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.
  ⇒ Please submit your results until 06.12. midnight.

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”
  \[ p(x) = \sum_{k=1}^{K} \pi_k N(x; \mu_k, \Sigma_k) \]

Recap: GMMs as Latent Variable Models
- Write GMMs in terms of latent variables \( z \)
  - Marginal distribution of \( x \)
    \[ p(x) = \sum_{z} p(x; z) = \sum_{z} p(z)p(x|z) = \sum_{k=1}^{K} \pi_k N(x; \mu_k, \Sigma_k) \]
  - Advantage of this formulation
    - We have represented the marginal distribution in terms of latent variables \( z \).
    - Since \( p(x) = \sum_{z} p(x; z) \), there is a corresponding latent variable \( z \), for each data point \( x \).
    - We are now able to work with the joint distribution \( p(x, z) \)
      instead of the marginal distribution \( p(x) \).
    ⇒ 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 $z$ from the marginal distribution $p(z)$.
  2. Generate a value $x$ from the conditional distribution $p(x|z)$.

Samples from the joint $p(x, z)$  
Samples from the marginal $p(x)$  
Evaluating the responsibilities $p(z|x)$

Recap: Gaussian Mixtures Revisited

- **Applying the latent variable view of EM**
  - Goal is to maximize the log-likelihood using the observed data $X$.
    
    \[
    \log p(X|\theta) = \log \left( \sum_z p(X, Z|\theta) \right)
    \]
  - Corresponding graphical model:

Suppose we are additionally given the values of the latent variables $Z$.

The corresponding graphical model for the complete data now looks like this:

⇒ Straightforward to marginalize.

Recap: Alternative View of EM

- In practice, however,...
  - We are not given the complete data set $(X, Z)$, but only the incomplete data $X$. All we can compute about $Z$ is the posterior distribution $p(Z|X, \theta)$.

  - Since we cannot use the complete-data log-likelihood, we consider instead its expected value under the posterior distribution of the latent variable:
    
    \[
    Q(\theta; \theta^{(t)}) = \sum_z p(Z|X, \theta^{(t)}) \log p(X, Z|\theta)
    \]
  - 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 $\theta^{(t+1)}$.

\[
\theta^{(t+1)} = \arg \max_{\theta} Q(\theta; \theta^{(t)})
\]

Recap: MAP-EM

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

  2. **E-step**: Evaluate $p(Z|X, \theta^{(t)})$
  3. **M-step**: Evaluate $\theta^{(t+1)}$ given by

    \[
    \theta^{(t+1)} = \arg \max_{\theta} Q(\theta; \theta^{(t)}) + \log p(\theta)
    \]

⇒ Suitable choices for the prior will remove the ML singularities!

Recap: General EM Algorithm

- **Algorithm**
  1. Choose an initial setting for the parameters $\theta^{(0)}$
  2. **E-step**: Evaluate $p(Z|X, \theta^{(t)})$
  3. **M-step**: Evaluate $\theta^{(t+1)}$ given by

    \[
    \theta^{(t+1)} = \arg \max_{\theta} Q(\theta; \theta^{(t)})
    \]

    where

    \[
    Q(\theta; \theta^{(t)}) = \sum_z p(Z|X, \theta^{(t)}) \log p(X, Z|\theta)
    \]

  4. While not converged, let $\theta^{(t+1)} \leftarrow \theta^{(t+1)}$ and return to step 2.

Gaussian Mixtures Revisited

- **Maximize the likelihood**
  - For the complete-data set $(X, Z)$, the likelihood has the form

    \[
    p(X, Z|\mu, \Sigma, \pi) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_{nk} \mathcal{N}(x_n|\mu_k, \Sigma_k)^{z_{nk}}
    \]

  - Taking the logarithm, we obtain

    \[
    \log p(X, Z|\mu, \Sigma, \pi) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \{ \log \pi_k + \log \mathcal{N}(x_n|\mu_k, \Sigma_k) \}
    \]

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

⇒ 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 $\pi_k$ are coupled by the summation constraint
    $$\sum_{j=1}^{K} \pi_j = 1$$
  - Solve with a Lagrange multiplier
    $$\log p(X, Z|\mu, \Sigma, \pi) + \lambda \left( \sum_{k=1}^{K} \pi_k - 1 \right)$$
  - Solution (after a longer derivation):
    $$\pi_k = \frac{1}{N} \sum_{n=1}^{N} z_{nk}$$
  - The complete-data log-likelihood can be maximized trivially in closed form.

