STA 4273H:
Statistical Machine Learning

Russ Salakhutdinov

Department of Statistics
rsalakhu@utstat.toronto.edu
http://www.utstat.utoronto.ca/~rsalakhu/
Sidney Smith Hall, Room 6002

Lecture 6

Projects

- Assignment 3 will be due in 2 weeks.
- You should think about your project.
- Project proposals will be due on Nov 5 (1-page summary of what you are going to do for your research project).
- Brief 5-minute presentations of projects will take place on Nov 19. You will have to use slides.

Approximate Inference

• When using probabilistic graphical models, we will be interested in evaluating the posterior distribution $p(\mathbf{Z}|\mathbf{X})$ of the latent variables $\mathbf{Z}$ given the observed data $\mathbf{X}$.

• For example, in the EM algorithm, we need to evaluate the expectation of the complete-data log-likelihood with respect to the posterior distribution over the latent variables.

• For more complex models, it may be infeasible to evaluate the posterior distribution, or compute expectations with respect to this distribution.

• This typically occurs when working with high-dimensional latent spaces, or when the posterior distribution has a complex form, for which expectations are not analytically tractable (e.g. Boltzmann machines).

• We will examine a range of deterministic approximation schemes, some of which scale well to large applications.

Computational Challenges

Remember: the big challenge is computing the posterior distribution. There are several main approaches:

- Analytical integration: If we use "conjugate" priors, the posterior distribution can be computed analytically (we saw this in case of Bayesian linear regression).

- Gaussian (Laplace) approximation: Approximate the posterior distribution with a Gaussian. Works well when there is a lot of data compared to the model complexity (as posterior is close to Gaussian).

- Monte Carlo integration: The dominant current approach is Markov Chain Monte Carlo (MCMC) -- simulate a Markov chain that converges to the posterior distribution. It can be applied to a wide variety of problems.

- Variational approximation: A cleverer way to approximate the posterior. It often works much faster, but not as general as MCMC.

Probabilistic Model

- Suppose that we have a fully Bayesian model in which all parameters are given prior distributions.
- The model may have latent variables and parameters, and we will denote the set of all latent variables and parameters by $\mathbf{Z}$.
- We will also denote the set of all observed variables by $\mathbf{X}$.
- For example, we may be given a set of N i.i.d data points, so that $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$ (as we saw in previous class).
- Our probabilistic model specifies the joint distribution $P(\mathbf{X}, \mathbf{Z})$.
- Our goal is to find approximate posterior distribution $P(\mathbf{Z}|\mathbf{X})$ and the model evidence $p(\mathbf{X})$.

Variational Bound

• As in our previous lecture, we can decompose the marginal log-probability as:

$$\log p(\mathbf{X}) = \mathcal{L}(q) + \mathrm{KL}(q||p),$$

where

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \mathrm{d}\mathbf{Z}$$

$$\mathrm{KL}(q||p) = -\int q(\mathbf{Z}) \ln \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} \mathrm{d}\mathbf{Z}.$$

• Note that parameters are now stochastic variables and are absorbed into **Z**.

• We can maximize the variational lower bound $\mathcal{L}(q)$ with respect to the distribution q(**Z**), which is equivalent to minimizing the KL divergence.

• If we allow any possible choice of q(**Z**), then the maximum of the lower bound occurs when:

$$q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}).$$

In this case KL divergence becomes zero.

Variational Bound

- As in our previous lecture, we can decompose the marginal log-probability as:

$$\log p(\mathbf{X}) = \mathcal{L}(q) + \mathrm{KL}(q||p),$$

- We will assume that the true posterior distribution is intractable.

- We can consider a restricted family of distributions q(**Z**) and then find the member of this family for which KL is minimized.

- Our goal is to restrict the family of distributions so that it contains only tractable distributions.

- At the same time, we want to allow the family to be sufficiently rich and flexible, so that it can provide a good approximation to the posterior.

- One option is to use parametric distributions q(**Z**|ω), governed by parameters ω.

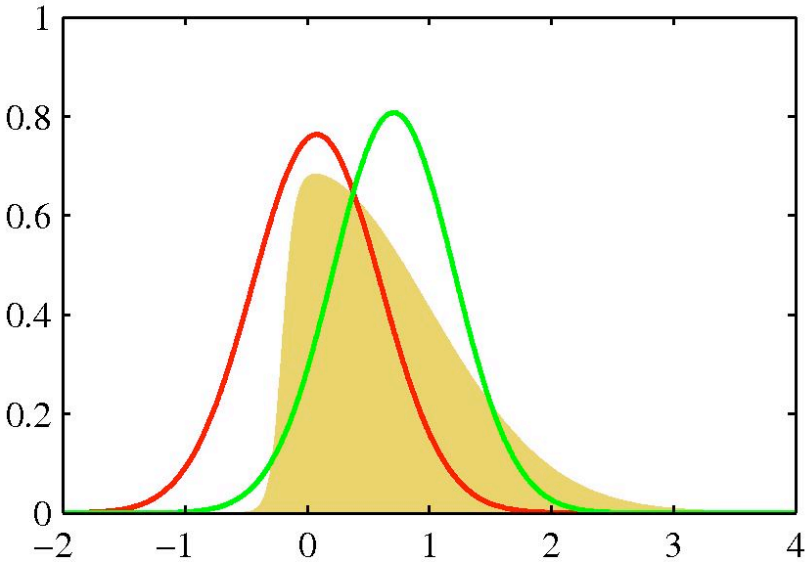
- The lower bound then becomes a function of ω, and can optimize the lower-bound to determine the optimal values for the parameters.

Example

- One option is to use parametric distributions q(**Z**|ω), governed by parameters ω.

$$\log p(\mathbf{X}) = \mathcal{L}(q) + \mathrm{KL}(q||p),$$

• An example, in which the variational distribution is Gaussian, and we optimize with respect to its mean and variance.



The original distribution (yellow), along with Laplace (red), and variational (green) approximations.

Mean-Field

- We now consider restricting the family of distributions.
- Partition the elements of **Z** into M disjoint groups, denoted by $\mathbf{Z}_i$, i=1,…,M.
- We assume that the q distribution factorizes with respect to these groups:

$$q(\mathbf{Z}) = \prod_{i=1}^{M} q_i(\mathbf{Z}_i).$$

- Note that we place no restrictions on the functional form of the individual factors q_i (we will often denote $q_i(\mathbf{Z}_i)$ as simply q_i).

- This approximation framework, developed in physics, is called **mean-field theory**.

Factorized Distributions

• Among all factorized distributions, we look for a distribution for which the variational lower bound is maximized.

• Denoting $q_i(\mathbf{Z}_i)$ as simply q_i, we have:

$$q(\mathbf{Z}) = \prod_{i=1}^{M} q_i(\mathbf{Z}_i).$$

$$\begin{aligned}
\mathcal{L}(q) &= \int q(\mathbf{Z}) \ln \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \mathrm{d}\mathbf{Z} \\
&= \int \prod_i q_i \left[\ln p(\mathbf{X}, \mathbf{Z}) - \sum_i \ln q_i \right] \mathrm{d}\mathbf{Z} \\
&= \int q_j \left[\int \ln p(\mathbf{X}, \mathbf{Z}) \prod_{i \neq j} q_i \, \mathrm{d}\mathbf{Z}_i \right] \mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j \, \mathrm{d}\mathbf{Z}_j + \text{const} \\
&= \int q_j \ln \tilde{p}(\mathbf{X}, \mathbf{Z}_j) \, \mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j \, \mathrm{d}\mathbf{Z}_j + \text{const}
\end{aligned}$$

where we denote a new distribution:

$$\tilde{p}(\mathbf{X}, \mathbf{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] + \text{const}.$$

Factorized Distributions

• Among all factorized distributions, we look for a distribution for which the variational lower bound is maximized.

• Denoting $q_i(\mathbf{Z}_i)$ as simply q_i, we have:

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \mathrm{d}\mathbf{Z}$$

$$= \int q_j \ln \tilde{p}(\mathbf{X}, \mathbf{Z}_j) \, \mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j \, \mathrm{d}\mathbf{Z}_j + \text{const}$$

where

$$\ln \tilde{p}(\mathbf{X}, \mathbf{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] + \text{const.}$$

• Here we take an expectation with respect to the q distribution over all variables $\mathbf{Z}_i$ for $i \neq j$, so that:

$$\mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] = \int \ln p(\mathbf{X}, \mathbf{Z}) \prod_{i \neq j} q_i \, \mathrm{d}\mathbf{Z}_i.$$

Maximizing Lower Bound

• Now suppose that we keep $\{q_{i\neq j}\}$ fixed, and optimize the lower bound with respect to all possible forms of the distribution $q_j(\mathbf{Z}_j)$.

• This optimization is easily done by recognizing that:

$$\mathcal{L}(q) = \int q_j \ln \tilde{p}(\mathbf{X}, \mathbf{Z}_j)\, d\mathbf{Z}_j - \int q_j \ln q_j\, d\mathbf{Z}_j + \text{const}$$

$$= -\text{KL}(q_j(\mathbf{Z}_j)||\tilde{p}(\mathbf{X}, \mathbf{Z}_j) + \text{const},$$

constant: does not depend on q.

$$\mathcal{L}(q) = \log p(\mathbf{X}) - \text{KL}(q||p)$$

so the minimum occurs when

$$q_j^*(\mathbf{Z}_j) = \tilde{p}(\mathbf{X}, \mathbf{Z}), \text{ or } \quad \ln q_j^*(\mathbf{Z}_j) = \mathbb{E}_{i\neq j}[\ln p(\mathbf{X}, \mathbf{Z})] + \text{const.}$$

• Observe: the log of the optimum solution for factor q_j is given by:

- Considering the log of the joint distribution over all hidden and visible variables
- Taking the expectation with respect to all other factors $\{q_i\}$ for $i \neq j$.

Maximizing Lower Bound

• Exponentiating and normalizing, we obtain:

$$q_j^*(\mathbf{Z}_j) = \frac{\exp\left(\mathbb{E}_{i\neq j}[\ln p(\mathbf{X},\mathbf{Z})]\right)}{\int \exp\left(\mathbb{E}_{i\neq j}[\ln p(\mathbf{X},\mathbf{Z})]\right)\mathrm{d}\mathbf{Z}_j}.$$

• The set of these equations for j=1,…,M represent the set of consistency conditions for the maximum of the lower bound subject to factorization constraint.

• To obtain a solution, we initialize all of the factors and then cycle through factors, replacing each in tern with a revised estimate.

• Convergence is guaranteed because the bound is convex with respect to each of the individual factors.

Factroized Gaussian

- Consider a problem of approximating a general distribution by a factorized distribution.
- To get some insight, let us look at the problem of approximating a Gaussian distribution using a factorized Gaussian distribution.
- Consider a Gaussian distribution over two correlated variables **z** = (z_1,z_2).

