Variational Algorithms for Approximate Bayesian Inference: Linear Regression, Mixture of Gaussians

Prof. Nicholas Zabaras
Center for Informatics and Computational Science
https://cics.nd.edu/
University of Notre Dame
Notre Dame, IN, USA

Email: nzabaras@gmail.com
URL: https://www.zabaras.com/

March 30, 2018

Contents

Following:

- Pattern Recognition and Machine Learning, Christopher M. Bishop, Chapter 10
- Machine Learning: A Probabilistic Perspective, Kevin Murphy, Chapter 21.

Variational Linear Regression

Variational Linear Regression

Consider the Bayesian linear regression model.

Recall that the likelihood function is given by

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

where $\boldsymbol{\phi}_n = \boldsymbol{\phi}(\boldsymbol{x}_n)$.
We use the following conjugate prior distributions (i.e. Gaussian-Gamma):

$$p(\boldsymbol{w}|\alpha) = \mathcal{N}(\boldsymbol{w}|0, \alpha^{-1}\boldsymbol{I})$$
$$p(\alpha) = \text{Gam}(\alpha|a_0, b_0)$$

To simplify discussion the noise precision parameter β is fixed and assumed to be known. The framework can be extended with this assumption being relaxed.

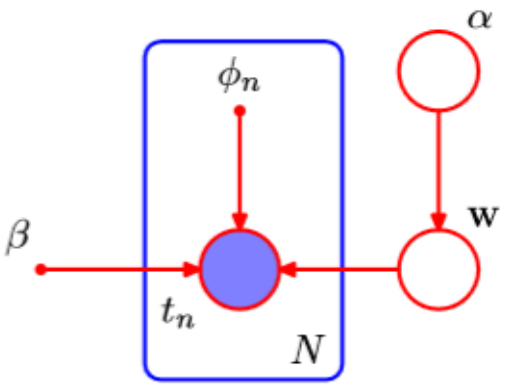
- Drugowitsch, J. (2008). Bayesian linear regression. Technical report, U. Rochester.

Variational Linear Regression

Collectively we have the joint distribution

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

which can be represented by the following graphical model



We assume a variational posterior of the following form to obtain the unknown hyperparameters

$$q(\boldsymbol{w}, \alpha) = q(\boldsymbol{w})q(\alpha)$$

Variational Linear Regression

Using the same framework as previously discussed, we obtain for $q(\alpha)$:

$$\ln q^*(\alpha) = \ln p(\alpha) + \mathbb{E}_{\boldsymbol{w}}[\ln p(\boldsymbol{w}|\alpha)] + const$$
$$= (a_0 - 1)\ln\alpha - b_0\alpha + \frac{M}{2}\ln\alpha - \frac{\alpha}{2}\mathbb{E}[\boldsymbol{w}^T\boldsymbol{w}] + const$$

As expected this gives a Gamma distribution

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

with

$$a_N = a_0 + \frac{M}{2}$$
$$b_N = b_0 + \frac{1}{2}\mathbb{E}[\boldsymbol{w}^T\boldsymbol{w}]$$

Variational Linear Regression

Similarly we obtain for $q(\boldsymbol{w})$:

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

$$= -\frac{\beta}{2}\sum_{n=1}^{N}\{\boldsymbol{w}^T\boldsymbol{\phi}_n - t_n\}^2 - \frac{1}{2}\mathbb{E}[\alpha]\boldsymbol{w}^T\boldsymbol{w} + const$$

$$= -\frac{1}{2}\boldsymbol{w}^T(\mathbb{E}[\alpha]\boldsymbol{I} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi})\boldsymbol{w} + \beta\boldsymbol{w}^T\boldsymbol{\Phi}^T\boldsymbol{t} + const$$

As expected this gives a normal distribution

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

with

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

$$\boldsymbol{S}_N = (\mathbb{E}[\alpha]\boldsymbol{I} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}$$

Recall that $\boldsymbol{\Phi}^T\boldsymbol{\Phi} = (\boldsymbol{\phi}_1 \quad .. \quad \boldsymbol{\phi}_N)\begin{pmatrix}\boldsymbol{\phi}_1^T \\ .. \\ \boldsymbol{\phi}_N^T\end{pmatrix} = \sum_{n=1}^{N}\boldsymbol{\phi}_n\boldsymbol{\phi}_n^T$

These are the same formulas obtained when α is treated as constant. In the VI approach, α is replaced by $\mathbb{E}[\alpha]$.

Variational Linear Regression

By standard properties for Gaussian and Gamma distributions we finally have

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

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

which collectively give

$$\mathbb{E}[\alpha] = \frac{a_N}{b_N} = \frac{a_0 + M/2}{b_0 + \mathbb{E}[\boldsymbol{w}^T\boldsymbol{w}]/2}$$

For the case of a_0=b_0=0 (infinitely broad prior over a): $\mathbb{E}[\alpha] = \frac{M}{\boldsymbol{m}_N^T\boldsymbol{m}_N + \mathrm{tr}(\boldsymbol{S}_N)}$. This is similar to the result obtained using the model evidence approximation.

It then remains a task to cycle between computing a_N, b_N, $\boldsymbol{m}_N$ and $\boldsymbol{S}_N$ until some convergence criterion is met.

These results are consistent with those obtained by maximizing the evidence using EM (except that the point estimate of α is now replaced by $\mathbb{E}[\alpha]$) and give identical results in the case of an infinitely broad prior for which $\mathbb{E}[\alpha] = \frac{M}{\boldsymbol{m}_N^T\boldsymbol{m}_N + \mathrm{tr}(\boldsymbol{S}_N)}$.

Predictive Distribution

The predictive distribution of t for a new input x is given as:

$$p(t|\boldsymbol{x},\boldsymbol{t}) = \int p(t|\boldsymbol{x},\boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{t})d\boldsymbol{w} \approx \int p(t|\boldsymbol{x},\boldsymbol{w})q(\boldsymbol{w})d\boldsymbol{w}$$

$$= \int \mathcal{N}(t|\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}),\beta^{-1})\mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_N,\boldsymbol{S}_N)\, d\boldsymbol{w} = \mathcal{N}(t|\boldsymbol{m}_N^T\boldsymbol{\phi}(\boldsymbol{x}),\sigma^2(\boldsymbol{x}))$$

Here we used an earlier result for linear Gaussian models. Given a marginal Gaussian distribution for **x** and a conditional Gaussian distribution for **y** given **x** in the form

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda}^{-1})$$
$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b},\mathbf{L}^{-1})$$

the marginal distribution of **y** is given by

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b},\mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathsf{T}})$$

Using the above result, the input dependent variance is given as:

$$\sigma^2(\boldsymbol{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{S}_N\,\boldsymbol{\phi}(\boldsymbol{x}), \text{ with } \boldsymbol{S}_N = (\mathbb{E}[\alpha]\boldsymbol{I} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}$$

This is the same form as that obtained when α is treated as constant but now α is replaced by $\mathbb{E}[\alpha]$.

Lower Bound

But how can we select M (degree of polynomial)? We can compute the lower bound

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

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

where

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

$$= -\frac{N}{2}\ln(2\pi) + \frac{N}{2}\ln(\beta) - \frac{\beta}{2}\{\boldsymbol{t}^T\boldsymbol{t} - 2\mathbb{E}[\boldsymbol{w}^T]\boldsymbol{\Phi}^T\boldsymbol{t} + Tr(\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T]\boldsymbol{\Phi}^T\boldsymbol{\Phi})\}$$

$$= \frac{N}{2}\ln\frac{\beta}{2\pi} - \frac{\beta}{2}\boldsymbol{t}^T\boldsymbol{t} + \beta\boldsymbol{m}_N^T\boldsymbol{\Phi}^T\boldsymbol{t} - \frac{\beta}{2}\mathrm{Tr}\,[\boldsymbol{\Phi}^T\boldsymbol{\Phi}(\boldsymbol{m}_N\boldsymbol{m}_N^T + \boldsymbol{S}_N)]$$

$$\mathbb{E}_{\boldsymbol{w},\alpha}[\ln p(\boldsymbol{w}|\alpha)] = -\frac{M}{2}\ln 2\pi + \frac{M}{2}\mathbb{E}[\ln\alpha] - \frac{\mathbb{E}[\alpha]}{2}\mathbb{E}[\boldsymbol{w}^T\boldsymbol{w}]$$

$$= -\frac{M}{2}\ln 2\pi + \frac{M}{2}(\psi(a_N) - \ln b_N) - \frac{a_N}{2b_N}[\boldsymbol{m}_N^T\boldsymbol{m}_N + \mathrm{Tr}\,(\boldsymbol{S}_N)]$$

Use expectation of the log of a Gamma distributed variable

$$\mathbb{E}_{\alpha}[\ln p(\alpha)] = a_0\ln b_0 + (a_0 - 1)\mathbb{E}[\ln\alpha] - b_0\,\mathbb{E}[\alpha] - \ln\Gamma(a_0)$$

$$= a_0\ln b_0 + (a_0 - 1)[\psi(a_N) - \ln b_N] - \frac{b_0 a_N}{b_N} - \ln\Gamma(a_0)$$

