Advanced Data Analytics
Lecture week 6: Variational Bayes

Ian T. Nabney

University of BRISTOL

Overview

- Understand the basis of variational inference applied to Bayesian modelling
- Application of variational methods to GMM

Further reading: Bishop sections 10.1 and 10.2.

Principles of variational methods

- We are picking up the story from the EM algorithm.
- Suppose we have a fully Bayesian model in which all parameters are given prior distributions. The set of all latent variables and parameters is denoted by $\mathbf{Z}$ and the set of all observed variables by $\mathbf{X}$.
- we can decompose the log marginal probability using

$$\ln p(\mathbf{X}) = \mathcal{L}(q) + \mathrm{KL}(q\|p) \tag{1}$$

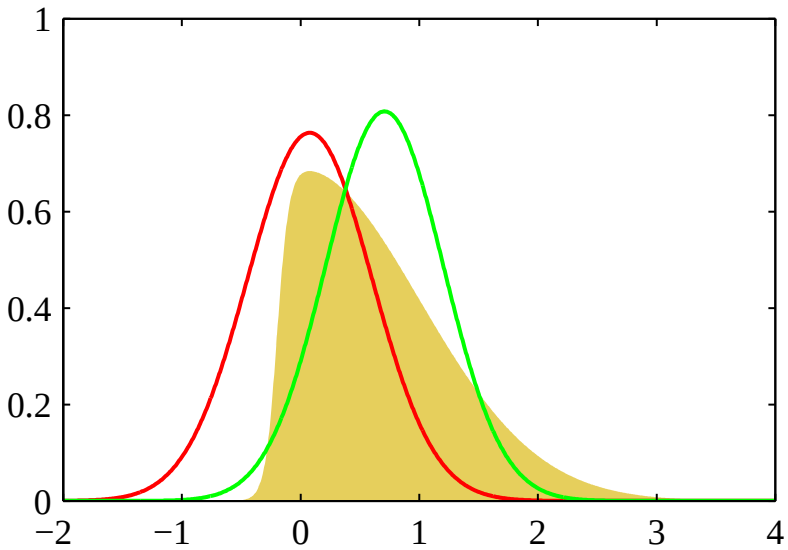
where we have defined

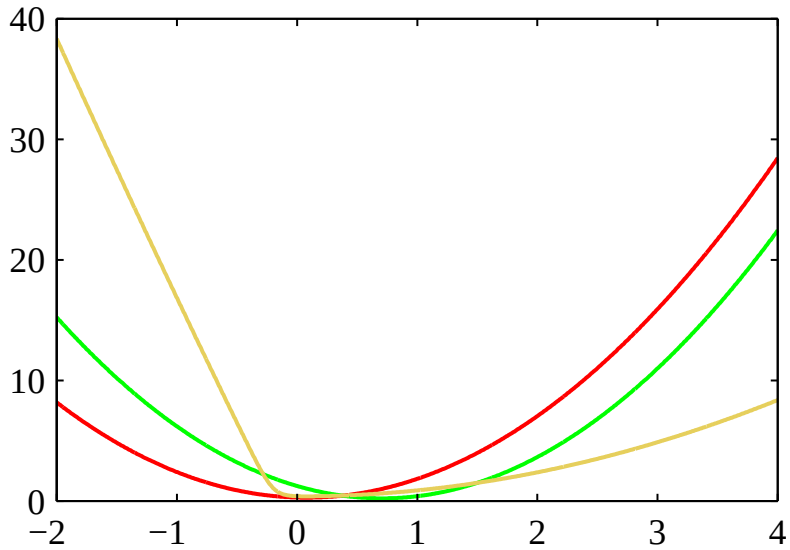
$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z} \tag{2}$$

$$\mathrm{KL}(q\|p) = -\int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}. \tag{3}$$

- We consider a restricted family of distributions $q(\mathbf{Z})$ and then seek the member of this family for which the KL divergence is minimized. Our goal is to restrict the family sufficiently that they comprise only tractable distributions, while at the same time allowing the family to be sufficiently rich and flexible that it can provide a good approximation to the true posterior distribution.