Gaussian Mixtures Revisited

- Continuing the estimation
  - The complete-data log-likelihood is therefore
    $$\mathbb{E}[\log p(X, Z|\mu, \Sigma, \pi)] = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \{ \log \pi_k + \log \mathcal{N}(x_n|\mu_k, \Sigma_k) \}$$
  - 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…

Mixtures of Bernoulli Distributions

- Discrete binary variables
  - Consider $D$ binary variables $x = (x_1, \ldots, x_D)^T$, each of them described by a Bernoulli distribution with parameter $\mu_i$, so that
    $$p(x|\mu) = \prod_{i=1}^{D} \mu_i^{x_i} (1 - \mu_i)^{1-x_i}$$
  - Mean and covariance are given by
    $$\mathbb{E}[x] = \mu$$
    $$\text{cov}[x] = \text{diag} \{ \mu (1 - \mu) \}$$
Mixtures of Bernoulli Distributions

- Mixtures of discrete binary variables
  - Now, consider a finite mixture of those distributions
    \[ p(\mathbf{x}|\mu, \pi) = \sum_{k=1}^{K} \pi_k p(\mathbf{x}|\mu_k) \]
    \[ = \sum_{k=1}^{K} \pi_k \prod_{i=1}^{D} \mu_k^i (1 - \mu_k)^{1-i} \]
  - Mean and covariance of the mixture are given by
    \[ \mathbb{E}[\mathbf{x}] = \sum_{k=1}^{K} \pi_k \mu_k \]
    \[ \text{cov}[\mathbf{x}] = \sum_{k=1}^{K} \pi_k \{ \Sigma_k + \mu_k \mu_k^T \} - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}]^T \]
    where \( \Sigma_k = \text{diag}(\mu_k(1 - \mu_k)) \).

- Log-likelihood for the model
  - Given a data set \( \mathbf{X} = \{ \mathbf{x}_1, \ldots, \mathbf{x}_N \} \),
    \[ \log p(\mathbf{X}|\mu, \pi) = \sum_{n=1}^{N} \log \left( \sum_{k=1}^{K} \pi_k p(\mathbf{x}_n|\mu_k) \right) \]
  - 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 \( \mathbf{x} \):
    \[ p(\mathbf{x}|\mu, \pi) = \prod_{k=1}^{K} p(\mathbf{x}|\mu_k)^{z_k} \]
  - Prior distribution for the latent variables
    \[ p(z|\pi) = \prod_{k=1}^{K} \pi_k^{z_k} \]
  - Again, we can verify that
    \[ p(\mathbf{x}|\mu, \pi) = \sum_{z} p(\mathbf{x}|z, \mu) p(z|\pi) = \sum_{k=1}^{K} \pi_k p(\mathbf{x}|\mu_k) \]

EM for Bernoulli Mixtures: E-Step

- Complete-data likelihood
  \[ p(\mathbf{X}, \mathbf{Z}|\mu, \pi) = \prod_{n=1}^{N} \prod_{k=1}^{K} [\pi_k p(\mathbf{x}_n|\mu_k)]^{z_{nk}} \]
  \[ = \prod_{n=1}^{N} \prod_{k=1}^{K} \left\{ \pi_k \prod_{i=1}^{D} \mu_k^i (1 - \mu_k)^{1-i} \right\}^{z_{nk}} \]

- Posterior distribution of the latent variables \( \mathbf{Z} \)
  \[ p(\mathbf{Z}|\mathbf{X}, \mu, \pi) = \frac{p(\mathbf{X}, \mathbf{Z}|\mu, \pi)}{p(\mathbf{X}|\mu, \pi)} \]
  \[ = \prod_{n=1}^{N} \prod_{k=1}^{K} \frac{\pi_k p(\mathbf{x}_n|\mu_k)^{z_{nk}}}{\sum_{k=1}^{K} \pi_k p(\mathbf{x}_n|\mu_k)} \]
  \( \Rightarrow \) Note: we again get the same form as for Gaussian mixtures
  \[ \gamma_{nk} = \frac{\pi_k p(\mathbf{x}_n|\mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k p(\mathbf{x}_n|\mu_k, \Sigma_k)} \]

