = 1)$

- If $\boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\alpha})$, the updated posterior for $\boldsymbol{\pi}$ given one observation is given by

$$\boldsymbol{\pi}|z \sim \mathrm{Dir}\left(\alpha_1 + \mathbb{I}(z = 1), \ldots, \alpha_K + \mathbb{I}(z = K)\right).$$

Dirichlet process as a prior distribution

- The DP generalizes the concept we saw before to arbitrary partitions.
- If $G \sim \mathrm{DP}(\alpha, H)$, then $p(\theta \in T_i) = H(T_i)$ and the posterior is

$$\begin{aligned} & p(G(T_1), \ldots, G(T_K) \mid \theta, \alpha, H) \\ & = \mathrm{Dir}(\alpha H(T_1) + \mathbb{I}(\theta \in T_1), \ldots, \alpha H(T_K) + \mathbb{I}(\theta = T_K)) \end{aligned}$$

- This holds for any set of partitions.
- If there are multiple samples $\overline{\theta}_i \sim G$, the new posterior is given by

$$G \mid \overline{\theta}_1, \ldots, \overline{\theta}_N, \alpha, H \sim \mathrm{DP}\left(\alpha + N, \frac{1}{\alpha + N}\left(\alpha H + \sum_{i=1}^{N} \delta_{\theta_i}\right)\right).$$

- The DP effectively defines a conjugate prior for arbitrary measurable spaces.

Stick breaking construction of the DP

- The stick-breaking construction provides a concrete way to build a DP.

- Let $\pi = \{\pi_k\}_{k=1}^{\infty}$ be an infinite sequence of mixture weights derived from the following process:

$$\beta_k \sim \text{Beta}(1, \alpha)$$

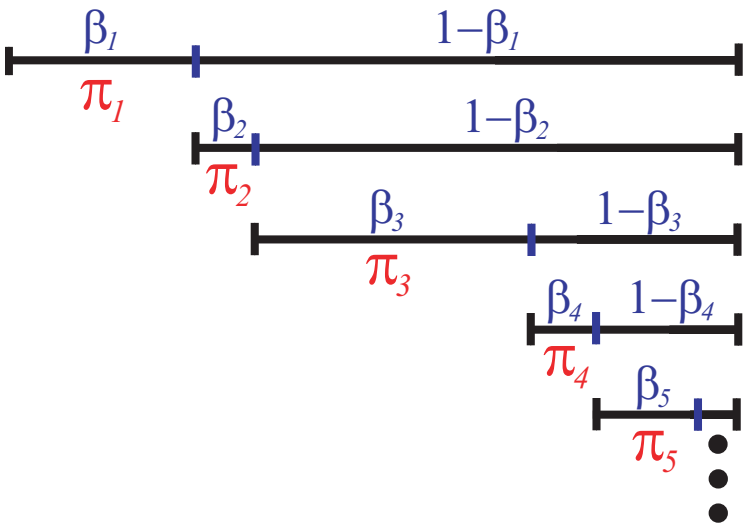
$$\pi_k = \beta_k \prod_{l=1}^{k-1} (1 - \beta_l) = \beta_k \left(1 - \sum_{l=1}^{k-1} \pi_l\right)$$

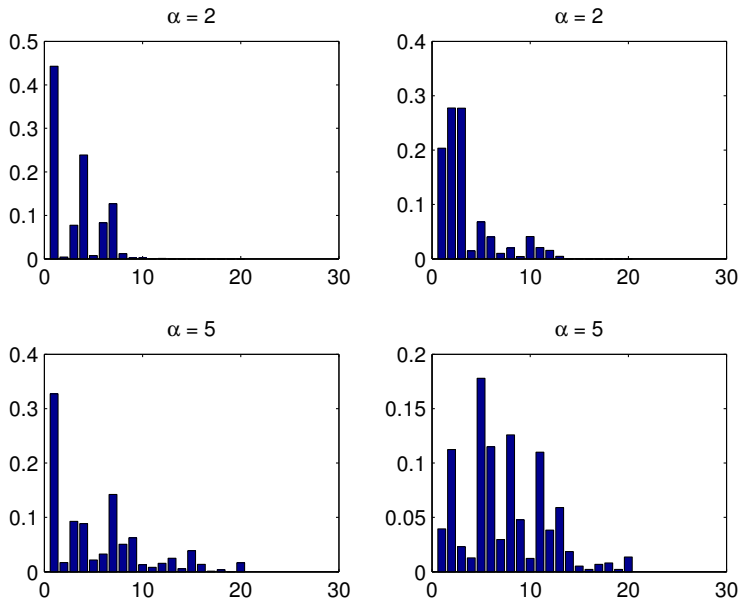
- This is often denoted by

$$\pi \sim \text{GEM}(\alpha)$$

where GEM stands for Griffiths, Engen and McCloskey (this term is due to (Ewens 1990)).

Illustration





Stick breaking construction of the DP

- We define

$$G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}(\theta),$$

where $\pi \sim \text{GEM}(\alpha)$ and $\theta_k \sim H$.

- Then one can show that $G \sim \text{DP}(\alpha, H)$.

- As a consequence of this construction, we see that samples from a DP are discrete with probability one.

- In other words, if you keep sampling it, you will get more and more repetitions of previously generated values.

Chinese restaurant processes (I)

- As the Dirichlet process assigns observations $\bar{\theta}_i$ to distinct values θ_k, it implicitly partitions the data.

- Let z_i indicate the subset, or cluster, associated with the i-th observation, so that $\bar{\theta}_i = \theta_{z_i}$.

- The predictive distribution of z_{N+1} given $\mathbf{z}_{1:N}$ is given as

$$p\left(z_{N+1} = z | \mathbf{z}_{1:N}, \alpha\right) = \frac{1}{\alpha + N}\left(\alpha \mathbb{I}\left(z = k^*\right) + \sum_{k=1}^{K} N_k \mathbb{I}(z = k)\right),$$

