RWTH AACHEN UNIVERSITY

Advanced Machine Learning
Lecture 10

Mixture Models II

30.11.2015

Bastian Leibe
RWTH Aachen
http://www.vision.rwth-aachen.de/

leibe@vision.rwth-aachen.de

Announcement

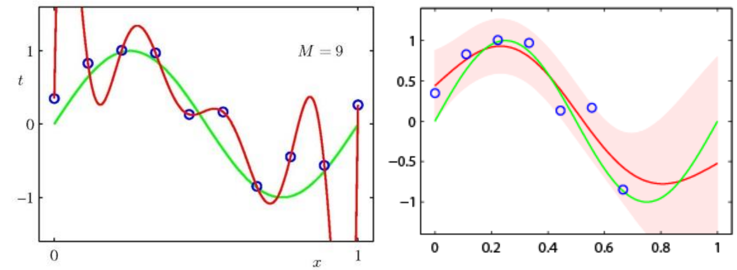
- Exercise sheet 2 online
  - Sampling
  - Rejection Sampling
  - Importance Sampling
  - Metropolis-Hastings
  - EM
  - Mixtures of Bernoulli distributions [today's topic]
  - Exercise will be on Wednesday, 07.12.

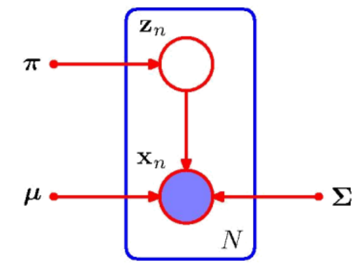
  $\Rightarrow$ *Please submit your results until 06.12. midnight.*

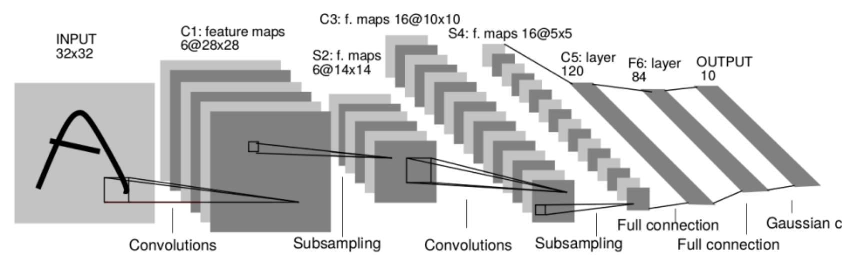
This Lecture: *Advanced Machine Learning*

- Regression Approaches
  - Linear Regression
  - Regularization (Ridge, Lasso)
  - Gaussian Processes
- Learning with Latent Variables
  - Probability Distributions
  - Approximate Inference
  - Mixture Models
  - EM and Generalizations
- Deep Learning
  - Neural Networks
  - CNNs, RNNs, RBMs, etc.

$$f : \mathcal{X} \to \mathbb{R}$$







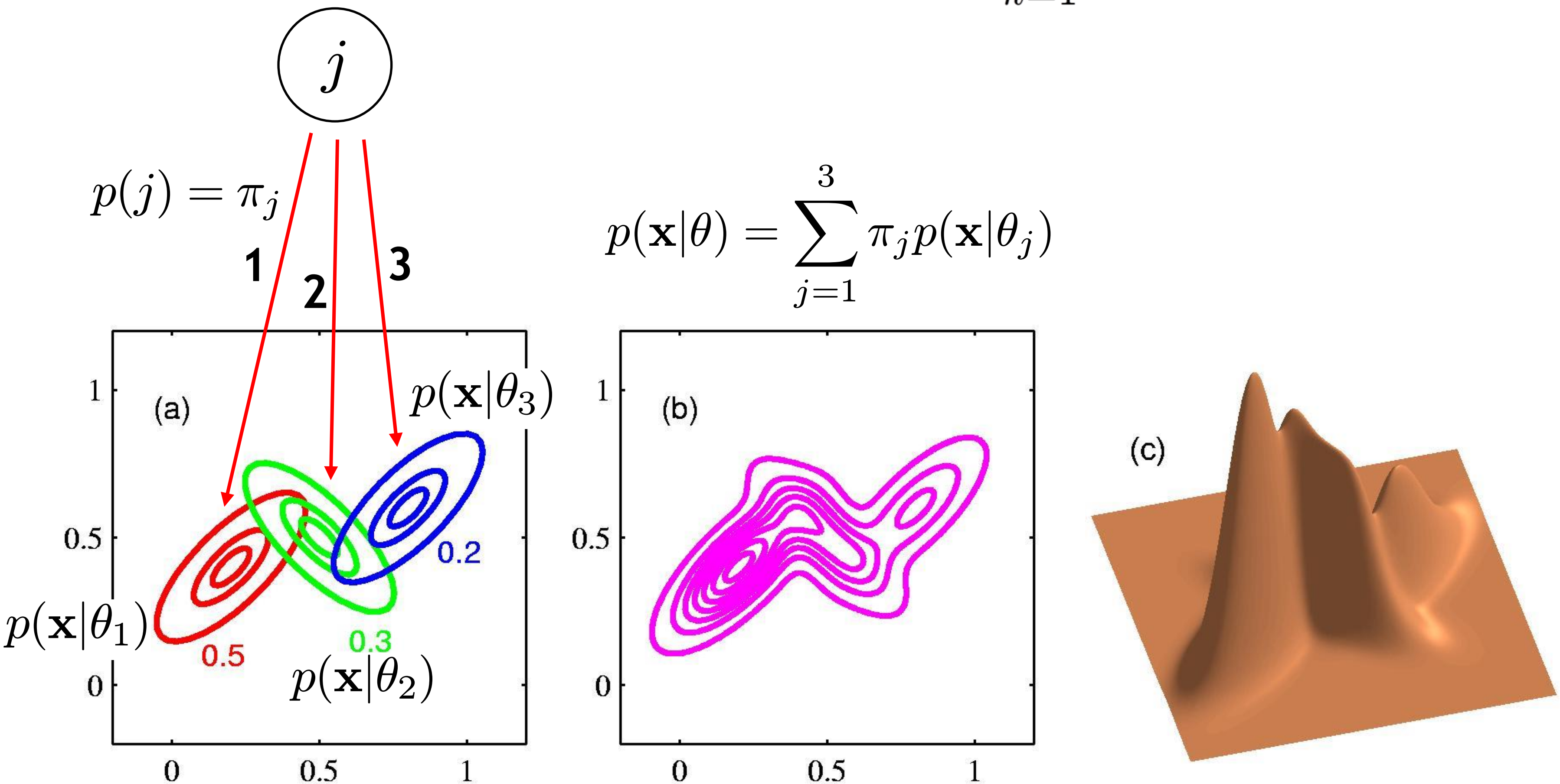
Topics of This Lecture

- The EM algorithm in general
  - Recap: General EM
  - Example: Mixtures of Bernoulli distributions
  - Monte Carlo EM
- Bayesian Mixture Models
  - Towards a full Bayesian treatment
  - Dirichlet priors
  - Finite mixtures
  - Infinite mixtures
  - Approximate inference (only as supplementary material)