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}),$$

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad \boldsymbol{\Lambda} = \begin{pmatrix} \beta_{11} & \beta_{12} \\ \beta_{12} & \beta_{22}. \end{pmatrix}$$

- Let us approximate this distribution using a factorized Gaussian of the form:

$$q(\mathbf{z}) = q_1(z_1)q_2(z_2).$$

Factroized Gaussian

- Remember:

$$\ln q_j^*(\mathbf{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] + \text{const.}$$

- Consider an expression for the optimal factor q_1:

$$\begin{aligned}
\ln q_1^*(z_1) &= \mathbb{E}_{q_2(z_2)}[\ln p(\mathbf{z})] + \text{const} \\
&= \mathbb{E}_{q_2(z_2)}\left[-\frac{\beta_{11}}{2}(z_1 - \mu_1)^2 - \beta_{12}(z_1 - \mu_1)(z_2 - \mu_2) \right] + \text{const} \\
&= -\frac{\beta_{11}}{2} z_1^2 + \beta_{11} z_1 \mu_1 - \beta_{12} z_1 (\mathbb{E}[z_2] - \mu_2) + \text{const.}
\end{aligned}$$

- Note that we have a quadratic function of z_1, and so we can identify $q_1(z_1)$ as a Gaussian distribution:

$$q_1^*(z_1) = \mathcal{N}(z_1 | m_1, \beta_{11}^{-1}), \quad m_1 = \mu_1 - \frac{\beta_{12}}{\beta_{11}}(\mathbb{E}[z_2] - \mu_2).$$

Factroized Gaussian

- By symmetry, we also obtain:

$$q_1^*(z_1) = \mathcal{N}(z_1|m_1, \beta_{11}^{-1}), \quad m_1 = \mu_1 - \frac{\beta_{12}}{\beta_{11}}(\mathbb{E}[z_2] - \mu_2).$$

$$q_2^*(z_2) = \mathcal{N}(z_2|m_2, \beta_{22}^{-1}), \quad m_2 = \mu_2 - \frac{\beta_{12}}{\beta_{22}}(\mathbb{E}[z_1] - \mu_1).$$

- There are two observations to make:

  - We did not assume that $q_i^*(z_i)$ is Gaussian, but rather we derived this result by optimizing variational bound over all possible distributions.

  - The solutions are coupled. The optimal $q_1^*(z_1)$ depends on expectation computed with respect to $q_2^*(z_2)$.

- One option is to cycle through the variables in turn and update them until convergence.

Factroized Gaussian

• By symmetry, we also obtain:

$$q_1^*(z_1) = \mathcal{N}(z_1|m_1, \beta_{11}^{-1}), \quad m_1 = \mu_1 - \frac{\beta_{12}}{\beta_{11}}(\mathbb{E}[z_2] - \mu_2).$$

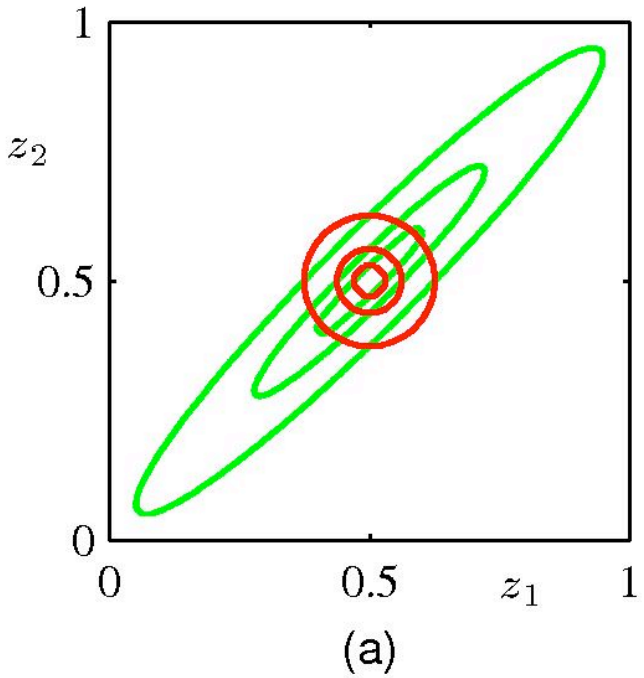
$$q_2^*(z_2) = \mathcal{N}(z_2|m_2, \beta_{22}^{-1}), \quad m_2 = \mu_2 - \frac{\beta_{12}}{\beta_{22}}(\mathbb{E}[z_1] - \mu_1).$$

• However, in our case, $\mathbb{E}[z_1] = \mu_1, \ \mathbb{E}[z_2] = \mu_2.$

• The green contours correspond to 1,2, and 3 standard deviations of the correlated Gaussian.

• The red contours correspond to the factorial approximation q(**z**) over the same two variables.

• Observe that a factorized variational approximation tends to give approximations that are too compact.



(a)

Alternative Form of KL Divergence

- We have looked at the variational approximation that minimizes KL(q||p).
- For comparison, suppose that we were minimizing KL(p||q).

$$\mathrm{KL}(p||q) = -\int p(\mathbf{Z}) \ln \frac{q(\mathbf{Z})}{p(\mathbf{Z})} \mathrm{d}\mathbf{Z}.$$

$$\mathrm{KL}(p||q) = -\int p(\mathbf{Z}) \left[\sum_{i=1}^{M} \ln q_i(\mathbf{Z}_i) \right] \mathrm{d}\mathbf{Z} + \int p(\mathbf{Z}) \ln \frac{1}{p(\mathbf{Z})} \mathrm{d}\mathbf{Z}.$$

constant: does not depend on q.

- It is easy to show that:

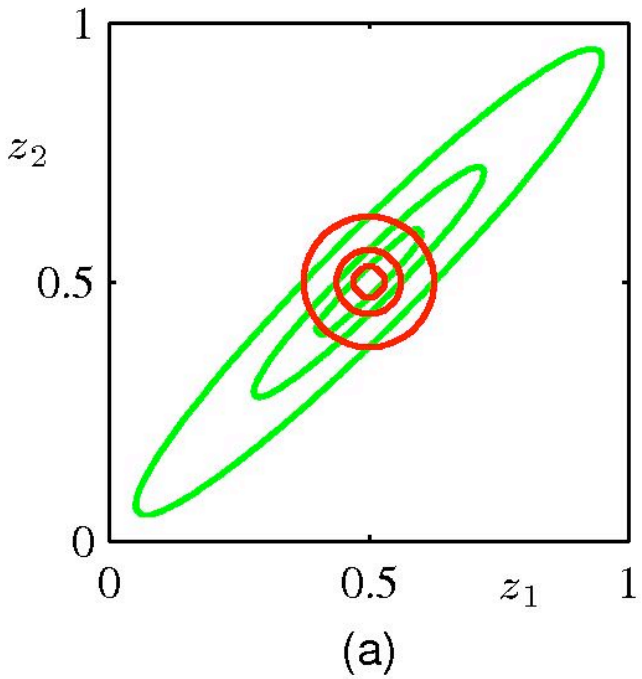
$$q_j^*(\mathbf{Z}_j) = \int p(\mathbf{Z}) \prod_{i \neq j} d\mathbf{Z}_i = p(\mathbf{Z}_j).$$

- The optimal factor is given by the marginal distribution of p(**Z**).

Comparison of two KLs

- Comparison of two the alternative forms for the KL divergence.

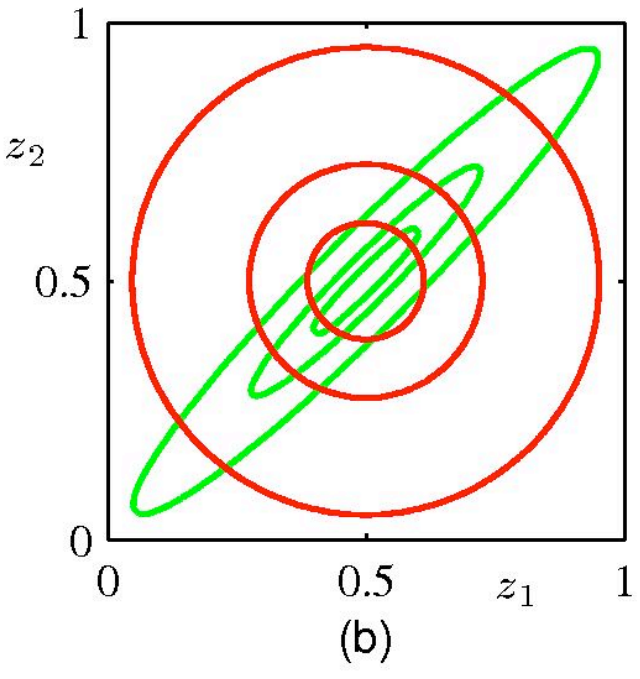
KL(q||p)



(a)

Approximation is too compact.

KL(p||q)



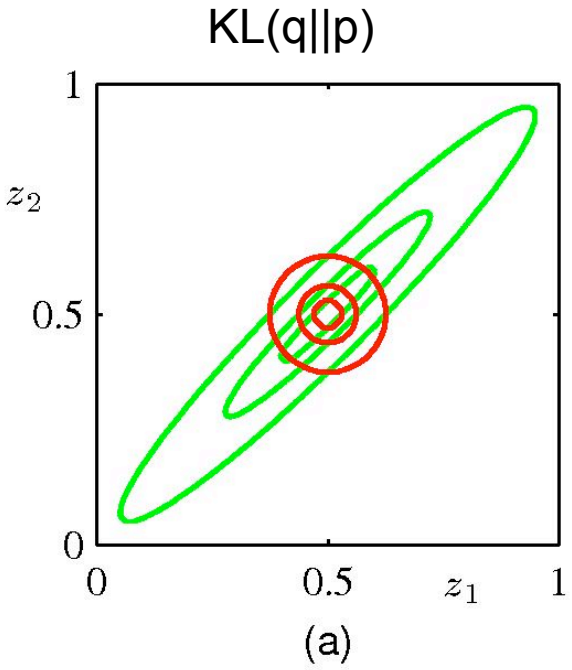
(b)

Approximation is too spread.

Comparison of two KLs

• The difference between these two approximations can be understood as follows:

$$\mathrm{KL}(q||p) = -\int q(\mathbf{Z}) \ln \frac{p(\mathbf{Z})}{q(\mathbf{Z})} \mathrm{d}\mathbf{Z}.$$



• There is a large positive contribution to the KL divergence from regions of **Z** space in which:

- p(**Z**) is near zero,
- unless q(**Z**) is also close to zero.

• Minimizing KL(q||p) leads to distributions q(**Z**) that avoid regions in which p(**Z**) is small.

Comparison of two KLs

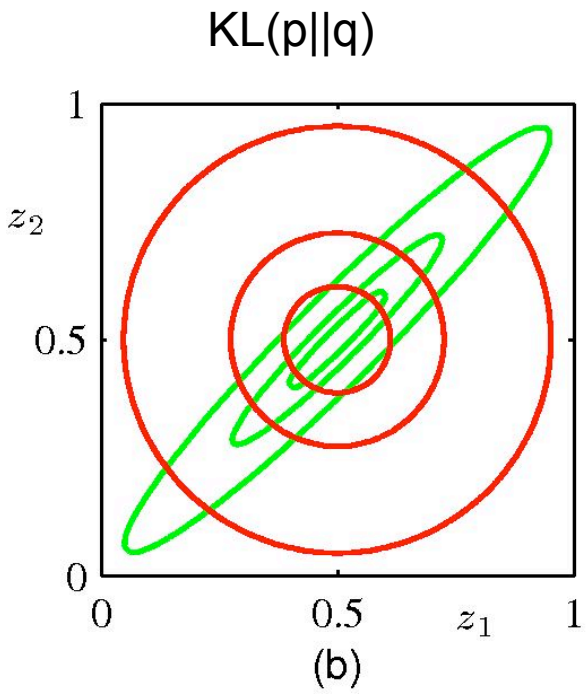
• Similar arguments apply for the alternative KL divergence:

$$\mathrm{KL}(p||q) = -\int p(\mathbf{Z}) \ln \frac{q(\mathbf{Z})}{p(\mathbf{Z})} \mathrm{d}\mathbf{Z}.$$

• There is a large positive contribution to the KL divergence from regions of **Z** space in which:

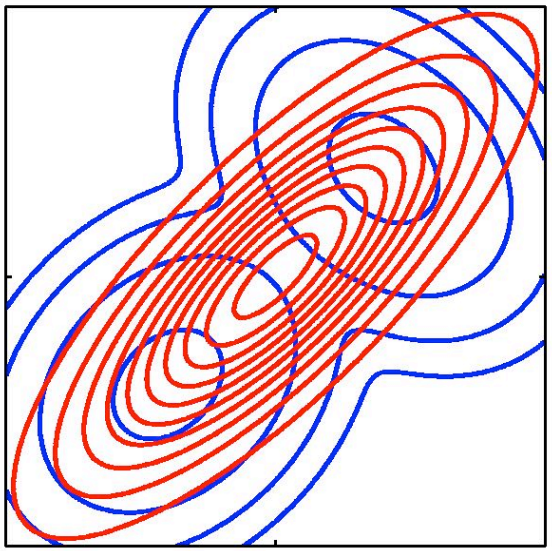
- q(**Z**) is near zero,
- unless p(**Z**) is also close to zero.

