# Mixture Models and Expectation-Maximization (EM)

*Prof. Nicholas Zabaras*
*University of Notre Dame*
*Notre Dame, IN, USA*

*Email: nzabaras@gmail.com*
*URL: http://www.zabaras.com/*

*November 1, 2017*

# Contents

- ❑ Gaussian Mixtures Revisited
- ❑ Mixtures of Bernoulli distributions
- ❑ MAP estimation
- ❑ Bayesian Linear Regression
- ❑ Variational Inference, Incremental EM
- ❑ Fitting with Missing Data

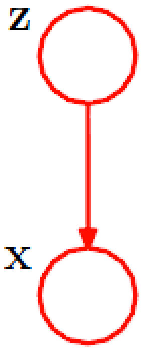
Note: Other topics related to EM can be found on this more extensive set of lecture notes.

- Bishop CM, *Pattern Recognition and Machine Learning*, Springer, 2006 (Chapter 8)
- Murphy, K., Machine Learning: A Probabilistic Perspective (Chapter 11)
- M. Jordan, An Introduction to Graphical Models, unpublished (Chapters 9 and 10)

# Gaussian Mixtures Revisited

# Gaussian Mixtures Revisited

- We will maximize the log of the joint distribution of latent and observed variables (complete data log likelihood), averaged with respect to the posterior distribution *p(**Z**|**X**)* of the latent variables <lnp(**X**,**Z**|θ)> – i.e. replace $z_{nk}$ with the $\gamma(z_{nk})$ (E step of the EM algorithm)



$$\gamma(z_{nk}) = \mathbb{E}[z_{nk}] \text{ (responsibilities)}$$

- This will give us the M step of the EM algorithm: Maximize

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

# Gaussian Mixtures Revisited

- Our original problem was to maximize the complete-data log likelihood:

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

- We change the problem statement by maximizing the log of the joint distribution of latent and observed variables, averaged with respect to the posterior distribution *p(**Z**|**X**)* of the latent variables <lnp(**X**,**Z**|θ)> – i.e. replace $z_{nk}$ with the $\gamma(z_{nk})$

$$\left\langle\ln p\left(\boldsymbol{X},\boldsymbol{Z}\middle|\theta\right)\right\rangle=\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma\left(z_{nk}\right)\left\{\ln\pi_k+\ln\mathcal{N}\left(\boldsymbol{x}\middle|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\right)\right\}$$

where

$$\gamma\left(z_{nk}\right)=\mathbb{E}\left[z_{nk}\right]\text{ (responsibilities)}$$

# Gaussian Mixtures Revisited: Summary

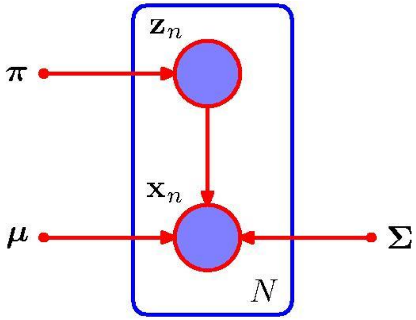
- Assume that for each $\mathbf{x}_n$ we are given the discrete variable (latent assignment variables) $z_n$

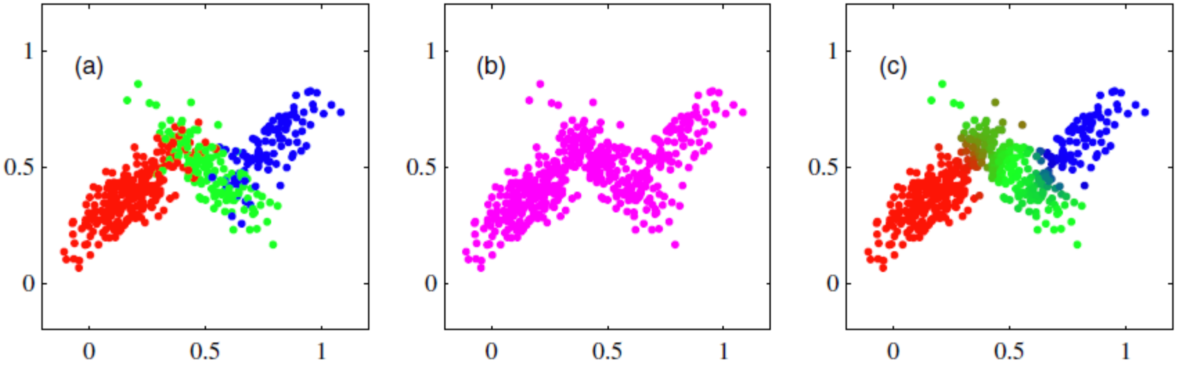
- Complete-data log-likelihood and expectation

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

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

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





# Gaussian Mixtures Revisited: Summary

- ❑ This leads to the EM algorithm for Gaussian mixtures discussed earlier.

- ❑ Choose initial values for $\boldsymbol{\mu}^{old}$, $\boldsymbol{\Sigma}^{old}$ and $\boldsymbol{\pi}^{old}$, and use these to evaluate the responsibilities (E step).

- ❑ Keep the responsibilities fixed and maximize

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

  with respect to $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ and $\boldsymbol{\pi}_k$ (M step).

- ❑ This leads to closed form solutions for $\boldsymbol{\mu}^{new}$, $\boldsymbol{\Sigma}^{new}$ and $\boldsymbol{\pi}^{new}$ identical as before (see proof of one of these next):

$$N_k = \sum_n \gamma(z_{nk}),\ \boldsymbol{\mu}_k^{new} = \frac{\sum_{n=1}^{N}\gamma(z_{nk})\boldsymbol{x}_n}{N_k},\ \boldsymbol{\Sigma}_k^{new} = \frac{\sum_{n=1}^{N}\gamma(z_{nk})(\boldsymbol{x}_n-\boldsymbol{\mu}_k)(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^T}{N_k},\ \pi_k^{new} = \frac{N_k}{N}$$

# Gaussian Mixtures Revisited: Summary

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

- We can write the rhs as follows:

$$-\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)+const$$

- Where the constant are terms independent of $\boldsymbol{\mu}_k$. Taking derivative wrt $\boldsymbol{\mu}_k$:

$$-\sum_{n=1}^{N}\gamma(z_{nk})\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{\mu}_k-\boldsymbol{x}_n)=0 \Rightarrow \boldsymbol{\mu}_k\sum_{n=1}^{N}\gamma(z_{nk}) = \sum_{n=1}^{N}\gamma(z_{nk})\boldsymbol{x}_n \Rightarrow \boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N}\gamma(z_{nk})\boldsymbol{x}_n}{N_k}$$

- Similarly, one can derive expressions for $\boldsymbol{\Sigma}^{new}$ and $\boldsymbol{\pi}^{new}$.

# EM Algorithm Vs. K-Means Algorithm

- ❑ K-Means does hard (unique) assignment of each point to a class. EM makes soft assignments based on posterior probabilities (responsibilities).
- ❑ K-Means is a certain limit of EM for Gaussian mixtures.
- ❑ Consider a Gaussian model with $\mathcal{N}(\boldsymbol{x}\,|\,\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi\varepsilon)^{D/2}} \exp\left\{ -\frac{1}{2\varepsilon} \left\| \boldsymbol{x} - \boldsymbol{\mu}_k \right\|^2 \right\}$ for all k (same ε)
- ❑ For a fixed ε and K-Gaussian mixture and assuming all $\pi_j \neq 0$, the responsibilities are:

$$\gamma(z_{nk}) = \frac{\exp\left\{ -\left\| \boldsymbol{x}_n - \boldsymbol{\mu}_k \right\|^2 \Big/ 2\varepsilon \right\} \pi_k}{\sum_j \exp\left\{ -\left\| \boldsymbol{x}_n - \boldsymbol{\mu}_j \right\|^2 \Big/ 2\varepsilon \right\} \pi_j}$$

- ❑ Consider: $\varepsilon \to 0$. Note that in this case and regardless of the $\pi_j \neq 0$ $\gamma(z_{nk}) \to r_{nk}$, where $r_{nk} = \begin{cases} 1 & if\ k = \arg\min_j \left\| \boldsymbol{x}_n - \boldsymbol{\mu}_j \right\|^2 \\ 0 & \text{otherwise} \end{cases}$ i.e. each data point is assigned to the closest mean.

# EM Algorithm Vs. K-Means Algorithm

- The EM re-estimation equation for the $\mu_k$ becomes in this case:

$$\boldsymbol{\mu}_k = \frac{1}{\sum_n \gamma(z_{nk})} \sum_{n=1}^{N} \gamma(z_{nk}) \boldsymbol{x}_n \to \frac{\sum_{n=1}^{N} r_{nk} \boldsymbol{x}_n}{\sum_{n=1}^{N} r_{nk}}$$

- The mixing coefficients $\pi_k$ are equal to the fraction of data points assigned to cluster k.

- The expected complete-data log-likelihood for $\varepsilon \to 0$ becomes:

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

$$\to -\frac{1}{2}\underbrace{\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\left\|\boldsymbol{x}_n - \boldsymbol{\mu}_k\right\|^2}_{J} + \text{constant}$$

- Thus maximizing $Q(\boldsymbol{\theta})$ is equivalent to mimimizing J in K-means

# Mixture of Bernoulli Distributions

# Mixture of Discrete Binary Variables

- Consider *D binary variables* $x_i$*, i=1,…,D each governed by a Bernoulli distribution with parameter* $\mu_i$ ($x_i$ independent given $\mu$):

$$p\left(\boldsymbol{x}|\boldsymbol{\mu}\right)=\prod_{k=1}^{D}\mu_k^{x_k}\left(1-\mu_k\right)^{1-x_k},$$

$$\boldsymbol{x}=\left\{x_1,\ldots,x_D\right\}^T,\ \boldsymbol{\mu}=\left\{\mu_1,\ldots,\mu_D\right\}^T$$

- The mean and covariance of this distribution are:

$$\mathbb{E}\left[\boldsymbol{x}\right]=\boldsymbol{\mu},\ \mathrm{cov}\left[\boldsymbol{x}\right]=diag\left\{\mu_i\left(1-\mu_i\right)\right\}$$

- Consider a mixture of these Bernoulli distributions

$$p\left(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\pi}\right)=\sum_{k=1}^{K}\pi_k p\left(\boldsymbol{x}|\boldsymbol{\mu}_k\right),\quad where\ \ p\left(\boldsymbol{x}|\boldsymbol{\mu}_k\right)=\prod_{i=1}^{D}\mu_{ki}^{x_i}\left(1-\mu_{ki}\right)^{1-x_i},$$

$$\boldsymbol{\pi}=\left\{\pi_1,\ldots,\pi_K\right\}^T,\ \boldsymbol{\mu}=\left\{\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_K\right\}^T$$

- Lazarsfeld, P. F. and N. W. Henry (1968). *Latent Structure Analysis*. Houghton Mifflin.
- McLachlan, G. J. and D. Peel (2000). *Finite Mixture Models*. Wiley.