Recap: Mixture of Gaussians

- **"Generative model"**

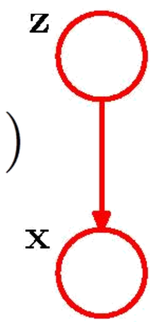
$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$



Recap: GMMs as Latent Variable Models

- Write GMMs in terms of latent variables $\mathbf{z}$
  - Marginal distribution of $\mathbf{x}$

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

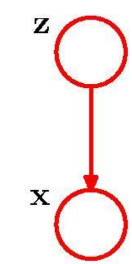


- Advantage of this formulation
  - We have represented the marginal distribution in terms of latent variables $\mathbf{z}$.
  - Since $p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z})$, there is a corresponding latent variable $\mathbf{z}_n$ for each data point $\mathbf{x}_n$.
  - We are now able to work with the joint distribution $p(\mathbf{x}, \mathbf{z})$ instead of the marginal distribution $p(\mathbf{x})$.

$\Rightarrow$ This will lead to significant simplifications…

Recap: Sampling from a Gaussian Mixture

- **MoG Sampling**
  - We can use ancestral sampling to generate random samples from a Gaussian mixture model.
    1. Generate a value $\hat{\mathbf{z}}$ from the marginal distribution $p(\mathbf{z})$.
    2. Generate a value $\hat{\mathbf{x}}$ from the conditional distribution $p(\mathbf{x}|\hat{\mathbf{z}})$.



Samples from the joint $p(\mathbf{x}, \mathbf{z})$ | Samples from the marginal $p(\mathbf{x})$ | Evaluating the responsibilities $\gamma(z_{nk})$

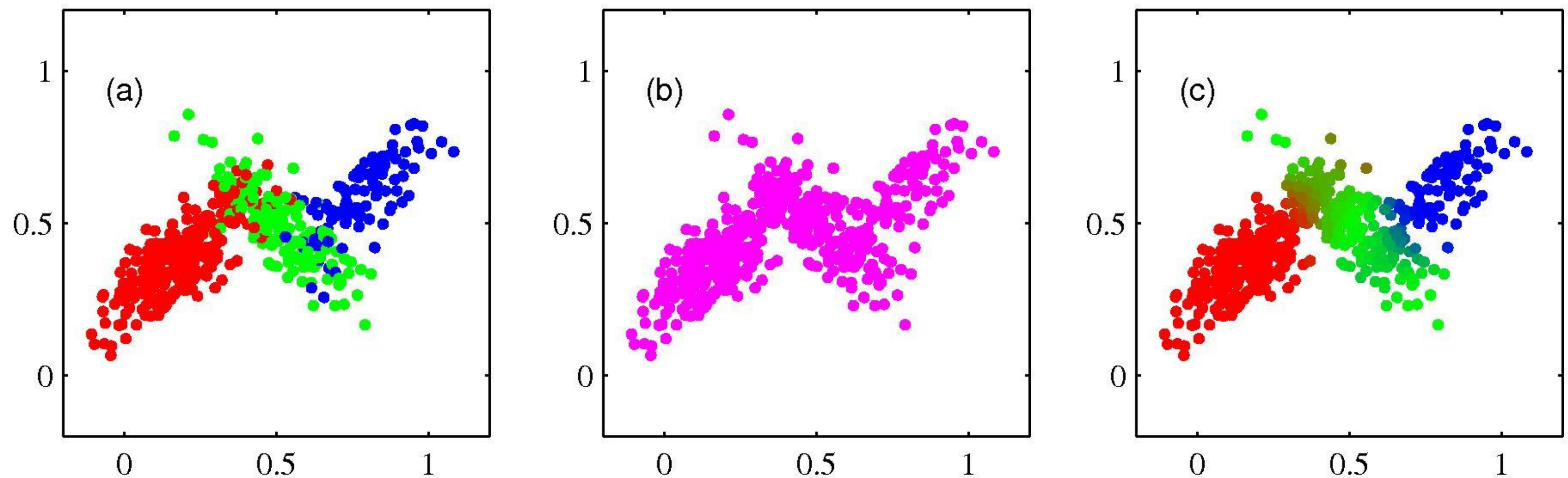


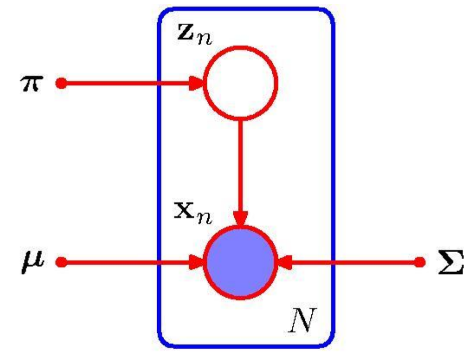
Image source: C.M. Bishop, 2006

Recap: Gaussian Mixtures Revisited

- **Applying the latent variable view of EM**
  - Goal is to maximize the log-likelihood using the observed data $\mathbf{X}$

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

  - Corresponding graphical model:



  - Suppose we are additionally given the values of the latent variables $\mathbf{Z}$.
  - The corresponding graphical model for the complete data now looks like this:

$\Rightarrow$ Straightforward to marginalize…

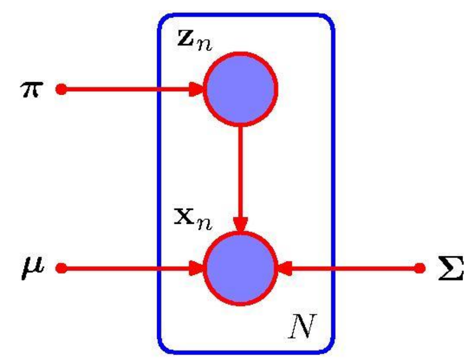


Image source: C.M. Bishop, 2006

Recap: Alternative View of EM

- In practice, however,...
  - We are not given the complete data set $\{\mathbf{X},\mathbf{Z}\}$, but only the incomplete data $\mathbf{X}$. All we can compute about $\mathbf{Z}$ is the posterior distribution $p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})$.
  - Since we cannot use the complete-data log-likelihood, we consider instead its expected value under the posterior distribution of the latent variable:

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

  - This corresponds to the E-step of the EM algorithm.
  - In the subsequent M-step, we then maximize the expectation to obtain the revised parameter set $\boldsymbol{\theta}^{\mathrm{new}}$.