• Minimizing KL(p||q) leads to distributions q(**Z**) that are nonzero in regions where p(**Z**) is nonzero.

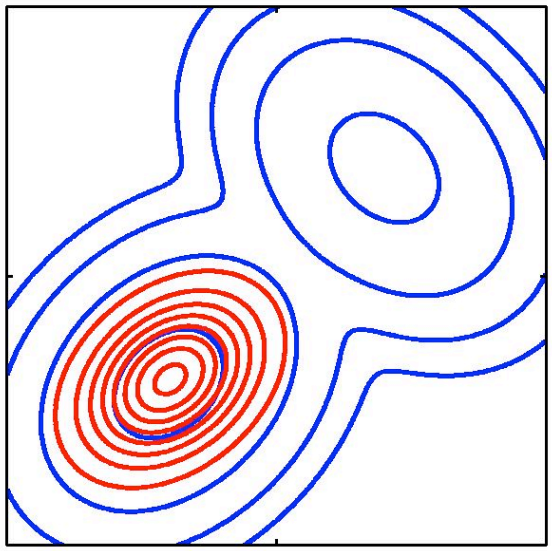


Approximating Multimodal Distribution

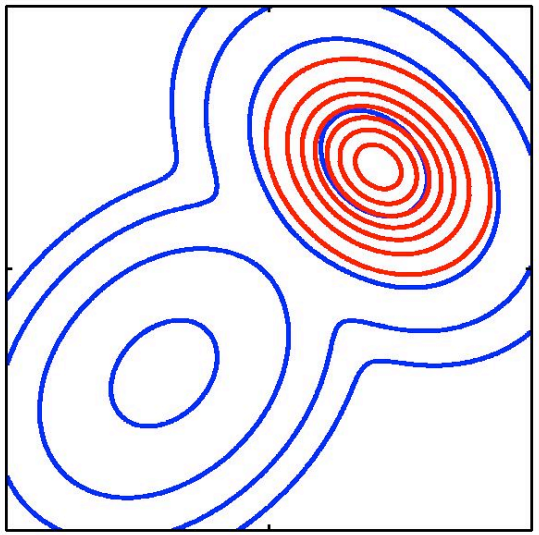
- Consider approximating multimodal distribution with a unimodal one.
- Blue contours show bimodal distribution p(**Z**), red contours show a single Gaussian distribution that best approximates q(**Z**) that best approximates p(**Z**).



KL(p||q)



KL(q||p)



KL(q||p)

- In practice, the true posterior will often be mutlimodal.
- KL(q||p) will tend to find a single mode, whereas KL(p||q) will average across all of the modes.

Alpha-family of Divergences

- The two forms of KL are members of the alpha-family divergences:

$$D_\alpha(p||q) = \frac{4}{1-\alpha^2}\left(1 - \int p(x)^{(1+\alpha)/2} q(x)^{(1-\alpha)/2} \mathrm{d}x\right), \quad -\infty < \alpha < \infty.$$

- Observe three points:
  - KL(p||q) corresponds to the limit $\alpha \to 1$.
  - KL(q||p) corresponds to the limit $\alpha \to$ -1.
  - D_α(p||q) $\geq$ 0, for all α, and D_α(p||q)=0 iff q(x) = p(x).

- Suppose p(x) is fixed and we minimize D_α(p||q) with respect to q distribution.
- For α < -1, the divergence is zero-forcing: q(x) will underestimate the support of p(x).
- For α > 1, the divergence is zero-avoiding: q(x) will stretch to cover all of p(x).
- For α = 0, we obtain a symmetric divergence which is related to Hellinger Distance:

$$D_H(p||q) = \frac{1}{2}\int \left(p(x)^{1/2} - q(x)^{1/2}\right)^2 \mathrm{d}x.$$

Univariate Gaussian

- Consider a factorized approximation using a Gaussian distribution over a single variable x.
- Given a dataset $\mathcal{D} = \{x_1, ..., x_N\}$, we would like to infer posterior distribution over the mean μ and precision τ.
- The likelihood term is given:

$$p(\mathcal{D}|\mu, \tau) = \left(\frac{\tau}{2\pi}\right)^{N/2} \exp\left(-\frac{\tau}{2} \sum_{n=1}^{N} (x_n - \mu)^2 \right).$$

- The conjugate prior is given by the Normal-Gamma prior:

$$p(\mu, \tau) = p(\mu|\tau)p(\tau), \qquad \begin{aligned} p(\tau) &= \text{Gam}(\tau|a_0, b_0), \\ p(\mu|\tau) &= \mathcal{N}(\mu|\mu_0, (\lambda_0\tau)^{-1}). \end{aligned}$$

- For this simple problem, the posterior also takes the form of Normal-Gamma distribution and hence has a closed form solution.
- However, let us consider a variational approximation to the posterior.

Approximating Mean

- We now consider a factorized variational approximation to the posterior:

$$q(\mu, \tau) = q(\mu)q(\tau).$$

- Note that the true posterior does not factorize this way!
- Remember: $\ln q_j^*(\mathbf{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] + \text{const.}$

$$p(\mathcal{D}|\mu, \tau) = \left(\frac{\tau}{2\pi}\right)^{N/2} \exp\left(-\frac{\tau}{2}\sum_{n=1}^{N}(x_n - \mu)^2\right).$$

- Hence:

$$p(\mu|\tau) = \mathcal{N}(\mu|\mu_0, (\lambda_0\tau)^{-1}).$$

$$\begin{aligned}\ln q^*(\mu) &= \mathbb{E}_{q(\tau)}[\ln p(\mathcal{D}|\mu, \tau) + \ln p(\mu|\tau)] + \text{const.} \\ &= -\frac{\mathbb{E}[\tau]}{2}\left(\sum_{n=1}^{N}(x_n - \mu)^2 + \lambda_0(\mu - \mu_0)^2\right) + \text{const.}\end{aligned}$$

- So:

$$q^*(\mu) = \mathcal{N}(\mu|\mu_N, \lambda_N^{-1}),$$

$$\mu_N = \frac{\lambda_0\mu_0 + N\bar{x}}{\lambda_0 + N},$$

$$\lambda_N = (\lambda_0 + N)\mathbb{E}[\tau].$$

Depends on expectation with respect to q(τ).

Approximating Mean

- We now consider a factorized variational approximation to the posterior:

$$q(\mu, \tau) = q(\mu)q(\tau).$$

$$\begin{aligned}\ln q^*(\mu) &= \mathbb{E}_{q(\tau)}[\ln p(\mathcal{D}|\mu, \tau) + \ln p(\mu|\tau)] + \text{const.} \\ &= -\frac{\mathbb{E}[\tau]}{2}\left(\sum_{n=1}^{N}(x_n - \mu)^2 + \lambda_0(\mu - \mu_0)^2\right) + \text{const.}\end{aligned}$$

$$q^*(\mu) = \mathcal{N}(\mu|\mu_N, \lambda_N^{-1}),$$

$$\mu_N = \frac{\lambda_0\mu_0 + N\bar{x}}{\lambda_0 + N},$$

$$\lambda_N = (\lambda_0 + N)\mathbb{E}[\tau].$$

- As N $\to \infty$, this gives Maximum Likelihood result $\mu_N = \bar{x},$ and the precision becomes infinite.

Approximating Precision

- We now consider a factorized variational approximation to the posterior:

$$q(\mu, \tau) = q(\mu)q(\tau).$$

- For optimal solution for the precision factor:

$$\ln q^*(\tau) = \mathbb{E}_{q(\mu)}[\ln p(\mathcal{D}|\mu, \tau) + \ln p(\mu|\tau) + \ln p(\tau)] + \text{const}$$

$$= \frac{1}{2}\ln\tau + \frac{N}{2}\ln\tau - \frac{\tau}{2}\mathbb{E}_{q(\mu)}\left[\sum_{n=1}^{N}(x_n - \mu)^2 + \lambda_0(\mu - \mu_0)^2\right]$$

$$+(a_0 - 1)\ln\tau - b_0\tau + \text{const.} \qquad p(\mu|\tau) = \mathcal{N}(\mu|\mu_0, (\lambda_0\tau)^{-1}).$$

$$\text{Gam}(\tau|a_0, b_0) = \frac{b_0^{a_0}}{\Gamma(a_0)}\tau^{a_0-1}\exp(-b_0\tau).$$

- Hence the optimal factor is a Gamma distribution: $q^*(\tau) = \text{Gam}(\tau|a_N, b_N)$,

$$a_N = a_0 + \frac{N+1}{2},$$

Depends on expectation with respect to q(μ).

$$b_N = b_0 + \frac{1}{2}\mathbb{E}_{q(\mu)}\left[\sum_{n=1}^{N}(x_n - \mu)^2 + \lambda_0(\mu - \mu_0)^2\right].$$

Approximating Precision

- We now consider a factorized variational approximation to the posterior:

$$q(\mu, \tau) = q(\mu)q(\tau).$$

$$\mathrm{Gam}(\tau|a_0, b_0) = \frac{b_0^{a_0}}{\Gamma(a_0)}\tau^{a_0-1}\exp(-b_0\tau).$$

- Hence the optimal factor is a Gamma distribution: $q^*(\tau) = \mathrm{Gam}(\tau|a_N, b_N)$,

$$a_N = a_0 + \frac{N+1}{2},$$

$$b_N = b_0 + \frac{1}{2}\mathbb{E}_{q(\mu)}\left[\sum_{n=1}^{N}(x_n - \mu)^2 + \lambda_0(\mu - \mu_0)^2\right].$$