Approximation comparison





The left-hand plot shows the original distribution (yellow) along with the Laplace (red) and variational (green) approximations, and the right-hand plot shows the negative logarithms of the corresponding curves.

Factorised distributions

- Another way in which to restrict the family of distributions $q(\mathbf{Z})$ is to partition the elements of $\mathbf{Z}$ into M disjoint groups that we denote by $\mathbf{Z}_i$.
- We then assume that the q distribution factorizes with respect to these groups, so that

$$q(\mathbf{Z}) = \prod_{i=1}^{M} q_i(\mathbf{Z}_i). \tag{4}$$

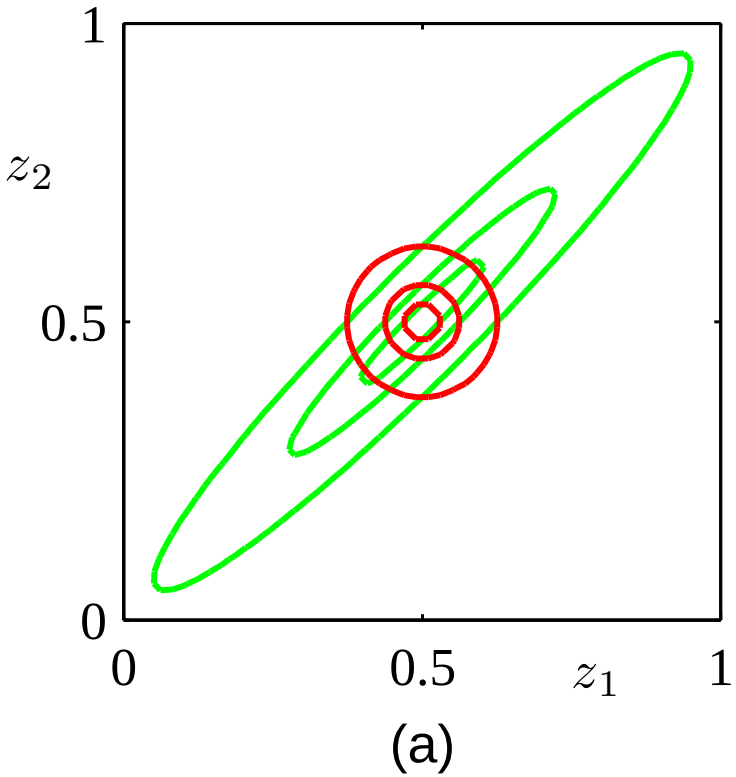
- We make a free form (variational) optimization of $\mathcal{L}(q)$ with respect to each of the distributions $q_i(\mathbf{Z}_i)$ in turn.

$$\ln q_j^\star(\mathbf{Z}_j) = \mathbb{E}_{i\neq j}[\ln p(\mathbf{X}, \mathbf{Z})] + c$$
$$\mathbb{E}_{i\neq j}[\ln p(\mathbf{X}, \mathbf{Z})] = \int \ln p(\mathbf{X}, \mathbf{Z}) \prod_{i\neq j} q_i \, \mathrm{d}\mathbf{Z}_i. \tag{5}$$