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

Recap: General EM Algorithm

- Algorithm
  1. Choose an initial setting for the parameters $\boldsymbol{\theta}^{\text{old}}$
  2. E-step: Evaluate $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$
  3. M-step: Evaluate $\boldsymbol{\theta}^{\text{new}}$ given by

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

     where

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

  4. While not converged, let $\boldsymbol{\theta}^{\text{old}} \leftarrow \boldsymbol{\theta}^{\text{new}}$ and return to step 2.

Recap: MAP-EM

- **Modification for MAP**
  - The EM algorithm can be adapted to find MAP solutions for models for which a prior $p(\boldsymbol{\theta})$ is defined over the parameters.
  - Only changes needed:

2. E-step: Evaluate $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\mathrm{old}})$

3. M-step: Evaluate $\boldsymbol{\theta}^{\mathrm{new}}$ given by

$$\boldsymbol{\theta}^{\mathrm{new}} = \arg\max_{\boldsymbol{\theta}} \; \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{old}}) + \log p(\boldsymbol{\theta})$$

$\Rightarrow$ Suitable choices for the prior will remove the ML singularities!

Gaussian Mixtures Revisited

- **Maximize the likelihood**
  - For the complete-data set $\{\mathbf{X},\mathbf{Z}\}$, the likelihood has the form

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

  - Taking the logarithm, we obtain

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

  - Compared to the incomplete-data case, the order of the sum and logarithm has been interchanged.

  $\Rightarrow$ Much simpler solution to the ML problem.
  - Maximization w.r.t. a mean or covariance is exactly as for a single Gaussian, except that it involves only the subset of data points that are “assigned” to that component.

Gaussian Mixtures Revisited

- **Maximization w.r.t. mixing coefficients**
  - More complex, since the π_k are coupled by the summation constraint

$$\sum_{j=1}^{K} \pi_j = 1$$

  - Solve with a Lagrange multiplier

$$\log p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right)$$

  - Solution (after a longer derivation):

$$\pi_k = \frac{1}{N} \sum_{n=1}^{N} z_{nk}$$

$\Rightarrow$ The complete-data log-likelihood can be maximized trivially in closed form.

Gaussian Mixtures Revisited

- In practice, we don't have values for the latent variables
  - Consider the expectation w.r.t. the posterior distribution of the latent variables instead.
  - The posterior distribution takes the form

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

and factorizes over n, so that the $\{\mathbf{z}_n\}$ are independent under the posterior.

Expected value of indicator variable z_{nk} under the posterior.

$$\begin{aligned}\mathbb{E}[z_{nk}] &= \frac{\sum_{z_{nk}} z_{nk} \left[\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right]^{z_{nk}}}{\sum_{z_{nj}} \left[\pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\right]^{z_{nj}}} \\ &= \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} = \gamma(z_{nk})\end{aligned}$$

Gaussian Mixtures Revisited

- Continuing the estimation
  - The complete-data log-likelihood is therefore

$$\mathbb{E}_{\mathbf{Z}}[\log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})] = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \left\{ \log \pi_k + \log \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

$\Rightarrow$ This is precisely the EM algorithm for Gaussian mixtures as derived before.

Summary So Far

- We have now seen a generalized EM algorithm
  - Applicable to general estimation problems with latent variables
  - In particular, also applicable to mixtures of other base distributions
  - In order to get some familiarity with the general EM algorithm, let's apply it to a different class of distributions…

Topics of This Lecture

- The EM algorithm in general
  - Recap: General EM
  - Example: Mixtures of Bernoulli distributions
  - Monte Carlo EM
- Bayesian Mixture Models
  - Towards a full Bayesian treatment
  - Dirichlet priors
  - Finite mixtures
  - Infinite mixtures
  - Approximate inference (only as supplementary material)

Mixtures of Bernoulli Distributions

- Discrete binary variables
  - Consider D binary variables $\mathbf{x} = (x_1,\ldots,x_D)^T$, each of them described by a Bernoulli distribution with parameter μ_i, so that

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{i=1}^{D} \mu_i^{x_i}(1-\mu_i)^{(1-x_i)}$$

  - Mean and covariance are given by

$$\begin{aligned} \mathbb{E}[\mathbf{x}] &= \boldsymbol{\mu} \\ \operatorname{cov}[\mathbf{x}] &= \operatorname{diag}\{\boldsymbol{\mu}(1-\boldsymbol{\mu})\} \end{aligned}$$

Diagonal covariance
$\Rightarrow$ variables independently modeled

Mixtures of Bernoulli Distributions

- **Mixtures of discrete binary variables**
  - Now, consider a finite mixture of those distributions

$$
\begin{aligned}
p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\pi}) &= \sum_{k=1}^{K} \pi_k p(\mathbf{x}|\boldsymbol{\mu}_k) \\
&= \sum_{k=1}^{K} \pi_k \prod_{i=1}^{D} \mu_{ki}^{x_i} (1-\mu_{ki})^{(1-x_i)}
\end{aligned}
$$

  - Mean and covariance of the mixture are given by

$$
\begin{aligned}
\mathbb{E}[\mathbf{x}] &= \sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_k \\
\mathrm{cov}[\mathbf{x}] &= \sum_{k=1}^{K} \pi_k \left\{ \boldsymbol{\Sigma}_k + \boldsymbol{\mu}_k \boldsymbol{\mu}_k^T \right\} - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}]^T
\end{aligned}
$$

where $\boldsymbol{\Sigma}_k = \mathrm{diag}\{\mu_{ki}(1 - \mu_{ki})\}$.

Covariance not diagonal
$\Rightarrow$ Model can capture dependencies between variables

Mixtures of Bernoulli Distributions

- Log-likelihood for the model
  - Given a data set $\mathbf{X} = \{\mathbf{x}_1,\ldots,\mathbf{x}_N\}$,

$$\log p(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\pi}) = \sum_{n=1}^{N} \log \left\{ \sum_{k=1}^{K} \pi_k p(\mathbf{x}_n|\boldsymbol{\mu}_k) \right\}$$

  - Again observation: summation inside logarithm $\Rightarrow$ difficult.
  - In the following, we will derive the EM algorithm for mixtures of Bernoulli distributions.
    - This will show how we can derive EM algorithms in the general case…

EM for Bernoulli Mixtures

- Latent variable formulation
  - Introduce latent variable $\mathbf{z} = (z_1,\ldots,z_K)^T$ with 1-of-K coding.
  - Conditional distribution of $\mathbf{x}$:

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

  - Prior distribution for the latent variables

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

  - Again, we can verify that

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

Recap: General EM Algorithm

- Algorithm
  1. Choose an initial setting for the parameters $\boldsymbol{\theta}^{\text{old}}$
  2. E-step: Evaluate $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$
  3. M-step: Evaluate $\boldsymbol{\theta}^{\text{new}}$ given by

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

  where

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

  4. While not converged, let $\boldsymbol{\theta}^{\text{old}} \leftarrow \boldsymbol{\theta}^{\text{new}}$ and return to step 2.