# Mixture of Bernoulli Distributions

$$p\left(\mathbf{x}\middle|\boldsymbol{\mu},\boldsymbol{\pi}\right)=\sum_{k=1}^{K}\pi_k\, p\left(\mathbf{x}\middle|\boldsymbol{\mu}_k\right), \quad where \quad p\left(\mathbf{x}\middle|\boldsymbol{\mu}_k\right)=\prod_{i=1}^{D}\mu_{ki}^{x_i}\left(1-\mu_{ki}\right)^{1-x_i},$$

$$\boldsymbol{\pi}=\left\{\pi_1,\ldots,\pi_K\right\}^T, \ \boldsymbol{\mu}=\left\{\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_K\right\}^T$$

- ❑ For any mixture distribution of the form $p(\mathbf{x})=\sum_{k=1}^{K}\pi_k\, p\left(\mathbf{x}\middle|k\right)$ with mean and covariance of p(**x**|k) being $\boldsymbol{\mu}_k$ and $\Sigma_k$, respectively, the mean and covariance of this mixture distribution are given as:

$$\mathbb{E}\left[\boldsymbol{x}\right]=\sum_{k=1}^{K}\pi_k\boldsymbol{\mu}_k$$

$$Use: \mathbb{E}\left[\boldsymbol{x}\right]=\sum_{k=1}^{K}p(z=k)\mathbb{E}\left[\boldsymbol{x}\,|\,z=k\right],$$

$$\mathrm{cov}\left[\boldsymbol{x}\right]=\sum_{k=1}^{K}p(z=k)\mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^T\,|\,z=k\right]-\mathbb{E}\left[\boldsymbol{x}\right]\mathbb{E}\left[\boldsymbol{x}\right]^T$$

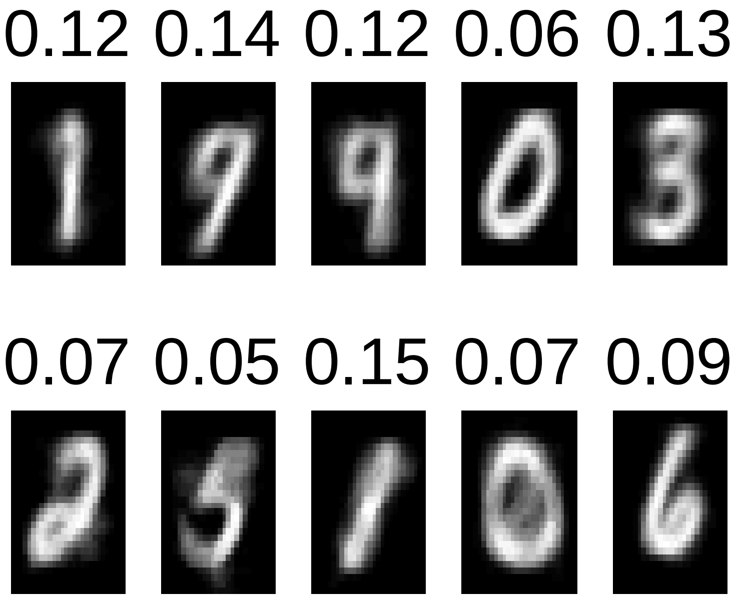
$$\mathrm{cov}\left[\boldsymbol{x}\right]=\mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^T\right]-\mathbb{E}\left[\boldsymbol{x}\right]\mathbb{E}\left[\boldsymbol{x}\right]^T=\sum_{k=1}^{K}\pi_k\mathbb{E}_k\left[\boldsymbol{x}\boldsymbol{x}^T\right]-\mathbb{E}\left[\boldsymbol{x}\right]\mathbb{E}\left[\boldsymbol{x}\right]^T$$

$$=\sum_{k=1}^{K}\pi_k\left\{\boldsymbol{\Sigma}_k+\boldsymbol{\mu}_k\boldsymbol{\mu}_k^T\right\}-\mathbb{E}\left[\boldsymbol{x}\right]\mathbb{E}\left[\boldsymbol{x}\right]^T$$

- ❑ The joint distribution is not factorized. *The mixture distribution captures correlations* between variables unlike the single product of Bernoullis model.

# Example: Mixture of Bernoullis and MLE

- We fit a mixture of Bernoullis to the MNIST handwritten digit data set using $K = 10$ and visualize the centroids (*MLE* of cluster means). The numbers on the top are MLE of the mixing weights.
- This discovers some digit classes, but it creates multiple clusters for some digits and no clusters for others. The reasons for this include:
  - The model is very simple (no notion of shape or a stroke).
  - Some digits exhibit a degree of visual variety (e.g. 7's with and without the cross bar).
  - We need $K \geq 10$ clusters for this data.
  - Using a large $K$, we create multiple versions of the same digit. One can use model selection to prevent this.
  - The likelihood function is not convex, so we may be stuck in a local optimum.
- One must be cautious trying to interpret any clusters that are discovered by the method.
- Using informative priors can help.



mixBerMnistEM
from Kevin Murphys' PMTK

# Mixture of Bernoulli Distributions

- If we are given a data set $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ then the log likelihood function for this model is given by

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

- Note the summation inside the log. The MLE again does not have a closed solution.

- To derive the EM algorithm, we introduce a latent variable ***z*** associated with each instance of ***x***. $\mathbf{z} = (z_1, \ldots, z_K)^T$ is a binary K-dimensional variable having a single component equal to 1, with all other components equal to 0:

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

- The prior for the latent variables is:

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

# Mixture of Bernoulli Distributions

- The complete-data log-likelihood is:

$$\ln p\left(\boldsymbol{X},\boldsymbol{Z}\middle|\boldsymbol{\mu},\boldsymbol{\pi}\right) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\left\{\ln \pi_k \right.$$

$$\left. +\sum_{i=1}^{D}\left[x_{ni}\ln\mu_{ki} + (1-x_{ni})\ln(1-\mu_{ki})\right]\right\}, \text{ where}: \boldsymbol{X} = \left\{\boldsymbol{x}_n\right\}, \boldsymbol{Z} = \left\{\boldsymbol{z}_n\right\}$$

- Take the expectation of the complete-data log likelihood with respect to the posterior distribution of the latent variables:

$$\mathbb{E}_{\boldsymbol{Z}}\left[\ln p\left(\boldsymbol{X},\boldsymbol{Z}\middle|\boldsymbol{\mu},\boldsymbol{\pi}\right)\right] = \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma_{nk}\left\{\ln \pi_k \right.$$

$$\left. +\sum_{i=1}^{D}\left[x_{ni}\ln\mu_{ki} + (1-x_{ni})\ln(1-\mu_{ki})\right]\right\}, \text{ where}: \gamma_{nk} = \mathbb{E}\left[z_{nk}\right]$$

*$\gamma(z_{nk})$* is the posterior probability, or responsibility, of component *k* given data point ***$\boldsymbol{x}_n$*.**

# Mixture of Bernoulli Distributions

- E-Step: Compute the Responsibilities using Bayes' rule

$$\gamma_{nk} = \mathbb{E}\left[z_{nk}\right] = \frac{\sum_{\boldsymbol{z}_n} z_{nk} \prod_{k'} \left(\pi_{k'} p(\boldsymbol{x}_n \mid \boldsymbol{\mu}_{k'})\right)^{z_{nk'}}}{\sum_{\boldsymbol{z}_n} \prod_{j} \left(\pi_j p(\boldsymbol{x}_n \mid \boldsymbol{\mu}_j)\right)^{z_{nj}}} = \frac{\pi_k p(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k)}{\sum_{j=1}^{K} \pi_j p(\boldsymbol{x}_n \mid \boldsymbol{\mu}_j)}$$

- Considering the sum in n,

$$\mathbb{E}_{\boldsymbol{Z}}\left[\ln p\left(\boldsymbol{X},\boldsymbol{Z} \middle| \boldsymbol{\mu},\boldsymbol{\pi}\right)\right] = \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}\left\{\ln \pi_k + \sum_{i=1}^{D}\left[x_{ni}\ln \mu_{ki} + (1-x_{ni})\ln(1-\mu_{ki})\right]\right\}$$

we note that the responsibilities come through the following terms:

$$N_k = \sum_{n=1}^{N} \gamma_{nk}\,, \quad \bar{\boldsymbol{x}}_k = \frac{1}{N_k}\sum_{n=1}^{N} \gamma_{nk}\,\boldsymbol{x}_n$$

- *$N_k$ is the effective number of data points associated with component k.*

# Mixture of Bernoulli Distributions

- □ M step: Maximize the expected complete-data log likelihood with respect to the parameters $\boldsymbol{\mu}_k$ and $\boldsymbol{\pi}$.

- □ If we set the derivative of

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

with respect to $\boldsymbol{\mu}_{ki}$ equal to zero and rearrange the terms, we obtain

$$\frac{\partial}{\partial\mu_{ki}}\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}\sum_{i=1}^{D}\left[x_{ni}\ln\mu_{ki}+(1-x_{ni})\ln(1-\mu_{ki})\right]=0\Rightarrow\sum_{n=1}^{N}\gamma_{nk}\left(\frac{x_{ni}}{\mu_{ki}}-\frac{1-x_{ni}}{1-\mu_{ki}}\right)=0\Rightarrow$$

$$\sum_{n=1}^{N}\frac{x_{ni}\gamma_{nk}-\mu_{ki}\gamma_{nk}}{\mu_{ki}\left(1-\mu_{ki}\right)}=0\Rightarrow\sum_{n=1}^{N}x_{ni}\gamma_{nk}=\mu_{ki}\sum_{n=1}^{N}\gamma_{nk}\Rightarrow\quad \mu_k=\bar{\boldsymbol{x}}_k\equiv\frac{1}{N_k}\sum_{n=1}^{N}\gamma_{nk}\,\boldsymbol{x}_n$$

- □ The mean of component k is equal to a weighted mean of the data. The weighting coefficients are given by the responsibilities that component k takes for data points.

# Mixture of Bernoulli Distributions

- For the maximization with respect to $\pi_k$, we enforce the constraint $\sum_k \pi_k = 1$.
- As for the mixture of Gaussians, we then obtain

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

- The mixing coefficient for component k is given by the effective fraction of points in the data set explained by that component.

# Degenerate Case: Initialization

- Note that the following holds for the mixture of Bernoulli distributions:

$$\mathbb{E}[x]=\sum_{k=1}^{K}\pi_k\boldsymbol{\mu}_k=\sum_{k=1}^{K}\underbrace{\pi_k\frac{1}{N_k}}_{1/N}\sum_{n=1}^{N}\gamma(z_{nk})\boldsymbol{x}_n=\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n\underbrace{\sum_{k=1}^{K}\gamma(z_{nk})}_{1}=\overline{\boldsymbol{x}}$$

- If we initialize the means by setting them to a common value $\boldsymbol{\mu}_k = \widehat{\boldsymbol{\mu}}, k = 1,\ldots,K$, then:

$$\gamma_{nk}=\frac{\pi_k p(\boldsymbol{x}_n\,|\,\boldsymbol{\mu}_k)}{\sum_{j=1}^{K}\pi_j p(\boldsymbol{x}_n\,|\,\boldsymbol{\mu}_j)}=\frac{\pi_k}{\sum_{j=1}^{K}\pi_j}=\pi_k\ (independent\ of\ n)$$

$$\boldsymbol{\mu}_k=\frac{1}{N_k}\sum_{n=1}^{N}\gamma_{nk}\boldsymbol{x}_n=\pi_k\frac{1}{N_k}N\overline{\boldsymbol{x}}=\overline{\boldsymbol{x}}$$

i.e. all means converge to the MLE estimate and will never be updated.

- This is a degenerate case that needs to be avoided with proper initialization.

# Mixture of Bernoulli Distributions

- In contrast to the mixture of Gaussians, there are no singularities when the likelihood function goes to infinity.

- This can be seen by noting that the likelihood function is bounded above. Indeed:

$$0 \le p(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k) \le 1,\ 0 \le \pi_k \le 1,\ \sum_k \pi_k = 1$$

- Then note that the max value of lnp(**X**|μ,π) is zero.

$$\ln p\left(\boldsymbol{X} \middle| \boldsymbol{\mu}, \boldsymbol{\pi}\right) = \sum_{n=1}^{N} \ln \underbrace{\left\{ \sum_{k=1}^{K} \pi_k\, p\left(\boldsymbol{x}_n \middle| \boldsymbol{\mu}_k\right) \right\}}_{\max\ value\ 1}$$

- The likelihood function can go to zero
  - These singularities would not be found provided that EM is not initialized to a pathological starting point.
  - Recall that the EM algorithm always increases the value of the likelihood function until a local maximum is found.

# Example: Modeling of Handwritten Digits

- We illustrate the Bernoulli mixture with modeling of handwritten digits.

- We convert the digit images to binary vectors by setting all elements whose values exceed 0.5 to 1 and the remaining elements to 0.

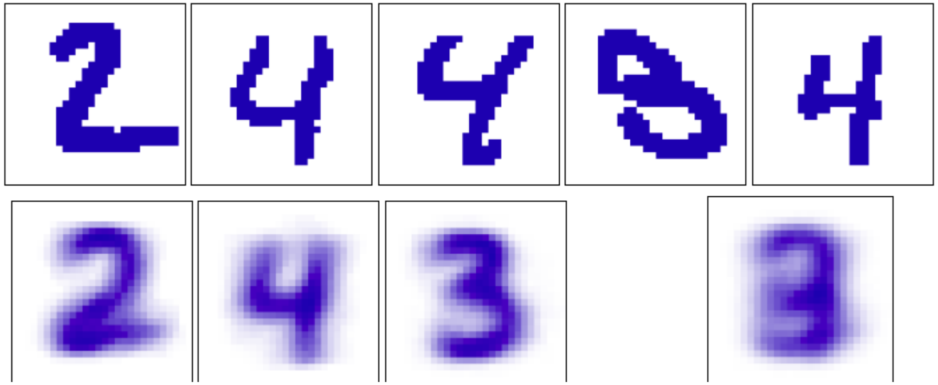
- We fit N = 600 digits, comprising the digits ‘2’, ‘3’, and ‘4’, with a *mixture of K = 3 Bernoulli distributions*. We run 10 iterations of EM.