Lower Bound

The final two terms in $\mathcal{L}(q)$ represent the entropies of the Gaussian and Gamma distributions:

$$-\mathbb{E}_{\boldsymbol{w}}[\ln q(\boldsymbol{w})] \quad = \frac{1}{2}\ln|\boldsymbol{S}_N| + \frac{M}{2}[1 + \ln 2\pi]$$

$$-\mathbb{E}_{\alpha}[\ln q(\alpha)] = \ln \Gamma(a_N) - (a_N - 1)\psi(a_N) - \ln b_N + a_N$$

We substitute in the Eqs. above the following expressions for the moments:

$$\mathbb{E}[\boldsymbol{w}] = \boldsymbol{m}_N$$

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

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

$$\mathbb{E}[ln\alpha] = \psi(a_N) - \ln b_N$$

to obtain the final expression for $\mathcal{L}(q)$:

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

Lower Bound Vs the Order of the Polynomial

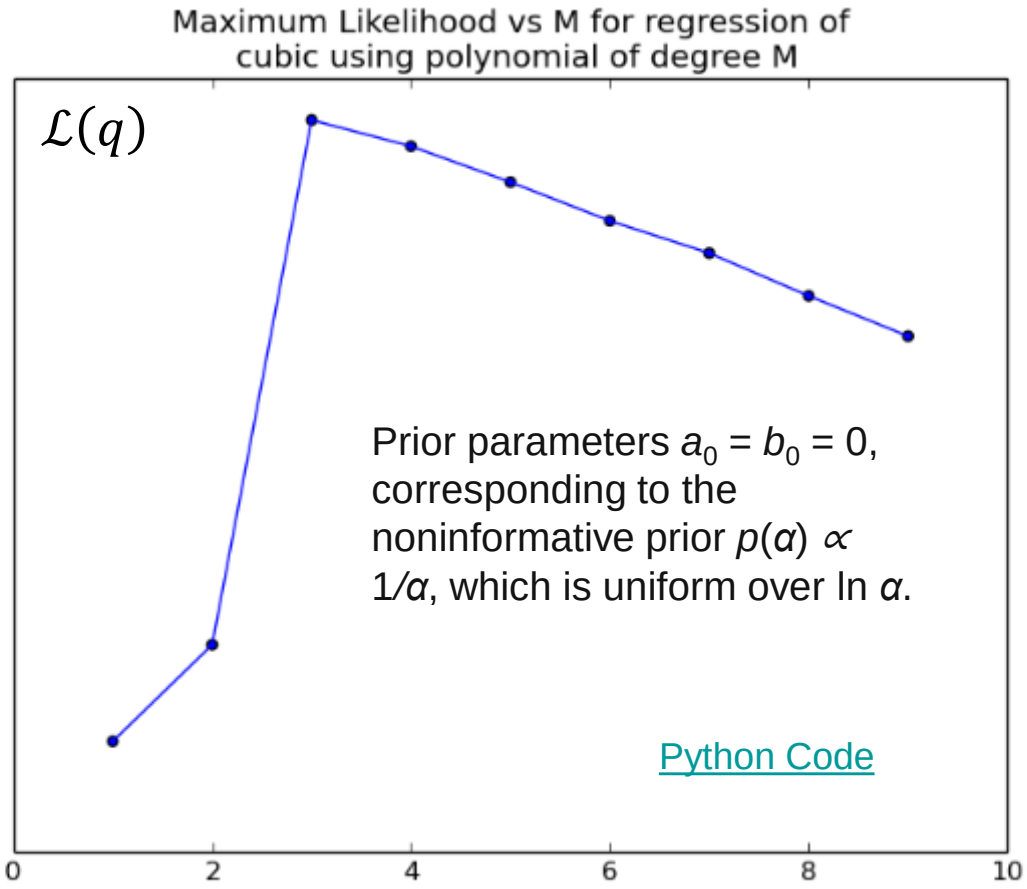
10 data points have been generated from a polynomial of degree 3 over [-5, 5] with additive Gaussian noise. The lower bound is maximized for M=3 corresponding to the true model form which the data was generated.

$\mathcal{L}$ represents lower bound on the log marginal likelihood ln$p$(**t**|M) for the model.

If we assign equal prior probabilities $p(M)$ to the different values of M, then we can interpret $\mathcal{L}$ as an approximation to $p(M|\mathbf{t})$.

Thus the variational framework assigns the highest probability to the model with $M = 3$.

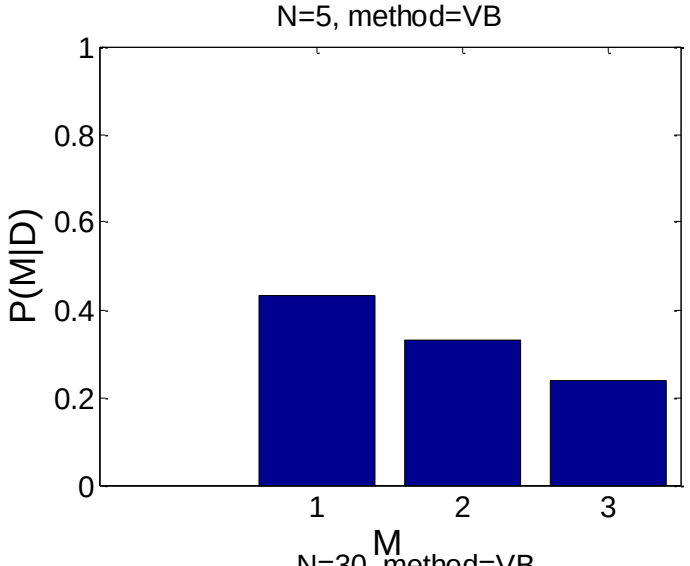
This should be contrasted with the MLE result, which assigns ever smaller residual error to models of increasing complexity until the residual error is driven to zero, causing MLE to over-fitted models.

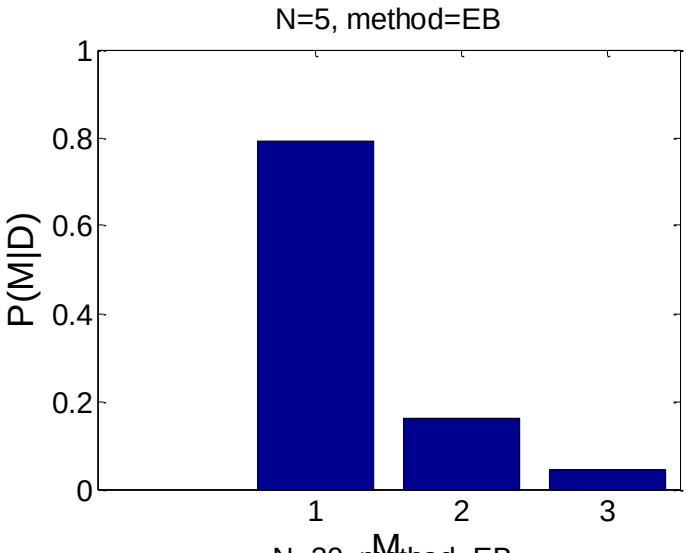


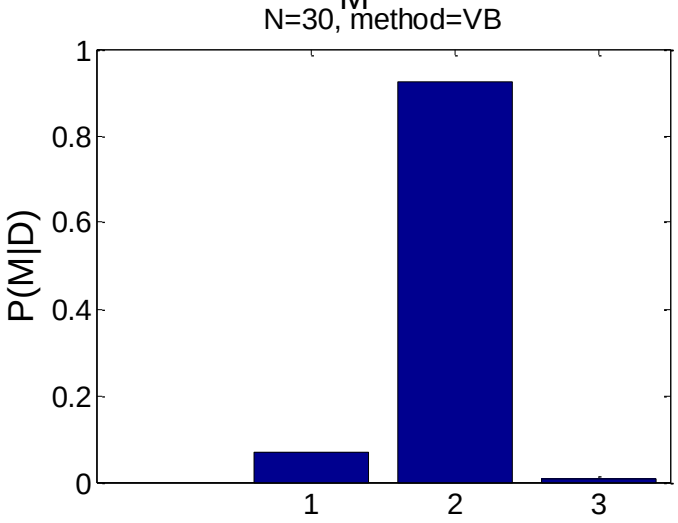
Prior parameters $a_0 = b_0 = 0$, corresponding to the noninformative prior $p(\alpha) \propto 1/\alpha$, which is uniform over ln α.