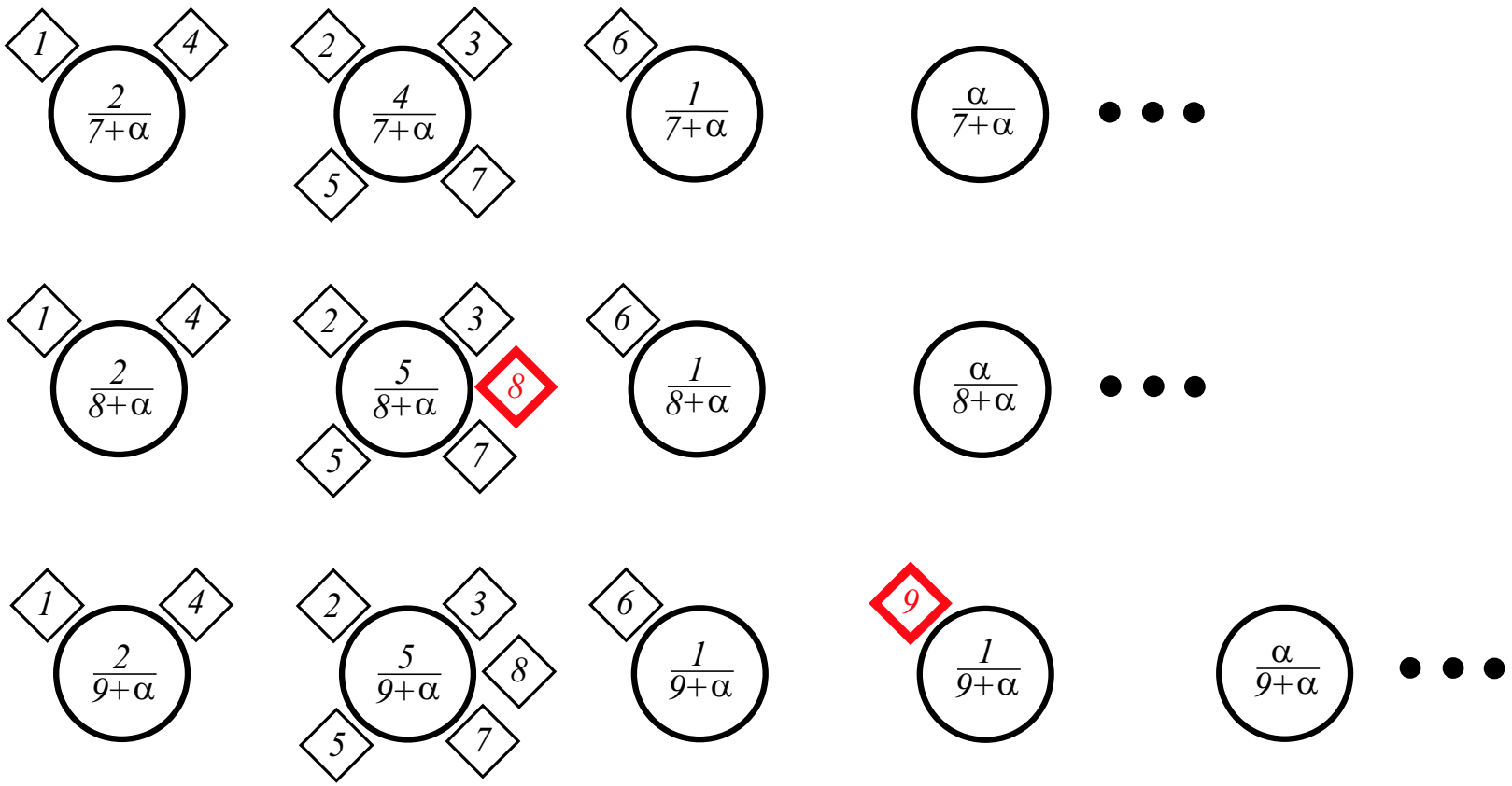
where N_k is the number of previous observations associated to the discrete variable $z = k$ and k^* represents a new cluster index that has not yet been used.

Chinese restaurant processes (II)

- Inspired by the seemingly infinite seating capacity of restaurants in San Francisco's Chinatown, Pitman and Dubins (2002) called this distribution over partitions the *Chinese restaurant process*.

- Dirichlet processes extend this construction by serving each table a different, independently chosen dish (parameter) θ_k.

Illustration



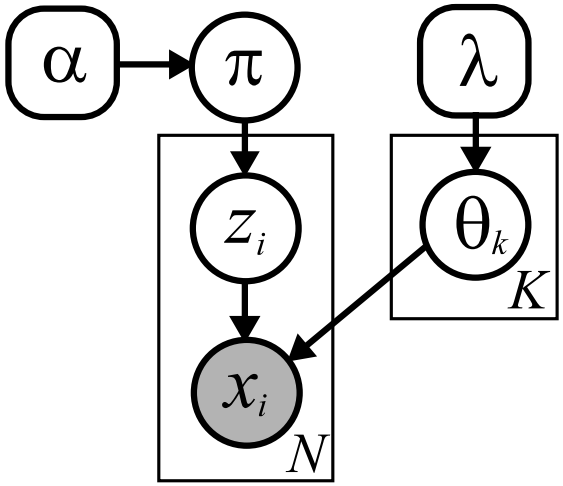
Taken from Erik B. Sudderth's PhD thesis (2006), figure 2.23.

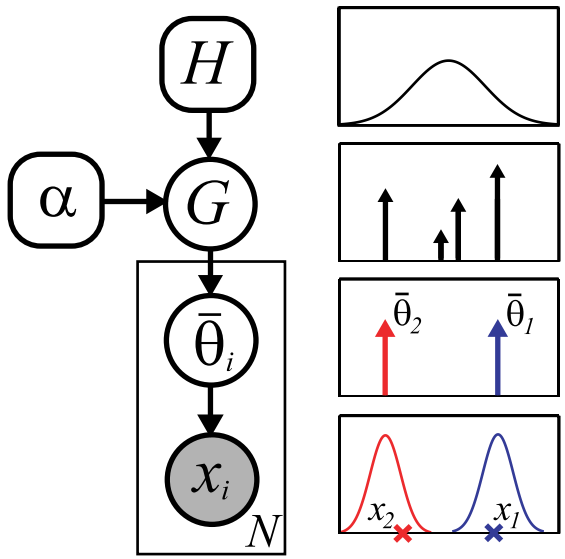
Applying Dirichlet processes to mixture modeling

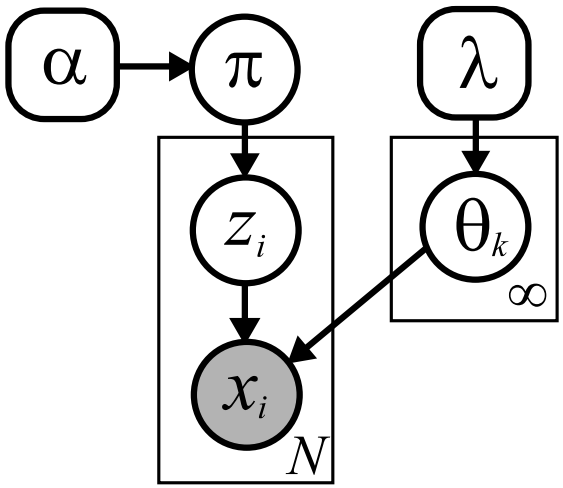
- The DP is not particularly useful as a model for data directly, since data vectors rarely repeat exactly.
- However, it is useful as a prior for the parameters of a stochastic data generating mechanism, such as a mixture model.
- We can write the model as follows