- The mixing coefficients were initialized to $\pi_k = 1/K$, and the parameters $\mu_{kj}$ *were set to random values chosen uniformly in the range (0.25, 0.75) and then normalized* to satisfy the constraint

$$\sum_j \mu_{kj} = 1$$

# Example: Modeling of Handwritten Digits

- *A mixture of 3 Bernoulli distributions* is able to find the 3 clusters in the data corresponding to the different digits.



MatLab Code

- On the top: examples from the data after converting the pixel values from grey scale to binary using a threshold of 0.5.

- On the bottom: the first three images show the parameters $\mu_{ki}$ for each of the three components in the mixture model.

- On the bottom, last image: we fit the data set using a single multivariate Bernoulli distribution with MLE. This amounts to averaging the counts in each pixel.

# Mixture of Bernoulli Distributions: Summary

- Bernoulli distributions over binary data vectors

$$p\left(\mathbf{x}\middle|\boldsymbol{\mu}\right)=\prod_{k=1}^{D}\mu_k^{x_k}\left(1-\mu_k\right)^{1-x_k}$$

- Mixture of Bernoullis can model variable correlations

- Bernoulli is member of the exponential family
  - The model is log-linear but the mixture is not. The complete-data log-likelihood however is.

- Simple EM algorithm to find MLE parameters
  - E-Step: Compute responsibilities $\gamma(z_{nk})\propto\pi_k\,p\left(\mathbf{x}_n\middle|\boldsymbol{\mu}_k\right)$
  - M-Step: Update parameters

$$\pi_k=\sum_n\gamma(z_k)/N,\ \boldsymbol{\mu}_k=\sum_{n=1}^{N}\gamma(z_k)\mathbf{x}_n/\left(N\pi_k\right)$$

# Mixture of Bernoulli Distributions: Extensions

- The conjugate prior for the parameters of a Bernoulli distribution is given by the beta distribution.

  - Recall that a beta prior is equivalent to introducing additional effective observations of **x.**

- We can introduce priors into the Bernoulli mixture model, and use EM to maximize the posterior probability distributions.

- Can extend the analysis of Bernoulli mixtures to multinomial binary variables having *M > 2* states by making use of the discrete distribution

$$p(\boldsymbol{x} \mid \boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$$

- We can then introduce Dirichlet prior $p(\pi|\alpha)$ and Beta priors $p(\mu_k|a_k, b_k)$.

# Mixture of Bernoulli Distributions: MAP

- The E-Step remains the same and in the M-step we need to maximize the following:

$$\mathcal{Q}\left(\boldsymbol{\theta},\boldsymbol{\theta}^{old}\right)+\ln p(\boldsymbol{\theta})=$$

$$\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}\left\{\ln\pi_k+\sum_{i=1}^{D}\left[x_{ni}\ln\mu_{ki}+(1-x_{ni})\ln(1-\mu_{ki})\right]\right\}$$

$$+\sum_{j=1}^{K}\sum_{i'=1}^{D}\left(\left(a_j-1\right)\ln\mu_{ji'}+\left(b_j-1\right)\ln\left(1-\mu_{ji'}\right)\right)+\sum_{l=1}^{K}(\alpha_l-1)\ln\pi_l$$

- Maximizing wrt to $\boldsymbol{\mu}_{ki}$ gives: $\mu_{ki}=\dfrac{N_k\bar{\boldsymbol{x}}_{ki}+a_k-1}{N_k+a_k-1+b_k-1}$, $\bar{\boldsymbol{x}}_{ki}=\dfrac{1}{N_k}\sum_{n=1}^{N}\gamma_{nk}\boldsymbol{x}_{ni}$

- Maximization wrt to $\pi_k$ using a Lagrange multiplier for $\sum_j \pi_j=1$ gives:

$$\pi_k=\frac{N_k+\alpha_k-1}{N+\alpha_0-K}$$

# MAP Estimation

- The overfitting of MLE may be severe. This can be addressed by performing MAP estimation. The new auxiliary function is the expected complete data log-likelihood plus the log prior:

$$\mathcal{Q}\ \ \boldsymbol{\theta},\boldsymbol{\theta}^{old}\ \ =\left[\sum_i\sum_k \gamma_{ik}\log\pi_{ik}+\sum_i\sum_k \gamma_{ik}\log p\ \ \boldsymbol{x}_i\,|\,\boldsymbol{\theta}_k\ \ \right]+\log p\ \ \pi\ \ +\sum_k \log p\ \ \boldsymbol{\theta}_k$$

- The E step is unchanged, but the M step needs to be modified.

- For the prior on the mixture weights, it is natural to use a Dirichlet prior, $\boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\alpha})$, since this is conjugate to the categorical distribution. The MAP estimate is given by

$$\pi_k=\frac{N_k+\alpha_k-1}{N+\sum_k \alpha_k-K}$$

- For a uniform prior, $\alpha_k = 1$, this reduces to MLE.

# MAP Estimation

- The prior on $\boldsymbol{\theta}_k$, $p(\boldsymbol{\theta}_k)$, depends on the form of the class conditional densities. We discuss the case of GMMs below.

- For simplicity, let us consider a conjugate prior of the form

$$p(\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) = \mathcal{NIW}(\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k|\mathbf{m}_0,\, \kappa_0,\, \nu_0,\, \mathbf{S}_0)$$

# MAP Estimation

- Using the results from an earlier lecture on Bayesian inference for Gaussian models, the MAP estimate is given by

$$\hat{\boldsymbol{\mu}}_k = \frac{\kappa_0 \boldsymbol{m}_0 + N_k \overline{\boldsymbol{x}}_k}{N_k + \kappa_0},\ N_k = \sum_i \gamma_{ik}\,,\ \overline{\boldsymbol{x}}_k = \frac{\sum_i \gamma_{ik} \boldsymbol{x}_i}{N_k}$$

$$\hat{\boldsymbol{\Sigma}}_k = \frac{\boldsymbol{S}_0 + \boldsymbol{S}_k + \frac{\kappa_0 N_k}{\kappa_0 + N_k}(\overline{\boldsymbol{x}}_k - \boldsymbol{m}_0)(\overline{\boldsymbol{x}}_k - \boldsymbol{m}_0)^T}{\nu_0 + N_k + D + 2},\ \boldsymbol{S}_k = \sum_{i=1}^{N} \gamma_{ik}(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_k)(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_k)^T$$

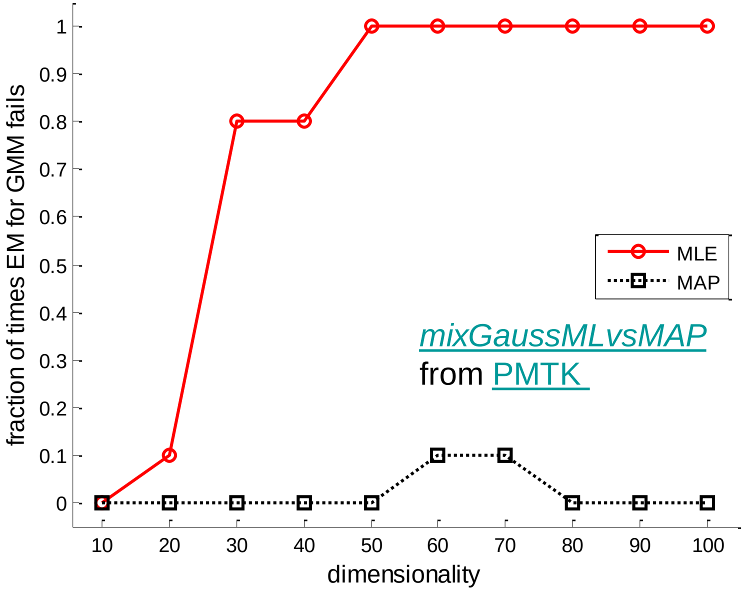
- We now illustrate the benefits of using MAP estimation instead of ML estimation in the context of GMMs. We apply EM to some synthetic data in *D* dimensions, using either ML or MAP estimation.

- We count the trial as a failure if there are numerical issues involving singular matrices. For each dimensionality, we conduct 5 random trials. The results are illustrated next using *N* = 100.

# MAP Estimation

- For *D* large, ML estimation crashes, whereas MAP works.
- When using MAP estimation, we need to specify the hyper-parameters. We can set $\kappa_0 = 0$, so that the $\boldsymbol{\mu}_k$ are unregularized, since the numerical problems only arise from $\boldsymbol{\Sigma}_k$. In this case, the MAP estimates simplify to

$$\hat{\boldsymbol{\mu}}_k = \overline{\boldsymbol{x}}_k,\ \hat{\boldsymbol{\Sigma}}_k = \frac{\boldsymbol{S}_0 + \boldsymbol{S}_k}{\nu_0 + N_k + D + 2}$$



*mixGaussMLvsMAP* from PMTK

- Fraley, C. and A. Raftery (2002). Model-based clustering, discriminant analysis, and density estimation. *J. of the Am. Stat. Assoc.* (97), 611–631 (see pp. 163)

- Using the pooled variance $s_j$ for each dimension j, we set:

$$\boldsymbol{S}_0 = \frac{1}{K^{1/D}} diag\left(s_1^2,...,s_D^2\right), s_j = \frac{1}{N}\sum_{i=1}^{N}\left(x_{ij} - \overline{x}_j\right)^2$$

- With the $1/K^{1/D}$ term, the volume of each ellipsoid is then given by

$$\left|\boldsymbol{S}_0\right| = \frac{1}{K}\left|diag\left(s_1^2,...,s_D^2\right)\right|$$

- The parameter $v_0$ controls how strongly we believe this prior. The weakest proper prior we can use, is to set $v_0 = D + 2$.

# Other Applications of EM

# EM for Bayesian Linear Regression

- Recall Bayesian Linear Regression:

$$p\left(\mathbf{t}\middle|\mathbf{w},\beta,\boldsymbol{X}\right)=\prod_{n=1}^{N}\mathcal{N}\left(t_n;\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n),\beta^{-1}\right)$$ Likelihood

$$p\left(\mathbf{w}\middle|\alpha\right)=\mathcal{N}\left(\mathbf{w};\alpha^{-1}\boldsymbol{I}\right)$$ Prior

$$p\left(\mathbf{t}\middle|\alpha,\beta,\boldsymbol{X}\right)=\int p\left(\mathbf{t}\middle|\mathbf{w},\beta\right)p\left(\mathbf{w}\,|\,\alpha\right)d\mathbf{w}$$ Marginal Likelihood

- Our goal is to maximize the evidence function *p(**t**|**α, β**)* with respect to *α* and *β.*
- Because ***w*** is marginalized out, we can regard it as a latent variable, and hence *we can optimize this marginal likelihood function using EM.*
- E step: compute the posterior distribution of **w** given the current setting of the parameters *α* and *β* and then use this to find the expected complete-data log likelihood.
- M step: maximize this quantity with respect to *α and β.*

# EM for Bayesian Linear Regression

- We have already derived the posterior distribution of **w** given by

$$p\left(\mathbf{w}\middle|\mathbf{t}\right)=\mathcal{N}\left(\mathbf{w};\mathbf{m}_N,\mathbf{S}_N\right),$$

$$\mathbf{m}_N=\mathbf{S}_N\left(\mathbf{S}_0^{-1}\mathbf{m}_0+\beta\mathbf{\Phi}^T\mathbf{t}\right)\quad \mathbf{S}_N^{-1}=\mathbf{S}_0^{-1}+\beta\mathbf{\Phi}^T\mathbf{\Phi},\quad \mathbf{S}_0^{-1}=\alpha^{-1}\mathbf{I}$$

- The complete-data log likelihood function is then given by

ln *p(**t**,**w**|α, β)=lnp(**t**|**w**, β)+lnp(**w**|α)* where:

$$p\left(\mathbf{t}\middle|\mathbf{w},\beta,\mathbf{X}\right)=\prod_{n=1}^{N}\mathcal{N}\left(t_n;\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n),\beta^{-1}\right)$$

$$p\left(\mathbf{w}\middle|\alpha\right)=\mathcal{N}\left(\mathbf{w};\alpha^{-1}\mathbf{I}\right)$$