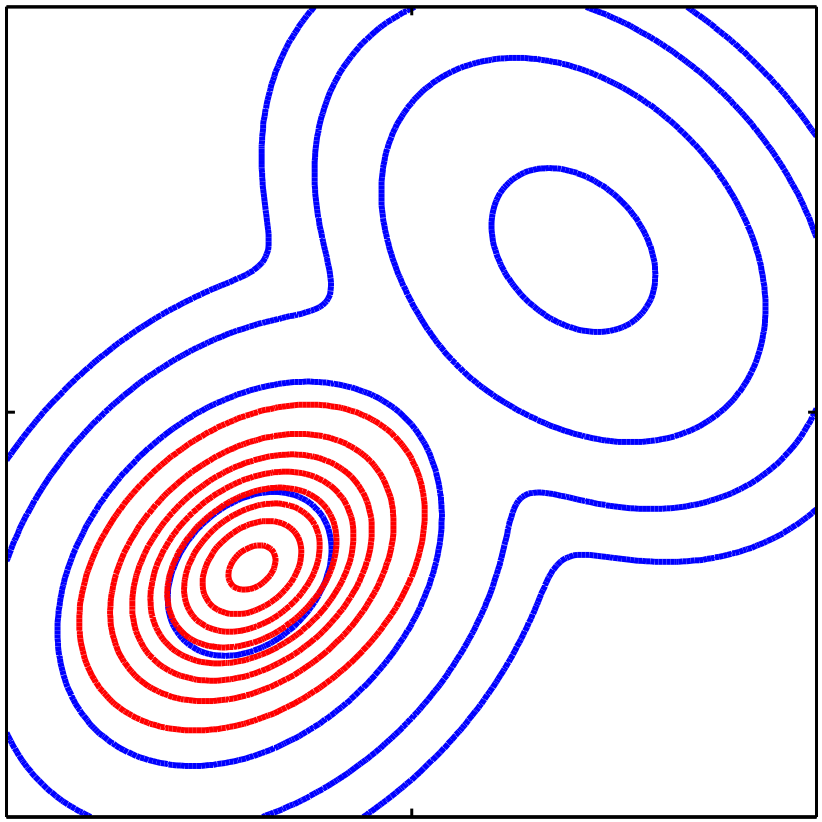
- A consistent solution is found by first initializing all of the factors $q_i(\mathbf{Z}_i)$ appropriately and then cycling through the factors and replacing each in turn with a revised estimate evaluated using the current estimates for all of the other factors.

Approximation properties

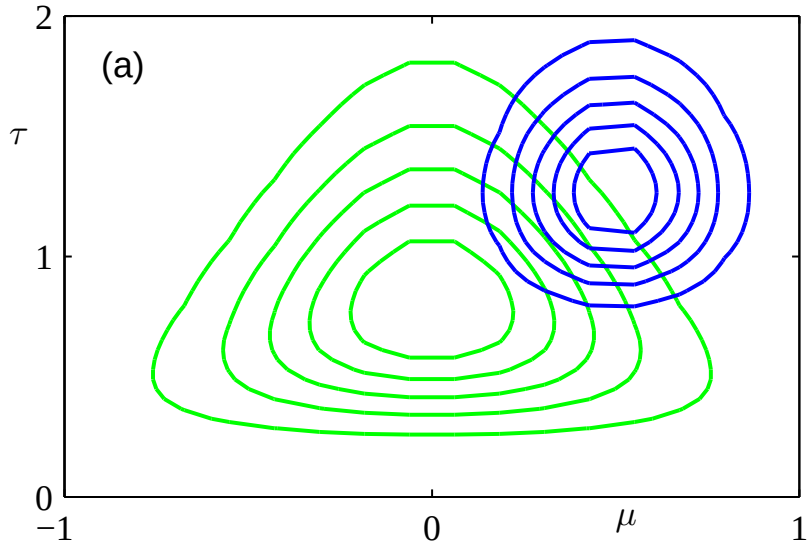
We consider dependence on KL divergence. The reverse divergence $\mathrm{KL}(p\|q)$ leads to the Expectation Propagation algorithm.