EM for Bernoulli Mixtures: E-Step

- **Complete-data likelihood**

$$
\begin{aligned}
p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \boldsymbol{\pi}) &= \prod_{n=1}^{N} \prod_{k=1}^{K} \left[\pi_k p(\mathbf{x}_n | \boldsymbol{\mu}_k)\right]^{z_{nk}} \\
&= \prod_{n=1}^{N} \prod_{k=1}^{K} \left\{ \pi_k \prod_{i=1}^{D} \mu_{ki}^{x_{ni}} (1 - \mu_{ki})^{(1-x_{ni})} \right\}^{z_{nk}}
\end{aligned}
$$

- **Posterior distribution of the latent variables Z**

$$
\begin{aligned}
p(\mathbf{Z} | \mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\pi}) &= \frac{p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \boldsymbol{\pi})}{p(\mathbf{X} | \boldsymbol{\mu}, \boldsymbol{\pi})} \\
&= \prod_{n=1}^{N} \prod_{k=1}^{K} \frac{\left[\pi_k p(\mathbf{x}_n | \boldsymbol{\mu}_k)\right]^{z_{nk}}}{\sum_{j=1}^{K} \pi_j p(\mathbf{x}_n | \boldsymbol{\mu}_j)}
\end{aligned}
$$

EM for Bernoulli Mixtures: E-Step

- **E-Step**
  - Evaluate the responsibilities

$$\gamma(z_{nk}) = \mathbb{E}[z_{nk}] = \sum_{z_{nk}} z_{nk} \frac{\left[\pi_k p(\mathbf{x}_n|\boldsymbol{\mu}_k)\right]^{z_{nk}}}{\sum_{j=1}^{K} \pi_j p(\mathbf{x}_n|\boldsymbol{\mu}_j)}$$

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

  - Note: we again get the same form as for Gaussian mixtures

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

Recap: General EM Algorithm

- Algorithm
  1. Choose an initial setting for the parameters $\boldsymbol{\theta}^{\text{old}}$
  2. E-step: Evaluate $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$
  3. M-step: Evaluate $\boldsymbol{\theta}^{\text{new}}$ given by

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

where

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

  4. While not converged, let $\boldsymbol{\theta}^{\text{old}} \leftarrow \boldsymbol{\theta}^{\text{new}}$ and return to step 2.

EM for Bernoulli Mixtures: M-Step

- **Complete-data log-likelihood**

$$\log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \Big\{ \log \pi_k + \sum_{i=1}^{D} \left[x_{ni} \log \mu_{ki} + (1 - x_{ni}) \log(1 - \mu_{ki}) \right] \Big\}$$

- **Expectation w.r.t. the posterior distribution of Z**

$$\underbrace{\mathbb{E}_{\mathbf{Z}}[\log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\pi})]}_{\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})} = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \Big\{ \log \pi_k + \sum_{i=1}^{D} \left[x_{ni} \log \mu_{ki} + (1 - x_{ni}) \log(1 - \mu_{ki}) \right] \Big\}$$

where $\gamma(z_{nk}) = \mathbb{E}[z_{nk}]$ are again the responsibilities for each $\mathbf{x}_n$.

EM for Bernoulli Mixtures: M-Step

- **Remark**
  - The $\gamma(z_{nk})$ only occur in two forms in the expectation:

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

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

- **Interpretation**
  - N_k is the effective number of data points associated with component k.
  - $\bar{\mathbf{x}}_k$ is the responsibility-weighted mean of the data points softly assigned to component k.

EM for Bernoulli Mixtures: M-Step

- **M-Step**
  - **Maximize the expected complete-data log-likelihood w.r.t the parameter $\boldsymbol{\mu}_k$.**

$$\frac{\partial}{\partial \boldsymbol{\mu}_k} \mathbb{E}_{\mathbf{Z}}[p(\mathbf{X},\mathbf{Z}|\boldsymbol{\mu},\boldsymbol{\pi})]$$

$$= \frac{\partial}{\partial \boldsymbol{\mu}_k} \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \left\{ \log \pi_k + [\mathbf{x}_n \log \boldsymbol{\mu}_k + (1-\mathbf{x}_n) \log(1-\boldsymbol{\mu}_k)] \right\}$$

$$= \frac{1}{\boldsymbol{\mu}_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n - \frac{1}{1-\boldsymbol{\mu}_k} \sum_{n=1}^{N} \gamma(z_{nk})(1-\mathbf{x}_n) \overset{!}{=} 0$$

$$\vdots$$

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

EM for Bernoulli Mixtures: M-Step

- **M-Step**
  - **Maximize the expected complete-data log-likelihood w.r.t the parameter π_k under the constraint $\sum_k \pi_k = 1$.**
  - **Solution with Lagrange multiplier λ**

$$\arg\max_{\pi_k} \mathbb{E}_{\mathbf{Z}}[p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\pi})] + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right)$$

$$\vdots$$

$$\pi_k = \frac{N_k}{N}$$

Discussion

- **Comparison with Gaussian mixtures**
  - In contrast to Gaussian mixtures, there are no singularities in which the likelihood goes to infinity.
  - This follows from the property of Bernoulli distributions that

$$0 \leq p(\mathbf{x}_n|\boldsymbol{\mu}_k) \leq 1$$

  - However, there are still problem cases when μ_{ki} becomes 0 or 1

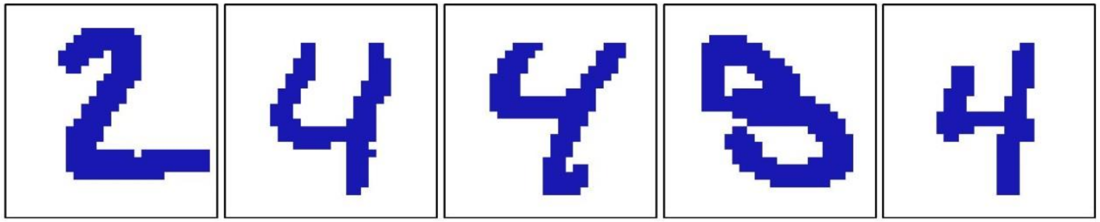
$$\mathbb{E}_{\mathbf{Z}}[\log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\pi})] = \ldots [x_{ni} \log \mu_{ki} + (1 - x_{ni}) \log(1 - \mu_{ki})]$$

  $\Rightarrow$ Need to enforce a range [MIN_VAL,1-MIN_VAL] for either μ_{ki} or γ.