- Taking the expectation wrt the posterior of ***w*** gives:

$$\mathbb{E}\left[\ln p\left(\mathbf{t},\mathbf{w}\middle|\alpha,\beta\right)\right]=\frac{M}{2}\ln\left(\frac{\alpha}{2\pi}\right)-\frac{\alpha}{2}\mathbb{E}\left[\mathbf{w}^T\mathbf{w}\right]+\frac{N}{2}\ln\left(\frac{\beta}{2\pi}\right)-\frac{\beta}{2}\sum_{n=1}^{N}\mathbb{E}\left[\left(t_n-\mathbf{w}^T\boldsymbol{\phi}_n\right)^2\right]$$

- M Step: Setting the derivatives wrt to α and β zero and using $\mathbb{E}\left[\mathbf{w}^T\mathbf{w}\right]=\mathbf{m}_N^T\mathbf{m}_N+Tr\left[\mathbf{S}_N\right]$ and $\mathbb{E}\left[\left(t_n-\mathbf{w}^T\boldsymbol{\phi}_n\right)^2\right]=\left(t_n-\mathbf{m}_N^T\boldsymbol{\phi}_n\right)^2+Tr\left[\boldsymbol{\phi}_n\boldsymbol{\phi}_n^T\mathbf{S}_N\right]$ we obtain:

$$\alpha^{-1}=\frac{1}{M}\left(\mathbf{m}_N^T\mathbf{m}_N+Tr\left[\mathbf{S}_N\right]\right),\quad \beta^{-1}=\frac{1}{N}\left(\left\|\mathbf{t}-\mathbf{\Phi}\mathbf{m}_N\right\|+Tr\left[\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{S}_N\right]\right)$$

# EM for Bayesian Linear Regression

- The re-estimation eqs

$$\alpha^{-1} = \frac{1}{M}\left(\boldsymbol{m}_N^T\boldsymbol{m}_N + Tr\left(\boldsymbol{S}_N\right)\right), \quad \beta^{-1} = \frac{1}{N}\left(\left\|\boldsymbol{t} - \boldsymbol{\Phi}\boldsymbol{m}_N\right\| + Tr\left[\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{S}_N\right]\right)$$

  seem slightly different from the corresponding result

$$\alpha = \frac{\gamma}{\boldsymbol{m}_N^T\boldsymbol{m}_N}, \quad \gamma = \sum_i \frac{\lambda_i}{\alpha + \lambda_i}$$

  derived by direct evaluation of the evidence function.

- Each involve inversion (or eigen decomposition) of an *M* × *M* matrix and hence have comparable computational cost per iteration.

# EM for Bayesian Linear Regression

- The two approaches of determining α converge to the same result (assuming they find the same local maximum of the evidence function). This can be verified by noting that the quantity γ is defined by

$$\gamma = M - \alpha \sum_i \frac{1}{\alpha + \lambda_i} = M - \alpha Tr\left[\boldsymbol{S}_N\right] \qquad \boldsymbol{S}_N^{-1} = \boldsymbol{S}_0^{-1} + \beta \boldsymbol{\Phi}^T \Phi\boldsymbol{,} \quad \boldsymbol{S}_0^{-1} = \alpha^{-1}\boldsymbol{I}$$

$$\beta \boldsymbol{\Phi}^T \boldsymbol{\Phi} \boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i$$

- At a stationary point of the evidence function, the re-estimation equation

$$\alpha = \frac{\gamma}{\boldsymbol{m}_N^T \boldsymbol{m}_N}$$

will be self-consistently satisfied and hence we can substitute for γ to give:

$$\alpha \boldsymbol{m}_N^T \boldsymbol{m}_N = \gamma = M - \alpha Tr\left[\boldsymbol{S}_N\right]$$

- Solving for α we obtain $\alpha^{-1} = \frac{1}{M}\left(\boldsymbol{m}_N^T \boldsymbol{m}_N + Tr\left[\boldsymbol{S}_N\right]\right)$

# The EM Algorithm Revisited

# The EM Algorithm in General

- Let **X** be the observed variables, **Z** denote all latent variables and **θ** the set of all parameters.
- Our goal is as before to maximize the likelihood:

$$p(\boldsymbol{X}\,|\,\boldsymbol{\theta}) = \int_{\boldsymbol{Z}} p(\boldsymbol{X},\boldsymbol{Z}\,|\,\boldsymbol{\theta})d\boldsymbol{Z}$$

- We assume that $p(\boldsymbol{X},\boldsymbol{Z}\,|\,\boldsymbol{\theta})$ is easier to compute than $p(\boldsymbol{X}\,|\,\boldsymbol{\theta})$.
- *Introduce an arbitrary distribution q(**Z**) over the latent variables*. One can then show:

$$\ln p\left(\boldsymbol{X}\,|\,\boldsymbol{\theta}\right) = \mathcal{L}\left(q,\boldsymbol{\theta}\right) + KL\left(q\|p\right)$$

where:

$$\mathcal{L}\left(q,\boldsymbol{\theta}\right) = \int q\left(\boldsymbol{Z}\right)\ln\left\{\frac{p\left(\boldsymbol{X},\boldsymbol{Z}\,|\,\boldsymbol{\theta}\right)}{q\left(\boldsymbol{Z}\right)}\right\}d\boldsymbol{Z}$$

$$KL\left(q\|p\right) = -\int q\left(\boldsymbol{Z}\right)\ln\left\{\frac{p\left(\boldsymbol{Z}\,|\,\boldsymbol{X},\boldsymbol{\theta}\right)}{q\left(\boldsymbol{Z}\right)}\right\}d\boldsymbol{Z},\ KL\left(q\|p\right)\geq 0,\ KL\left(q\|p\right)=0\ if\ \ q(\boldsymbol{Z})=p(\boldsymbol{Z}\,|\,\boldsymbol{X},\boldsymbol{\theta})$$

# The EM Algorithm in General

- For the proof of the identity shown earlier, note that:

$$\mathcal{L}(q,\boldsymbol{\theta}) = \int q(\mathbf{Z}) \ln\left\{ \frac{p(\mathbf{X},\mathbf{Z}\,|\,\boldsymbol{\theta})}{q(\mathbf{Z})} \right\} d\mathbf{Z}$$

$$= \int q(\mathbf{Z}) \ln\big(p(\mathbf{X},\mathbf{Z}\,|\,\boldsymbol{\theta})\big) d\mathbf{Z} - \int q(\mathbf{Z}) \ln q(\mathbf{Z}) d\mathbf{Z}$$

$$= \int q(\mathbf{Z}) \Big[ \ln\big(p(\mathbf{Z}\,|\,\mathbf{X},\boldsymbol{\theta})\big) + \ln p(\mathbf{X}\,|\,\boldsymbol{\theta}) \Big] d\mathbf{Z} - \int q(\mathbf{Z}) \ln q(\mathbf{Z}) d\mathbf{Z}$$

$$= \int q(\mathbf{Z}) \ln\big(p(\mathbf{Z}\,|\,\mathbf{X},\boldsymbol{\theta})\big) d\mathbf{Z} + \ln p(\mathbf{X}\,|\,\boldsymbol{\theta}) - \int q(\mathbf{Z}) \ln q(\mathbf{Z}) d\mathbf{Z}$$

$$= \int q(\mathbf{Z}) \ln\left( \frac{p(\mathbf{Z}\,|\,\mathbf{X},\boldsymbol{\theta})}{q(\mathbf{Z})} \right) d\mathbf{Z} + \ln p(\mathbf{X}\,|\,\boldsymbol{\theta}) = \ln p(\mathbf{X}\,|\,\boldsymbol{\theta}) - KL(q\,||\,p)$$

# Lower Bound on Model Evidence p(X|θ)

- KL(q||p) is Kullback-Leibler distance between q and the posterior p(**Z**|**X**,**q**)

$$KL\left(q\|p\right)=-\int q\left(\boldsymbol{Z}\right)\ln\left\{\frac{p\left(\boldsymbol{Z}\mid\boldsymbol{X},\boldsymbol{\theta}\right)}{q\left(\boldsymbol{Z}\right)}\right\}d\boldsymbol{Z},\; KL\left(q\|p\right)\geq 0,\; KL\left(q\|p\right)=0\; if\;\; q(\boldsymbol{Z})=p(\boldsymbol{Z}\mid\boldsymbol{X},\boldsymbol{\theta})$$

- From

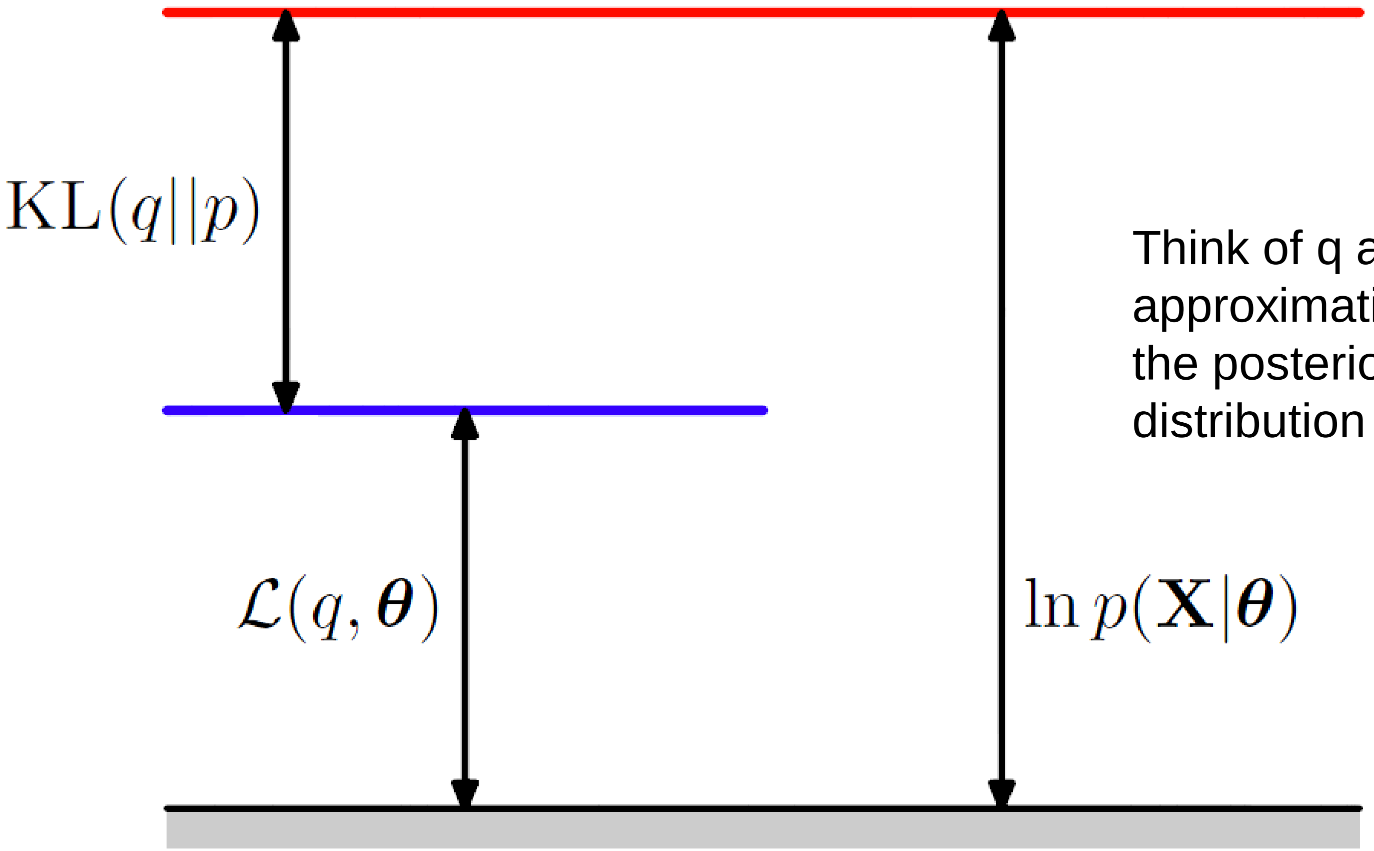
$$\ln p\left(\boldsymbol{X}\mid\boldsymbol{\theta}\right)=\mathcal{L}\left(q,\boldsymbol{\theta}\right)+KL\left(q\|p\right)$$

it follows that $\mathcal{L}\left(q,\boldsymbol{\theta}\right)$ *is a lower bound of lnp(**X**|**q**):*

$$\ln p\left(\boldsymbol{X}\mid\boldsymbol{\theta}\right)=\mathcal{L}\left(q,\boldsymbol{\theta}\right)+KL\left(q\|p\right)\geq\mathcal{L}\left(q,\boldsymbol{\theta}\right)$$

- Maximizing $\mathcal{L}\left(q,\boldsymbol{\theta}\right)$ over *q(**Z**)* would give the true posterior but this is not computationally tractable.