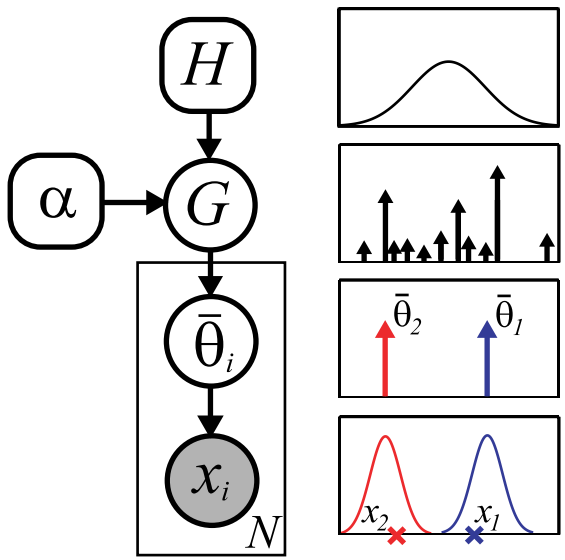
$$\begin{aligned}
\boldsymbol{\pi} &\sim \text{GEM}(\alpha) \\
z_i &\sim \boldsymbol{\pi} \\
\boldsymbol{\theta}_k &\sim H(\lambda) \\
\mathbf{x}_i &\sim F(\boldsymbol{\theta}_{z_i})
\end{aligned}$$

Finite vs infinite mixture models





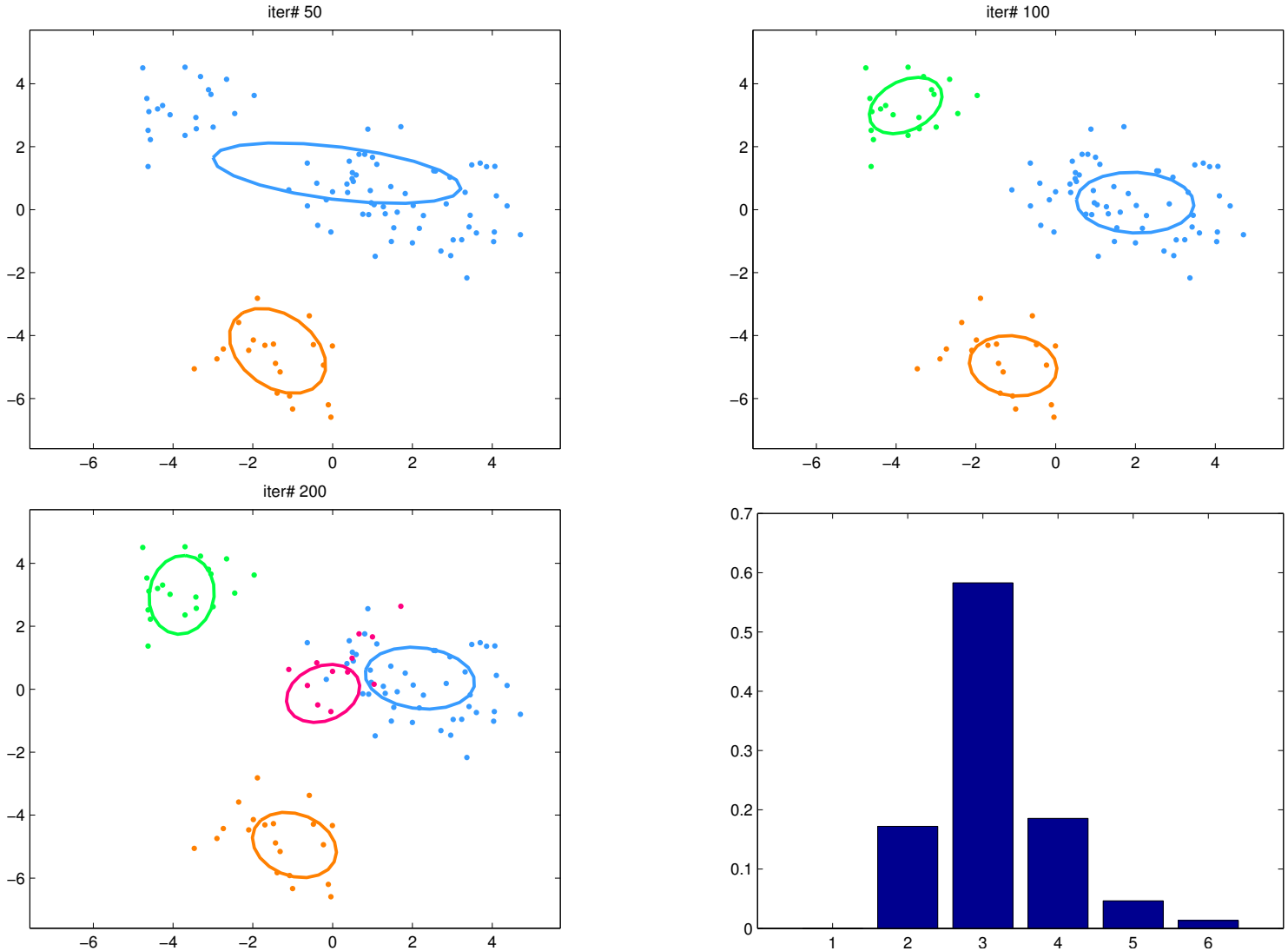




Fitting a DP mixture model

Given a dataset $\mathbf{X}$, we can uncover $\mathbf{Z}$ either by Gibbs sampling (Murphy, 2013) or variational inference (Blei and Jordan, 2005).

Example



100 data points in 2d are clustered using a DP mixture fit with collapsed Gibbs sampling. Samples from the posterior after 50,100, 200 samples.

Posterior over K, based on 200 samples, discarding the first 50 as burnin.

Contents

Introduction

- The original ME model can be viewed as a tree-structured architecture, based on the principle of divide and conquer.

- It has three main components:
  - several experts that are either regression functions or classifiers.
  - a gate that makes soft partitions of the input space and defines those regions where the individual expert opinions are trustworthy.
  - a probabilistic model to combine the experts and the gate.

Model (I)

- The basic expression for the ME is given as

$$p(y|\mathbf{x},\theta) = \sum_{k=1}^{K} \pi_k(\mathbf{x}|\theta) p_k(y|\mathbf{x},\theta),$$

  where $\pi_k(\mathbf{x})$ is known as *gating* functions and the individual component densities $p_k(\mathbf{y}|\mathbf{x})$ are called *experts*.