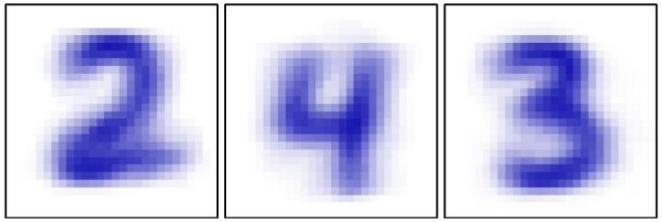
- **General remarks**
  - Bernoulli mixtures are used in practice in order to represent binary data.
  - The resulting model is also known as latent class analysis.

Example: Handwritten Digit Recognition

- Binarized digit data (examples from set of 600 digits)



- Means of a 3-component Bernoulli mixture (10 EM iter.)



- Comparison: ML result of single multivariate Bernoulli distribution

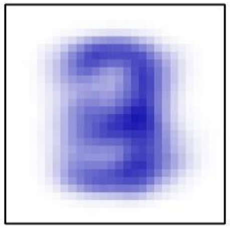


Image source: C.M. Bishop, 2006

Topics of This Lecture

- The EM algorithm in general
  - Recap: General EM
  - Example: Mixtures of Bernoulli distributions
  - Monte Carlo EM
- Bayesian Mixture Models
  - Towards a full Bayesian treatment
  - Dirichlet priors
  - Finite mixtures
  - Infinite mixtures
  - Approximate inference (only as supplementary material)

Monte Carlo EM

- **EM procedure**
  - M-step: Maximize expectation of complete-data log-likelihood

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \int p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \mathrm{d}\mathbf{Z}$$

  - For more complex models, we may not be able to compute this analytically anymore…

- **Idea**
  - Use sampling to approximate this integral by a finite sum over samples $\{\mathbf{Z}^{(l)}\}$ drawn from the current estimate of the posterior

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) \sim \frac{1}{L} \sum_{l=1}^{L} \log p(\mathbf{X}, \mathbf{Z}^{(l)}|\boldsymbol{\theta}^{\text{old}})$$

  - This procedure is called the Monte Carlo EM algorithm.

Topics of This Lecture

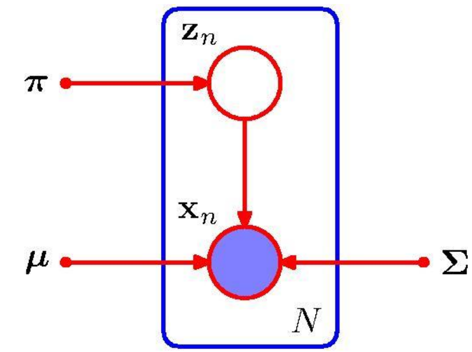
- The EM algorithm in general
  - Recap: General EM
  - Example: Mixtures of Bernoulli distributions
  - Monte Carlo EM
- Bayesian Mixture Models
  - Towards a full Bayesian treatment
  - Dirichlet priors
  - Finite mixtures
  - Infinite mixtures
  - Approximate inference (only as supplementary material)

RWTH AACHEN UNIVERSITY

Towards a Full Bayesian Treatment…

- Mixture models
  - We have discussed mixture distributions with K components

$$p(\mathbf{X}|\boldsymbol{\theta}) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$



  - So far, we have derived the ML estimates ⇒ **EM**
  - Introduced a prior $p(\boldsymbol{\theta})$ over parameters ⇒ **MAP-EM**

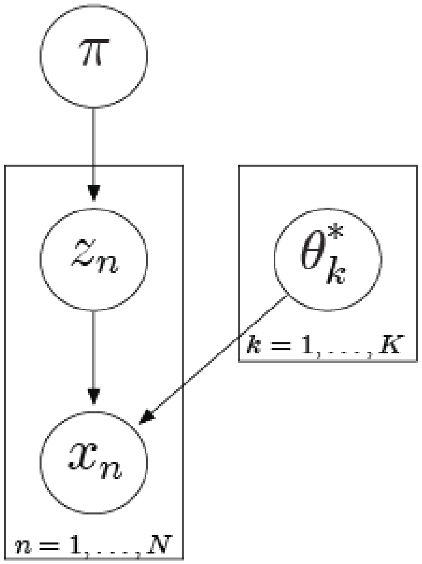
  - One question remains open: how to set K ?

⇒ Let's also set a prior on the number of components…

Bayesian Mixture Models

- **Let's be Bayesian about mixture models**
  - Place priors over our parameters
  - Again, introduce variable $\mathbf{z}_n$ as indicator which component data point $\mathbf{x}_n$ belongs to.

$$\mathbf{z}_n | \boldsymbol{\pi} \sim \text{Multinomial}(\boldsymbol{\pi})$$
$$\mathbf{x}_n | \mathbf{z}_n = k, \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)$$



  - This is similar to the graphical model we've used before, but now the π and $\theta_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ are also treated as random variables.
  - *What would be suitable priors for them?*

Bayesian Mixture Models

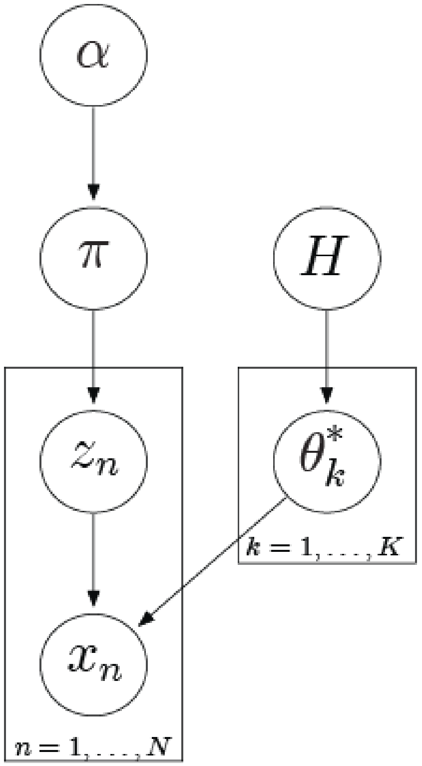
- **Let's be Bayesian about mixture models**
  - Place priors over our parameters
  - Again, introduce variable $\mathbf{z}_n$ as indicator which component data point $\mathbf{x}_n$ belongs to.

$$\mathbf{z}_n | \boldsymbol{\pi} \sim \text{Multinomial}(\boldsymbol{\pi})$$
$$\mathbf{x}_n | \mathbf{z}_n = k, \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)$$

  - Introduce conjugate priors over parameters

$$\boldsymbol{\pi} \sim \text{Dirichlet}(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K})$$
$$\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \sim H = \mathcal{N} - \mathcal{IW}(0, s, d, \phi)$$

"Normal - Inverse Wishart"



Bayesian Mixture Models

- **Full Bayesian Treatment**
  - Given a dataset, we are interested in the cluster assignments

$$p(\mathbf{Z}|\mathbf{X}) = \frac{p(\mathbf{X}|\mathbf{Z})p(\mathbf{Z})}{\sum_{\mathbf{Z}} p(\mathbf{X}|\mathbf{Z})p(\mathbf{Z})}$$

where the likelihood is obtained by marginalizing over the parameters θ

$$\begin{aligned} p(\mathbf{X}|\mathbf{Z}) &= \int p(\mathbf{X}|\mathbf{Z},\boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \\ &= \int \prod_{n=1}^{N} \prod_{k=1}^{K} p(\mathbf{x}_n|z_{nk},\boldsymbol{\theta}_k)p(\boldsymbol{\theta}_k|H)\mathrm{d}\boldsymbol{\theta} \end{aligned}$$

- **The posterior over assignments is intractable!**
  - Denominator requires summing over all possible partitions of the data into K groups!