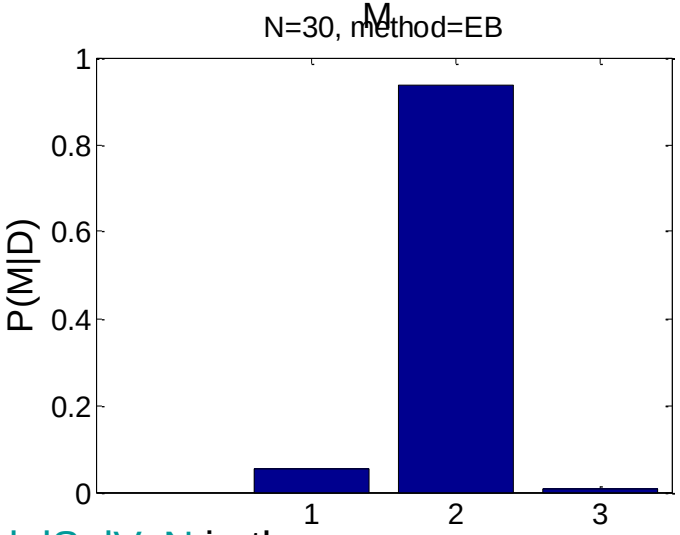
Python Code

Lower Bound Vs the Order of the Polynomial









Run linregEbModelSelVsN in the PMTK3 toolbox

Variational Linear Regression with Gam($\beta|c_0, d_0$)

We now extend the variational treatment of Bayesian linear regression to include a gamma hyperprior Gam($\beta|c_0, d_0$) over β and solve variationally, by assuming a factorized variational distribution of the form $q(\mathbf{w})q(\alpha)q(\beta)$.

We modify the joint distribution of all variables as:

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

The formulae for $p(\alpha)$ remain the same:

$$q^*(\alpha) = \mathrm{Gam}(\alpha|a_N, b_N)\,, a_N = a_0 + \frac{M}{2},\ b_N = b_0 + \frac{1}{2}\mathbb{E}[\boldsymbol{w}^T\boldsymbol{w}]$$

For $q^*(\boldsymbol{w})$ we have:

$$\ln q^*(\boldsymbol{w}) = \ln p(\boldsymbol{t}|\boldsymbol{w},\beta) + \mathbb{E}_\alpha[\ln p(\boldsymbol{w}|\alpha)] + const$$
$$= -\frac{\mathbb{E}[\beta]}{2}\sum_{n=1}^{N}\{\boldsymbol{w}^T\boldsymbol{\phi}_n - t_n\}^2 - \frac{1}{2}\mathbb{E}[\alpha]\boldsymbol{w}^T\boldsymbol{w} + const$$
$$= -\frac{1}{2}\boldsymbol{w}^T(\mathbb{E}[\alpha]\boldsymbol{I} + \mathbb{E}[\beta]\boldsymbol{\Phi}^T\boldsymbol{\Phi})\boldsymbol{w} + \mathbb{E}[\beta]\boldsymbol{w}^T\boldsymbol{\Phi}^T\boldsymbol{t} + const$$

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

$$\boldsymbol{m}_N = \mathbb{E}[\beta]\boldsymbol{S}_N\boldsymbol{\Phi}^T\boldsymbol{t}$$
$$\boldsymbol{S}_N = (\mathbb{E}[\alpha]\boldsymbol{I} + \mathbb{E}[\beta]\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}$$

Variational Linear Regression with $Gam(\beta|c_0, d_0)$

For $q^*(\beta)$

$$\ln q^*(\beta) = \mathbb{E}[\ln p(\boldsymbol{t}|\boldsymbol{w}, \beta)] + \ln p(\beta) + const$$

$$= \frac{N}{2} ln\beta - \frac{\beta}{2}\mathbb{E}_{\boldsymbol{w}}\left[\sum_{n=1}^{N}\{\boldsymbol{w}^T\boldsymbol{\phi}_n - t_n\}^2\right] + (c_0 - 1)ln\beta - d_0\beta + const$$

We recognize the log of a Gamma distribution with:

$$q^*(\beta) = \mathrm{Gam}(\beta|c_N, d_N), c_N = c_0 + \frac{N}{2},$$

$$d_N = d_0 + \frac{1}{2}\mathbb{E}[\textstyle\sum_{n=1}^{N}\{\boldsymbol{w}^T\boldsymbol{\phi}_n - t_n\}^2] = d_0 + \frac{1}{2}(Tr(\boldsymbol{\Phi}^T\boldsymbol{\Phi})\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T] + \boldsymbol{t}^T\boldsymbol{t}) - \boldsymbol{t}^T\,\boldsymbol{\Phi}\mathbb{E}[\boldsymbol{w}]$$

$$= d_0 + \frac{1}{2}\left(\|\boldsymbol{t} - \boldsymbol{\Phi}\boldsymbol{m}_N\|^2 + Tr(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{S}_N)\right)$$

Where we used:

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

$\mathbb{E}[\boldsymbol{w}]=\boldsymbol{m}_N$, with $\boldsymbol{m}_N = \mathbb{E}[\beta]\boldsymbol{S}_N\boldsymbol{\Phi}^T\boldsymbol{t}, \boldsymbol{S}_N = (\mathbb{E}[\alpha]\boldsymbol{I} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}$

Thus $\mathbb{E}[\beta] = \frac{c_N}{d_N}$

Variational Linear Regression with $Gam(\beta|c_0, d_0)$

The lower bound also needs to be modified. Starting with the modified log-likelihood and using $\mathbb{E}[\boldsymbol{w}]=\boldsymbol{m}_N$, $\mathbb{E}[\beta] = \frac{c_N}{d_N}, \mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T] = \boldsymbol{m}_N\boldsymbol{m}_N^T + \boldsymbol{S}_N$ and $\mathbb{E}[ln\beta] = \psi(c_N) - lnd_N$:

$$\mathbb{E}_\beta\left[\mathbb{E}_{\boldsymbol{w}}[\ln p(\boldsymbol{t}|\boldsymbol{w})]\right] = \frac{N}{2}(\mathbb{E}[\beta] - \ln(2\pi)) - \frac{\mathbb{E}[\beta]}{2} - \mathbb{E}[\|\boldsymbol{t} - \boldsymbol{\Phi w}\|^2]$$
$$= \frac{N}{2}(\psi(c_N) - lnd_N - \ln(2\pi)) - \frac{c_N}{2d_N}\left(\|\boldsymbol{t} - \boldsymbol{\Phi w}\|^2 + Tr(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{S}_N)\right)$$

Next using $\mathbb{E}[\beta] = \frac{c_N}{d_N}$ and $\mathbb{E}[ln\beta] = \psi(c_N) - lnd_N$, we consider the term corresponding to log prior over β:

$$\mathbb{E}[lnp(\beta)] = (c_0 - 1)\mathbb{E}[ln\beta] - d_0\mathbb{E}[\beta] + c_0\ ln\ d_0 - ln\Gamma(c_0)$$
$$= (c_0 - 1)(\psi(c_N) - lnd_N) - \frac{d_0 c_N}{d_N} + c_0\ ln\ d_0 - ln\Gamma(c_0)$$

Finally we compute the negative entropy of the posterior over β:

$$-\mathbb{E}[lnq^*(\beta)] = (c_N - 1)\psi(c_N) + lnd_N - c_N - ln\Gamma(c_N)$$

Finally the predictive distribution is given as:

$$p(t|\boldsymbol{x}, \boldsymbol{t}) \approx \mathcal{N}(t|\boldsymbol{m}_N^T\boldsymbol{\phi}(\boldsymbol{x}), \sigma^2(\boldsymbol{x})),\ \sigma^2(\boldsymbol{x}) = \frac{1}{\mathbb{E}[\beta]} + \boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{S}_N\ \boldsymbol{\phi}(\boldsymbol{x})$$

Variational Inference for a Mixture of Gaussians

Variational Mixture of Gaussians

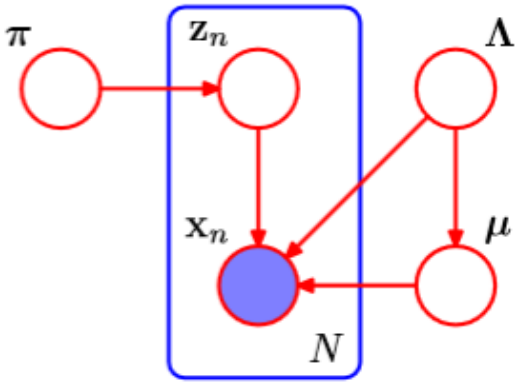
Suppose we have a set of observations $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and let the latent variables be $\boldsymbol{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$. Each $\mathbf{z}_n$ is a 1-of-K binary vector with elements z_{nk}, k=1,…,K.

The number of latent variables increases with the size of the data set whereas the number of all other parameters in the model below remain constant.

The starting point is the likelihood function for the Gaussian mixture model (see graphical model below).

In order to formulate a variational treatment of this model, we write down the joint distribution of all random variables

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



- Attias, H. (1999b). Inferring parameters and structure of latent variable models by variational Bayes. In K. B. Laskey and H. Prade (Eds.), *Uncertainty in Artificial Intelligence: Proceedings of the Fifth Conference*, pp. 21–30. Morgan Kaufmann.

Variational Mixture of Gaussians

In order to formulate a variational treatment of this model, it is first convenient to write down the joint distribution of all random variables

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

$p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ is the likelihood for observations **X**, given the model parameters:

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

$p(\boldsymbol{Z}|\boldsymbol{\pi})$ is the conditional distribution of **Z**, given the mixing coefficients $\boldsymbol{\pi}$:

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

$p(\boldsymbol{\pi})$ is the prior distribution for the mixing coefficients $\boldsymbol{\pi}$:

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

This is a conjugate prior (α_0 here is interpreted as the effective prior number of observations associated with each component).