- More specifically,

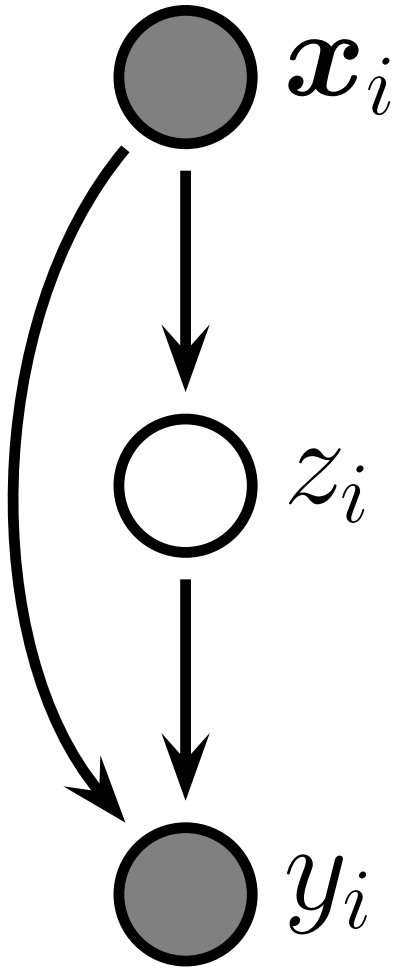
$$p\left(y|\mathbf{x},\theta\right) = \sum_{k=1}^{K} p\left(z=k|\mathbf{x},\theta\right) p\left(y|\mathbf{x},z=k,\theta\right).$$

- Different components can model the distribution in different regions (they are 'experts' at making predictions in their own regions).

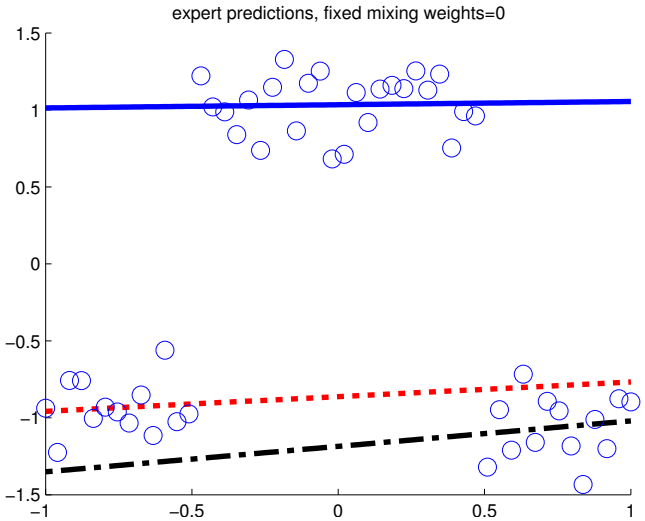
Model (II)

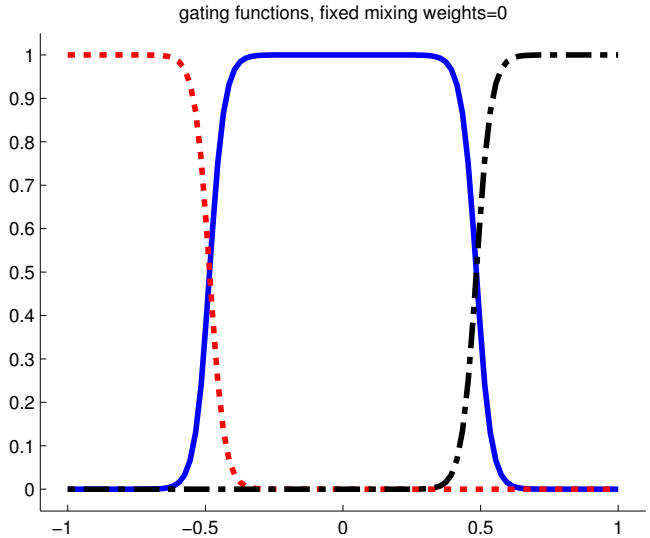
- The gating functions determine which components are dominant in which region.
- The gating functions $\pi_k(\mathbf{x})$ must satisfy the constraints for mixing coefficients,
  - $0 \leqslant \pi_k(\mathbf{x}) \leqslant 1$.
  - $\sum_k \pi_k(\mathbf{x}) = 1$.

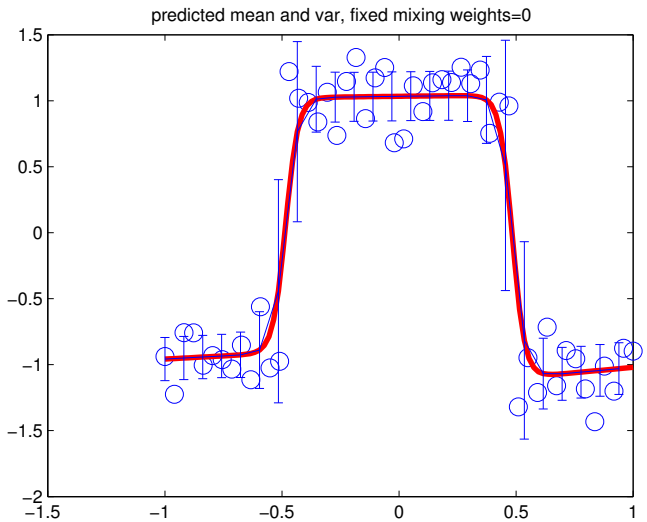
Graphical model



Linear models as experts



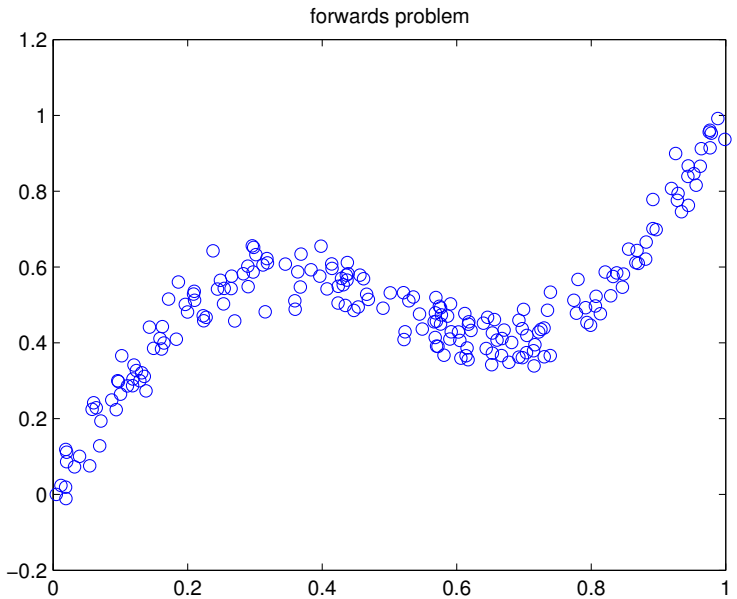




Some data fit with three separate regression lines; the gating functions for three different “experts” and the conditionally weighted average of the three expert predictions.