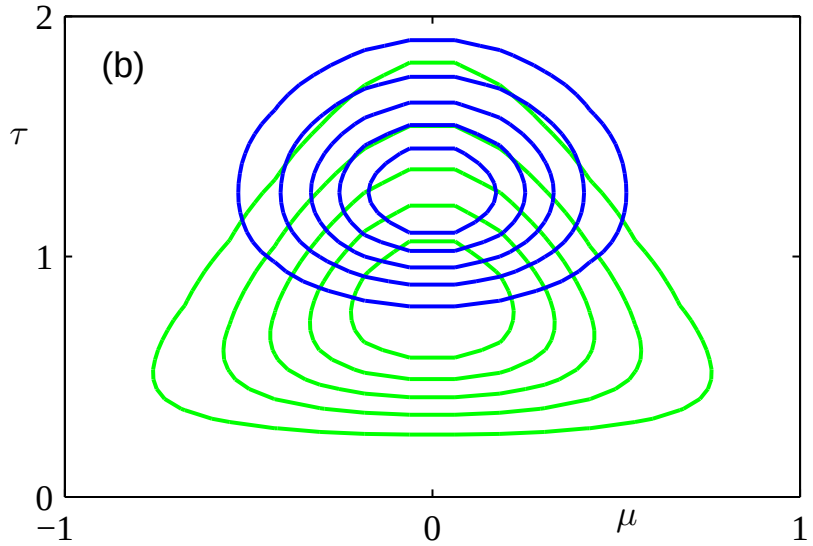


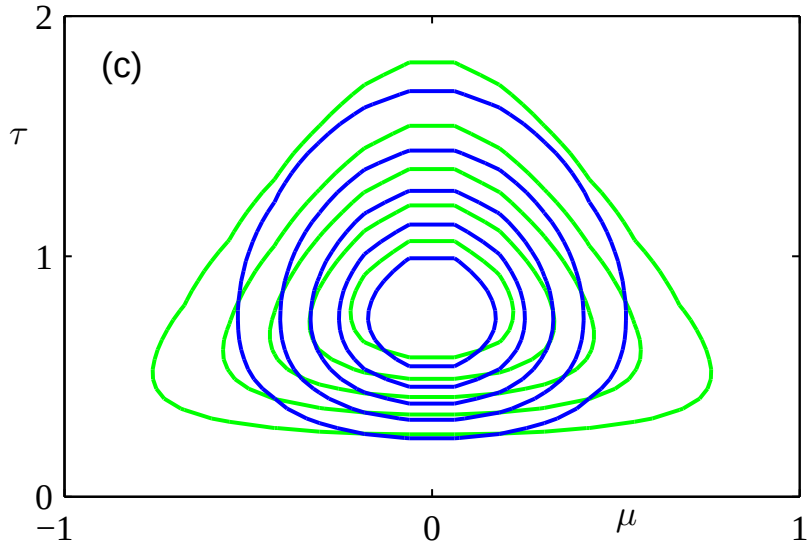
(a)