Variational Mixture of Gaussians

$p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = p(\boldsymbol{\mu}|\boldsymbol{\Lambda})p(\boldsymbol{\Lambda})$ is prior distribution for the parameters governing the mean and precision of each Gaussian component.

Using a Gaussian-Wishart prior for each component governing the mean and precision (conjugate prior when both the mean and precision are unknown):

$$p(\boldsymbol{\mu}|\boldsymbol{\Lambda})p(\boldsymbol{\Lambda}) = \prod_{k=1}^{K} \mathcal{N}(\mu_k|\boldsymbol{m}_0, (\beta_0\boldsymbol{\Lambda}_k)^{-1})\, \mathcal{W}(\boldsymbol{\Lambda}_k|\boldsymbol{W}_0, \nu_0)$$

We do not know α_0, $\boldsymbol{m}_0$, $\mathbf{W}_0$, β_0, ν_0 parameters, although a sensible choice would be as follows:

$\boldsymbol{m}_0 = 0$
$\boldsymbol{W}_0 = \boldsymbol{I}$
$\nu_0 = D + 1$ (where D is the dimensionality of the data)
$0 < \alpha_0, \beta_0 \ll 1$

This choice of initial parameters has been chosen with the mind set of placing more importance on the observed data rather than the prior beliefs.

Variational Mixture of Gaussians

We now consider a variational distribution that factorizes between the latent variables and model parameters:

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

This is the only assumption we make in order to solve the Bayesian mixture model

Computing q*(Z)

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

These factors will be determined by optimization of the variational distribution.

We start with $\mathbf{q}(\boldsymbol{Z})$

From our general results, the log of the distribution minimizing the KL divergence of the factorized distribution above for latent variables $\boldsymbol{Z}$ is given by:

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

This can be decomposed:

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

By substituting $p(\boldsymbol{Z}|\boldsymbol{\pi}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\pi_k^{z_{nk}}$ and $p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})^{z_{nk}}$, we obtain:

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

Where we define: $\ln \rho_{nk} = \mathbb{E}[\ln \pi_k] + \frac{1}{2}\mathbb{E}[\ln |\boldsymbol{\Lambda}_k|] - \frac{D}{2}\ln(2\pi) - \frac{1}{2}\mathbb{E}_{\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k}[(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Lambda}_k (\boldsymbol{x}_n - \boldsymbol{\mu}_k)]$

Computing q*(Z)

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

Taking the exponential and requiring that the resulting distribution is normalized, we obtain

$$q^*(\boldsymbol{Z}) \propto \prod_{n=1}^{N}\prod_{k=1}^{K} \rho_{nk}^{z_{nk}}$$

$$q^*(\boldsymbol{Z}) = \prod_{n=1}^{N}\prod_{k=1}^{K} r_{nk}^{z_{nk}}$$

where

$$r_{nk} = \frac{\rho_{nk}}{\sum_{j=1}^{K} \rho_{nj}}$$

Note that the form of $q^*(\boldsymbol{Z})$ is the same as the form of the prior $p(\boldsymbol{Z}|\boldsymbol{\pi}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \pi_k^{z_{nk}}$.

Also note that for the discrete distribution, we have the standard result $\mathbb{E}[z_{nk}] = r_{nk}$ from which it can be see that r_{nk} are playing the roles of responsibilities.

Computing $q^*(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$

Since $q^*(\boldsymbol{Z})$ depends on the distribution of other variables we have coupled update equations and therefore as before we have to solve the update equations iteratively.

Before proceeding to consider $q^*(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ it is instructive to define the following quantities:

effective number of observations $$N_k = \sum_{n=1}^{N} r_{nk}$$

effective sample mean $$\bar{\boldsymbol{x}}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} \boldsymbol{x}_n$$

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

These are analogous to what is used in the EM algorithm for Gaussian mixtures.

A useful result from these definitions (expand the l.h.s. in the definition of $\boldsymbol{S}_k$) follows:

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

Computing $q^*(\boldsymbol{\pi})$

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

We now consider $q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$

The log of the distribution minimizing the KL divergence of the factorized distribution above for parameters $\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}$ is given by:

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

The variational posterior clearly factorizes into the following form

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

Using, $p(\boldsymbol{\pi}) = C(\alpha_0) \prod_{k=1}^{K} \pi_k^{\alpha_0 - 1}$ and $p(\boldsymbol{Z}|\boldsymbol{\pi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}}$, and using $\mathbb{E}[z_{nk}] = r_{nk}$, keeping the terms with $\boldsymbol{\pi}$, for $q(\boldsymbol{\pi})$ we have

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

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

Computing $q^*(\boldsymbol{\mu}, \boldsymbol{\Lambda})$

Let us now keep the terms that depend on $\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k$:

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

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

Using $p(\boldsymbol{\mu}|\boldsymbol{\Lambda})p(\boldsymbol{\Lambda}) = \prod_{k=1}^{K} \mathcal{N}(\mu_k|\boldsymbol{m}_0, (\beta_0\boldsymbol{\Lambda}_k)^{-1})\, \mathcal{W}(\boldsymbol{\Lambda}_k|\boldsymbol{W}_0, \nu_0)$, we derive:

$$\ln q^*(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) = \ln \mathcal{N}(\boldsymbol{\mu}_k|\boldsymbol{m}_0, (\beta_0\boldsymbol{\Lambda}_k)^{-1}) + \ln \mathcal{W}(\boldsymbol{\Lambda}_k|\boldsymbol{W}_0, \nu_0) + \sum_{n=1}^{N} \mathbb{E}_Z[z_{nk}] \ln \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}) + const$$

$$= -\frac{\beta_0}{2}(\boldsymbol{\mu}_k - \boldsymbol{m}_0)^T\boldsymbol{\Lambda}_k(\boldsymbol{\mu}_k - \boldsymbol{m}_0) + \frac{1}{2}\ln|\boldsymbol{\Lambda}_k| - \frac{1}{2}\mathrm{tr}(\boldsymbol{\Lambda}_k\boldsymbol{W}_0^{-1}) + \frac{\nu_0 - D - 1}{2}\ln|\boldsymbol{\Lambda}_k|$$
$$-\frac{1}{2}\sum_{n=1}^{N} \mathbb{E}_Z[z_{nk}](\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T\boldsymbol{\Lambda}_k(\boldsymbol{x}_n - \boldsymbol{\mu}_k) + \frac{1}{2}\sum_{n=1}^{N} \mathbb{E}_Z[z_{nk}]\ \ln|\boldsymbol{\Lambda}_k| + const$$

Using the product rule of probability, we express $\ln q^*(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) = \ln q^*(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k) + \ln q^*(\boldsymbol{\Lambda}_k)$. We look first for terms that depend on $\boldsymbol{\mu}_k$.

$$\ln q^*(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k) = -\frac{1}{2}\boldsymbol{\mu}_k^T(\beta_0 + \sum_{n=1}^{N} \mathbb{E}_Z[z_{nk}])\boldsymbol{\Lambda}_k\boldsymbol{\mu}_k + \boldsymbol{\mu}_k^T\boldsymbol{\Lambda}_k(\beta_0\boldsymbol{m}_0 + \sum_{n=1}^{N} \mathbb{E}_Z[z_{nk}]\boldsymbol{x}_n) + const =$$
$$-\frac{1}{2}\boldsymbol{\mu}_k^T(\beta_0 + N_k)\boldsymbol{\Lambda}_k\boldsymbol{\mu}_k + \boldsymbol{\mu}_k^T\boldsymbol{\Lambda}_k(\beta_0\boldsymbol{m}_0 + N_k\bar{\boldsymbol{x}}_k) + const$$

$$q^*(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k) = \mathcal{N}(\boldsymbol{\mu}_k|\boldsymbol{m}_k, (\beta_k\boldsymbol{\Lambda}_k)^{-1}),\ where\ \beta_k = \beta_0 + N_k, \boldsymbol{m}_k = \frac{1}{\beta_k}(\beta_0\boldsymbol{m}_0 + N_k\bar{\boldsymbol{x}}_k)$$