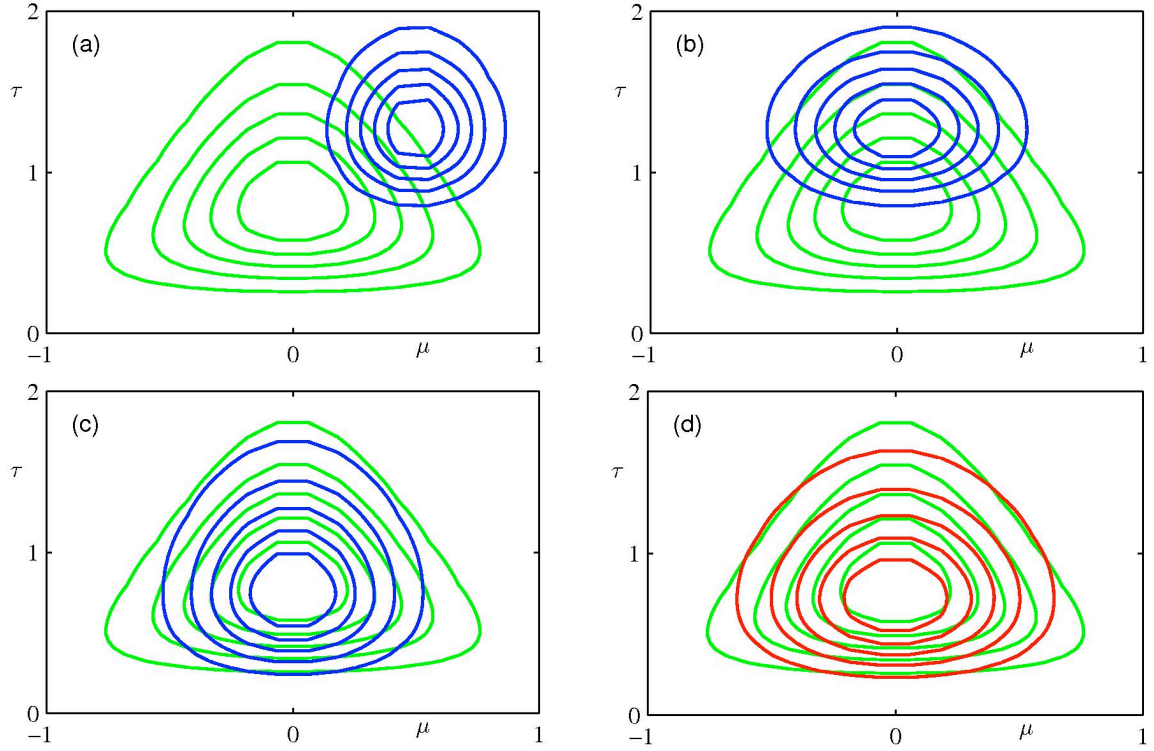
- As N $\to \infty$, the variational posterior q*(τ) has a mean given by the inverse of the maximum likelihood estimator for the variance of the data, and a variance that goes to zero.

- Note that we did not assume specific functional forms for the optimal q distributions. They were derived by optimizing variational bound over all possible distributions.

Iterative Procedure

• The optimal distributions for mean and precision terms depend on moments evaluated with respect to the other distributions.

• One option is to cycle through the mean and precision in turn and update them until convergence.



Variational inference for the mean and precision.

Green contours represent the true posterior, blue contours represent variational approximation.

Mixture of Gaussians

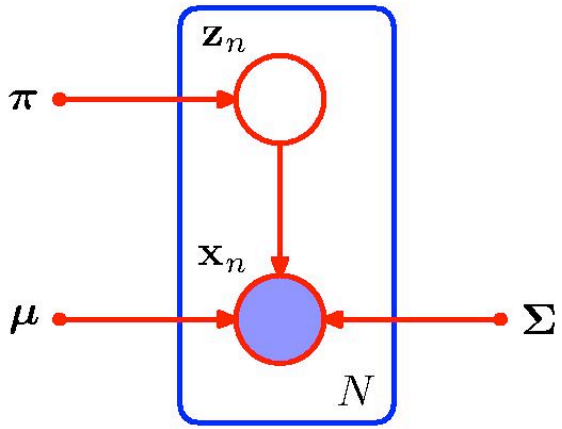
• We will look at the Bayesian mixture of Gaussians model and apply the variational inference to approximate the posterior.

• Note: Many models, corresponding to much more sophisticated distributions, can be solved by straightforward extensions of this analysis.

• Remember the Gaussian mixture model:

$$p(\mathbf{Z}|\boldsymbol{\pi}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\pi_k^{z_{nk}},$$

$$p(\mathbf{X}|\mathbf{Z},\boldsymbol{\mu},\boldsymbol{\Lambda}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\left[\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k^{-1})\right]^{z_{nk}}.$$



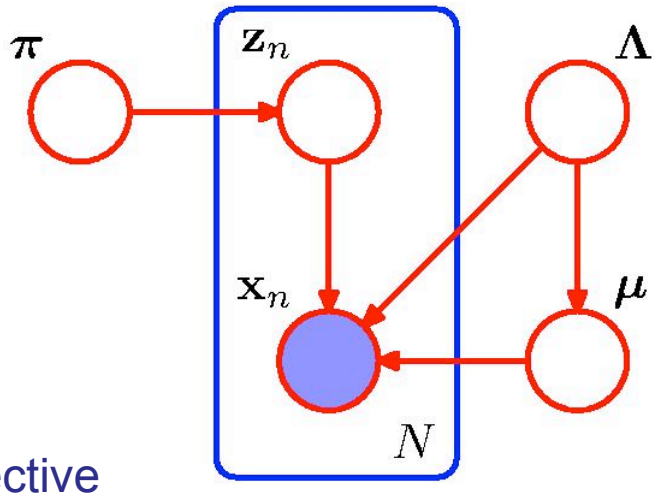
• The log-likelihood takes form:

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

Bayesian Mixture of Gaussians

• We next introduce priors over parameters π, μ, and Λ.

$$p(\pi) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}_0) = \frac{\Gamma(\sum_{k=1}^{K}\alpha_{k0})}{\prod_{k=1}^{K}\Gamma(\alpha_{k0})}\prod_{k=1}^{K}\pi_k^{\alpha_{k0}-1}.$$



• From now on, we assume that each component has the same parameter: $\alpha_{k0} = \alpha_0$, for k=1,…,K.

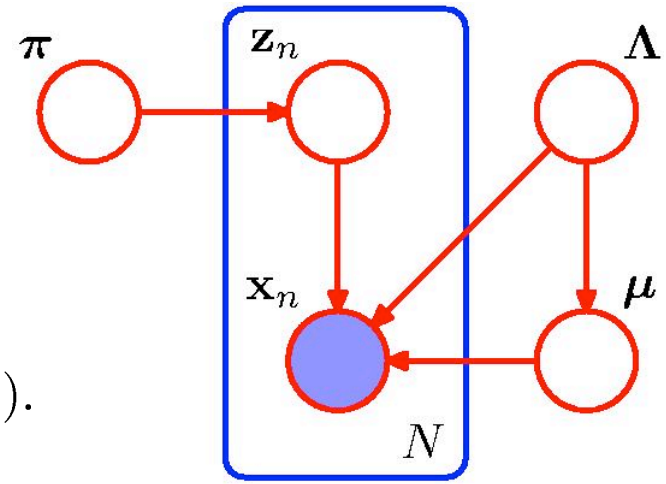
• Note that parameters α_o can be viewed as the effective prior number of observations associated with each component in the mixture,

• If α_o is small, the posterior will be influenced primarily by the data, rather than by the prior.

Bayesian Mixture of Gaussians

- We next introduce priors over parameters π, μ, and Λ.

$$p(\pi) = \mathrm{Dir}(\boldsymbol{\pi}|\alpha_0) = \frac{1}{\mathcal{Z}(\boldsymbol{\alpha}_0)} \prod_{k=1}^{K} \pi_k^{\alpha_0 - 1}.$$



- We also place Gaussian-Wishart prior:

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = p(\boldsymbol{\mu}|\boldsymbol{\Lambda})p(\boldsymbol{\Lambda}) =$$

$$\prod_{k=1}^{K} \mathcal{N}\left(\mu_k|\boldsymbol{m}_0, (\beta_0\boldsymbol{\Lambda}_k)^{-1}\right) \mathcal{W}(\boldsymbol{\Lambda}_k|\boldsymbol{W}_0, \upsilon).$$

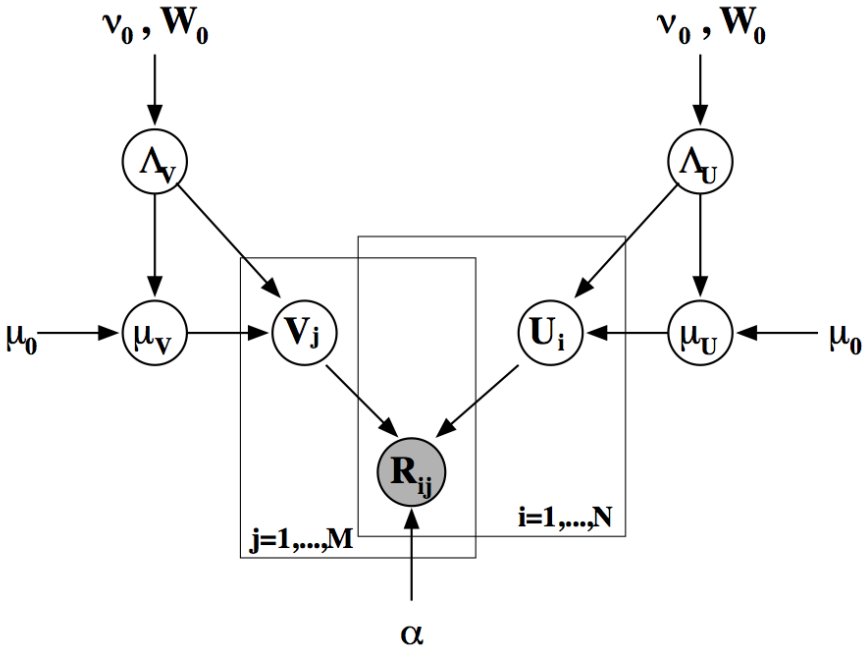
- Typically, we would choose, $\mathbf{m_0}$ = 0 (by symmetry), and $\mathbf{W_0}$ = I, and

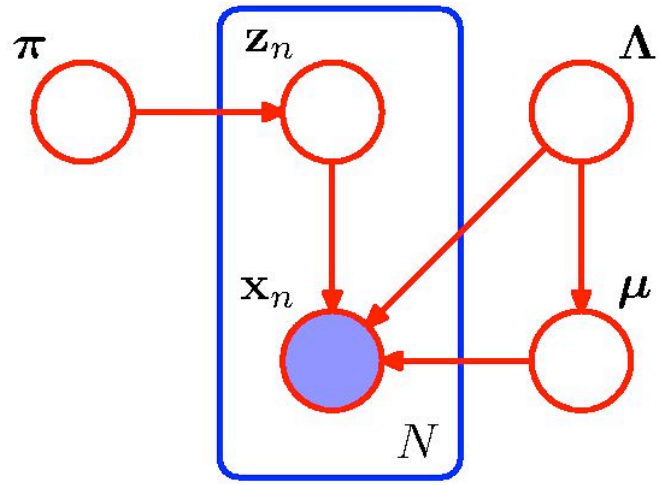
$$\mathcal{W}(\boldsymbol{\Lambda}|\boldsymbol{W}_0, \upsilon) = \frac{1}{\mathcal{Z}(\boldsymbol{W}, \upsilon)} |\boldsymbol{\Lambda}|^{(\upsilon - D - 1)/2} \exp\left(-\frac{1}{2}\mathrm{Tr}(\boldsymbol{W}_0^{-1}\boldsymbol{\Lambda})\right).$$

- Notice the distinction between latent variables and parameters.

Bayesian Matrix Factorization

• One of the popular matrix factorization model used in collaborative filtering / recommender systems.