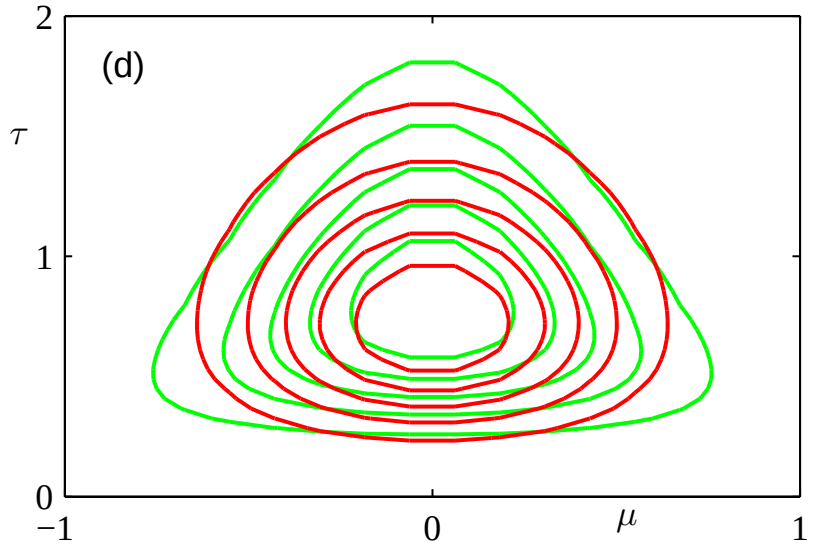


Example: univariate Gaussian









University of BRISTOL

Variational mixture of Gaussians

- The starting point is the latent variable model for a GMM, but with precision matrices $\boldsymbol{\Lambda}$ in place of covariances.
- We choose conjugate priors over the parameters $\boldsymbol{\mu}$, $\boldsymbol{\Lambda}$ and $\boldsymbol{\pi}$.
- We choose a Dirichlet distribution over the mixing coefficients $\boldsymbol{\pi}$

$$p(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}_0) = C(\boldsymbol{\alpha}_0) \prod_{k=1}^{K} \pi_k^{\alpha_0 - 1} \tag{6}$$

The parameter α_0 can be interpreted as the effective prior number of observations associated with each component of the mixture. If the value of α_0 is small, then the posterior distribution will be influenced primarily by the data rather than by the prior.

- We choose an independent Gaussian-Wishart prior governing the mean and precision of each Gaussian component, given by

$$\begin{aligned} p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) &= p(\boldsymbol{\mu}|\boldsymbol{\Lambda}) p(\boldsymbol{\Lambda}) \\ &= \prod_{k=1}^{K} \mathcal{N}\left(\boldsymbol{\mu}_k | \mathbf{m}_0, (\beta_0 \boldsymbol{\Lambda}_k)^{-1}\right) \mathcal{W}(\boldsymbol{\Lambda}_k | \mathbf{W}_0, \nu_0) \end{aligned} \tag{7}$$

Typically we would choose $\mathbf{m}_0 = \mathbf{0}$ by symmetry.

Variational distribution

- The joint distribution of all of the random variables is given by

$$p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})p(\mathbf{Z}|\boldsymbol{\pi})p(\boldsymbol{\pi})p(\boldsymbol{\mu}|\boldsymbol{\Lambda})p(\boldsymbol{\Lambda}) \tag{8}$$

- We use a variational distribution which factorizes between the latent variables and the parameters so that

$$q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\mathbf{Z})q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}). \tag{9}$$

- This is the only assumption that we need to make in order to obtain a tractable practical solution to our Bayesian mixture model.
- The functional form of the factors $q(\mathbf{Z})$ and $q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ is determined automatically by optimization of the variational distribution.

Update equation for $q(\mathbf{Z})$

- Using the general result (5),the log of the optimized factor is given by

$$\ln q^{\star}(\mathbf{Z}) = \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda}}[\ln p(\mathbf{X},\mathbf{Z},\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda})] + \text{const.} \tag{10}$$

- We now make use of the decomposition (8) to obtain

$$\ln q^{\star}(\mathbf{Z}) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} \ln \rho_{nk} + \text{const} \tag{11}$$

where we have defined

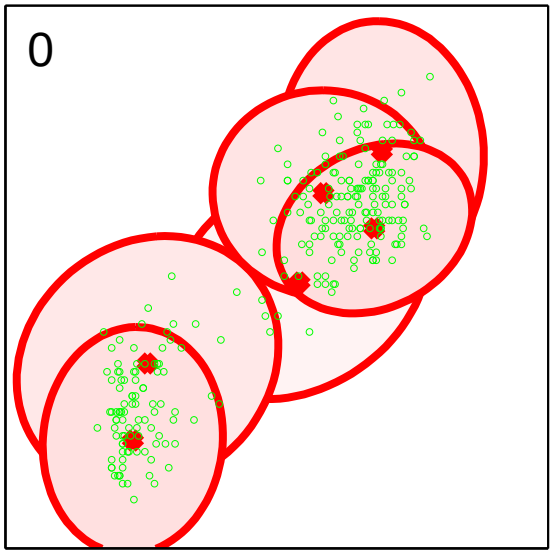
$$\begin{aligned} \ln \rho_{nk} \quad = \quad & \mathbb{E}[\ln \pi_k] + \frac{1}{2}\mathbb{E}\left[\ln |\boldsymbol{\Lambda}_k|\right] - \frac{D}{2}\ln(2\pi) \\ & -\frac{1}{2}\mathbb{E}_{\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k}\left[(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\Lambda}_k(\mathbf{x}_n - \boldsymbol{\mu}_k)\right] \end{aligned} \tag{12}$$