Application to inverse problems (I)

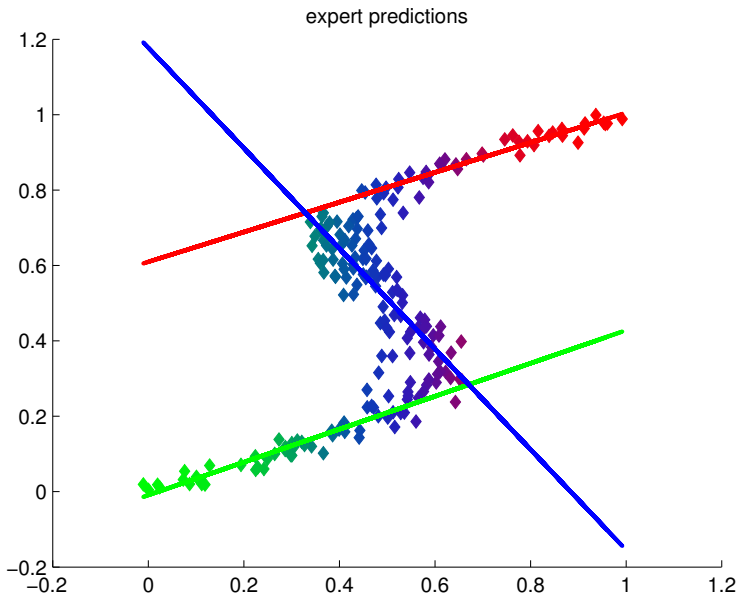
- Mixtures of experts are useful in solving **inverse problems**.
- These are problems where we have to invert a many-to-one mapping.
- This figure shows an example of a function $y = f(x)$: for every value x along the horizontal axis, there is a unique response y.



- This is sometimes called the **forwards model**.

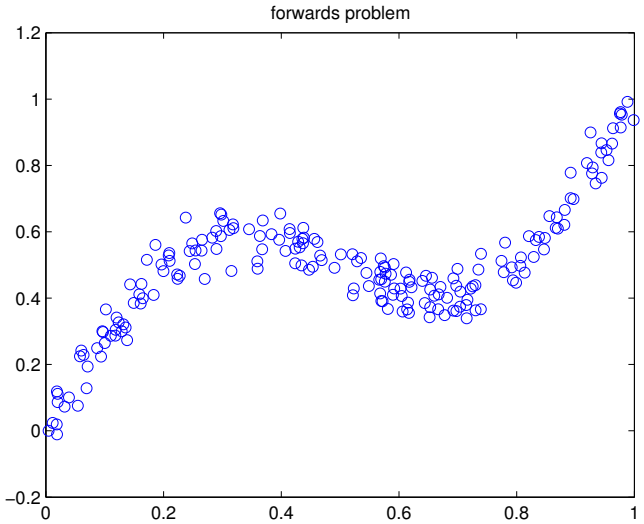
Application to inverse problems (II)

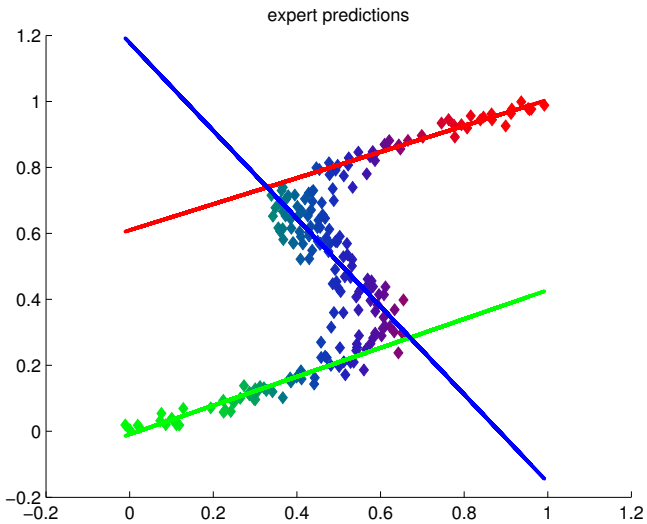
- Now consider the problem of computing $x = f^{-1}(y)$.
- The corresponding inverse model is shown in the following figure

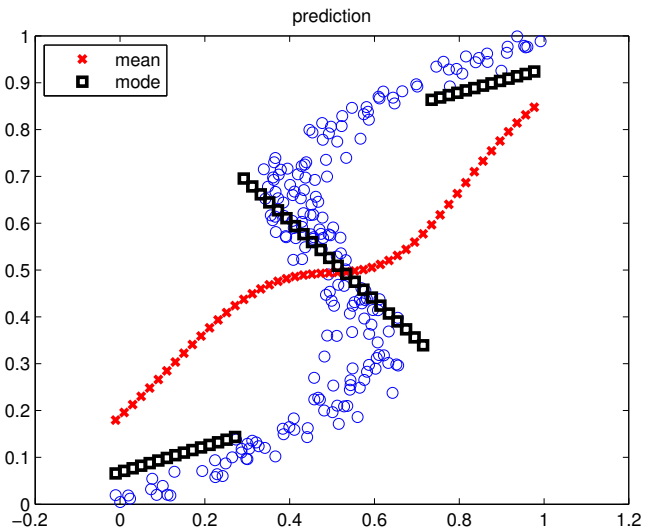


- The figure is obtained by simply interchanging the x and y axes.
- For some values along the horizontal axis, there are multiple possible outputs, so the inverse is not uniquely defined. For example, if $y = 0.6$, then x could be 0.2 or 0.8.
- Consequently, the predictive distribution, $p(x|y, \theta)$, is multimodal.
- Mixtures of experts are useful in solving **inverse problems**.

Fit of a mixture of linear experts to the data







- The posterior mean does not yield good predictions.
- However, the posterior mode, where the mode is input dependent, provides a reasonable approximation

Contents

First mixture of Gaussian processes

- The first mixture of GPs was proposed by Volker Tresp in 2001 (Tresp, 2001).
- There are $\{f_k^{\mu}(\mathbf{x})\}_{k=1}^K$ Gaussian process regression models used as experts.
- There are $\{f_k^{z}(\mathbf{x})\}_{k=1}^K$ used to compute the gating function $\pi_k(\mathbf{x})$.
- In particular, if there are K components and $\pi_i(\mathbf{x}) = p(z = i|\mathbf{x})$, then

$$p(z = i|\mathbf{x}) = \frac{\exp(f_i^z(\mathbf{x}))}{\sum_{j=1}^K \exp(f_j^z(\mathbf{x}))}$$

First mixture of Gaussian processes

- Furthermore, Tresp (2001) considers heteroscedastic GPs for the experts, such that

$$p_k(y|\mathbf{x}) = \mathcal{N}(y|f_k^{\mu}(\mathbf{x}), \exp(2f_k^{\sigma}(\mathbf{x}))).$$

- The ME is then given as

$$p(y|\mathbf{x}) = \sum_{k=1}^{K} p(z=k|\mathbf{x})\mathcal{N}(y|f_k^{\mu}(\mathbf{x}), \exp(2f_k^{\sigma}(\mathbf{x}))).$$

- The conditional expected value for y given $\mathbf{x}$ follows as

$$\mathbb{E}[y|\mathbf{x}] = \sum_{k=1}^{K} p(z=k|\mathbf{x}) f_k^{\mu}(\mathbf{x}).$$

Inference using the EM algorithm

- Tresp (2001) proposes an Expectation-Maximization (EM) algorithm for inference over the gating function and the experts.