Computing $q^*(\boldsymbol{\Lambda})$

To compute $\ln q^*(\boldsymbol{\Lambda}_k)$, we use $\ln q^*(\boldsymbol{\Lambda}_k) = \ln q^*(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) - \ln q^*(\boldsymbol{\mu}_k | \boldsymbol{\Lambda}_k)$. *We use*: $q^*(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k) = \mathcal{N}(\boldsymbol{\mu}_k|\boldsymbol{m}_k, (\beta_k\boldsymbol{\Lambda}_k)^{-1})$ to derive:

$$\ln q^*(\boldsymbol{\Lambda}_k) = -\frac{\beta_0}{2}(\boldsymbol{\mu}_k - \boldsymbol{m}_0)^T\boldsymbol{\Lambda}_k(\boldsymbol{\mu}_k - \boldsymbol{m}_0) + \frac{1}{2}\ln|\boldsymbol{\Lambda}_k| - \frac{1}{2}\mathrm{tr}(\boldsymbol{\Lambda}_k\boldsymbol{W}_0^{-1}) + \frac{\nu_0 - D - 1}{2}\ln|\boldsymbol{\Lambda}_k|$$
$$-\frac{1}{2}\sum_{n=1}^{N}\mathbb{E}_Z[z_{nk}](\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T\boldsymbol{\Lambda}_k(\boldsymbol{x}_n - \boldsymbol{\mu}_k) + \frac{1}{2}\sum_{n=1}^{N}\mathbb{E}_Z[z_{nk}]\ln|\boldsymbol{\Lambda}_k| + const$$
$$+\frac{\beta_k}{2}(\boldsymbol{\mu}_k - \boldsymbol{m}_k)^T\boldsymbol{\Lambda}_k(\boldsymbol{\mu}_k - \boldsymbol{m}_k) - \frac{1}{2}\ln|\boldsymbol{\Lambda}_k| + const$$

We keep terms that depend only on $\boldsymbol{\Lambda}_k$.

$$\ln q^*(\boldsymbol{\Lambda}_k) = -\frac{\beta_0}{2}(\boldsymbol{\mu}_k - \boldsymbol{m}_0)^T\boldsymbol{\Lambda}_k(\boldsymbol{\mu}_k - \boldsymbol{m}_0) - \frac{1}{2}\mathrm{tr}(\boldsymbol{\Lambda}_k\boldsymbol{W}_0^{-1}) + \frac{\nu_0 - D - 1}{2}\ln|\boldsymbol{\Lambda}_k| -$$
$$\frac{1}{2}\sum_{n=1}^{N}\mathbb{E}_Z[z_{nk}](\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T\boldsymbol{\Lambda}_k(\boldsymbol{x}_n - \boldsymbol{\mu}_k) + \frac{1}{2}\sum_{n=1}^{N}\mathbb{E}_Z[z_{nk}]\ln|\boldsymbol{\Lambda}_k| + \frac{\beta_k}{2}(\boldsymbol{\mu}_k - \boldsymbol{m}_k)^T\boldsymbol{\Lambda}_k(\boldsymbol{\mu}_k - \boldsymbol{m}_k) + const$$

$$= \frac{\nu_0 + \sum_{n=1}^{N}\mathbb{E}_Z[z_{nk}] - D - 1}{2}\ln|\boldsymbol{\Lambda}_k| - \frac{1}{2}\mathrm{tr}(\boldsymbol{\Lambda}_k\boldsymbol{W}_0^{-1} + \boldsymbol{\Lambda}_k\beta_0(\boldsymbol{\mu}_k - \boldsymbol{m}_0)(\boldsymbol{\mu}_k - \boldsymbol{m}_0)^T +$$
$$+ \boldsymbol{\Lambda}_k\sum_{n=1}^{N}\mathbb{E}_Z[z_{nk}](\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T - \boldsymbol{\Lambda}_k\beta_k(\boldsymbol{\mu}_k - \boldsymbol{m}_k)(\boldsymbol{\mu}_k - \boldsymbol{m}_k)^T)$$

$$= \frac{\nu_k - D - 1}{2}\ln|\boldsymbol{\Lambda}_k| - \frac{1}{2}\mathrm{tr}(\boldsymbol{\Lambda}_k\boldsymbol{W}_k^{-1}) + const$$

where $\boldsymbol{W}_k^{-1} = \boldsymbol{W}_0^{-1} + \beta_0\boldsymbol{m}_0\boldsymbol{m}_0^T + \sum_{n=1}^{N} r_{nk}\,\boldsymbol{x}_n\boldsymbol{x}_n^T - \beta_k\boldsymbol{m}_k\boldsymbol{m}_k^T$ and $\nu_k = \nu_0 + N_k$.

Computing $q^*(\boldsymbol{\Lambda})$

Thus: $\mathrm{q}^*(\boldsymbol{\Lambda}_k) = \mathcal{W}(\boldsymbol{\Lambda}_k|\boldsymbol{W}_k, \nu_k)$.

Here we defined

$$\nu_k = \nu_0 + N_k$$

$$\boldsymbol{W}_k^{-1} = \boldsymbol{W}_0^{-1} + \beta_0 \boldsymbol{m}_0 \boldsymbol{m}_0^T + \sum_{n=1}^{N} r_{nk}\, \boldsymbol{x}_n \boldsymbol{x}_n^T - \beta_k \boldsymbol{m}_k \boldsymbol{m}_k^T$$

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

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

We used $\boldsymbol{m}_k = \frac{1}{\beta_k}(\beta_0 \boldsymbol{m}_0 + N_k \bar{\boldsymbol{x}}_k)$ and $N_k \boldsymbol{S}_k = \sum_{n=1}^{N} r_{nk} \boldsymbol{x}_n \boldsymbol{x}_n^T - N_k \bar{\boldsymbol{x}}_k \bar{\boldsymbol{x}}_k^T, \beta_k = \beta_0 + N_k$.

Computing $q^*(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$

Finally for $\mathrm{q}^*(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$ we have shown that it takes a Gaussian-Wishart distribution of the form:

$$\mathrm{q}^*(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) = \mathcal{N}(\boldsymbol{\mu}_k | \boldsymbol{m}_k, (\beta_k \boldsymbol{\Lambda}_k)^{-1}) \mathcal{W}(\boldsymbol{\Lambda}_k | \boldsymbol{W}_k, \nu_k)$$

where we have defined:

$$\beta_k = \beta_0 + N_k$$

$$\boldsymbol{m}_k = \frac{1}{\beta_k}(\beta_0 \boldsymbol{m}_0 + N_k \bar{\boldsymbol{x}}_k)$$

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

$$\nu_k = \nu_0 + N_k$$

These equations are analogous to the M-step in the EM algorithm (involve the same sums over the data set).

Computing the responsibilities

Let us return terms in the expression for the responsibilities. Its terms need to be evaluated:

$$\ln \rho_{nk} = \mathbb{E}[\ln \pi_k] + \frac{1}{2}\mathbb{E}[\ln |\boldsymbol{\Lambda}_k|] - \frac{D}{2}\ln \pi - \frac{1}{2}\mathbb{E}_{\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k}[(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T \Lambda_k (\boldsymbol{x}_n - \boldsymbol{\mu}_k)]$$

Using our results $q^*(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) = \mathcal{N}(\boldsymbol{\mu}_k|\boldsymbol{m}_k, (\beta_k\boldsymbol{\Lambda}_k)^{-1})\mathcal{W}(\boldsymbol{\Lambda}_k|\boldsymbol{W}_k, \nu_k)$, we can write:

$$\mathbb{E}_{\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k}\left[(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Lambda}_k (\boldsymbol{x}_n - \boldsymbol{\mu}_k)\right] = \int\int \left(Tr\left[\boldsymbol{\Lambda}_k (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T\right] q^*(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k) d\boldsymbol{\mu}_k\right) q^*(\boldsymbol{\Lambda}_k) d\boldsymbol{\Lambda}_k$$

From the moments of $q^*(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k) = \mathcal{N}(\boldsymbol{\mu}_k|\boldsymbol{m}_k, (\beta_k\boldsymbol{\Lambda}_k)^{-1})$, we can compute:

$$\mathbb{E}[\boldsymbol{\mu}_k] = \boldsymbol{m}_k, \quad \mathbb{E}[\boldsymbol{\mu}_k\boldsymbol{\mu}_k^T] = \boldsymbol{m}_k\boldsymbol{m}_k^T + \beta_k^{-1}\boldsymbol{\Lambda}_k^{-1}$$

Thus:

$$\mathbb{E}_{\boldsymbol{\mu}_k}\left[(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T\right] = \boldsymbol{x}_n\boldsymbol{x}_n^T - \boldsymbol{x}_n\boldsymbol{m}_k^T - \boldsymbol{m}_k\boldsymbol{x}_n^T + \boldsymbol{m}_k\boldsymbol{m}_k^T + \beta_k^{-1}\boldsymbol{\Lambda}_k^{-1} = (\boldsymbol{x}_n - \boldsymbol{m}_k)(\boldsymbol{x}_n - \boldsymbol{m}_k)^T + \beta_k^{-1}\boldsymbol{\Lambda}_k^{-1}$$

Using the results for the mean of the Wishart:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k}\left[(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Lambda}_k (\boldsymbol{x}_n - \boldsymbol{\mu}_k)\right] &= \int Tr\left\{\boldsymbol{\Lambda}_k\left[(\boldsymbol{x}_n - \boldsymbol{m}_k)(\boldsymbol{x}_n - \boldsymbol{m}_k)^T + \beta_k^{-1}\boldsymbol{\Lambda}_k^{-1}\right]\right\}\mathcal{W}(\boldsymbol{\Lambda}_k|\boldsymbol{W}_k, \nu_k) d\boldsymbol{\Lambda}_k \\
&= \int\left(Tr\left\{\boldsymbol{\Lambda}_k\left[(\boldsymbol{x}_n - \boldsymbol{m}_k)(\boldsymbol{x}_n - \boldsymbol{m}_k)^T\right]\right\} + \beta_k^{-1}D\right)\mathcal{W}(\boldsymbol{\Lambda}_k|\boldsymbol{W}_k, \nu_k) d\boldsymbol{\Lambda}_k \\
&= \nu_k(\boldsymbol{x}_n - \boldsymbol{m}_k)^T \boldsymbol{W}_k (\boldsymbol{x}_n - \boldsymbol{m}_k) + \beta_k^{-1}D
\end{aligned}$$