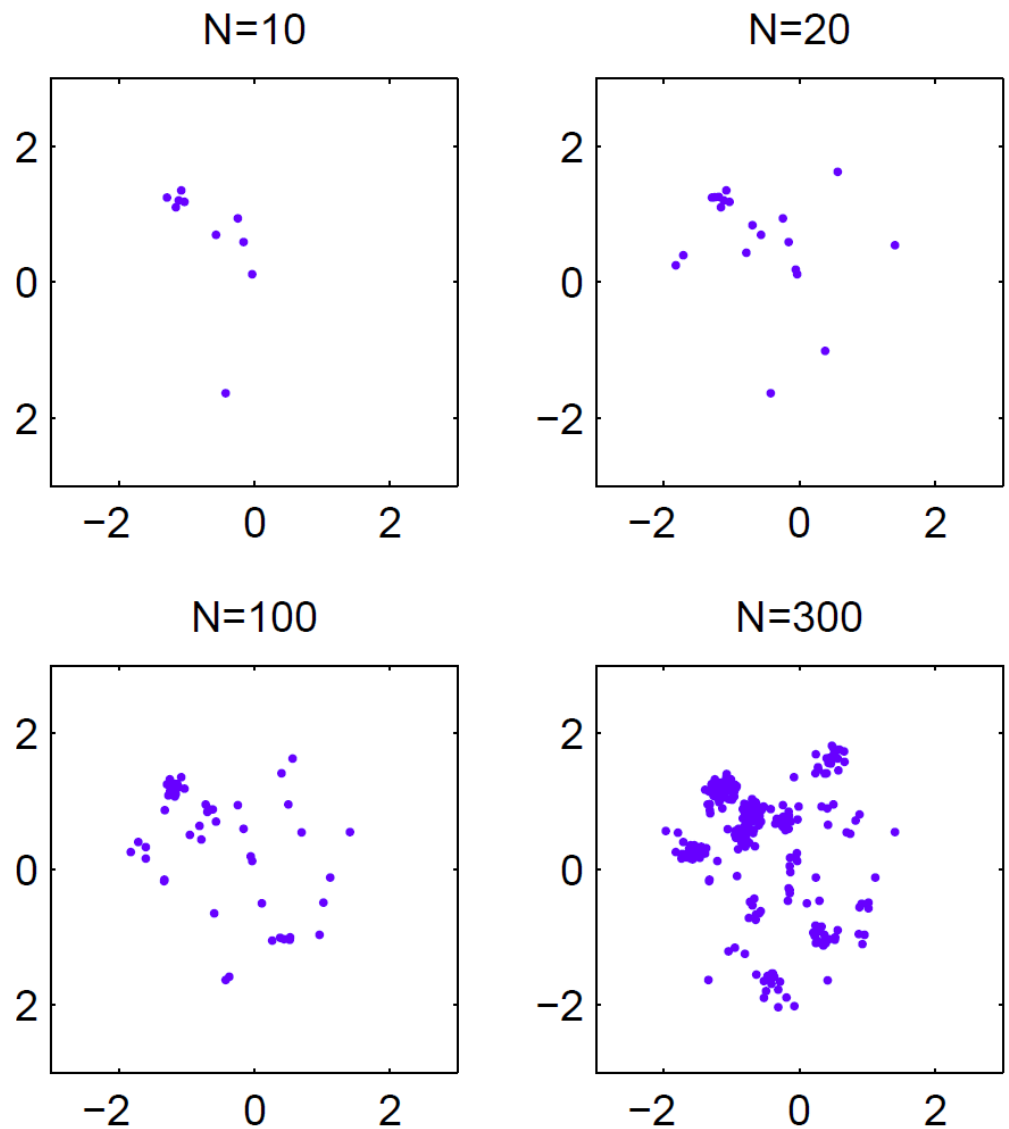
$\Rightarrow$ Need efficient approximate inference methods to solve this...

Bayesian Mixture Models

- Let's examine this model more closely
  - Role of Dirichlet priors?
  - How can we perform efficient inference?
  - What happens when K goes to infinity?

- This will lead us to an interesting class of models…
  - Dirichlet Processes
  - Possible to express infinite mixture distributions with their help
  - Clustering that automatically adapts the number of clusters to the data and *dynamically creates new clusters on-the-fly*.

Sneak Preview: Dirichlet Process MoG



Samples drawn from DP mixture

⇒ More structure appears as more points are drawn

Recap: The Dirichlet Distribution

- **Dirichlet Distribution**
  - **Conjugate prior for the Categorical and the Multinomial distrib.**

$$\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1} \quad \textbf{with} \quad \alpha_0 = \sum_{k=1}^{K} \alpha_k$$

  - **Symmetric version (with concentration parameter α)**

$$\mathrm{Dir}(\boldsymbol{\mu}|\alpha) = \frac{\Gamma(\alpha)}{\Gamma(\alpha/K)^K} \prod_{k=1}^{K} \mu_k^{\alpha/K - 1}$$

  - **Properties** **(symmetric version)**

$$\mathbb{E}[\mu_k] = \frac{\alpha_k}{\alpha_0} = \frac{1}{K}$$

$$\mathrm{var}[\mu_k] = \frac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0^2(\alpha_0 + 1)} = \frac{K-1}{K^2(\alpha+1)}$$

$$\mathrm{cov}[\mu_j \mu_k] = -\frac{\alpha_j \alpha_k}{\alpha_0^2(\alpha_0 + 1)} = -\frac{1}{K^2(\alpha+1)}$$