# Variational Lower Bound


$\mathrm{KL}(q||p)$
$\mathcal{L}(q,\boldsymbol{\theta})$
$\ln p(\mathbf{X}|\boldsymbol{\theta})$

Think of q as an approximation to the posterior distribution $p(\boldsymbol{Z}\,|\,\boldsymbol{X},\boldsymbol{\theta})$

Working with lnp(**X|q**) is intractable – e.g. it is log of sum of Gaussians- Work instead with $\mathcal{L}(q,\boldsymbol{\theta})$

*Approximate* $\ln p(\boldsymbol{X}\,|\,\boldsymbol{\theta})$ *with its lower bound* $\mathcal{L}(q,\boldsymbol{\theta})$

*Make* $\mathcal{L}(q,\boldsymbol{\theta})$ *as big as possible or equivalently*

*make* $KL(q\,||\,p)$ *as small as possible*

# The EM Algorithm in General

- Maximizing $\mathcal{L}(q,\boldsymbol{\theta})$ over a free form *q* would give the true posterior but this is not computationally tractable

$$q(\boldsymbol{Z}) = p(\boldsymbol{Z} \mid \boldsymbol{X}, \theta)$$

- The EM algorithm is a two-stage iterative optimization technique for finding maximum likelihood solutions. We can use the decomposition

$$\ln p(\boldsymbol{X} \mid \boldsymbol{\theta}) = \mathcal{L}(q,\boldsymbol{\theta}) + KL(q \| p)$$

to define the EM algorithm and to demonstrate that it does indeed maximize the log likelihood.

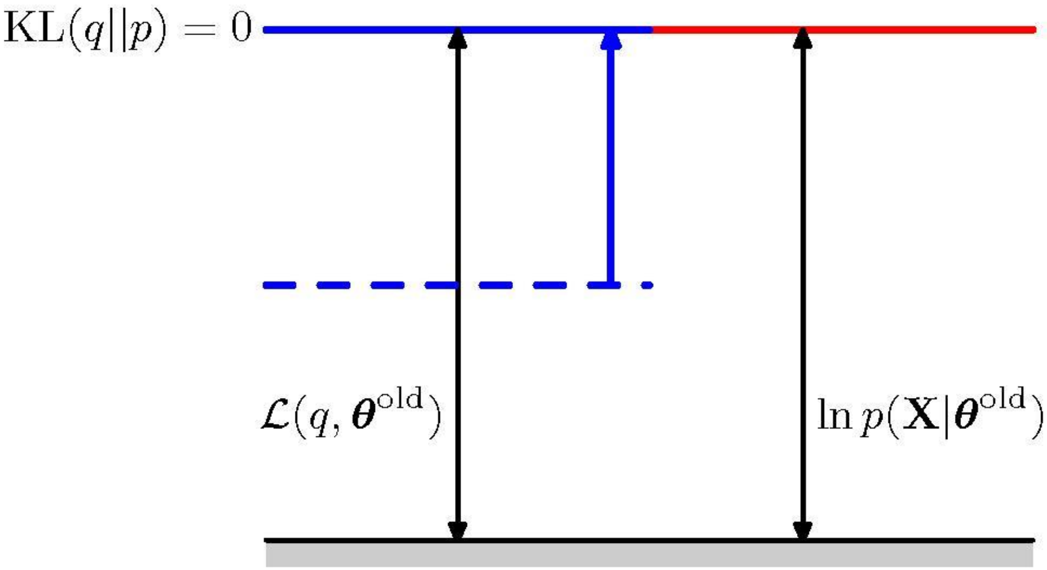
# The EM Algorithm in General

- ❑ Suppose that the current value of the parameter vector is $\theta^{old}$.

- ❑ In the E step, the lower bound *$\mathcal{L}(q, \theta^{old})$ is maximized with respect to $q(\mathbf{Z})$ while holding $\theta^{old}$ fixed*.

- ❑ The solution to this maximization problem is easily seen by noting that the value of *$\ln p(\mathbf{X}|\theta^{old})$ does not depend on $q(\mathbf{Z})$ and so the largest value of $\mathcal{L}(q, \theta^{old})$ will occur when the KL divergence vanishes, i.e. when $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \theta^{old})$.*

- ❑ In this case, the lower bound will equal the log likelihood.


$\mathrm{KL}(q||p) = 0$
$\mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{old}})$
$\ln p(\mathbf{X}|\boldsymbol{\theta}^{\mathrm{old}})$

# The EM Algorithm in General

- The lower bound then becomes:

$$\mathcal{L}(q,\theta)=\sum_{z}\underbrace{p\left(\mathbf{Z}\mid\mathbf{X},\boldsymbol{\theta}_{old}\right)}_{\text{play the role of responsibilities}}\ln\left\{\frac{p\left(\mathbf{X},\mathbf{Z}\mid\boldsymbol{\theta}\right)}{p\left(\mathbf{Z}\mid\mathbf{X},\boldsymbol{\theta}_{old}\right)}\right\}$$

- This as a function of θ is the expected complete-data log likelihood up to an additive constant.

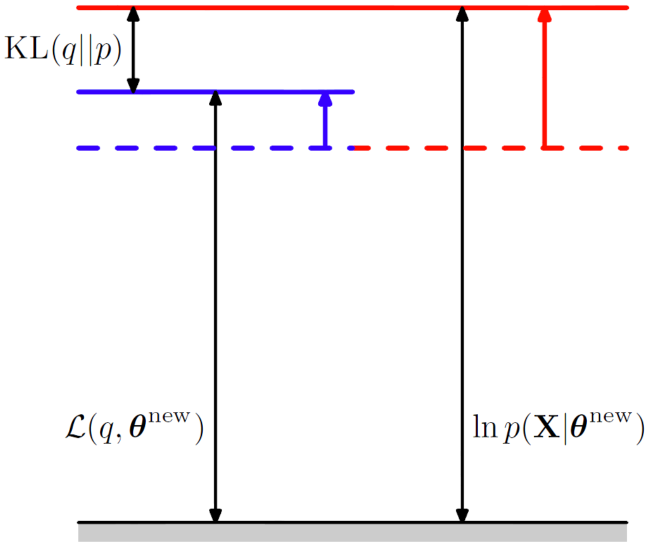
# The EM Algorithm in General

- ❑ In the subsequent M step, the distribution q(**Z**) is held fixed and $\mathcal{L}(q, \boldsymbol{\theta})$ is maximized with respect to **θ** to give some new value $\boldsymbol{\theta}^{new}$.

- ❑ This causes $\mathcal{L}$ to increase (unless it is already at a maximum), which will cause the corresponding log likelihood function to increase.

- ❑ Because q is determined using the old parameter values rather than the new values and is held fixed during the M step, it will not equal the new posterior distribution $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{new})$, and hence there will be a nonzero KL divergence.

- ❑ The increase in the log likelihood function is therefore greater than the increase in the lower bound, as shown.


$\mathrm{KL}(q||p)$
$\mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{new}})$
$\ln p(\mathbf{X}|\boldsymbol{\theta}^{\mathrm{new}})$

# The EM Algorithm in General

- We can show that the maximization in the M step is that of the expected value of the complete data log likelihood.

- Indeed, if we substitute $q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X},\theta)$

  in the lower bound $\mathcal{L}(q,\theta) = \int q(\boldsymbol{Z}) \ln\left\{\frac{p(\boldsymbol{X},\boldsymbol{Z}|\theta)}{q(\boldsymbol{Z})}\right\} d\boldsymbol{Z}$

  we see that the lower bound after the E step becomes

$$\mathcal{L}(q,\boldsymbol{\theta}) = \int q(\boldsymbol{Z}) \ln\left\{\frac{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})}\right\} d\boldsymbol{Z} = \int p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{old}) \ln \frac{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})}{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{old})} d\boldsymbol{Z}$$

$$= \int p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{old}) \ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta}) d\boldsymbol{Z} - \int p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{old}) \ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta}^{old})$$

$$= Q(\boldsymbol{\theta},\boldsymbol{\theta}^{old}) + const$$

  where the constant is the entropy of q and therefore independent of θ.

- $\mathcal{Q}$(**q**,**q**$^{old}$) is the expectation of the complete data log likelihood wrt posterior of the latent variables.

# The EM Algorithm in General

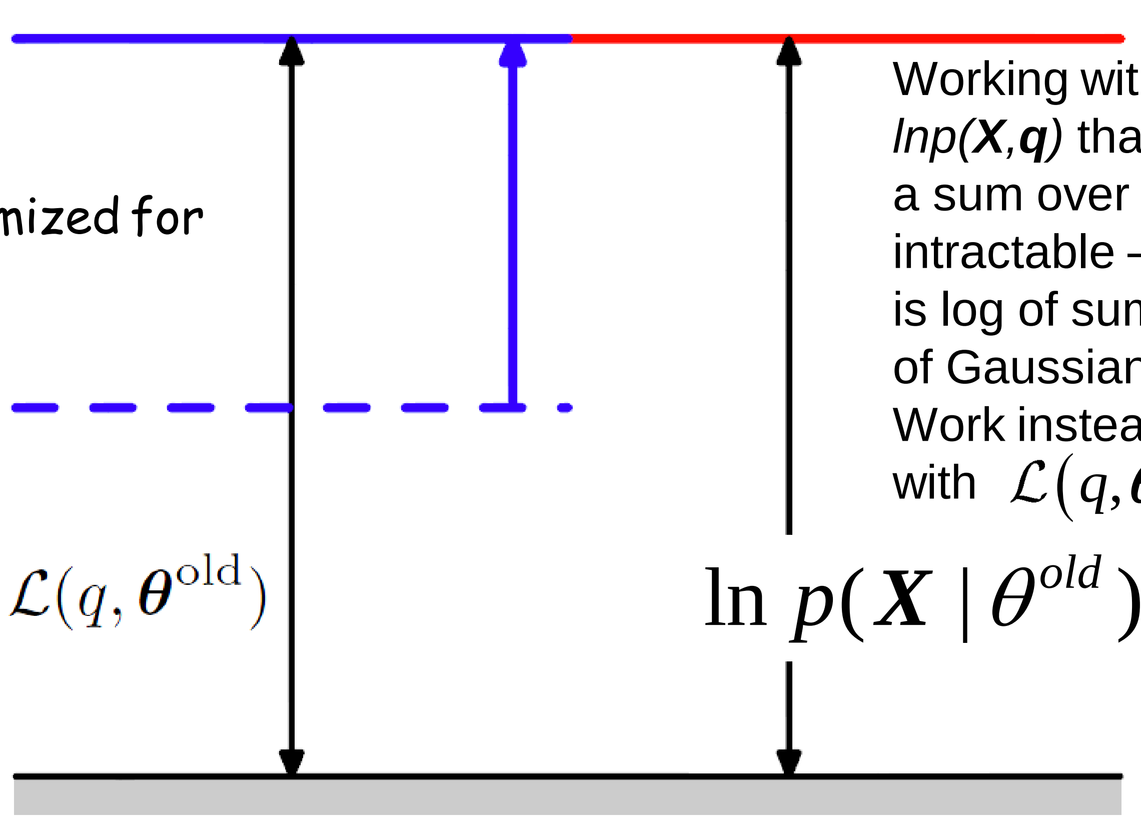
- Thus in the M step, we maximize the expectation of the complete-data log likelihood, as we saw earlier in the case of mixtures of Gaussians.

$$\begin{aligned}\mathcal{L}(q,\boldsymbol{\theta}) &= \int p\left(\boldsymbol{Z}\middle|\boldsymbol{X},\boldsymbol{\theta}^{old}\right)\ln p\left(\boldsymbol{X},\boldsymbol{Z}\middle|\boldsymbol{\theta}\right)d\boldsymbol{Z} - \int p\left(\boldsymbol{Z}\middle|\boldsymbol{X},\boldsymbol{\theta}^{old}\right)\ln p\left(\boldsymbol{X},\boldsymbol{Z}\middle|\boldsymbol{\theta}^{old}\right)\\ &= Q(\boldsymbol{\theta},\boldsymbol{\theta}^{old}) + const\end{aligned}$$

- Note that *the variable* ***θ*** *over which we are optimizing appears only inside the logarithm.*

- *If p(**Z**,**X**|**θ**) is from the exponential family, then the log cancels the exponential* leading to an M step that will be simpler than the maximization of p(**X**|**θ**).