Computing the Responsibilities

We now look at the term $\mathbb{E}[\ln \pi_k]$:

$$\ln \rho_{nk} = \mathbb{E}[\ln \pi_k] + \frac{1}{2}\mathbb{E}[\ln |\boldsymbol{\Lambda}_k|] - \frac{D}{2}\ln \pi - \frac{1}{2}\mathbb{E}_{\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k}[(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T \Lambda_k (\boldsymbol{x}_n - \boldsymbol{\mu}_k)]$$

$$\ln \tilde{\pi}_k \equiv \mathbb{E}[\ln \pi_k] = \psi(\alpha_k) - \psi(\hat{\alpha}), \ \hat{a} = \sum_k \alpha_k$$

Recall that $q^*(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha})$ where $\alpha_k = \alpha_0 + N_k$ and a useful expression for the mean of the log. Here $\psi(\alpha_k)$ is the digamma function defined as:

$$\psi(\alpha) = \frac{d}{d\alpha}\ln\Gamma(\alpha)$$

The proof for this can also be found in an earlier lecture on the Dirichlet distribution.

Computing the Responsibilities

$$\ln \rho_{nk} = \mathbb{E}[\ln \pi_k] + \frac{1}{2}\mathbb{E}[\ln |\Lambda_k|] - \frac{D}{2}\ln \pi - \frac{1}{2}\mathbb{E}_{\mu_k,\Lambda_k}[(x_n - \mu_k)^T \Lambda_k (x_n - \mu_k)]$$

Finally we need to compute the term $\mathbb{E}[\ln |\Lambda_k|]$: $\mathbb{E}_{\Lambda_k}\left[\ln|\Lambda_k|\right] = \int \ln|\Lambda_k| \mathcal{W}(\Lambda_k | W_k, \nu_k) d\Lambda_k$

We use the following interesting result from the reference below: $\frac{|\Lambda_k|}{|W_k|} \sim \sum_{i=1}^{D} \chi^2_{\nu_k+1-i}$

Using this we obtain:

$$\mathbb{E}\left[\ln|\Lambda_k|\right] = \sum_{i=1}^{D}\mathbb{E}\left[\ln \chi^2_{\nu_k+1-i}\right] + \ln|W_k| = \sum_{i=1}^{D}\left\{\ln 2 + \psi\left(\frac{\nu_k+1-i}{2}\right)\right\} + \ln|W_k| = \sum_{i=1}^{D}\psi\left(\frac{\nu_k+1-i}{2}\right) + D\ln 2 + \ln|W_k|$$

Here we used that $\mathbb{E}\left[\ln \chi^2_{\nu_k}\right] = \ln 2 + \psi\left(\frac{\nu_k}{2}\right)$. This can be derived directly by recalling that

$$\chi^2_\nu \equiv Gamma\left(\frac{\nu}{2}, 2\right) \text{ and } \mathbb{E}[\ln X] = \psi\left(\frac{\nu}{2}\right) + \ln 2 \text{ where } X \sim Gamma\left(\frac{\nu}{2}, 2\right)$$

Note here in the relation of χ^2_ν and $Gamma\left(\frac{\nu}{2}, 2\right) = \frac{1}{\Gamma\left(\frac{\nu}{2}\right)2^{\frac{\nu}{2}}} x^{\frac{\nu}{2}-1} e^{-\frac{x}{2}}$, the parametrization of Gamma is on shape and scale parameters.

- Goodman, N. R. The Distribution of the Determinant of a Complex Wishart Distributed Matrix. Ann. Math. Statist. 34 (1963), no. 1, 178--180.
- http://en.wikipedia.org/wiki/Exponential_family#Moment_generating_function_of_the_sufficient_statistic

Computing the Responsibilities

Finally the various terms in our expression for the (unnormalized) responsibilities:

$$\ln \rho_{nk} = \mathbb{E}[\ln \pi_k] + \frac{1}{2}\mathbb{E}[\ln |\boldsymbol{\Lambda}_k|] - \frac{D}{2}\ln \pi - \frac{1}{2}\mathbb{E}_{\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k}[(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T \Lambda_k (\boldsymbol{x}_n - \boldsymbol{\mu}_k)]$$

are as follows:

$$\ln \tilde{\pi}_k \equiv \mathbb{E}[\ln \pi_k] = \psi(\alpha_k) - \psi(\hat{\alpha}),\ \hat{a}_k = \sum_k \alpha_k$$

$$ln\tilde{\boldsymbol{\Lambda}}_k \equiv \mathbb{E}[\ln |\boldsymbol{\Lambda}_k|] = \sum_{i=1}^{D}\left(\psi \frac{\nu_k + 1 - i}{2}\right) + D \ln 2 + \ln|\boldsymbol{W}_k|$$

$$\mathbb{E}_{\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k}[(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Lambda}_k (\boldsymbol{x}_n - \boldsymbol{\mu}_k)] = D\beta_k^{-1} + \nu_k (\boldsymbol{x}_n - \boldsymbol{m}_k)^T \boldsymbol{W}_k (\boldsymbol{x}_n - \boldsymbol{m}_k)$$

Here $\psi(\alpha_k)$ is the digamma function defined as:

$$\psi(\alpha) = \frac{d}{d\alpha} ln\Gamma(\alpha)$$

The following result is finally obtained for the normalized responsibilities:

$$r_{nk} \propto \tilde{\pi}_k \tilde{\boldsymbol{\Lambda}}_k^{\frac{1}{2}} exp\left(-\frac{D}{2\beta_k} - \frac{\nu_k}{2}(\boldsymbol{x}_n - \boldsymbol{m}_k)^T \boldsymbol{W}_k (\boldsymbol{x}_n - \boldsymbol{m}_k)\right)$$

This is similar to the corresponding result in maximum likelihood EM.

$$r_{nk} \propto \pi_k |\boldsymbol{\Lambda}_k|^{1/2} exp\left(-\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{m}_k)^T \boldsymbol{\Lambda}_k (\boldsymbol{x}_n - \boldsymbol{m}_k)\right)$$

Role of the Prior $q^*(\boldsymbol{\pi}) = \mathcal{D}ir(\boldsymbol{\pi}|\boldsymbol{\alpha})$

Note that the variational posterior has the same functional form as the joint distribution:

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

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

Using $q^*(\boldsymbol{\pi}) = \mathcal{D}ir(\boldsymbol{\pi}|\boldsymbol{\alpha})$ where $\alpha_k = \alpha_0 + N_k$, the expected values of the mixing coefficients are given below:

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

Components that take essentially no responsibility for explaining the data points have $r_{nk} \approx 0$ and hence $N_k \approx 0$. From $\alpha_k = \alpha_0 + N_k$, we see that $\alpha_k \approx \alpha_0$ and the other parameters (β_k, $\mathbf{m}_k$, $\mathbf{W}_k$, ν_k) revert to their prior values.

Consider a component for which $N_k \approx 0$ and $\alpha_k \approx \alpha_0$.

- If the prior is broad so that $\alpha_0 \rightarrow 0$, then $\mathbb{E}[\pi_k] \rightarrow 0$ and the component plays no role in the model.
- If the prior tightly constrains the mixing coefficients, $\alpha_0 \rightarrow \infty$, then $\mathbb{E}[\pi_k] \rightarrow 1/K$.

Summary of the Algorithm

1) Initialize parameters $\boldsymbol{m}_0, \boldsymbol{W}_0, \nu_0, \alpha_0, \beta_0$

2) E-step: Update the responsibilities r_{nk} needed in the approximate posterior $q^*(\boldsymbol{Z}) = \prod_{n=1}^{N} \prod_{k=1}^{K} r_{nk}^{z_{nk}}$

3) M-step: Update the parameters N_k, $\bar{\boldsymbol{x}}_{\boldsymbol{k}}$, $\boldsymbol{S}_k$, α_k, β_k, $\boldsymbol{m}_k$, $\boldsymbol{W}_k^{-1}$, ν_k needed in $q^*(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$

4) Cycle between the E and M steps until convergence (Next slide)

- Svensén, M. and C. M. Bishop (2004). Robust Bayesian mixture modelling. *Neurocomputing* **64**, 235–252.
- Corduneanu, A. and C. M. Bishop (2001). Variational Bayesian model selection for mixture distributions. In T. Richardson and T. Jaakkola (Eds.), *Proceedings Eighth International Conference on Artificial Intelligence and Statistics*, pp. 27–34. Morgan Kaufmann.