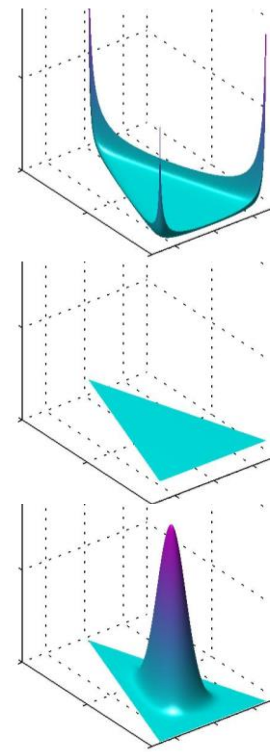
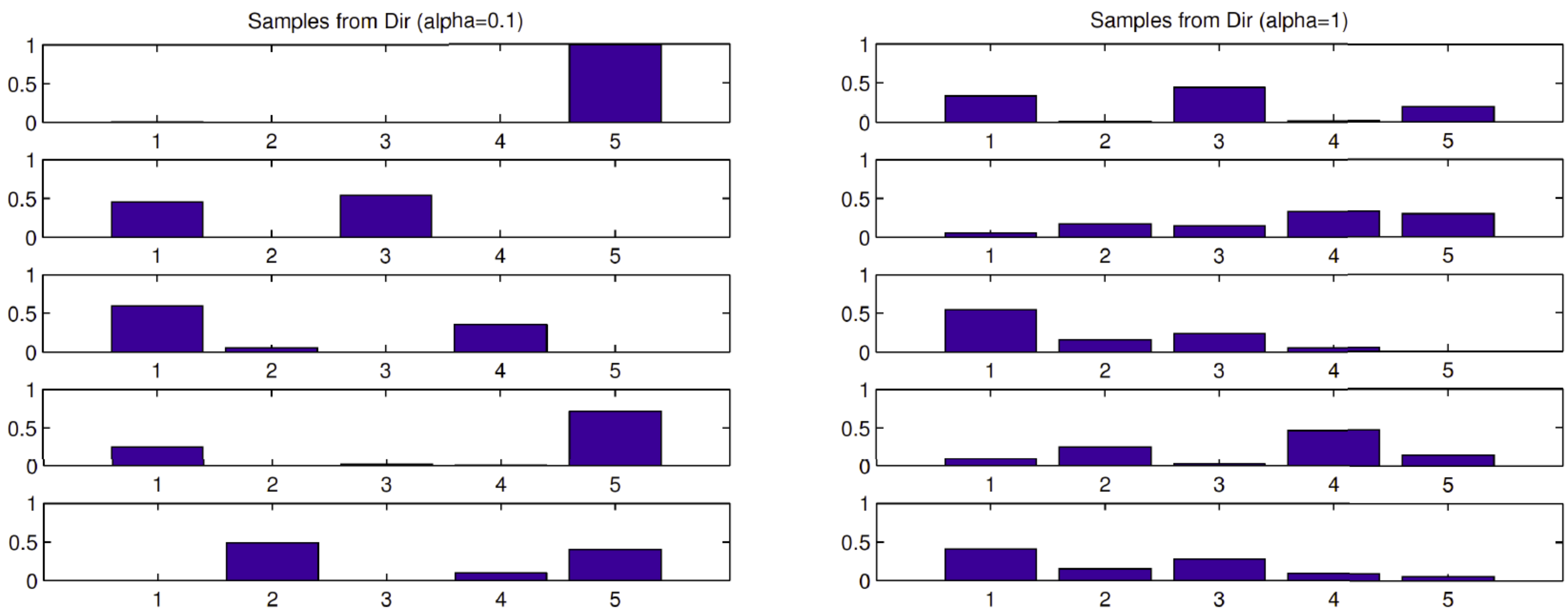


Image source: C. Bishop, 2006

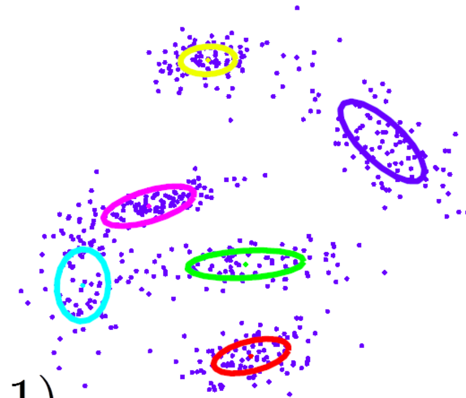
Dirichlet Samples



$\mathrm{Dir}(\theta \mid 0.1, 0.1, 0.1, 0.1, 0.1)$ $\quad$ $\mathrm{Dir}(\theta \mid 1.0, 1.0, 1.0, 1.0, 1.0)$

- **Effect of concentration parameter α**
  - **Controls sparsity of the resulting samples**

Mixture Model with Dirichlet Priors



- **Finite mixture of K components**

$$p(\mathbf{x}_n|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k p(\mathbf{x}_n|\theta_k)$$
$$= \sum_{k=1}^{K} p(z_{nk}=1|\pi_k) p(\mathbf{x}_n|\theta_k, z_{nk}=1)$$

  - **The distribution of latent variables $\mathbf{z}_n$ given π is multinomial**

$$p(\mathbf{z}|\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{N_k}, \quad N_k \stackrel{\text{def}}{=} \sum_{n=1}^{N} z_{nk}$$

  - **Assume mixing proportions have a given symmetric conjugate Dirichlet prior**

$$p(\boldsymbol{\pi}|\alpha) = \frac{\Gamma(\alpha)}{\Gamma(\alpha/K)^K} \prod_{k=1}^{K} \pi_k^{\alpha/K-1}$$

Mixture Model with Dirichlet Priors

- **Integrating out the mixing proportions π:**

$$
\begin{aligned}
p(\mathbf{z}|\alpha) &= \int p(\mathbf{z}|\boldsymbol{\pi})p(\boldsymbol{\pi}|\alpha)\mathrm{d}\boldsymbol{\pi} \\
&= \int \prod_{k=1}^{K} \pi_k^{N_k} \cdot \frac{\Gamma(\alpha)}{\Gamma(\alpha/K)^K} \prod_{k=1}^{K} \pi_k^{\alpha/K-1} \mathrm{d}\boldsymbol{\pi} \\
&= \int \frac{\Gamma(\alpha)}{\Gamma(\alpha/K)^K} \prod_{k=1}^{K} \pi_k^{N_k+\alpha/K-1} \mathrm{d}\boldsymbol{\pi}
\end{aligned}
$$

  - **This is again a Dirichlet distribution (reason for conjugate priors)**

$$
= \frac{\Gamma(\alpha)}{\Gamma(\alpha/K)^K} \frac{\prod_{k=1}^{K}\Gamma(N_k+\alpha/K)}{\Gamma(N+\alpha)} \int \frac{\Gamma(N+\alpha)}{\prod_{k=1}^{K}\Gamma(N_k+\alpha/K)} \prod_{k=1}^{K} \pi_k^{N_k+\alpha/K-1} \mathrm{d}\boldsymbol{\pi}
$$

Completed Dirichlet form $\rightarrow$ integrates to 1

Mixture Models with Dirichlet Priors

- **Integrating out the mixing proportions π (cont'd)**

$$p(\mathbf{z}|\alpha) = \frac{\Gamma(\alpha)}{\Gamma(\alpha/K)^K} \frac{\prod_{k=1}^K \Gamma(N_k + \alpha/K)}{\Gamma(N+\alpha)}$$

$$= \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)} \prod_{k=1}^K \frac{\Gamma(N_k + \alpha/K)}{\Gamma(\alpha/K)}$$

- **Conditional probabilities**
  - **Let's examine the conditional of $\mathbf{z}_n$ given all other variables**

$$p(z_{nk} = 1|\mathbf{z}_{-n}, \alpha) = \frac{p(z_{nk} = 1, \mathbf{z}_{-n}|\alpha)}{p(\mathbf{z}_{-n}|\alpha)}$$

where $\mathbf{z}_{-n}$ denotes all indizes except n.