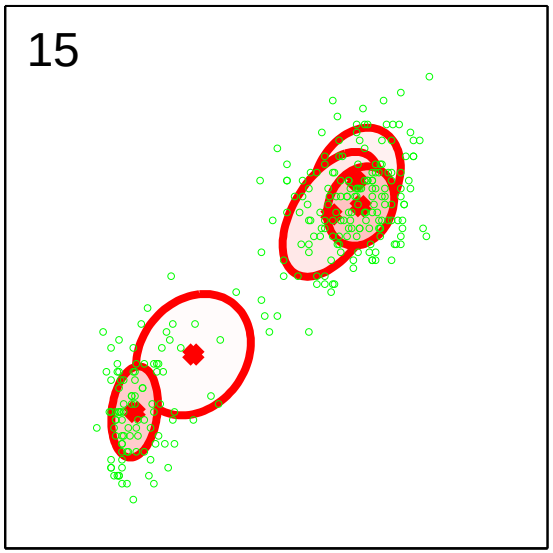
- Requiring the distribution to be normalised,

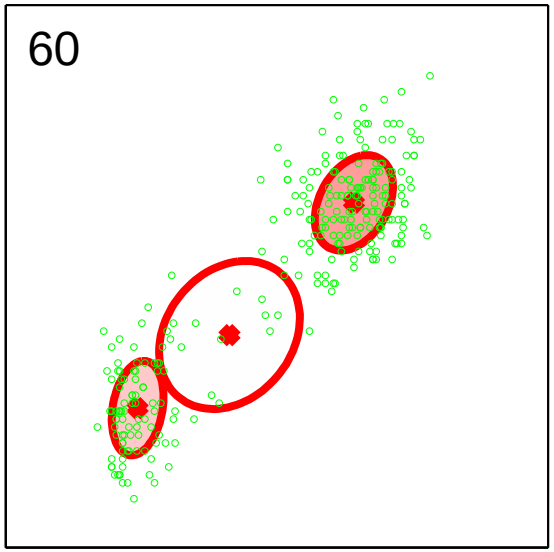
$$q^{\star}(\mathbf{Z}) = \prod_{n=1}^{N}\prod_{k=1}^{K} r_{nk}^{z_{nk}} \quad \text{where} \quad r_{nk} = \frac{\rho_{nk}}{\sum_{j=1}^{K}\rho_{nj}}. \tag{13}$$

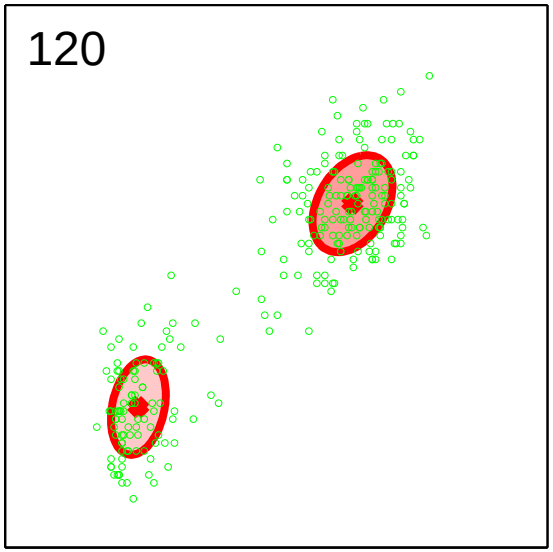
So the values r_{nk} play the role of responsibilities.