- The negative log of the likelihood, including the log of the priors over the different Gaussian processes is given as

$$
\begin{aligned}
&-\sum_{n=1}^{N}\log\sum_{k=1}^{K}p\left(z=k|\mathbf{x}_n\right)\mathcal{N}\left(y_n|f_k^{\mu}\left(\mathbf{x}_n\right),\exp\left(2f_k^{\sigma}\left(\mathbf{x}_n\right)\right)\right)\\
&+\frac{1}{2}\sum_{i=1}^{K}\left(\mathbf{f}_i^{z,k}\right)^{\top}\left(\boldsymbol{\Sigma}_i^{z,k}\right)^{-1}\mathbf{f}_i^{z,k}+\frac{1}{2}\sum_{i=1}^{K}\left(\mathbf{f}_i^{\mu,k}\right)^{\top}\left(\boldsymbol{\Sigma}_i^{\mu,k}\right)^{-1}\mathbf{f}_i^{\mu,k}\\
&+\frac{1}{2}\sum_{i=1}^{K}\left(\mathbf{f}_i^{\sigma,k}\right)^{\top}\left(\boldsymbol{\Sigma}_i^{\sigma,k}\right)^{-1}\mathbf{f}_i^{\sigma,k},
\end{aligned}
$$

where $\mathbf{f}_i^{\mu,k}=\left(f_i^{\mu}\left(\mathbf{x}_1\right),\ldots,f_i^{\mu}\left(\mathbf{x}_N\right)\right)^{\top}$.

E step

- In the E step, we compute the posterior distribution $p(z = k|y_k, \mathbf{x}_k)$.

- It follows the same expression that we use for the responsabilities in a Gaussian mixture model, this is

$$\hat{p}\left(z = i|\mathbf{x}_n, y_n\right) = \frac{p\left(z = i|\mathbf{x}_n\right)\mathcal{N}\left(y_n|\hat{f}_i^{\mu}\left(\mathbf{x}_n\right), \exp\left(2\hat{f}_i^{\sigma}\left(\mathbf{x}_n\right)\right)\right)}{\sum_{j=1}^{K} p\left(z = j|\mathbf{x}_n\right)\mathcal{N}\left(y_n|\hat{f}_j^{\mu}\left(\mathbf{x}_n\right), \exp\left(2\hat{f}_j^{\sigma}\left(\mathbf{x}_n\right)\right)\right)}$$

M step

- In the M step, the Gaussian processes at the data points are updated.

- We obtain

$$\hat{\mathbf{f}}_i^{\mu,k} = \boldsymbol{\Sigma}_i^{\mu,k} \left(\boldsymbol{\Sigma}_i^{\mu,k} + \boldsymbol{\Psi}_i^{\mu,k} \right)^{-1} \mathbf{y}^k,$$

  where $\boldsymbol{\Psi}_i^{\mu,k}$ is a diagonal matrix with entries

$$\left(\boldsymbol{\Psi}_i^{\mu,k} \right)_{m,m} = \exp \left(2 \hat{f}_i^{\sigma} \left(\mathbf{x}_n \right) \right) / \hat{p} \left(z = i | x_n, y_n \right).$$

- To update the other Gaussian processes iterative Fisher scoring steps have to be used (Tresp, 2001).

Infinite mixture of Gaussian process experts

- Rasmussen and Ghahramani (2001) proposed a mixture of Gaussian process experts where the gating network is based on a Dirichlet process.
- To make the gating network input dependent, they use a kernel function to compute a parameter (the occupation number) in the Dirichlet process.
- The authors use Gibbs sampling for inference.

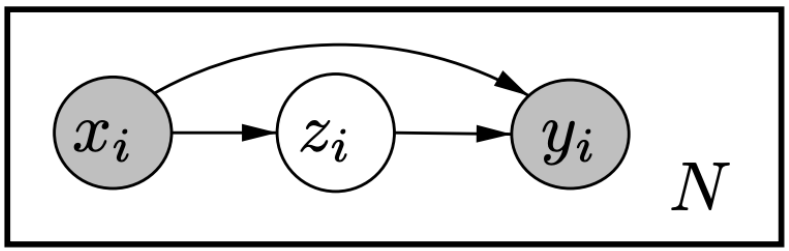
Gating function: discriminative vs generative

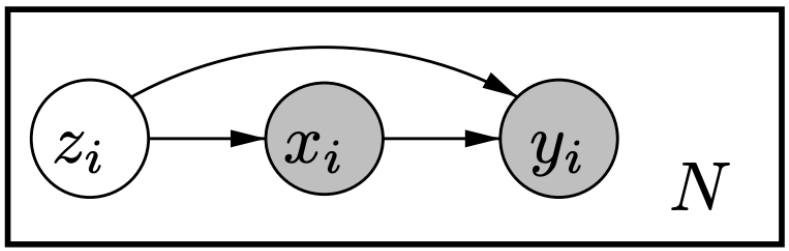
- In the mixture of experts, we have modeled the gating function as $p(z = k|\mathbf{x})$, e.g. as a conditional distribution.
- In ML, these types of models are usually referred to as discriminative.
- Alternatively, we can model both $p(z = k)$ and $p(\mathbf{x}|z = k)$, and compute the posterior distribution $p(z = k|\mathbf{x})$ to use as the gating function,

$$p(z = k|\mathbf{x}) = \frac{p(z = k)p(\mathbf{x}|z = k)}{\sum_j p(z = j)p(\mathbf{x}|z = j)}.$$

- The distribution of input locations is now given by a mixture model, with components for each expert.
- Conditioned on the input locations, the posterior responsibilities for each mixture component behave like a gating network.
- This approach is known as generative.

Gating function: discriminative vs generative





Advantages generative approach: deals with partially specified data and infer inverse functional mappings (from **x** to y).

Variational Mixture of GP Experts: experts

- Yuan and Neubauer (2009) proposed to use variational inference for a finite mixture of Gaussian process experts.

- A local Gaussian process expert is specified by the following linear model given the expert indicator $t = l$ (where $l = 1 : L$)

$$P\left(y|\mathbf{x}, t = l, \mathbf{v}_l, \boldsymbol{\theta}_l, \mathcal{I}_l, \gamma_l\right) = \mathcal{N}\left(y|\mathbf{v}_l^\top \boldsymbol{\phi}_l(\mathbf{x}), \gamma_l^{-1}\right),$$

where $\phi_l(\mathbf{x}) = \left[k_l\left(\mathbf{x}, \mathbf{x}_{\mathcal{I}_{l_1}}\right), k_l\left(\mathbf{x}, \mathbf{x}_{\mathcal{I}_{l_2}}\right), \ldots, k_l\left(\mathbf{x}, \mathbf{x}_{\mathcal{I}_{l_M}}\right)\right]^\top$ and $\{\mathcal{I}_{l_j}\}_{j=1}^M$ refer to indexes of the elements of the active set $\mathcal{I}_l$.