Mixture Models with Dirichlet Priors

$$p(\mathbf{z}|\alpha) = \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)} \prod_{k=1}^{K} \frac{\Gamma(N_k + \alpha/K)}{\Gamma(\alpha/K)}$$

- **Conditional probabilities**

$$p(z_{nk} = 1|\mathbf{z}_{-n}, \alpha) = \frac{p(z_{nk} = 1, \mathbf{z}_{-n}|\alpha)}{p(\mathbf{z}_{-n}|\alpha)}$$

$$= \frac{\frac{\Gamma(\alpha)}{\Gamma(N+\alpha)} \frac{\Gamma(N_k+\alpha/K)}{\Gamma(\alpha/K)} \prod_{j=1, j\neq k}^{K} \frac{\Gamma(N_j+\alpha/K)}{\Gamma(\alpha/K)}}{\frac{\Gamma(\alpha)}{\Gamma(N_{-n}+\alpha)} \frac{\Gamma(N_{-n,k}+\alpha/K)}{\Gamma(\alpha/K)} \prod_{j=1, j\neq k}^{K} \frac{\Gamma(N_j+\alpha/K)}{\Gamma(\alpha/K)}}$$

$$= \frac{\Gamma(N_{-n} + \alpha)}{\Gamma(N + \alpha)} \frac{\Gamma(N_k + \alpha/K)}{\Gamma(N_{-n,k} + \alpha/K)}$$

Mixture Models with Dirichlet Priors

- **Conditional probabilities**

$$\Gamma(n+1) = n\Gamma(n)$$

$$p(z_{nk}=1|\mathbf{z}_{-n},\alpha) = \frac{p(z_{nk}=1,\mathbf{z}_{-n}|\alpha)}{p(\mathbf{z}_{-n}|\alpha)}$$

$$= \frac{\frac{\Gamma(\alpha)}{\Gamma(N+\alpha)} \frac{\Gamma(N_k+\alpha/K)}{\Gamma(\alpha/K)} \prod_{j=1,j\neq k}^{K} \frac{\Gamma(N_j+\alpha/K)}{\Gamma(\alpha/K)}}{\frac{\Gamma(\alpha)}{\Gamma(N_{-n}+\alpha)} \frac{\Gamma(N_{-n,k}+\alpha/K)}{\Gamma(\alpha/K)} \prod_{j=1,j\neq k}^{K} \frac{\Gamma(N_j+\alpha/K)}{\Gamma(\alpha/K)}}$$

$$= \frac{\Gamma(N_{-n}+\alpha)}{\Gamma(N+\alpha)} \frac{\Gamma(N_k+\alpha/K)}{\Gamma(N_{-n,k}+\alpha/K)}$$

$$= \frac{1}{N-1+\alpha} \frac{N_{-n,k}+\alpha/K}{1}$$

$$= \frac{N_{-n,k}+\alpha/K}{N-1+\alpha}$$

Finite Dirichlet Mixture Models

- Conditional probabilities: Finite K

$$p(z_{nk} = 1 | \mathbf{z}_{-n}, \alpha) = \frac{N_{-n,k} + \alpha/K}{N - 1 + \alpha}, \qquad N_{-n,k} \stackrel{\text{def}}{=} \sum_{i=1, i \neq n}^{N} z_{ik}$$

- This is a very interesting result. *Why?*
  - We directly get a numerical probability, no distribution.
  - The probability of joining a cluster mainly depends on the number of existing entries in a cluster.

  $\Rightarrow$ The more populous a class is, the more likely it is to be joined!
  - In addition, we have a base probability of also joining as-yet empty clusters.
  - This result can be directly used in Gibbs Sampling…

Infinite Dirichlet Mixture Models

- Conditional probabilities: Finite K

$$p(z_{nk} = 1 | \mathbf{z}_{-n}, \alpha) = \frac{N_{-n,k} + \alpha / K}{N - 1 + \alpha}, \qquad N_{-n,k} \stackrel{\text{def}}{=} \sum_{i=1, i \neq n}^{N} z_{ik}$$

- Conditional probabilities: Infinite K
  - Taking the limit as $K \to \infty$ yields the conditionals

$$p(z_{nk} = 1 | \mathbf{z}_{-n}, \alpha) = \begin{cases} \frac{N_{-n,k}}{N-1+\alpha} & \text{if } k \text{ represented} \\ \frac{\alpha}{N-1+\alpha} & \text{if all } k \text{ not represented} \end{cases}$$

  - Left-over mass α $\Rightarrow$ countably infinite number of indicator settings

Discussion

- Infinite Mixture Models
  - What we have just seen is a first example of a Dirichlet Process.
  - DPs allow us to work with models that have an infinite number of components.
  - This will raise a number of issues
    - How to represent infinitely many parameters?
    - How to deal with permutations of the class labels?
    - How to control the effective size of the model?
    - How to perform efficient inference?

  ⇒ More background needed here!

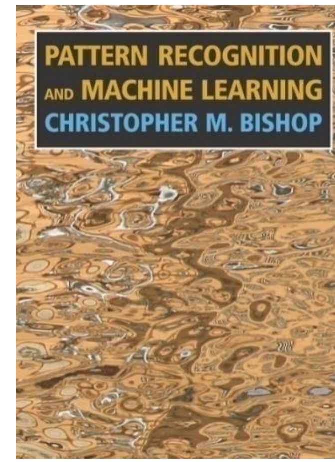
  - DPs are a very interesting class of models, but would take us too far here.
  - If you're interested in learning more about them, take a look at the Advanced ML slides from Winter 2012.

Next Lecture...

References and Further Reading

- More information about EM estimation is available in Chapter 9 of Bishop's book (recommendable to read).

Christopher M. Bishop
Pattern Recognition and Machine Learning
Springer, 2006



- Additional information
  - Original EM paper:
    - A.P. Dempster, N.M. Laird, D.B. Rubin, „Maximum-Likelihood from incomplete data via EM algorithm”, In Journal Royal Statistical Society, Series B. Vol 39, 1977
  - EM tutorial:
    - J.A. Bilmes, "A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models", TR-97-021, ICSI, U.C. Berkeley, CA,USA