Example









University of BRISTOL

Consideration of mixing coefficients

- Components that take essentially no responsibility for explaining the data points have $r_{nk} \simeq 0$ and hence $N_k = \sum_{n=1}^{N} r_{nk} \simeq 0$. The other parameters revert to their prior values.
- In principle such components are fitted slightly to the data points, but for broad priors this effect is too small to be seen numerically.
- For the variational Gaussian mixture model the expected values of the mixing coefficients in the posterior distribution are given by

$$\mathbb{E}[\pi_k] = \frac{\alpha_k + N_k}{K\alpha_0 + N}. \tag{14}$$

  Consider a component for which $N_k \simeq 0$ and $\alpha_k \simeq \alpha_0$. If the prior is broad so that $\alpha_0 \to 0$, then $\mathbb{E}[\pi_k] \to 0$ and the component plays no role in the model, whereas if the prior tightly constrains the mixing coefficients so that $\alpha_0 \to \infty$, then $\mathbb{E}[\pi_k] \to 1/K$.
- The figure was generated with $\alpha_0 = 10^{-3}$. If instead we choose $\alpha_0 = 1$ we obtain three components with nonzero mixing coefficients, and for $\alpha = 10$ all six components have nonzero mixing coefficients.

Benefits of Bayesian approach

- For anything other than very small data sets, the dominant computational cost of the variational algorithm for Gaussian mixtures arises from the evaluation of the responsibilities, together with the evaluation and inversion of the weighted data covariance matrices. These computations mirror precisely those that arise in the maximum likelihood EM algorithm, and so there is little computational overhead in using this Bayesian approach.
- The singularities that arise in maximum likelihood when a Gaussian component 'collapses' onto a specific data point are absent in the Bayesian treatment. Indeed, these singularities are removed if we simply introduce a prior and then use a MAP estimate instead of maximum likelihood.
- There is no over-fitting if we choose a large number K of components in the mixture.
- The variational treatment opens up the possibility of determining the optimal number of components in the mixture without resorting to techniques such as cross-validation.

Variational lower bound

- We can also evaluate the lower bound (2) for this model. In practice, it is useful to be able to monitor the bound during the re-estimation in order to test for convergence since we cannot compute the log likelihood of the model directly.
- For the variational mixture of Gaussians, the lower bound (2) is given by