# E-Step


$\mathrm{KL}(q||p) = 0$
$\mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{old}})$
$\ln p(\boldsymbol{X} \mid \theta^{old})$

E Step : $\mathcal{L}(q,\boldsymbol{\theta})$ is maximized for $q(\boldsymbol{Z}) = p\left(\boldsymbol{Z} \middle| \boldsymbol{X}, \boldsymbol{\theta}^{old}\right)$

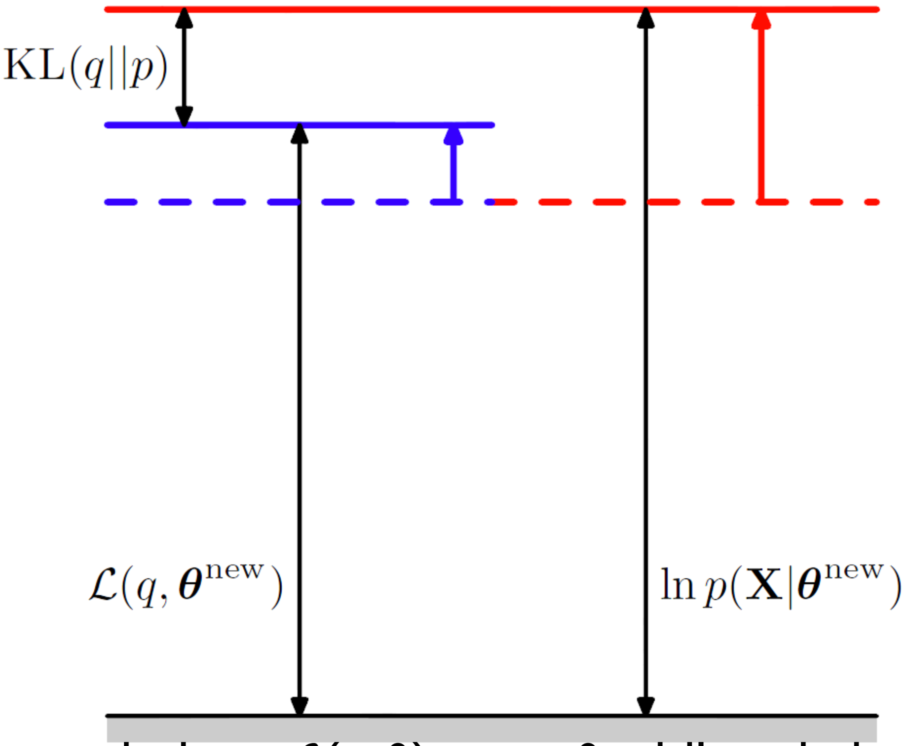
Working with *lnp(**X**,**q**)* that involves a sum over ***Z*** is intractable – it is log of sum of Gaussians. Work instead with $\mathcal{L}(q,\boldsymbol{\theta})$

- ❑ E-step: Maximizes $\mathcal{L}(q,\theta)$ w.r.t. q for fixed θ

$$\mathcal{L}(q,\boldsymbol{\theta}) = \ln p(\boldsymbol{X} \mid \boldsymbol{\theta}) - KL\left(q(\boldsymbol{Z}) \,||\, p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta})\right)$$

- ❑ At every step, the EM algorithm increases this lower bound on the log probability on the data (log-likelihood function)

# M-Step


$\mathrm{KL}(q||p)$
$\mathcal{L}(q,\boldsymbol{\theta}^{\mathrm{new}})$
$\ln p(\mathbf{X}|\boldsymbol{\theta}^{\mathrm{new}})$

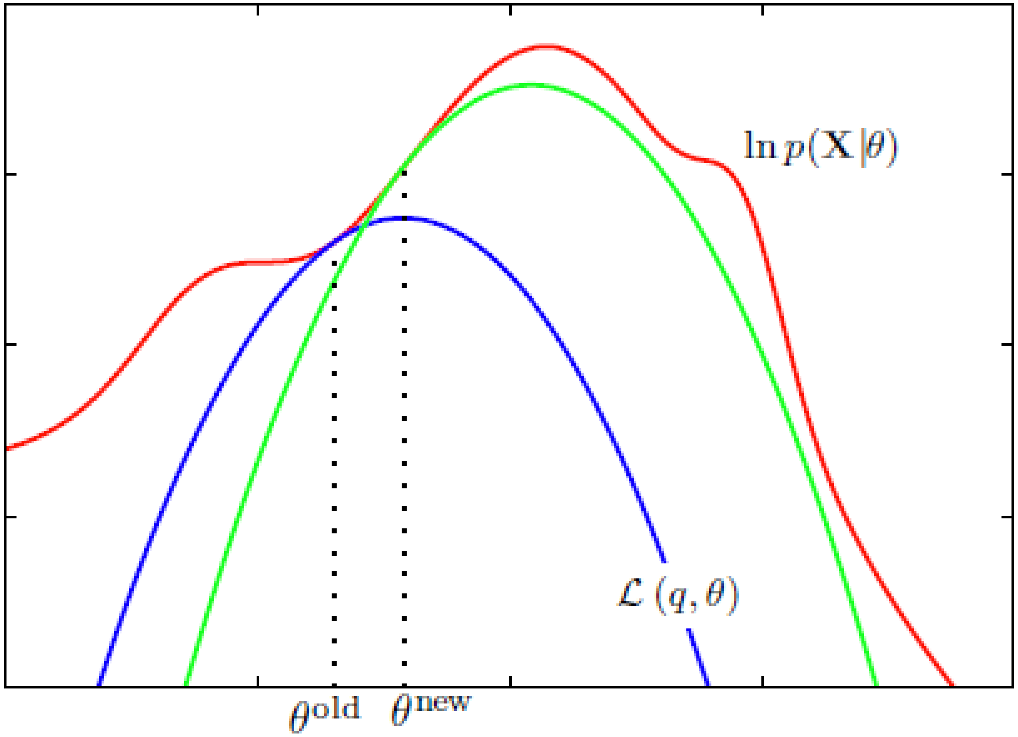
- The M-step, maximizes $\mathcal{L}$(q,θ) w.r.t. θ while q is kept fixed (function of $\theta^{old}$).

$$\mathcal{L}(q,\theta)=\int q(\boldsymbol{Z})\ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})d\boldsymbol{Z}-\int q(\boldsymbol{Z})\ln q(\boldsymbol{Z})d\boldsymbol{Z}$$

- The $\ln p(\boldsymbol{X}|\theta_{new})$ goes up at least as much as $\mathcal{L}$ ($q,\theta^{new}$) creating *KL(q||p).*
- *$\mathcal{L}$ maximized for* $\boldsymbol{\theta}=\arg\max_{\theta}\int q(\boldsymbol{Z})\ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})d\boldsymbol{Z}$

# EM in the Space of Parameters

- Can view the EM algorithm in the space of parameters.
- The red curve is the incomplete data log likelihood function whose value we wish to maximize.
- We start with some initial parameter value $\theta^{(old)}$, and in the first E step we evaluate the posterior distribution over latent variables, which gives rise to a lower bound $\mathcal{L}(q, \theta^{(old)})$ whose value equals the log likelihood at $\theta^{(old)}$ as shown by the blue curve.



emLogLikelihoodMax
from PMTK

# EM in the Space of Parameters

- Note that the bound $\mathcal{L}(q,\theta)$ with $q(\mathbf{Z})=p(\mathbf{Z}|\mathbf{X},\theta^{(old)})$ is tangent to the log likelihood $\ln p(\mathbf{X}|\theta)$ at $\theta^{(old)}$ i.e. that both curves have the same gradient.

- This is obvious after noting that KL(q||p) is at its minimum (i.e. 0) when $q(\mathbf{Z})=p(\mathbf{Z}|\mathbf{X},\theta^{(old)})$.

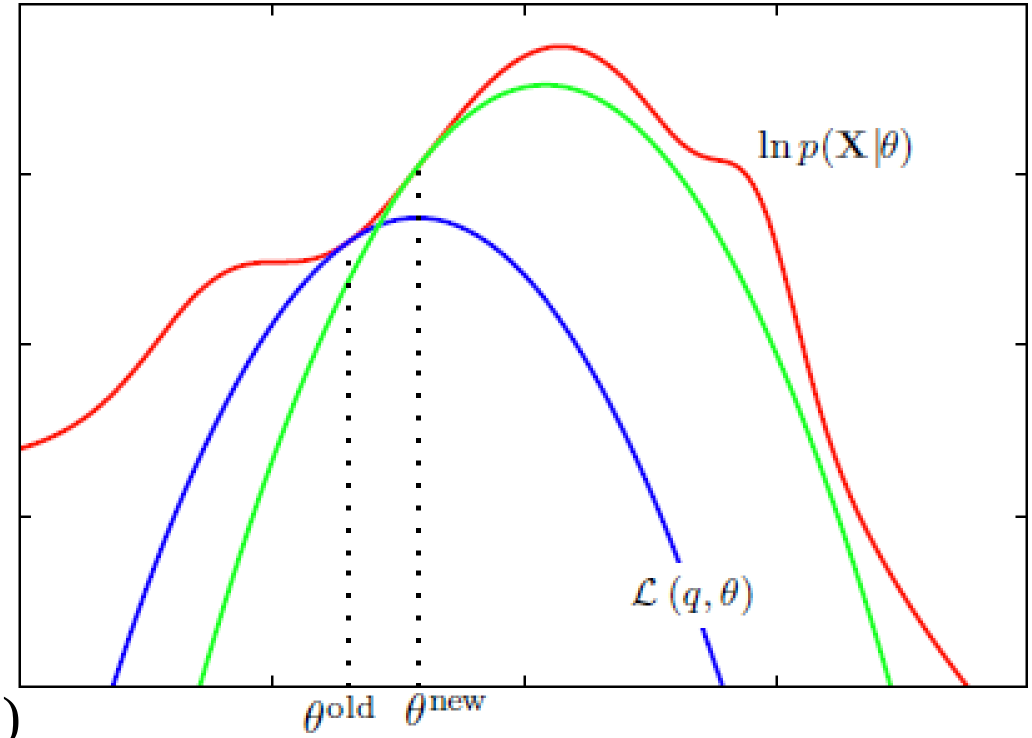
- This means that:

$$\frac{\partial}{\partial\theta}KL(q\,||\,p)=0$$

since $p(\mathbf{Z}|\mathbf{X},\theta)$ depends on $\theta$.