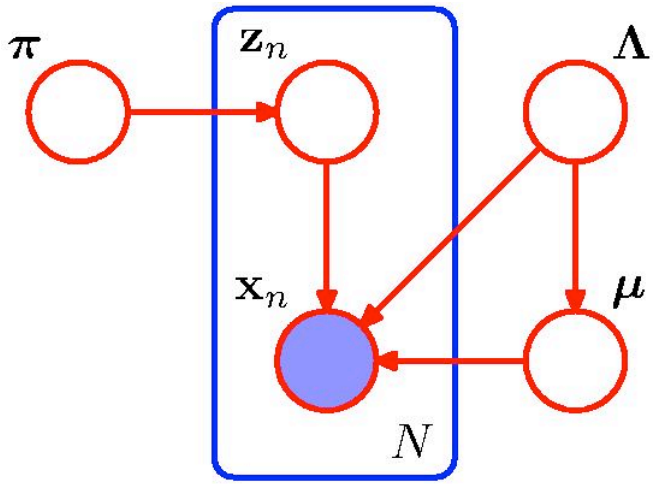


• As pointed before: Many models, corresponding to much more sophisticated distributions, can be solved by straightforward extensions of presented analysis.

Variational Distribution

• We can write down the joint distribution over all random variables:

$$p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) =$$
$$p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) p(\mathbf{Z}|\boldsymbol{\pi}) p(\boldsymbol{\pi}) p(\boldsymbol{\mu}|\boldsymbol{\Lambda}) p(\boldsymbol{\Lambda}).$$



• Consider a variational distribution that factorizes between the latent variables and model parameters:

$$q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\mathbf{Z}) q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}).$$

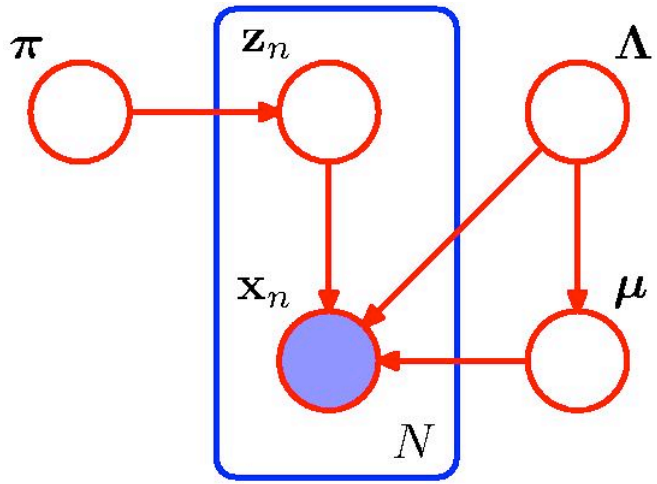
• Remarkably, **this is the only assumption** we need in order to obtain a tractable practical solution to our Bayesian mixture model.

• The functional form of the factors will be determined automatically by optimization of the variational distribution.

Variational Distribution

• Using our general result, we can obtain the optimal factor for q(**Z**):

$$\ln q^*(\mathbf{Z}) = \mathbb{E}_{q(\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda})}[\ln p(\mathbf{X},\mathbf{Z},\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda}] + \text{const.}$$



• Using decomposition of $p(\mathbf{X},\mathbf{Z},\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda})$ and retaining terms that depend on **Z**:

$$\ln q^*(\mathbf{Z}) = \mathbb{E}_{q(\boldsymbol{\pi})}[\ln p(\mathbf{Z}|\boldsymbol{\pi})] + \mathbb{E}_{q(\boldsymbol{\mu},\boldsymbol{\Lambda})}[\ln p(\mathbf{X}|\mathbf{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})] + \text{const.}$$

• Substituting, we obtain:

$$\ln q^*(\mathbf{Z}) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} \ln \rho_{nk} + \text{const.}$$

$$p(\mathbf{Z}|\boldsymbol{\pi}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \pi_k^{z_{nk}},$$

$$\ln \rho_{nk} = \mathbb{E}[\ln \pi_k] + \frac{1}{2}\mathbb{E}[\ln |\Lambda_k|] - \frac{D}{2}\ln(2\pi) - \frac{1}{2}\mathbb{E}\left[(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \Lambda_k (\mathbf{x}_n - \boldsymbol{\mu}_k)\right] + \text{const.}$$

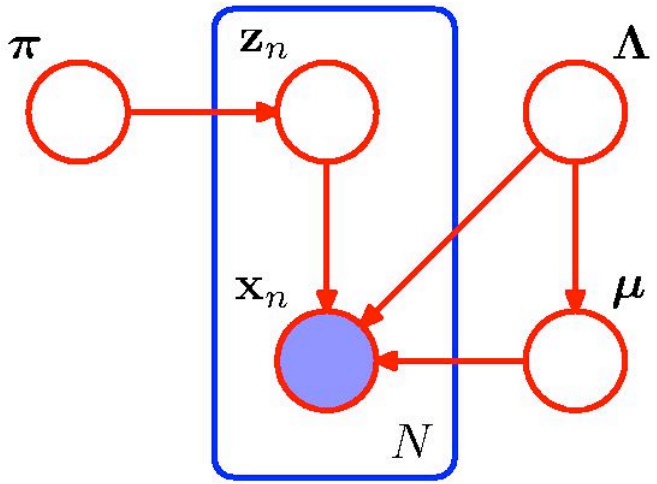
$$p(\mathbf{X}|\mathbf{Z},\boldsymbol{\mu},\boldsymbol{\Lambda}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\left[\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k^{-1})\right]^{z_{nk}}.$$

where expectations are taken with respect to $q(\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda})$.

Variational Distribution

- Using our general result, we can obtain the optimal factor for q(**Z**):

$$\ln q^*(\mathbf{Z}) = \mathbb{E}_{q(\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda})}[\ln p(\mathbf{X},\mathbf{Z},\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda}] + \text{const.}$$



- So far, we have:

$$\ln q^*(\mathbf{Z}) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} \ln \rho_{nk} + \text{const.}$$

- Exponentiating and normalizing, we have:

$$q^*(\mathbf{Z}) = \prod_{n=1}^{N}\prod_{k=1}^{K} r_{nk}^{z_{nk}}, \qquad r_{nk} = \frac{\rho_{nk}}{\sum_{j=1}^{K}\rho_{nj}}.$$

plays the role of responsibility.

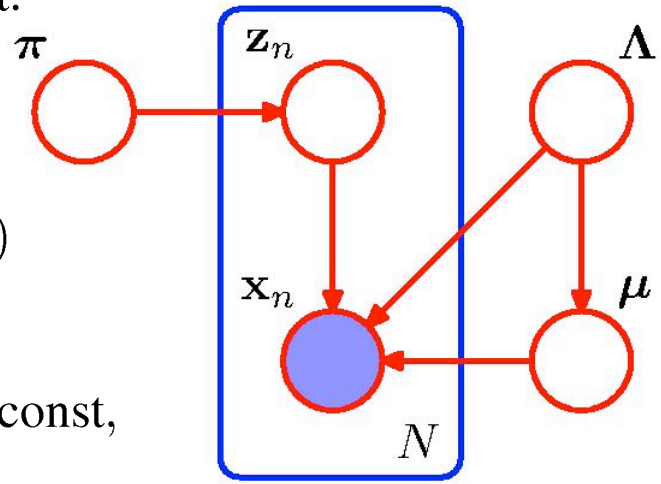
- The optimal solution takes the same function form as the prior p(**Z**|π) (multinomial), and $\mathbb{E}_{q^*(\mathbf{Z})}[z_{nk}] = r_{nk}.$

- Note that the optimal solution depends on moments evaluated with respect to distributions of other variables.

Variational Distribution

- Using our general result, consider the optimal factor for q(π,μ,Λ):

$$\ln q^*(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \mathbb{E}_{q(\mathbf{Z})}[\ln p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}] + \text{const.}$$



- Substituting, we obtain:

$$\ln q^*(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \sum_{n=1}^{N}\sum_{k=1}^{K} \mathbb{E}[z_{nk}] \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})$$

$$+\mathbb{E}[\ln p(\mathbf{Z}|\boldsymbol{\pi})] + \ln p(\boldsymbol{\pi}) + \sum_{k=1}^{K} \ln p(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) + \text{const},$$

where expectations are taken with respect to $q(\mathbf{Z})$.

$$p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) =$$

$$p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) p(\mathbf{Z}|\boldsymbol{\pi}) p(\pi) p(\boldsymbol{\mu}|\boldsymbol{\Lambda}) p(\boldsymbol{\Lambda}).$$

- The result decomposes into a sum of terms involving only π and {μ_k,Λ_k}, k=1,..,K.

$$q^*(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q^*(\boldsymbol{\pi}) \prod_{k=1}^{K} q^*(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) = q^*(\boldsymbol{\pi}) \prod_{k=1}^{K} q^*(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k) q^*(\boldsymbol{\Lambda}_k).$$

Variational Distribution

- Substituting we obtain:

$$\ln q^*(\boldsymbol{\pi}) = \mathbb{E}[\ln p(\mathbf{Z}|\boldsymbol{\pi})] + \ln p(\boldsymbol{\pi}) + \text{const},$$

- So the optimal $q^*(\pi)$ takes form:

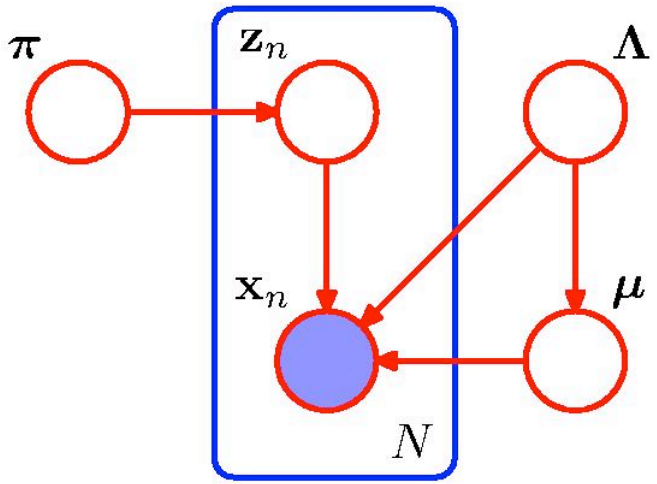
$$\ln q^*(\boldsymbol{\pi}) = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk} \ln \pi_k + (\alpha_0 - 1)\sum_{k=1}^{K} \ln \pi_k + \text{const}.$$

- Exponentiating, we have a Dirichlet distribution:

$$q^*(\boldsymbol{\pi}) = \text{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}),$$

where α has components:

$$\alpha_k = \alpha_0 + N_k, \quad N_k = \sum_{n=1}^{N} r_{nk}.$$



$$p(\mathbf{Z}|\boldsymbol{\pi}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \pi_k^{z_{nk}},$$

$$p(\pi) = \text{Dir}(\boldsymbol{\pi}|\alpha_0) = \frac{1}{\mathcal{Z}(\boldsymbol{\alpha}_0)} \prod_{k=1}^{K} \pi_k^{\alpha_0 - 1}.$$

Variational Distribution

• It will be convenient to define three statistics of the observed dataset with respect to the responsibilities:

$$N_k = \sum_{n=1}^{N} r_{nk},$$

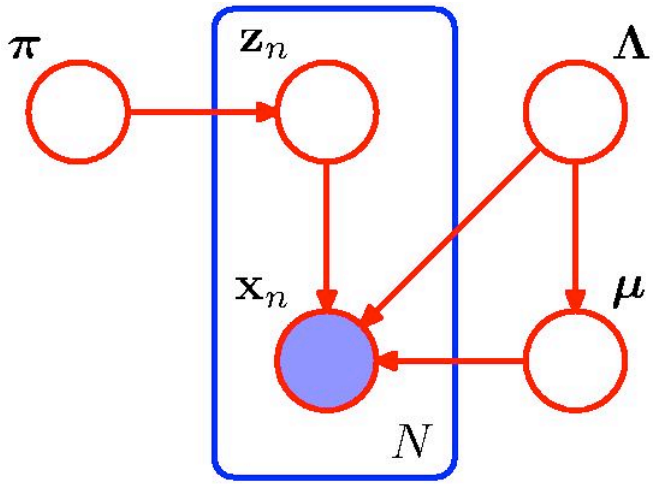
Effective number of points in component k.

$$\bar{\mathbf{x}}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} \mathbf{x}_n,$$

The mean of component k.

$$\boldsymbol{S}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} (\mathbf{x}_n - \bar{\mathbf{x}}_k)(\mathbf{x}_n - \bar{\mathbf{x}}_k)^T.$$

Covariance of component k.