Recap: General EM Algorithm

- Algorithm
  1. Choose an initial setting for the parameters \( \theta^{\text{old}} \)
  2. E-step: Evaluate \( Q(\theta, \theta^{\text{old}}) \)
  3. M-step: Evaluate \( \theta^{\text{new}} \) given by
    \[ \theta^{\text{new}} = \arg \max_{\theta} Q(\theta, \theta^{\text{old}}) \]
    where
    \[ Q(\theta, \theta^{\text{old}}) = \sum_{Z} p(\mathbf{Z} | \mathbf{X}, \theta^{\text{old}}) \log p(\mathbf{X}, \mathbf{Z}|\theta) \]
  4. While not converged, let \( \theta^{\text{old}} = \theta^{\text{new}} \) and return to step 2.

EM for Bernoulli Mixtures: E-Step

- E-step
  - Evaluate the responsibilities
    \[ \gamma(\tau_{nk}) = \mathbb{E}[\tau_{nk}] = \sum_{k=1}^{K} \frac{\pi_k p(\mathbf{x}_n|\mu_k)^{z_{nk}}}{\sum_{j=1}^{K} \pi_j p(\mathbf{x}_n|\mu_j)} \]
    \[ = \frac{\pi_k p(\mathbf{x}_n|\mu_k)}{\sum_{j=1}^{K} \pi_j p(\mathbf{x}_n|\mu_j)} \]
    \( \Rightarrow \) Note: we again get the same form as for Gaussian mixtures
    \[ \gamma_{nk} = \frac{\pi_k p(\mathbf{x}_n|\mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k p(\mathbf{x}_n|\mu_k, \Sigma_k)} \]
Recap: General EM Algorithm

- **Algorithm**
  1. Choose an initial setting for the parameters \( \theta^{(0)} \)
  2. **E-step:** Evaluate \( p(Z|X, \theta^{(t)}) \)
  3. **M-step:** Evaluate \( \theta^{(t+1)} \) given by
    \[
    Q(\theta, \theta^{(t)}) = \sum p(Z|X, \theta^{(t)}) \log p(X, Z|\theta)
    \]
    where
    \[
    Q(\theta, \theta^{(t)}) = \sum_{k=1}^{K} \pi_k \sum_{n=1}^{N} \log \pi_k + \sum_{n=1}^{N} (1 - \gamma_{nk}) \log (1 - \pi_k)
    \]
  4. While not converged, let \( \theta^{(t+1)} = \theta^{(t)} \) and return to step 2.

EM for Bernoulli Mixtures: M-Step

- **Remark**
  - The \( \gamma(z_{nk}) \) only occur in two forms in the expectation:
    \[
    N_k = \sum_{n=1}^{N} \gamma(z_{nk})
    \]
    \[
    X_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) x_n
    \]
- **Interpretation**
  - \( N_k \) is the effective number of data points associated with component \( k \).
  - \( X_k \) 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 \( \mu_k \).
    \[
    \frac{\partial}{\partial \mu_k} \mathbb{E}_Z[p(X, Z|\pi, \mu)]
    = \sum_{n=1}^{N} \gamma(z_{nk}) \log \pi_k + (1 - \gamma(z_{nk})) \log (1 - \pi_k)
    - \sum_{n=1}^{N} \gamma(z_{nk}) \frac{x_n}{\pi_k} + \frac{(1 - x_n)}{1 - \pi_k}
    \]
  - \( \mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) x_n = X_k \)

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
    \[ 0 \leq p(x_n|\mu_k) \leq 1 \]
    - However, there are still problem cases when \( \mu_k \) becomes 0 or 1
    - Need to enforce a range \([\text{MIN}_k, 1 - \text{MIN}_k]\) for each \( \mu_k \) or \( \gamma \).
- **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)