Summary of the Algorithm

E-Step

Compute the following:

$$r_{nk} = \frac{\rho_{nk}}{\sum_{j=1}^{K} \rho_{nk}}$$

$$\ln \tilde{\pi}_k \equiv \mathbb{E}[\ln \pi_k] = \psi(\alpha_k) - \psi(\hat{\alpha}) \text{ , } \hat{a}_k = \sum_k \alpha_k$$

$$ln\tilde{\Lambda}_k$$

$$\equiv \mathbb{E}[\ln |\boldsymbol{\Lambda}_k|] = \sum_{i=1}^{D} \left(\psi \frac{\nu_k + 1 - i}{2} \right) + D \ln 2 + \ln |\boldsymbol{W}_k|$$

$$\mathbb{E}_{\mu_k, \Lambda_k}[(x_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Lambda}_k (\boldsymbol{x}_n - \boldsymbol{\mu}_k)]$$
$$= D\beta_k^{-1} + \nu_k (\boldsymbol{x}_n - \boldsymbol{m}_k)^T \boldsymbol{W}_k (\boldsymbol{x}_n - \boldsymbol{m}_k)$$

$$r_{nk}$$
$$\propto \tilde{\pi}_k \tilde{\boldsymbol{\Lambda}}_k^{\frac{1}{2}} exp\left(-\frac{D}{2\beta_k} - \frac{\nu_k}{2} (\boldsymbol{x}_n - \boldsymbol{m}_k)^T \boldsymbol{W}_k (\boldsymbol{x}_n - \boldsymbol{m}_k) \right)$$

M-Step

Compute the following:

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

$$\bar{\boldsymbol{x}}_{\boldsymbol{k}} = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} \boldsymbol{x}_n$$

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

$$\alpha_k = \alpha_0 + N_k$$

$$\beta_k = \beta_0 + N_k$$

$$\boldsymbol{m}_k = \frac{1}{\beta_k} (\beta_0 \boldsymbol{m}_0 + N_k \bar{\boldsymbol{x}}_k)$$

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

$$\nu_k = \nu_0 + N_k$$

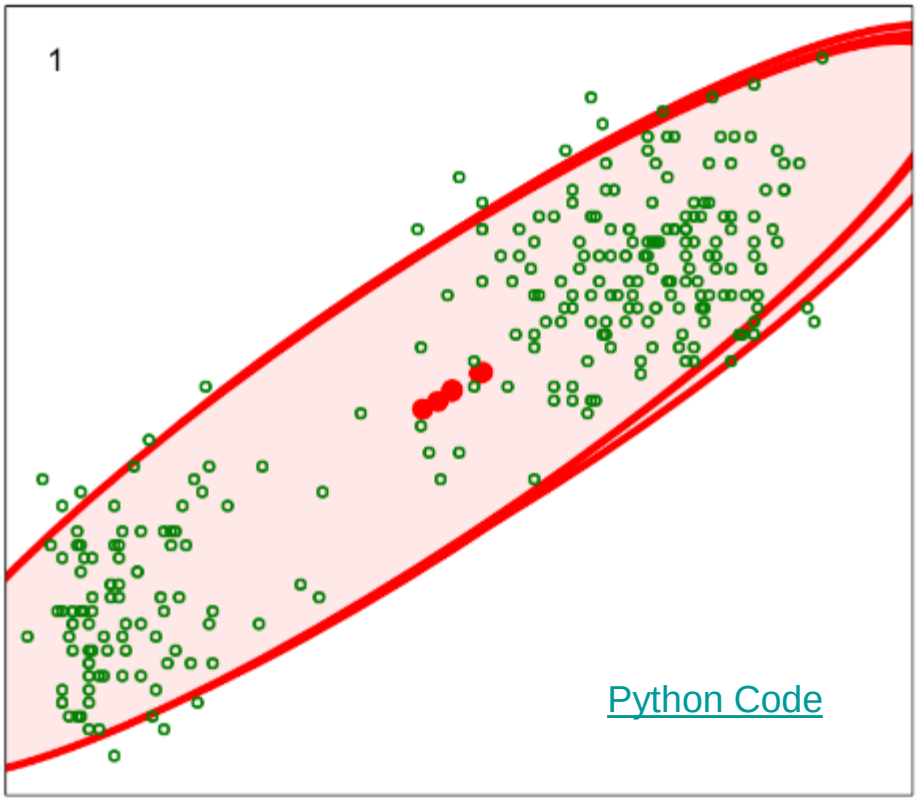
Variational Mixture of Gaussians: Example

Here, $\alpha_0 = 10^{-3}$ (favors sparsity)

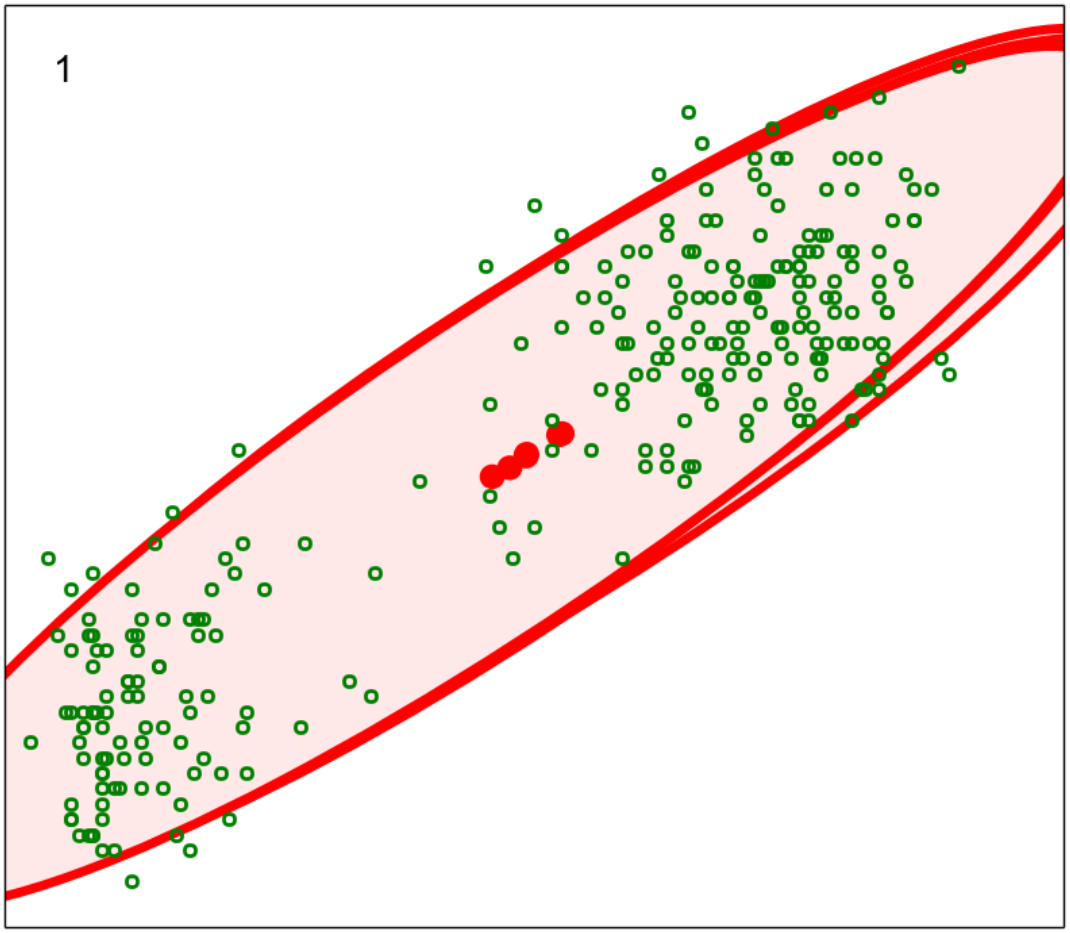
Variational mixture of $K = 6$ Gaussians (Old Faithful data set).

The ellipses denote the one standard-deviation density contours for each of the components, and the density of red ink inside each ellipse corresponds to the mean value of the mixing coefficient for each component.

The number in the top left of each diagram shows the number of iterations of VI.