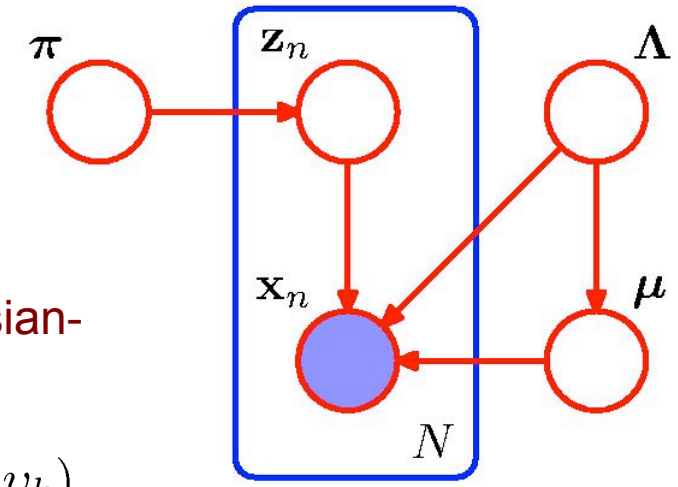


• These are analogous to quantities evaluated in the maximum likelihood EM for the Gaussian mixture models.

Variational Distribution

- Substituting we obtain:

$$\ln q^*(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \sum_{k=1}^{K} \sum_{n=1}^{N} \mathbb{E}[z_{nk}] \ln \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}) + \sum_{k=1}^{K} \ln p(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) + \text{const},$$



- It is easy to verify that optimal $q^*(\mu_k, \Lambda_k)$ is a Gaussian-Wishart distribution:

$$q^*(\boldsymbol{\mu}_k, \Lambda_k) = \mathcal{N}\left(\mu_k | \boldsymbol{m}_k, (\beta_k \boldsymbol{\Lambda}_k)^{-1}\right) \mathcal{W}(\boldsymbol{\Lambda}_k | \boldsymbol{W}_k, \upsilon_k),$$

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = p(\boldsymbol{\mu}|\boldsymbol{\Lambda}) p(\boldsymbol{\Lambda}) = \prod_{k=1}^{K} \mathcal{N}\left(\mu_k | \boldsymbol{m}_0, (\beta_0 \boldsymbol{\Lambda}_k)^{-1}\right) \mathcal{W}(\boldsymbol{\Lambda}_k | \boldsymbol{W}_0, \upsilon).$$

$$\boldsymbol{m}_k = \left(\frac{\beta_0}{\beta_0 + N_k} \boldsymbol{m}_0 + \frac{N_k}{\beta_0 + N_k} \bar{\mathbf{x}}_k\right),$$

$$\boldsymbol{W}_k^{-1} = \boldsymbol{W}_0^{-1} + N_k \mathbf{S}_k + \frac{\beta_0 N_k}{\beta_0 + N_k} (\bar{\mathbf{x}}_k - \boldsymbol{m}_0)(\bar{\mathbf{x}}_k - \boldsymbol{m}_0)^T,$$

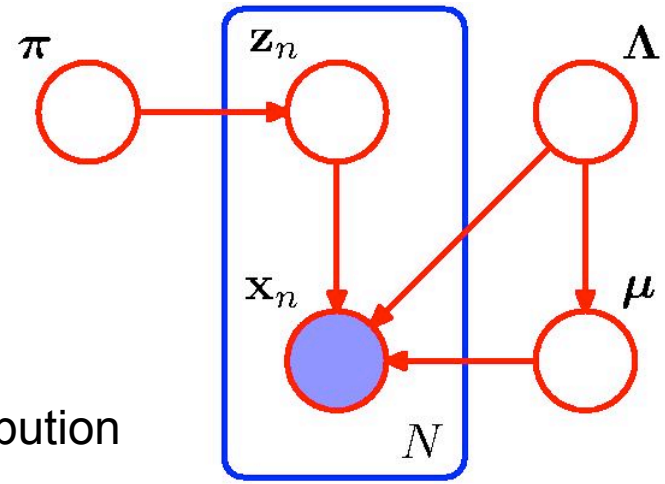
$$\beta_k = \beta_0 + N_k, \quad \upsilon_k = \upsilon_0 + N_k + 1.$$

These update equations are quite intuitive.

But they depend on responsibilities!

Iterative Optimization

• The optimization of variational posterior amounts to cycling between two stages, analogous to the E and M steps of the maximum likelihood EM.



• In the variational E-step, we use the current distribution over parameters q(π,μ,Λ) to evaluate responsibilities:

$$\mathbb{E}[z_{nk}] = r_{nk}.$$

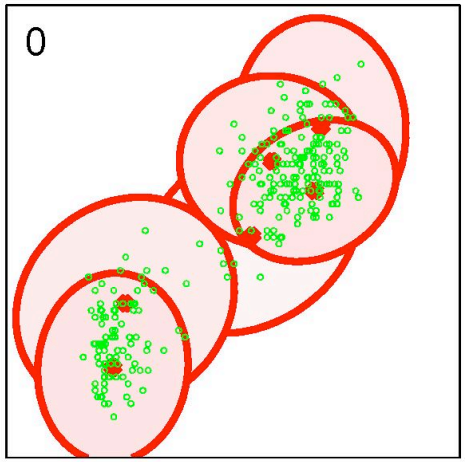
• In the variational M-step, we use the current distribution over latents q(**Z**), or responsibilities, to compute the variational distribution over parameters.

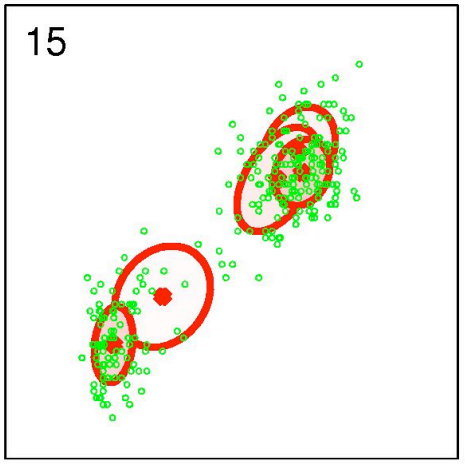
• Each step improves (does not decrease) the variational lower bound on the log-probability of the data.

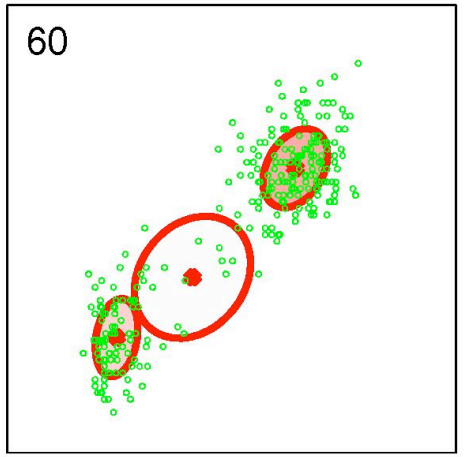
• The variational posterior has the same function form as the corresponding factor in the joint distribution.

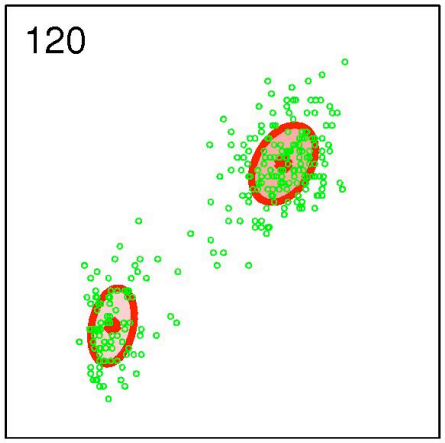
Example

• Variational Bayesian mixture of K=6 Gaussians. Components whose expected mixing coefficients are numerically indistinguishable from zero are not shown.









$\alpha_0 = 0.001$

The posterior over latents is given by Dirichlet:

$$q^*(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}), \quad \alpha_k = \alpha_0 + N_k.$$

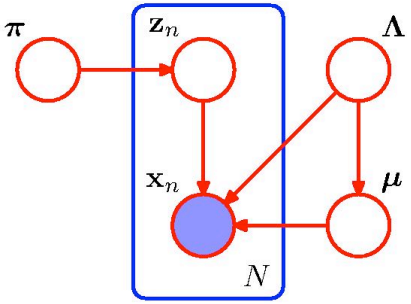
$$\mathbb{E}[\pi_k] = \frac{\alpha_0 + N_k}{K\alpha_0 + N}. \qquad p(\pi) = \mathrm{Dir}(\boldsymbol{\pi}|\alpha_0) = \frac{1}{\mathcal{Z}(\boldsymbol{\alpha}_0)} \prod_{k=1}^{K} \pi_k^{\alpha_0 - 1}.$$

Consider a component for which $N_k \simeq 0$ and $\alpha_o \simeq 0$.

If the prior is broad, so that $\alpha_o \to 0$, the $E[\pi_k] \to 0$, and the component plays no role in the model.

Variational Lower Bound

- It is straightforward to evaluate the variational lower bound for this model.
- We can monitor during the re-estimation in order to test for convergence.
- It also provides a valuable check on the mathematical updates and software implementation.
- At each step of the iterative procedure, the variational lower bound should not decrease.
- For the variational mixture of Gaussian model, the lower bound can be evaluated as:



$$\mathcal{L}(q) = \sum_{\mathbf{Z}} \iiint q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \ln \frac{p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})}{q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})} \mathrm{d}\boldsymbol{\pi} \mathrm{d}\boldsymbol{\mu} \mathrm{d}\boldsymbol{\Lambda}$$

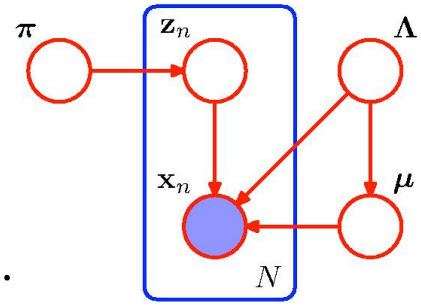
$$= \mathbb{E}[\ln p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] - \mathbb{E}[\ln q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})]$$

$$= \mathbb{E}[\ln p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] + \mathbb{E}[\ln p(\mathbf{Z}|\boldsymbol{\pi})] + \mathbb{E}[\ln p(\boldsymbol{\pi})] + \mathbb{E}[\ln p(\boldsymbol{\mu}, \boldsymbol{\Lambda})]$$
$$- \mathbb{E}[\ln q(\mathbf{Z})] - \mathbb{E}[\ln q(\boldsymbol{\pi})] - \mathbb{E}[\ln q(\boldsymbol{\mu}, \boldsymbol{\Lambda})].$$

- The various terms in the bound can be easily evaluated.