![Digit Image]

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

![Mixture Means]

- Comparison: ML result of single multivariate Bernoulli distribution

![Single Distribution]

Monte Carlo EM

- EM procedure
  - **M-step**: Maximize expectation of complete-data log-likelihood
    \[ Q(\theta, \theta^{(k)}) = \int p(Z|X, \theta^{(k)}) \log p(X, Z|\theta) dZ \]
  - 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 \( \{Z^{(l)}\} \) drawn from the current estimate of the posterior
    \[ Q(\theta, \theta^{(k)}) \approx \frac{1}{L} \sum_{l=1}^{L} \log p(X, Z^{(l)}|\theta^{(k)}) \]
    - This procedure is called the Monte Carlo EM algorithm.

Towards a Full Bayesian Treatment...

- Mixture models
  - We have discussed mixture distributions with \( K \) components
    \[ p(X|\theta) = \sum_{z} p(X, Z|\theta) \]
  - So far, we have derived the ML estimates
  - Introduced a prior \( p(\theta) \) over parameters
  - One question remains open: how to set \( K \)?
  - Let’s also set a prior on the number of components...

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)

Bayesian Mixture Models

- Let’s be Bayesian about mixture models
  - Place priors over our parameters
  - Again, introduce variable \( z_n \) as indicator which component data point \( x_n \) belongs to.
    \[ z_n|\pi \sim \text{Multinomial}(\pi) \]
    \[ x_n|z_n = k; \mu_k, \Sigma_k \sim \mathcal{N}(\mu_k, \Sigma_k) \]
  - This is similar to the graphical model we’ve used before, but now the \( \pi \) and \( \theta_k = [\mu_k, \Sigma_k] \) 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 $z_n$ as indicator which component data point $x_n$ belongs to.
    - Introduce conjugate priors over parameters
  - Conjugate priors

Bayesian Mixture Models

- Full Bayesian Treatment
  - Given a dataset, we are interested in the cluster assignments
    $$ p(Z|X) = \frac{p(X|Z)p(Z)}{\sum_{Z} p(X|Z)p(Z)} $$
  - Where the likelihood is obtained by marginalizing over the parameters $\theta$
    $$ p(X|Z) = \int p(X|\theta)p(\theta) d\theta $$
    $$ = \prod_{n=1}^{N} \prod_{k=1}^{K} p(x_n|z_{nk}, \theta_k)p(\theta_k|H)d\theta $$
  - The posterior over assignments is intractable!
    - Denominator requires summing over all possible partitions of the data into $K$ groups!
    - 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.

Recap: The Dirichlet Distribution

- Dirichlet Distribution
  - Conjugate prior for the Categorical and the Multinomial distribution.
    $$ \text{Dir}(\mu|\alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k-1} \text{ with } \alpha_0 = \sum_{k=1}^{K} \alpha_k $$
  - Symmetric version (with concentration parameter $\alpha$)
    $$ \text{Dir}(\mu|\alpha) = \frac{\Gamma(\alpha)}{\Gamma(\alpha/K)^K} \prod_{k=1}^{K} \mu_k^{\alpha/K-1} $$
  - Properties
    $$ \mathbb{E}[\mu_k] = \frac{\alpha_k}{\alpha_0} $$
    $$ \text{var}[\mu_k] = \frac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0^2(\alpha_0 + 1)} $$
    $$ \text{cov}[\mu_j, \mu_k] = -\frac{\alpha_{jk}}{\alpha_0^2(\alpha_0 + 1)} $$
- Effect of concentration parameter $\alpha$
  - Controls sparsity of the resulting samples

Dirichlet Samples

- Samples from Dir(1.0, 0.1, 1.0, 0.1, 1.0, 1.0)
- Samples from Dir(0.1, 0.1, 0.1, 1.0, 0.0, 1.0)
Mixture Model with Dirichlet Priors

- Finite mixture of $K$ components

\[ p(x_n | \theta) = \sum_{k=1}^{K} \pi_k p(x_n | \theta_k) \]

\[ = \sum_{k=1}^{K} p(z_{nk} = 1 | \pi_k) p(x_n | \theta_k, z_{nk} = 1) \]