- From $\mathcal{L}(q,\boldsymbol{\theta})=\ln p(\boldsymbol{X}\,|\,\boldsymbol{\theta})-KL(q\,||\,p)$ we conclude that:

$$\frac{\partial}{\partial\theta}\mathcal{L}\left(q,\boldsymbol{\theta}^{(old)}\right)=\frac{\partial}{\partial\theta}\ln p(\boldsymbol{X}\,|\,\boldsymbol{\theta}^{(old)})$$

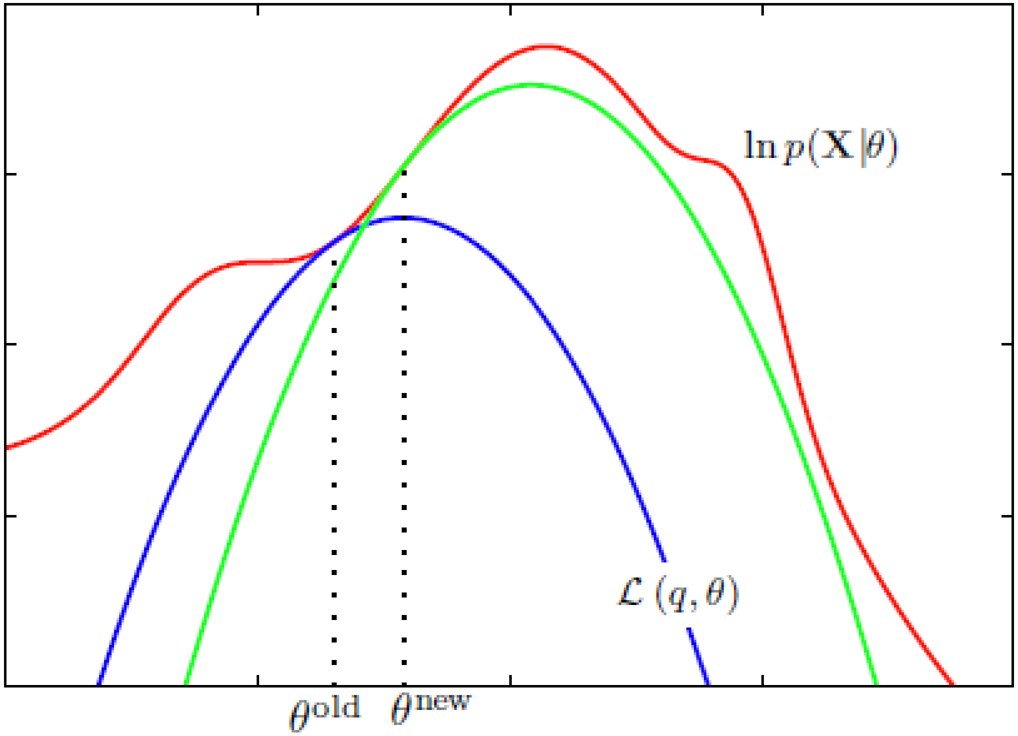


# EM in the Space of Parameters

- *The lower bound $\mathcal{L}(q,\theta)$ is a convex function having a unique maximum* (for mixture components from the exponential family).

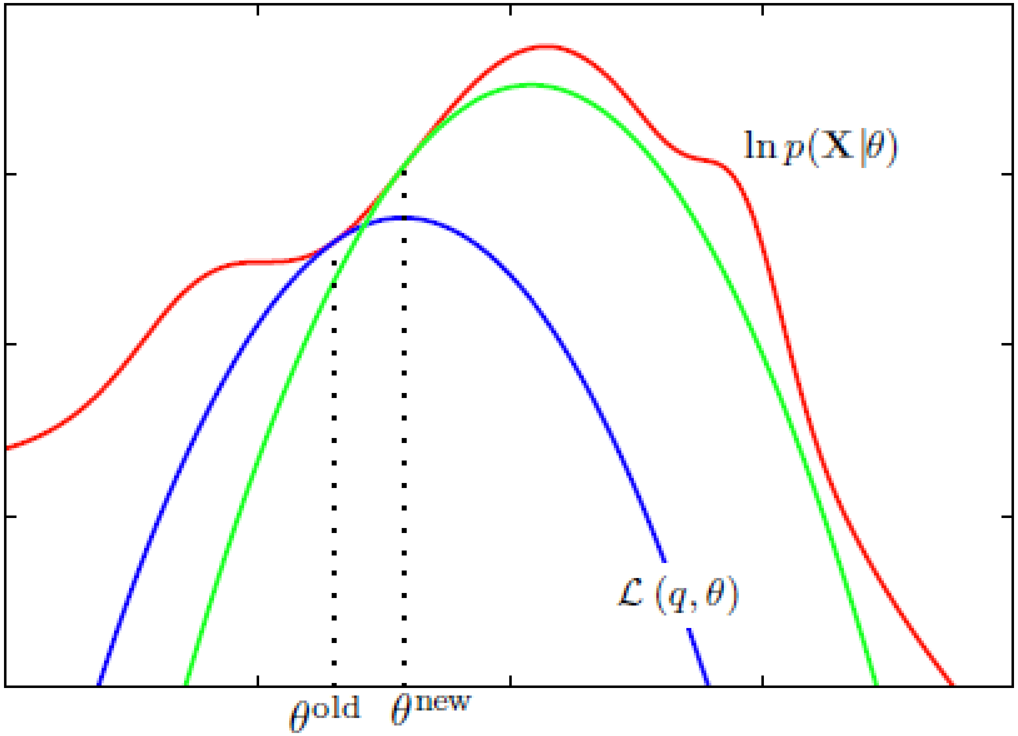
- In the M step, the bound $\mathcal{L}(q,\theta)$ is maximized giving the value $\theta^{(new)}$ which gives a larger value of log likelihood than $\theta^{(old)}$

- *The subsequent E step then constructs a bound $\mathcal{L}(q,\theta^{(new)})$ that is tangential at $\theta^{(new)}$ as shown by the green curve.*



# EM in General: Parameter Space Representation

- E-Step resets bound $\mathcal{L}(q,\theta)$ on $\ln p(\mathbf{X}|\theta)$ at $\theta=\theta^{old}$, it is
  - Tight at $\theta=\theta^{old}$,
  - Tangetial at $\theta=\theta^{old}$,
  - Convex (easy) in $\theta$ for exponential family mixture components



# EM In General

- Consider an i.i.d. data set, **X** that comprises of N data points {**x**$_n$}. **Z** comprises N corresponding latent variables {**z**$_n$}, n = 1, . . . , N.

- From the independence assumption, we have

$$p(\boldsymbol{X},\boldsymbol{Z}) = \prod_n p\left(\boldsymbol{x}_n,\boldsymbol{z}_n\right)$$

and, by marginalizing over {**z**$_n$ } we have

$$p(\boldsymbol{X}) = \prod_n p\left(\boldsymbol{x}_n\right)$$

- Using the sum and product rules, we see that the posterior probability that is evaluated in the E step takes the form

$$p(\boldsymbol{Z}\,|\,\boldsymbol{X},\theta) = \frac{p(\boldsymbol{X},\boldsymbol{Z}\,|\,\theta)}{\sum_{\boldsymbol{Z}} p(\boldsymbol{X},\boldsymbol{Z}\,|\,\theta)} = \frac{\prod_{n=1}^{N} p(\boldsymbol{x}_n,\boldsymbol{z}_n\,|\,\theta)}{\sum_{\boldsymbol{Z}}\prod_{n=1}^{N} p(\boldsymbol{x}_n,\boldsymbol{z}_n\,|\,\theta)} = \prod_{n=1}^{N} p(\boldsymbol{z}_n\,|\,\boldsymbol{x}_n,\theta)$$

# EM In General

$$p(\boldsymbol{Z} \mid \boldsymbol{X}, \theta) = \prod_{n=1}^{N} p(\boldsymbol{z}_n \mid \boldsymbol{x}_n, \theta)$$

- *Thus the posterior distribution of the latent variables also factorizes with respect to n.*

- For the Gaussian mixture model*: the responsibility that each of the mixture components takes for a particular* $\boldsymbol{x}_n$ *depends only on the value of* $\boldsymbol{x}_n$ *and on* $\boldsymbol{\theta}$*, not on the values of the other data points.*

- We have seen that both the E and the M steps of the EM algorithm are increasing the value of a well-defined bound on the log likelihood function and that the complete EM cycle will change the model parameters in such a way as to cause the log likelihood to increase (unless it is already at a maximum, in which case the parameters remain unchanged).

# Using EM to Maximize p(θ|X)

- We *can also use the EM algorithm to maximize the posterior distribution p(θ|X)* for models in which we have introduced a prior p(**θ**) over the parameters.

- Note that as a function of **θ**, we have p(**θ**|**X**) = p(**θ**,**X**)/p(**X**) and so

$$\ln p(\boldsymbol{\theta}|\mathbf{X}) = \ln p(\boldsymbol{\theta},\mathbf{X}) - \ln p(\mathbf{X})$$

- Making use of the decomposition $\ln p(\boldsymbol{X}\,|\,\theta) = \mathcal{L}(q,\theta) + KL(q\|p)$

$$\ln p(\boldsymbol{\theta}|\mathbf{X}) = \mathcal{L}(q, \boldsymbol{\theta}) + KL(q\|p) + \ln p(\boldsymbol{\theta}) - \ln p(\mathbf{X})$$
$$\geq \mathcal{L}(q, \boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) - \ln p(\mathbf{X})$$

  where ln p(**X**) is a constant.

- We can again optimize the right-hand side alternately with respect to q and **θ**.

# Using EM to Maximize p(θ|X)

$$\ln p(\theta|\mathbf{X}) = \mathcal{L}(q, \theta) + KL(q||p) + \ln p(\theta) - \ln p(\mathbf{X})$$
$$\geq \mathcal{L}(q, \theta) + \ln p(\theta) - \ln p(\mathbf{X})$$

$$\mathcal{L}(q,\theta) = \int p\left(\boldsymbol{Z}\middle|\boldsymbol{X},\boldsymbol{\theta}^{old}\right)\ln p\left(\boldsymbol{X},\boldsymbol{Z}\middle|\boldsymbol{\theta}\right)d\boldsymbol{Z} - \int p\left(\boldsymbol{Z}\middle|\boldsymbol{X},\boldsymbol{\theta}^{old}\right)\ln p\left(\boldsymbol{X},\boldsymbol{Z}\middle|\boldsymbol{\theta}^{old}\right)$$
$$= Q(\boldsymbol{\theta},\boldsymbol{\theta}^{old}) + const$$

- ❑ Since q appears only in $\mathcal{L}(q, \theta)$, optimization with respect to q gives rise to the *same E step equations as for the standard EM*.
- ❑ *The M-step equations are modified through the introduction of the prior ln p(θ), which requires only a small modification to the standard MLE M-step equations.*

# Generalizations of the EM Algorithm

# Incremental EM Algorithm

- Consider e.g. for the case of a Gaussian mixture performing an update for data point m in which the old and new values of the responsibilities are denoted $\gamma^{old}(z_{mk})$ and $\gamma^{new}(z_{mk})$.

- In the M step, the required sufficient statistics can be updated incrementally. For instance, for the means the sufficient statistics are defined by

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

from which we obtain (see proof next)

$$\boldsymbol{\mu}_k^{new} = \boldsymbol{\mu}_k^{old} + \left(\frac{\gamma^{new}(z_{mk}) - \gamma^{old}(z_{mk})}{N_k^{new}}\right)\left(\mathbf{x}_m - \boldsymbol{\mu}_k^{old}\right)$$

$$N_k^{new} = N_k^{old} + \gamma^{new}(z_{mk}) - \gamma^{old}(z_{mk})$$

- The results for the mixing coefficients and covariances are:

$$\pi_k^{new} = \pi_k^{old} - \frac{\gamma^{old}(z_{mk})}{N} + \frac{\gamma^{new}(z_{mk})}{N}$$

$$\boldsymbol{\Sigma}_k^{new} = \Sigma_k^{old} - \frac{\gamma^{old}(z_{mk})}{N_k^{new}}\left(\left(\mathbf{x}_m - \boldsymbol{\mu}_k^{old}\right)\left(\mathbf{x}_m - \boldsymbol{\mu}_k^{old}\right)^T - \Sigma_k^{old}\right) + \frac{\gamma^{new}(z_{mk})}{N_k^{new}}\left(\left(\mathbf{x}_m - \boldsymbol{\mu}_k^{new}\right)\left(\mathbf{x}_m - \boldsymbol{\mu}_k^{new}\right)^T - \Sigma_k^{old}\right)$$

# Incremental EM Algorithm

- ❑ Start with $N_k^{old}=\sum_n \gamma^{old}(z_{nk})$ and obtain $N_k^{new}$ by updating $\gamma^{new}(z_{mk})$ of the data point $\mathbf{x}_m$:

$$N_k^{new}=\sum_{n\neq m}\gamma^{old}(z_{nk})+\gamma^{new}(z_{mk})=N_k^{old}-\gamma^{old}(z_{mk})+\gamma^{new}(z_{mk})$$