Predictive Density

- In general, we will be interested in the predictive density for a new value $\hat{\mathbf{x}}$ of the observed variable.
- We will also have a corresponding latent variable $\hat{\mathbf{z}}$.
- The predictive density takes form:



$$p(\hat{\mathbf{x}}|\mathbf{X}) = \sum_{\hat{\mathbf{z}}} \iiint p(\hat{\mathbf{x}}|\hat{\mathbf{z}}, \boldsymbol{\mu}, \mathbf{\Lambda}) p(\hat{\mathbf{z}}|\boldsymbol{\pi}) p(\boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{\Lambda}|\mathbf{X}) \mathrm{d}\boldsymbol{\pi} \mathrm{d}\boldsymbol{\mu} \mathrm{d}\mathbf{\Lambda}.$$

probability of a new data point given latent component and parameters.

probability of a latent component that is summed out.

posterior probability over parameters conditioned on the entire dataset.

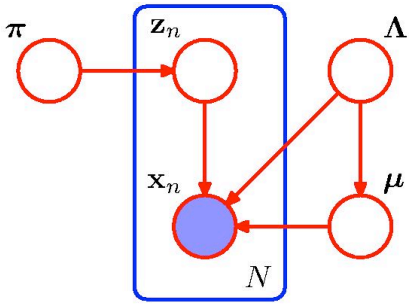
- Summing out latent variable we obtain:

$$p(\hat{\mathbf{x}}|\mathbf{X}) = \sum_{k} \iiint \pi_k \mathcal{N}(\hat{\mathbf{x}}|\boldsymbol{\mu}_k, \mathbf{\Lambda}_k) p(\boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{\Lambda}|\mathbf{X}) \mathrm{d}\boldsymbol{\pi} \mathrm{d}\boldsymbol{\mu} \mathrm{d}\mathbf{\Lambda}.$$

Predictive Density

- Predictive density takes form:

$$p(\hat{\mathbf{x}}|\mathbf{X}) = \sum_k \iiint \pi_k \mathcal{N}(\hat{\mathbf{x}}|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}|\mathbf{X}) \mathrm{d}\boldsymbol{\pi} \mathrm{d}\boldsymbol{\mu} \mathrm{d}\boldsymbol{\Lambda}.$$



- We now approximate the true posterior distribution p(π,μ,Λ) with its variational approximation q(π)q(μ,Λ).

- This gives an approximation:

$$p(\hat{\mathbf{x}}|\mathbf{X}) \simeq \hat{p}(\hat{\mathbf{x}}|\mathbf{X}) = \sum_k \iiint \pi_k \mathcal{N}(\hat{\mathbf{x}}|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) q(\boldsymbol{\pi}) q(\boldsymbol{\mu}, \boldsymbol{\Lambda}) \mathrm{d}\boldsymbol{\pi}_k \mathrm{d}\boldsymbol{\mu}_k \mathrm{d}\boldsymbol{\Lambda}_k.$$

- The integration can now be performed analytically, giving a mixture of Student's t-distribution:

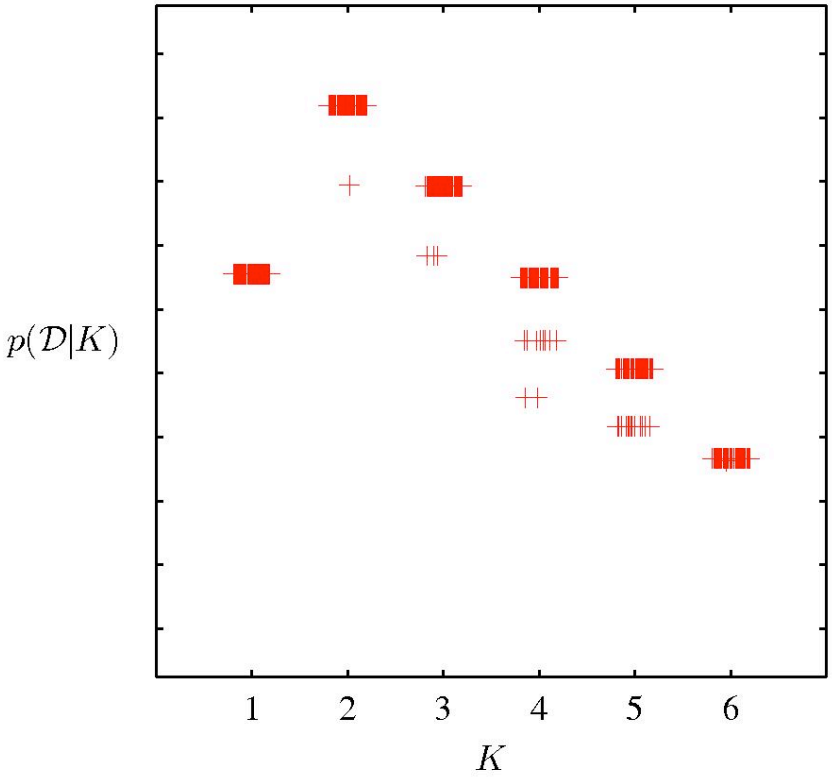
$$\hat{p}(\hat{\mathbf{x}}|\mathbf{X}) = \frac{1}{\hat{\alpha}} \sum_k \alpha_k \mathrm{St}(\hat{\mathbf{x}}|\boldsymbol{\mu}_k, \boldsymbol{L}_k, \upsilon_k + 1 - D),$$

where $\hat{\alpha} = \sum_k \alpha_k$, and the precision is given by: $\boldsymbol{L}_k = \frac{(\upsilon_k + 1 - D)\beta_k}{1 + \beta_k} \boldsymbol{W}_k.$

- As N becomes large, the predictive distribution reduces to a mixture of Gaussians.

Determining the Number of Components

• Plot of the variational lower bound (including multimodility factor K!) vs. the number K of components.



• For each value of K, the model is trained from 100 different random starts.

• Maximum likelihood would increase monotonically with K (assuming no singular solutions).

Induced Factorizations

- In our variational mixture of Gaussians model, we assumed a particular factorization:

$$q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\mathbf{Z})q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}).$$

- The optimal solutions for various factors exhibit additional factorizations:

$$q^*(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q^*(\boldsymbol{\pi}) \prod_{k=1}^{K} q^*(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k)q^*(\boldsymbol{\Lambda}_k),$$

- We call these induced factorizations.

- In a numerical implementation of the variational approach it is important to take into account additional factorizations.

- These additional factorizations can be detected using d-separation.

Induced Factorizations

• Let us partition the latent variables into A,B,C, and assume the the following factorization that would approximate the true posterior:

$$q(A, B, C) = q(A, B)q(C).$$

• The solution for q(A,B) takes form:

$$\begin{aligned} \ln q^*(A, B) &= \mathbb{E}_C[\ln p(\mathbf{X}, A, B, C)] + \text{const} \\ &= \mathbb{E}_C[\ln p(A, B|C, \mathbf{X})] + \text{const}. \end{aligned}$$

• We now test whether the resulting solution factorizes between A and B:

$$q^*(A, B) = q^*(A)q^*(B).$$

• This will happen iff:

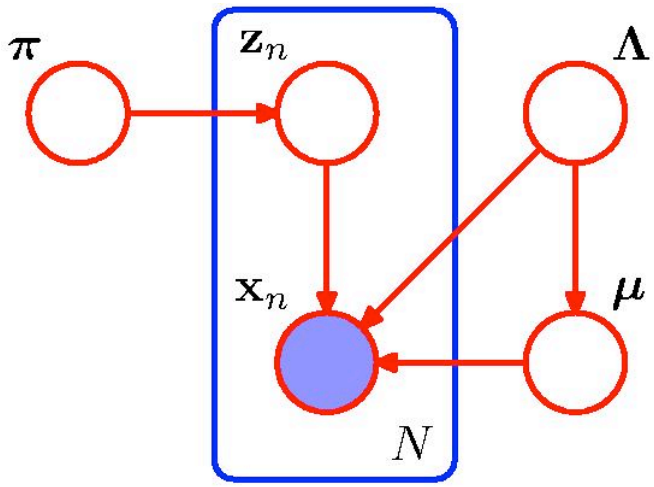
$$\ln p(A, B|C, \mathbf{X}) = \ln p(A|C, \mathbf{X}) + \ln(B|C, \mathbf{X}).$$

• Or if A is independent of B conditioned on C and **X**.

Induced Factorizations

• In case of Bayesian mixture of Gaussians, we can immediately see that the variational posterior over parameters must factorize between π and (μ,Λ).

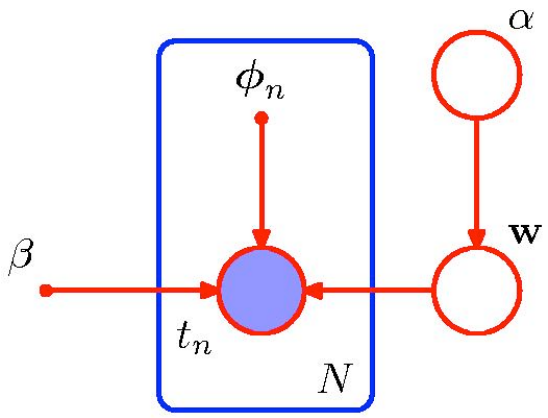
$$q^*(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q^*(\boldsymbol{\pi}) \prod_{k=1}^{K} q^*(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k) q^*(\boldsymbol{\Lambda}_k),$$



• All paths that connecting μ or Λ must pass through one of the nodes z_n, all of which are in our conditioning set.

Variational Linear Regression

- We now look at the Bayesian linear regression model as another example.
- Interating over parameters and hyperparameters is often intractable.
- We can find a tractable approximation using variational methods.



- Bayesian Linear Regression model:

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{i=1}^{N} \mathcal{N}(t_n|\mathbf{w}^T\boldsymbol{\phi}_n, \beta^{-1}),$$

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$$

$$\mathrm{Gam}(\alpha|a_0, b_0) = \frac{b_0^{a_0}}{\Gamma(a_0)}\alpha^{a_0-1}\exp(-b_0\alpha).$$

where $\boldsymbol{\phi}_n = \boldsymbol{\phi}(\mathbf{x}_n)$.