- The distribution of latent variables $z_n$ given $\pi$ is multinomial

\[ p(z | \pi) = \prod_{k=1}^{K} \pi_k^{N_k}, \quad N_k = \sum_{n=1}^{N} z_{nk} \]

- Assume mixing proportions have a given symmetric conjugate Dirichlet prior

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

- Integrating out the mixing proportions $\pi$:

\[ p(z | \alpha) = \int p(z | \pi) p(\pi | \alpha) \, d\pi \]

\[ = \int \prod_{k=1}^{K} \pi_k^{N_k} \cdot \frac{\Gamma(\alpha)}{\Gamma(\alpha/K)^K} \prod_{k=1}^{K} \pi_k^{\alpha/K - 1} \, d\pi \]

\[ = \frac{\Gamma(\alpha)}{\Gamma(\alpha/K)^K} \prod_{k=1}^{K} N_k^{\alpha/K - 1} \]

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

\[ = \frac{\Gamma(\alpha)}{\Gamma(\alpha/K)^K} \prod_{k=1}^{K} \frac{\Gamma(N_k + \alpha/K)}{\Gamma(N_k + \alpha)} \int \frac{\Gamma(N + \alpha)}{\prod_{k=1}^{K} \Gamma(N_k + \alpha/K)} \prod_{k=1}^{K} N_k^{\alpha/K - 1} \, d\pi \]

Completed Dirichlet form $\rightarrow$ integrates to 1

Mixture Models with Dirichlet Priors

- Integrating out the mixing proportions $\pi$ (cont'd)

\[ p(z_{nk} = 1 | x_{-n}, \alpha) = \frac{p(z_{nk} = 1, z_{-n} | \alpha)}{p(z_{-n} | \alpha)} \]

where $z_{-n}$ denotes all indices except $n$.

Mixture Models with Dirichlet Priors

- Conditional probabilities

\[ p(z_{nk} = 1 | x_{-n}, \alpha) = \frac{\Gamma(N + 1) - \Gamma(N)}{\Gamma(N + 1)} \]

- Conditional probabilities: Finite $K$

\[ p(z_{nk} = 1 | x_{-n}, \alpha) = \frac{N - n + \alpha/\alpha}{N^{\alpha}} \]

\[ = \frac{\Gamma(N - n + \alpha)}{\Gamma(N)} \cdot \frac{\Gamma(N + \alpha)}{\Gamma(N + \alpha)} \]

\[ = \frac{\Gamma(N - n + \alpha)}{\Gamma(N)} \cdot \frac{\Gamma(N + \alpha)}{\Gamma(N + \alpha)} \]

\[ = \frac{\Gamma(N - n + \alpha)}{\Gamma(N)} \cdot \frac{\Gamma(N + \alpha)}{\Gamma(N + \alpha)} \]

\[ = \frac{1}{N^{\alpha}} \cdot \frac{N - n + \alpha}{N^{\alpha}} \]

\[ = \frac{N - n + \alpha}{N^{\alpha}} \]

Finite Dirichlet Mixture Models

- Conditional probabilities: Finite $K$

\[ p(z_{nk} = 1 | x_{-n}, \alpha) = \frac{N - n + \alpha}{N - 1 + \alpha} \]

\[ = \frac{N - n + \alpha}{N - 1 + \alpha} \]

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.

- The more populous 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 $K$
  \[ p(z_{nk} = 1|z_{n\cdot}, \alpha) = \frac{N_{nk} + \alpha/K}{N - 1 + \alpha}, \quad N_{nk} \text{ def } \sum_{i=1}^{N} z_{nk} \]

- Conditional probabilities: Infinite $K$
  - Taking the limit as $K \to \infty$ yields the conditionals
  \[ p(z_{nk} = 1|z_{n\cdot}, \alpha) = \begin{cases} \frac{N_{nk}}{N-1+\alpha} & \text{if } k \text{ represented} \\ \frac{\alpha}{N-1+\alpha} & \text{if all } k \text{ not represented} \end{cases} \]
  - **Left-over mass** $\alpha \Rightarrow$ 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.

References and Further Reading

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