- The prior for $\mathbf{v}_l$ is given as $\mathcal{N}\left(\mathbf{v}_l|\mathbf{0}, \mathbf{U}_l^{-1}\right)$, where $\mathbf{U}_l = \mathbf{K}_l + \sigma_{hl}^2\mathbf{I}$ and $\mathbf{K}_l$ is the kernel matrix computed from the elements in the active set.

- $\boldsymbol{\theta}_l$ refers to the hyperparameters of the kernel.

Variational Mixture of GP Experts: experts

- The combination of $\mathcal{N}\left(y|\mathbf{v}_l^\top \phi_l(\mathbf{x}), \gamma_l^{-1}\right)$ and $\mathcal{N}\left(\mathbf{v}_l|\mathbf{0}, \mathbf{U}_l^{-1}\right)$ is equivalent to the subset of regressors approximation for sparse GPs.

- The prior of γ_l is set as a Gamma distribution.

Variational Mixture of GP Experts: gating function

- The gating functions are modeled as GMMs (generative approach)

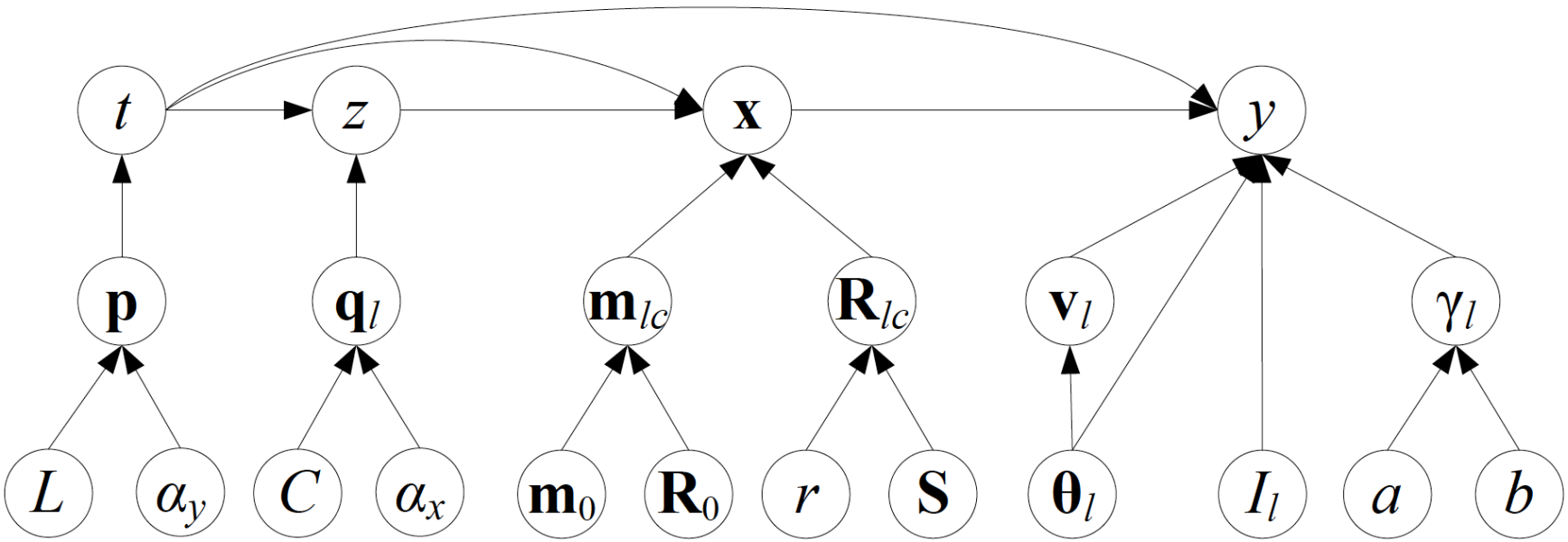
$$
\begin{aligned}
p(\mathbf{x}|t=l) &= \sum_{c=1}^{C} p(z=c|t=l,\mathbf{q}_l)\mathcal{N}(\mathbf{x}|\mathbf{m}_{lc},\mathbf{R}_{lc}^{-1}) \\
&= \sum_{c=1}^{C} q_{lc}\mathcal{N}(\mathbf{x}|\mathbf{m}_{lc},\mathbf{R}_{lc}^{-1}) \\
p(\mathbf{t}|\mathbf{p}) &= \mathrm{Cat}(\mathbf{t}|\mathbf{p}), \\
p(\mathbf{p}) &= \mathrm{Dir}\left(\mathbf{p}|\frac{\alpha_y}{L},\cdots,\frac{\alpha_y}{L}\right).
\end{aligned}
$$

- Additionally, the following priors are used

$$
\begin{aligned}
p(\mathbf{q}_l) &= \mathrm{Dir}\left(\mathbf{q}_l|\frac{\alpha_x}{C},\cdots,\frac{\alpha_x}{C}\right) \\
p(\mathbf{m}_{lc}) &= \mathcal{N}\left(\mathbf{m}_{lc}|\mathbf{m}_0,\mathbf{R}_0^{-1}\right) \\
p(\mathbf{R}_{lc}) &= \mathcal{W}\left(\mathbf{R}_{lc}|r,\mathbf{S}\right),
\end{aligned}
$$

where $\mathcal{W}(\cdot)$ is a Wishart distribution.

Variational Mixture of GP Experts: graphical model



Variational Mixture of GP Experts: mean field

- The authors use mean-field variational inference to compute the posterior distribution over the parameters $\mathbf{\Psi}$, expert indicators $\mathbf{T} = \{t_{1:N}\}$, and cluster indicators $\mathbf{Z} = \{z_{1:N}\}$.

- They use $\mathbf{\Omega} = \{\mathbf{\Psi}, \mathbf{T}, \mathbf{Z}\}$.

- The approximated posterior has the form

$$Q(\mathbf{\Omega}) = \prod_{l,c} Q(\mathbf{m}_{lc})\, Q(\mathbf{R}_{lc}) \prod_{l} Q(\mathbf{q}_l)\, Q(\mathbf{v}_l)\, Q(\gamma_l)\, Q(\mathbf{p}) \prod_{n} Q(t_n, z_n)\,.$$

- Expressions for all the posterior distributions above can be computed in closed form.

Variational inference for the infinite mixtures of GPs

- Sun and Xu (2011) proposed to use variational inference for computing posterior distributions in the model proposed by Rasmussen and Ghahramani (2001)

- They use a generative model for the gating functions similarly to Yuan and Neubauer (2009).

Variational inference for the infinite mixtures of GPs

- In practice, the model is pretty similar to the one proposed by Yuan and Neubauer (2009).
- The differences are:
  - Instead of using a GMM, they use a single Gaussian distribution for $p(\mathbf{x}|t = l)$. So instead of

$$p(\mathbf{x}|t = l) = \sum_{c=1}^{C} p(z = c|t = l, \mathbf{q}_l)\mathcal{N}(\mathbf{x}|\mathbf{m}_{lc}, \mathbf{R}_{lc}^{-1}),$$

    they use

$$p(\mathbf{x}|t = l) = \mathcal{N}(\mathbf{x}|\mathbf{m}_l, \mathbf{R}_l^{-1}).$$

  - Instead of using a Dirichlet distribution for $p(\mathbf{p})$, they use a Dirichlet process

$$p(\mathbf{p}) \sim DP(\alpha, H),$$

    where the base measure H is a Normal-Wishart distribution, similar to Yuan and Neubauer (2009).