- We next place a conjugate Gamma prior over α:

$$p(\alpha) = \mathrm{Gam}(\alpha|a_0, b_0).$$

Variational Linear Regression

- The joint distribution takes form:

$$p(\mathbf{t}, \mathbf{w}, \alpha) = p(\mathbf{t}|\mathbf{w})p(\mathbf{w}|\alpha)p(\alpha).$$

- Our goal is to find an approximation to the posterior:

$$p(\mathbf{w}, \alpha|\mathbf{t}).$$

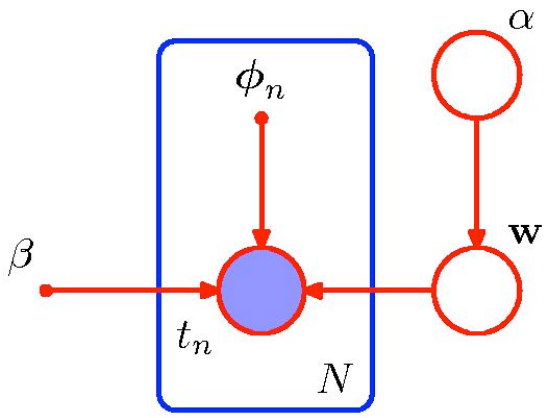
- Using variational framework, we assume approximate posterior factorizes:

$$q(\mathbf{w}, \alpha) = q(\mathbf{w})q(\alpha).$$

- Consider distribution over α:

$$\ln q^*(\alpha) = \ln p(\alpha) + \mathbb{E}_{\mathbf{w}}[\ln p(\mathbf{w}|\alpha)] + \text{const}$$

$$= (a_0 - 1)\ln\alpha - b_0\alpha + \frac{M}{2}\ln\alpha - \frac{\alpha}{2}\mathbb{E}_{\mathbf{w}}[\mathbf{w}^T\mathbf{w}] + \text{const}$$



$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$$

$$p(\alpha) = \text{Gam}(\alpha|a_0, b_0).$$

$$\text{Gam}(\alpha|a_0, b_0) = \frac{b_0^{a_0}}{\Gamma(a_0)}\alpha^{a_0-1}\exp(-b_0\alpha).$$

Variational Linear Regression

- The distribution over α:

$$\ln q^*(\alpha) = \ln p(\alpha) + \mathbb{E}_{\mathbf{w}}[\ln p(\mathbf{w}|\alpha)] + \text{const}$$

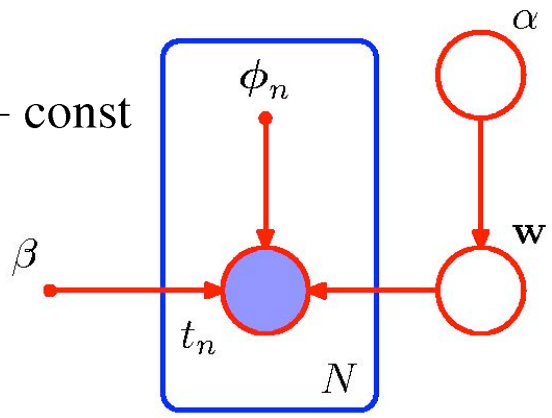
$$= (a_0 - 1)\ln\alpha - b_0\alpha + \frac{D}{2}\ln\alpha - \frac{\alpha}{2}\mathbb{E}_{\mathbf{w}}[\mathbf{w}^T\mathbf{w}] + \text{const}$$

- We can easily recognize this as the log of a Gamma distribution:

$$q^*(\alpha) = \text{Gam}(\alpha|a_N, b_N),$$

$$a_N = a_0 + \frac{M}{2},$$

$$b_N = b_0 + \frac{1}{2}\mathbb{E}[\mathbf{w}^T\mathbf{w}].$$



$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$$

$$p(\alpha) = \text{Gam}(\alpha|a_0, b_0).$$

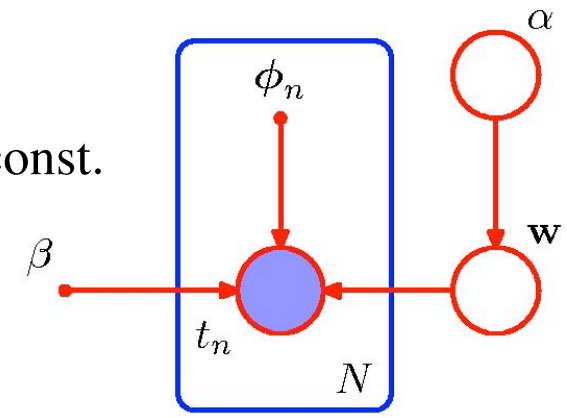
$$\text{Gam}(\alpha|a_0, b_0) = \frac{b_0^{a_0}}{\Gamma(a_0)}\alpha^{a_0-1}\exp(-b_0\alpha).$$

Variational Linear Regression

- The optimal factor for model parameters takes form:

$$\ln q^*(\mathbf{w}) = \ln p(\mathbf{t}|\mathbf{w}) + \mathbb{E}_\alpha[\ln p(\mathbf{w}|\alpha)] + \text{const}$$

$$= -\frac{\beta}{2}\sum_n \left(\mathbf{w}^T\boldsymbol{\phi}_n - t_n\right)^2 - \frac{1}{2}\mathbb{E}_\alpha[\alpha]\mathbf{w}^T\mathbf{w} + \text{const}.$$



$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$$

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{i=1}^{N} \mathcal{N}(t_n|\mathbf{w}^T\boldsymbol{\phi}_n, \beta^{-1}),$$

- Hence the distribution $q^*(w)$ is Gaussian (due to quadratic form):

$$q^*(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\boldsymbol{m}_N, \boldsymbol{S}_N),$$

$$\boldsymbol{m}_N = \beta\boldsymbol{S}_N\Phi^T\mathbf{t},$$

$$\boldsymbol{S}_N = \left(\mathbb{E}_\alpha[\alpha]I + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right).$$

Same form when alpha was treated as a fixed parameter.

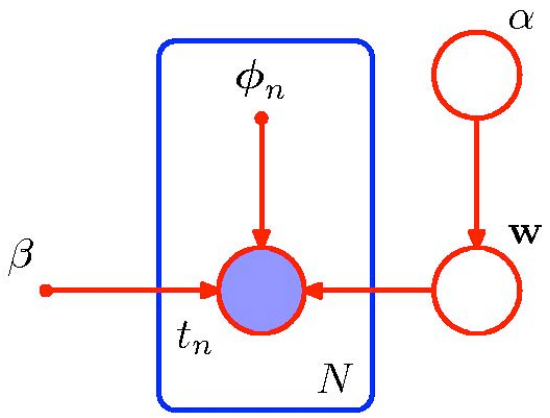
- The difference is that fixed α is replaced by its expectation under the variational approximation $q(\alpha)$.

Variational Lower Bound

- Once we have identified the optimal q distributions, it is easy to compute:

$$\mathbb{E}[\alpha] = \frac{a_N}{b_N},$$

$$\mathbb{E}[\mathbf{w}\mathbf{w}^T] = \boldsymbol{m}_N \boldsymbol{m}_N^T + \boldsymbol{S}_N.$$



$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$$

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{i=1}^{N} \mathcal{N}(t_n|\mathbf{w}^T\boldsymbol{\phi}_n, \beta^{-1}),$$

$$p(\alpha) = \mathrm{Gam}(\alpha|a_0, b_0).$$

- We can also easily evaluate the variational lower bound on the log-probability of the data:

$$\mathcal{L}(q) = \mathbb{E}[\ln p(\mathbf{t}, \mathbf{w}, \alpha)] - \mathbb{E}[\ln q(\mathbf{w}, \alpha)]$$

Expected complete-data log-likelihood

Negative entropy of approximate q distribution

$$\mathcal{L}(q) = \mathbb{E}_{\mathbf{w}}[\ln p(\mathbf{t}|\mathbf{w})] + \mathbb{E}_{\mathbf{w},\alpha}[\ln p(\mathbf{w}|\alpha)] + \mathbb{E}_{\alpha}[\ln p(\alpha)]$$
$$-\mathbb{E}_{\mathbf{w}}[\ln q(\mathbf{w})] - \mathbb{E}_{\alpha}[\ln q(\alpha)].$$

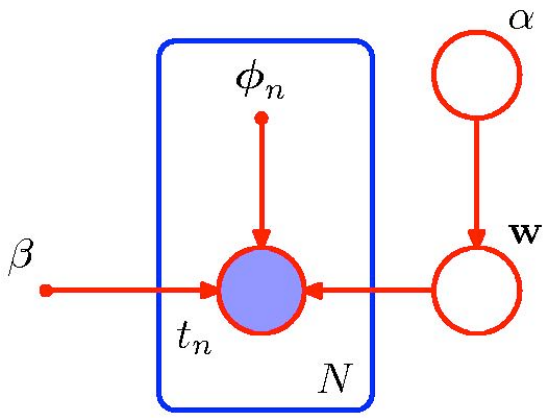
Predictive Distribution

• We can also easily evaluate the predictive distribution over t given a new input x.

$$p(t|\mathbf{x},\mathbf{t}) = \int p(t|\mathbf{x},\mathbf{w})p(\mathbf{w}|\mathbf{t})\mathrm{d}\mathbf{w}$$

$$\simeq \int p(t|\mathbf{x},\mathbf{w})q(\mathbf{w}|\mathbf{t})\mathrm{d}\mathbf{w}$$

$$= \mathcal{N}(t|\mu_N^T\boldsymbol{\phi}(\mathbf{x}), \sigma^2(\mathbf{x}),$$



$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$$

$$p(\mathbf{t}|\mathbf{X},\mathbf{w},\beta) = \prod_{i=1}^{N} \mathcal{N}(t_n|\mathbf{w}^T\boldsymbol{\phi}_n, \beta^{-1}),$$

$$p(\alpha) = \mathrm{Gam}(\alpha|a_0, b_0).$$

where the input-dependent variance is given by:

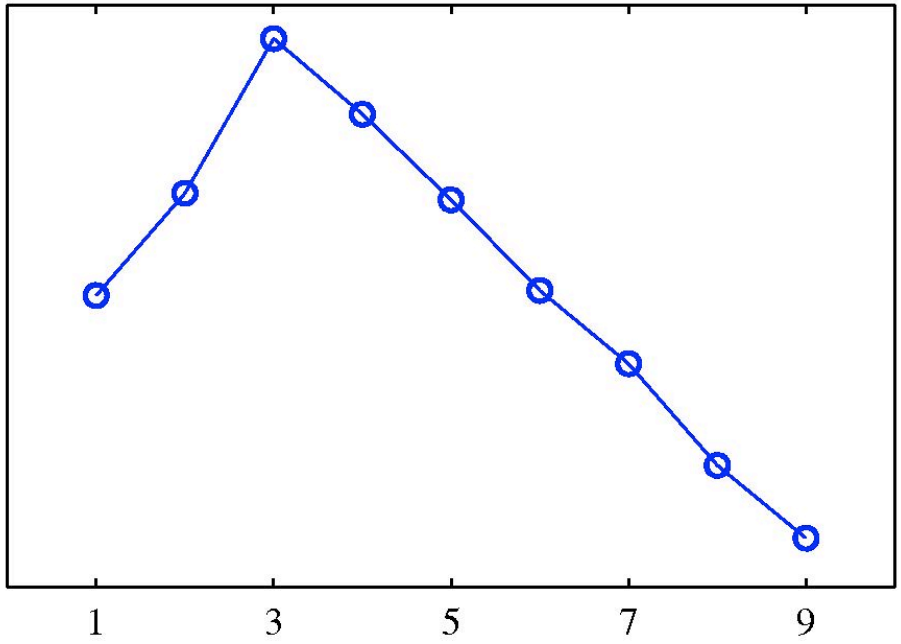
$$\sigma^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^T\boldsymbol{S}_N\boldsymbol{\phi}(\mathbf{x}).$$

• This takes exactly the same form as we considered before with fixed α.

• The difference is that fixed α is replaced by its expectation under the variational approximation q(α).

Predictive Distribution

- Plot of the lower bound vs. the order M of the polynomail:



- So far we have looked at Bayesian models where we are interested in approximating the posterior. The same variational framework can also be applied when learning other models, including undirected models with latent variables (e.g. Deep Boltzmann Machines).