- ❑ Similarly start with $\boldsymbol{\mu}_k^{old}=\frac{1}{N_k}\sum_{n=1}^{N}\gamma^{old}(z_{nk})\boldsymbol{x}_n$ and obtain $\boldsymbol{\mu}_k^{new}$ by updating the responsibilities $\gamma^{new}(z_{mk})$ of the data point $\mathbf{x}_m$:

$$\boldsymbol{\mu}_k^{new}=\frac{1}{N_k^{new}}\left(\sum_{n\neq m}^{N}\gamma^{old}(z_{nk})\boldsymbol{x}_n+\gamma^{new}(z_{mk})\boldsymbol{x}_m\right)=\frac{1}{N_k^{new}}\left(\sum_{n}^{N}\gamma^{old}(z_{nk})\boldsymbol{x}_n-\gamma^{old}(z_{mk})\boldsymbol{x}_m+\gamma^{new}(z_{mk})\boldsymbol{x}_m\right)$$

$$=\frac{1}{N_k^{new}}\left(N_k^{old}\boldsymbol{\mu}_k^{old}-\gamma^{old}(z_{mk})\boldsymbol{x}_m+\gamma^{new}(z_{mk})\boldsymbol{x}_m\right)$$

$$=\frac{1}{N_k^{new}}\left(\left(N_k^{new}-\gamma^{new}(z_{mk})+\gamma^{old}(z_{mk})\right)\boldsymbol{\mu}_k^{old}-\gamma^{old}(z_{mk})\boldsymbol{x}_m+\gamma^{new}(z_{mk})\boldsymbol{x}_m\right)$$

$$=\boldsymbol{\mu}_k^{old}+\left(\frac{\gamma^{new}(z_{mk})-\gamma^{old}(z_{mk})}{N_k^{new}}\right)\left(\boldsymbol{x}_m-\boldsymbol{\mu}_k^{old}\right)$$

# Incremental EM Algorithm

- *Thus both the E step and the M step take fixed time that is independent of the total number of data points.*

- Because the parameters are revised after each data point, rather than waiting until after the whole data set is processed, *this incremental version can converge faster than the batch version.*

- Each E or M step in this incremental algorithm is increasing the value of $\mathcal{L}(q, \boldsymbol{\theta})$ and, as we have shown above, if the algorithm converges to a local (or global) maximum of $\mathcal{L}(q, \boldsymbol{\theta})$ , this will correspond to a local (or global) maximum of the log likelihood function $\ln p(\mathbf{X}|\boldsymbol{\theta})$.

# EM For Missing Data Problems

# Fitting Models with Missing Data

- We want to fit a joint density model by MLE but we have holes in our data matrix due to missing data (NaNs). Let $O_{ij}$ = 1 if component *j* of data case *i* is observed, and let $O_{ij}$ = 0 otherwise. Let $\mathbf{X}_v$ be the visible data, and $\mathbf{X}_h$ be the missing (hidden) data:

$$\boldsymbol{X}_v = \left\{ x_{ij} : O_{ij} = 1 \right\}, \boldsymbol{X}_h = \left\{ x_{ij} : O_{ij} = 0 \right\}$$

- Our goal is to compute

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} p(\boldsymbol{X}_v | \boldsymbol{\theta}, \boldsymbol{O})$$

- Under the *missing at random assumption*, we have

$$p\left(\boldsymbol{X}_v | \boldsymbol{\theta}, \boldsymbol{O}\right) = \prod_{i=1}^{N} p\left(\boldsymbol{x}_{iv} | \boldsymbol{\theta}\right)$$

Here $\mathbf{x}_{iv}$ is a vector created from row *i* and the columns *{j* : $O_{ij}$ = 1*}*.

- Hence the log-likelihood has the form

$$\log p\left(\boldsymbol{X}_v | \boldsymbol{\theta}, \boldsymbol{O}\right) = \sum_i \log p\left(\boldsymbol{x}_{iv} | \boldsymbol{\theta}\right), \ where: p\left(\boldsymbol{x}_{iv} | \boldsymbol{\theta}\right) = \sum_{\boldsymbol{x}_{ih}} p\left(\boldsymbol{x}_{iv}, \boldsymbol{x}_{ih} | \boldsymbol{\theta}\right)$$

- We finally obtain our familiar form of LVMs:

$$\log p\left(\boldsymbol{X}_v | \boldsymbol{\theta}, \boldsymbol{O}\right) = \sum_i \log \sum_{\boldsymbol{x}_{ih}} p\left(\boldsymbol{x}_{iv}, \boldsymbol{x}_{ih} | \boldsymbol{\theta}\right)$$

# Fitting Models with Missing Data

- We want to fit an MVN by MLE based on those rows of the data matrix that are fully observed. If there are no such rows, we can use some ad-hoc imputation procedures, and then compute an initial MLE.

- **E step:** Once we have $\boldsymbol{\theta}^{t-1}$, we can compute the expected complete data log likelihood at iteration *t* as follows:

$$Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{t-1}\right)=\mathbb{E}\left[\sum_{i=1}^{N}\log\mathcal{N}\left(\boldsymbol{x}_i\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma}\right)|\,\mathcal{D},\boldsymbol{\theta}^{t-1}\right]=$$

$$=-\frac{N}{2}\log|2\pi\boldsymbol{\Sigma}|-\frac{1}{2}\sum_i\mathbb{E}\left[\left(\boldsymbol{x}_i-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}_i-\boldsymbol{\mu}\right)\right]$$

$$=-\frac{N}{2}\log|\boldsymbol{\Sigma}|-\frac{ND}{2}\log\left(2\pi\right)-\frac{1}{2}tr\left(\boldsymbol{\Sigma}^{-1}\mathbb{E}\left[S(\boldsymbol{\mu})\right]\right)$$

where:

$$\mathbb{E}\left[S(\boldsymbol{\mu})\right]=\sum_i\mathbb{E}\left[\left(\boldsymbol{x}_i-\boldsymbol{\mu}\right)\left(\boldsymbol{x}_i-\boldsymbol{\mu}\right)^T\right]=\sum_i\left(\mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i^T\right]+\boldsymbol{\mu}\boldsymbol{\mu}^T-2\boldsymbol{\mu}\mathbb{E}\left[\boldsymbol{x}_i\right]^T\right)$$

- To simplify the notation, we drop the conditioning of the expectation on $\mathcal{D}$ and $\boldsymbol{\theta}^{t-1}$ . We need to compute to expected sufficient statistics.

- We use the results for the conditionals of a MVN from an earlier lecture.

$$\boldsymbol{x}_{ih}\,|\,\boldsymbol{x}_{iv},\theta\sim\mathcal{N}\left(\boldsymbol{m}_i,\boldsymbol{V}_i\right)$$

$$\boldsymbol{m}_i=\boldsymbol{\mu}_h+\boldsymbol{\Sigma}_{hv}\boldsymbol{\Sigma}_{vv}^{-1}\left(\boldsymbol{x}_{iv}-\boldsymbol{\mu}_v\right),\boldsymbol{V}_i=\boldsymbol{\Sigma}_{hh}-\boldsymbol{\Sigma}_{hv}\boldsymbol{\Sigma}_{vv}^{-1}\boldsymbol{\Sigma}_{vh}$$

# Fitting Models with Missing Data

- Hence the expected sufficient statistics are

$$\mathbb{E}\left[\boldsymbol{x}_i\right]=\left(\mathbb{E}\left[\boldsymbol{x}_{ih}\right];\boldsymbol{x}_{iv}\right)=\left(\boldsymbol{m}_i;\boldsymbol{x}_{iv}\right),\ \mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i^T\right]=\mathbb{E}\left[\begin{pmatrix}\boldsymbol{x}_{ih}\\ \boldsymbol{x}_{iv}\end{pmatrix}\begin{pmatrix}\boldsymbol{x}_{ih} & \boldsymbol{x}_{iv}\end{pmatrix}\right]=\begin{pmatrix}\mathbb{E}\left[\boldsymbol{x}_{ih}\boldsymbol{x}_{ih}^T\right] & \mathbb{E}\left[\boldsymbol{x}_{ih}\right]\boldsymbol{x}_{iv}^T\\ \boldsymbol{x}_{iv}\mathbb{E}\left[\boldsymbol{x}_{ih}\right]^T & \boldsymbol{x}_{iv}\boldsymbol{x}_{iv}^T\end{pmatrix}$$

$$\mathbb{E}\left[\boldsymbol{x}_{ih}\boldsymbol{x}_{ih}^T\right]=\mathbb{E}\left[\boldsymbol{x}_{ih}\right]\mathbb{E}\left[\boldsymbol{x}_{ih}\right]^T+\boldsymbol{V}_i=\boldsymbol{m}_i\boldsymbol{m}_i^T+\boldsymbol{V}_i$$

- To simplify the notation we assume that the unobserved variables come before the observed variables in the node ordering.

- M-Step: By solving $\nabla Q(\boldsymbol{\theta},\boldsymbol{\theta}^{(t-1)})=0$, we can show that *the M step is equivalent to plugging the ESS into the MLE equations:*

$$\boldsymbol{\mu}^t=\frac{1}{N}\sum_i\mathbb{E}\left[\boldsymbol{x}_i\right],\ \boldsymbol{\Sigma}^t=\frac{1}{N}\sum_i\mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i^T\right]-\boldsymbol{\mu}^t\left(\boldsymbol{\mu}^t\right)^T$$

- EM is *not* equivalent to simply replacing variables by their expectations and plugging into the standard MLE formula; that ignores the posterior variance and results in incorrect estimates. *Instead we must compute the expectation of the sufficient statistics and plug that into the usual equation for the MLE.*

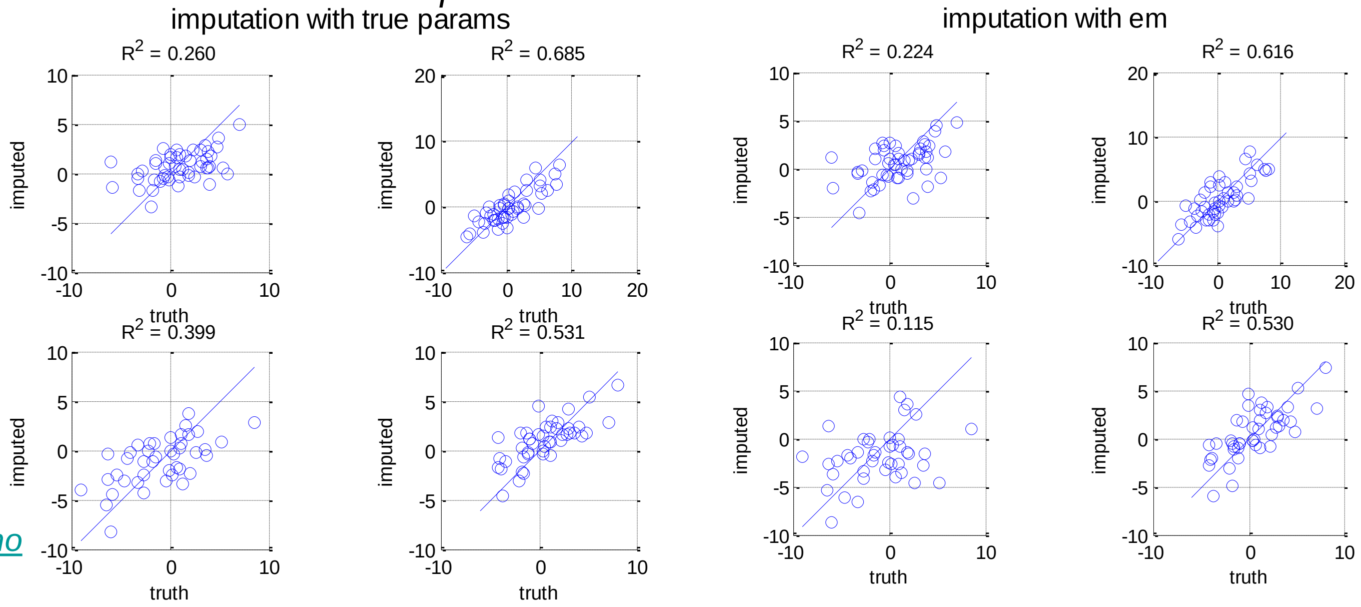
- We can now easily modify the algorithm to perform MAP estimation.

# Fitting Models with Missing Data

- Consider the imputation problem with *N* = 100 10-dim data cases, with 50% missing data. We fit the parameters using EM. Call the resulting parameters $\hat{\theta}$ . *We make predictions as* $\mathbb{E}\left[\boldsymbol{x}_{ih}|\boldsymbol{x}_{iv},\hat{\theta}\right]$

- The results obtained using the learned parameters are as good as with the true parameters. Performance improves with more data, or with less missing data.

- One can also fit a mixture of Gaussians in the presence of partially observed data vectors $\mathbf{x}_i$.



*gaussImputationDemo* from PMTK