Variational inference for the infinite mixtures of GPs

- Sun and Xu (2011) use the stick breaking construction for the Dirichlet process.
- They use mean field variational inference for computing the posterior distributions.
- The posterior distributions for the terms β_l in the stick breaking construction are approximated as in Blei and Jordan (2005).

Contents

Linear operators

- Since differentiation is a linear operator, the derivative of a Gaussian process is another Gaussian process.
- We can use GPs to either make predictions about derivatives, or to make inference based on derivative information.
- We can make inference based on the joint Gaussian distribution of function values and partial derivatives.

Kernel functions and inference

- The covariance between function values and partial derivatives and the covariance between partial derivatives are given as

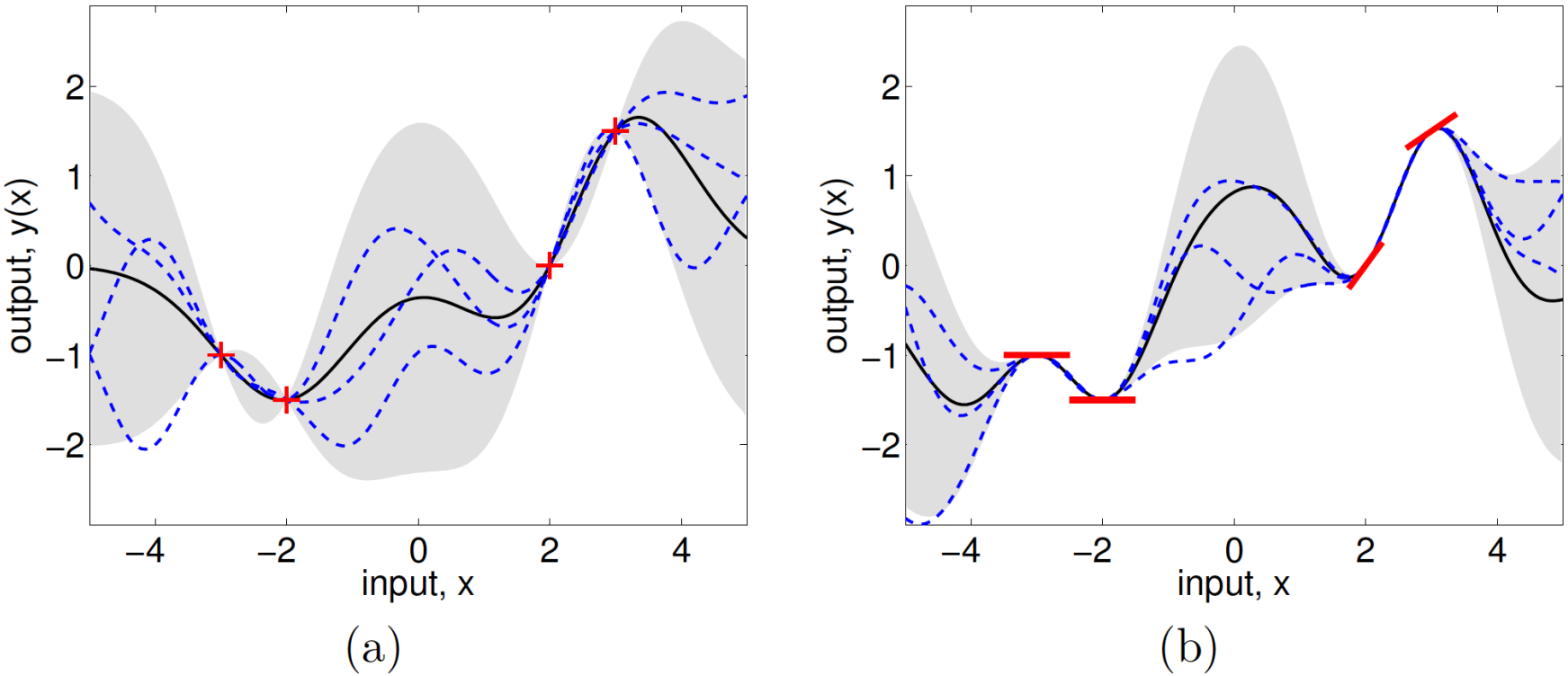
$$\operatorname{cov}\left(f_i, \frac{\partial f_j}{\partial x_{dj}}\right) = \frac{\partial k\left(\mathbf{x}_i, \mathbf{x}_j\right)}{\partial x_{dj}}, \quad \operatorname{cov}\left(\frac{\partial f_i}{\partial x_{di}}, \frac{\partial f_j}{\partial x_{ej}}\right) = \frac{\partial^2 k\left(\mathbf{x}_i, \mathbf{x}_j\right)}{\partial x_{di} \partial x_{ej}}$$

- Observed function values and derivatives may often have different noise levels

- We add a diagonal contribution with differing hyperparameters, one for the function and one for the derivative.

- Inference and predictions are done as usual.

Function and derivative evaluations



In panel (a) we show four data points in a one dimensional noise-free regression problem, together with three functions sampled from the posterior and the 95% confidence region in light grey. In panel (b) the same observations have been augmented by noise-free derivative information, indicated by small tangent segments at the data points. The covariance function is the squared exponential with unit process variance and unit length-scale. (Rasmussen and Williams, 2006)

References I

David M. Blei and Michael I. Jordan. Variational inference for dirichlet process mixtures. *Bayesian Analysis*, 1:121–144, 2005.

Kevin P. Murphy. *Machine learning : a probabilistic perspective*. MIT Press, Cambridge, Mass. [u.a.], 2013.

Carl Edward Rasmussen and Zoubin Ghahramani. Infinite mixtures of gaussian process experts. In *In Advances in Neural Information Processing Systems 14*, pages 881–888. MIT Press, 2001.

CE. Rasmussen and CKI. Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, USA, January 2006.

S. Sun and X. Xu. Variational inference for infinite mixtures of gaussian processes with applications to traffic flow prediction. *IEEE Transactions on Intelligent Transportation Systems*, 12(2):466–475, 2011. doi: 10.1109/TITS.2010.2093575.

Volker Tresp. Mixtures of gaussian processes. In *Advances in Neural Information Processing Systems 13*, pages 654–660. MIT Press, 2001.

Chao Yuan and Claus Neubauer. Variational mixture of gaussian process experts. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21, pages 1897–1904. Curran Associates, Inc., 2009. URL `https://proceedings.neurips.cc/paper/2008/file/f4b9ec30ad9f68f89b29639786cb62ef-Paper.pdf`.