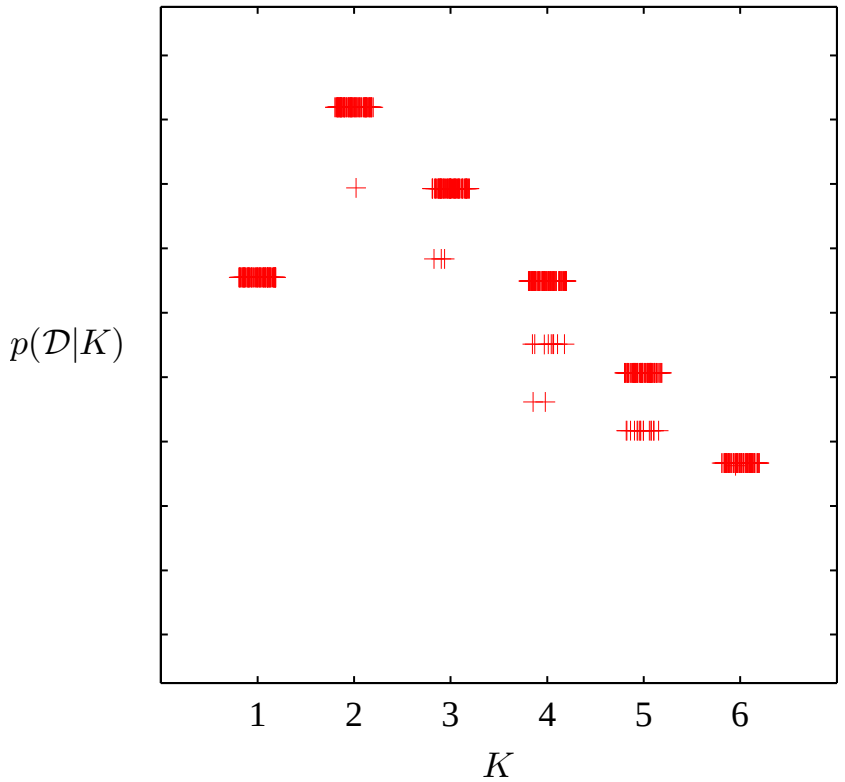
$$\begin{aligned}
\mathcal{L} &= \sum_{\mathbf{Z}} \iiint q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})}{q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})} \right\} \mathrm{d}\boldsymbol{\pi}\, \mathrm{d}\boldsymbol{\mu}\, \mathrm{d}\boldsymbol{\Lambda} \\
&= \mathbb{E}[\ln p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] - \mathbb{E}[\ln q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] \\
&= \mathbb{E}[\ln p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] + \mathbb{E}[\ln p(\mathbf{Z}|\boldsymbol{\pi})] + \mathbb{E}[\ln p(\boldsymbol{\pi})] + \mathbb{E}[\ln p(\boldsymbol{\mu}, \boldsymbol{\Lambda})] \\
&\quad - \mathbb{E}[\ln q(\mathbf{Z})] - \mathbb{E}[\ln q(\boldsymbol{\pi})] - \mathbb{E}[\ln q(\boldsymbol{\mu}, \boldsymbol{\Lambda})] \qquad (15)
\end{aligned}$$

Determining the number of components

- For any given setting of the parameters in a Gaussian mixture model (except for specific degenerate settings), there will exist other parameter settings for which the density over the observed variables will be identical by permuting the components.
- If we have a mixture model comprising K components, then each parameter setting will be a member of a family of $K!$ equivalent settings.
- In a Bayesian setting we marginalize over all possible parameter values.
- If, however, we wish to compare different values of K, then we need to take account of this multimodality. A simple approximate solution is to add a term $\ln K!$ onto the lower bound when used for model comparison and averaging.

Choosing the number of components

- Plot of the variational lower bound $\mathcal{L}$ versus the number K of components in the Gaussian mixture model, for the Old Faithful data.
- For each value of K, the model is trained from 100 different random starts; the results are plotted with small random horizontal perturbations so that they can be distinguished.
- Some solutions find suboptimal local maxima, but that this happens infrequently.



Optimising mixing coefficients

- An alternative approach to determining a suitable value for K is to treat the mixing coefficients $\boldsymbol{\pi}$ as parameters and make point estimates of their values by maximizing the lower bound with respect to $\boldsymbol{\pi}$ instead of maintaining a probability distribution over them.
- This leads to the re-estimation equation

$$\pi_k = \frac{1}{N} \sum_{n=1}^{N} r_{nk} \tag{16}$$

  and this maximization is interleaved with the variational updates for the q distribution over the remaining parameters.
- Components that provide insufficient contribution to explaining the data will have their mixing coefficients driven to zero during the optimization, and so they are effectively removed from the model.
- This allows us to make a single training run in which we start with a relatively large initial value of K, and allow surplus components to be pruned out of the model.

Conclusions

- The variational approach with factorised distribution is a powerful way of applying Bayesian inference flexibly and efficiently to many models.
- The mathematics is demanding but the practical benefits are substantial. There are some limitations: for example, there is no way to compute a model likelihood exactly.
- It is possible to extend the approach still further to (potentially) infinite mixtures: see Teh, Yee Whye, Michael I. Jordan, Matthew J. Beal, and David M. Blei. "Hierarchical dirichlet processes." Journal of the american statistical association 101, no. 476 (2006): 1566-1581.
- A related technique, Minimum Message Length, has been applied to selecting input variables for the GTM.