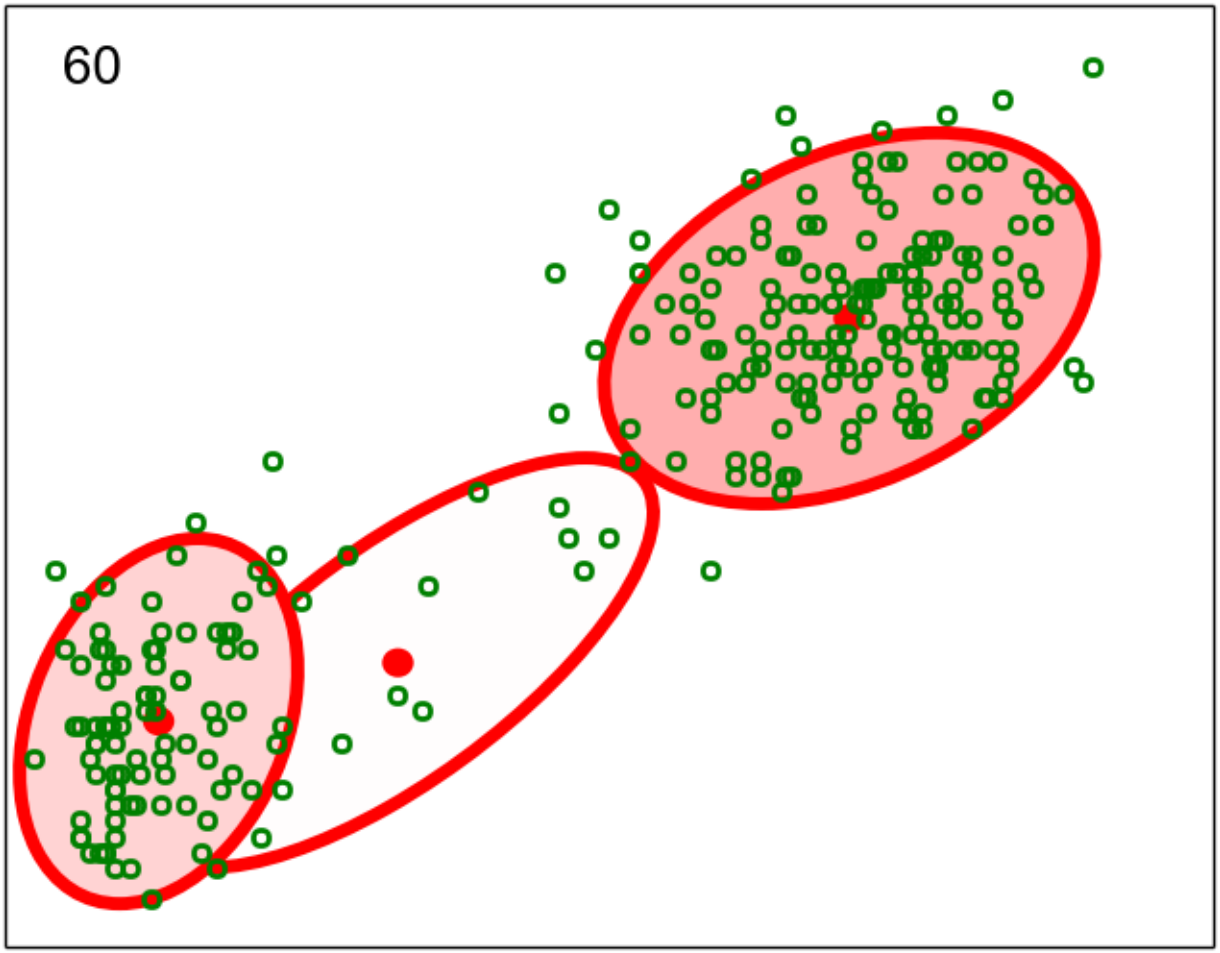


Variational Mixture of Gaussians: Example

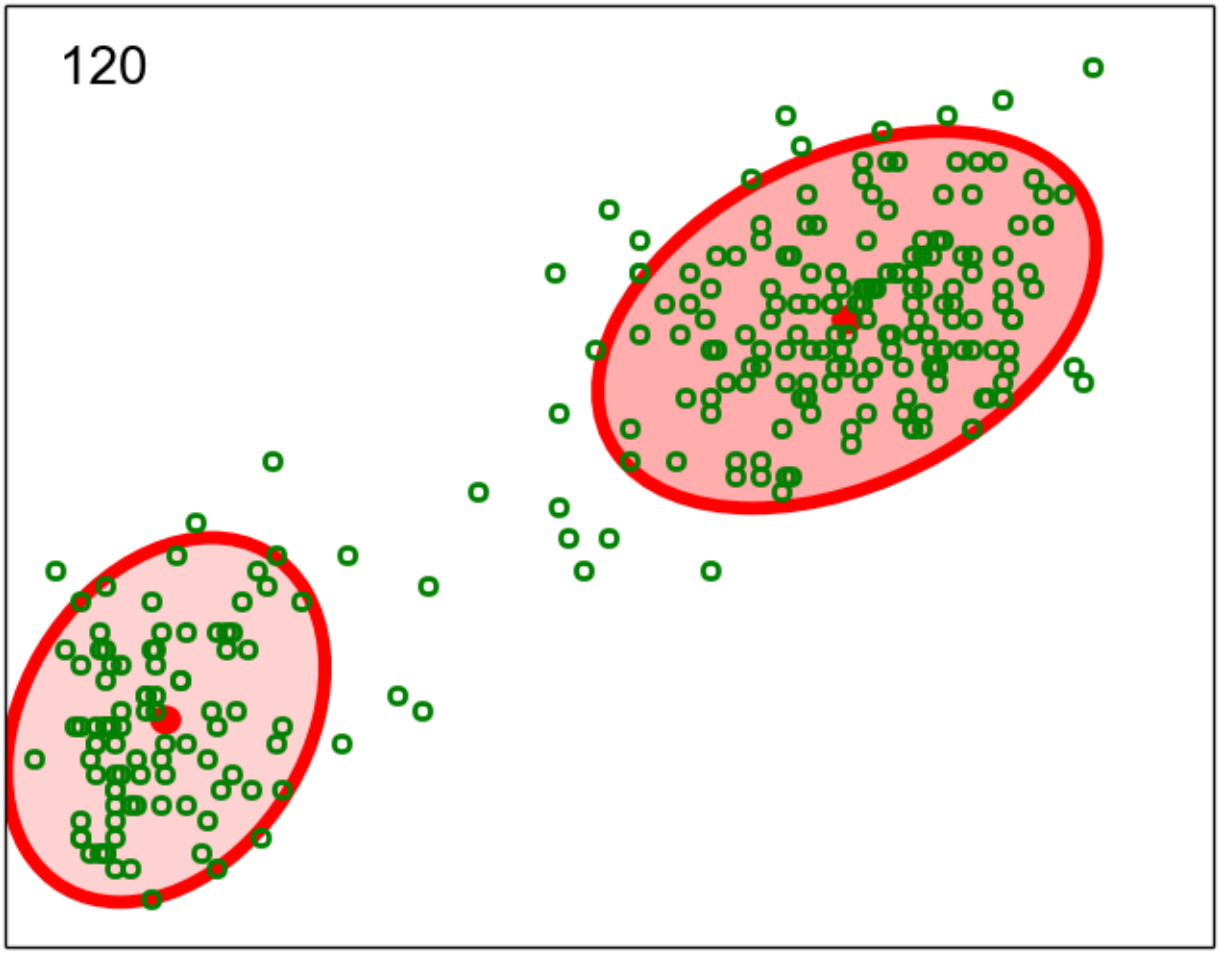


Variational Mixture of Gaussians: Example

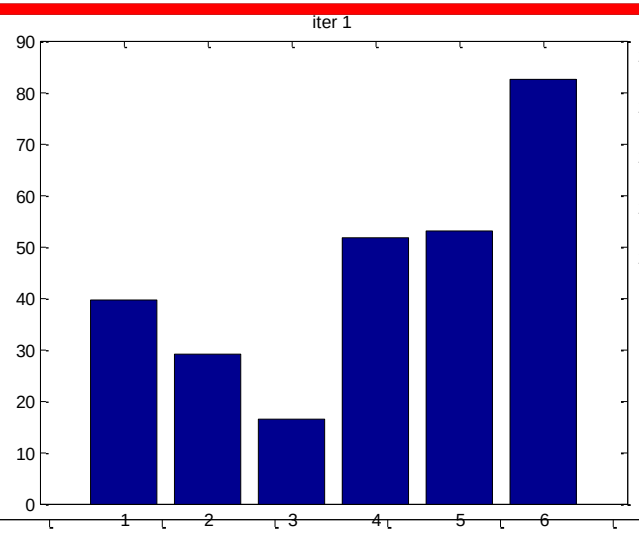
Variational Mixture of Gaussians: Example

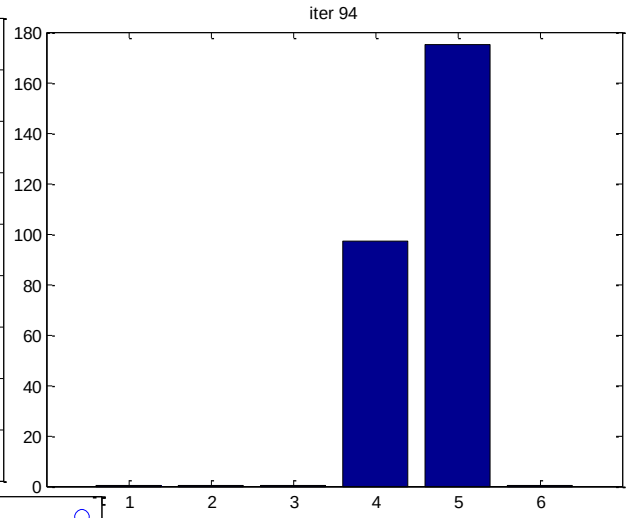


Variational Mixture of Gaussians: Example



Variational Mixture of Gaussians: Example



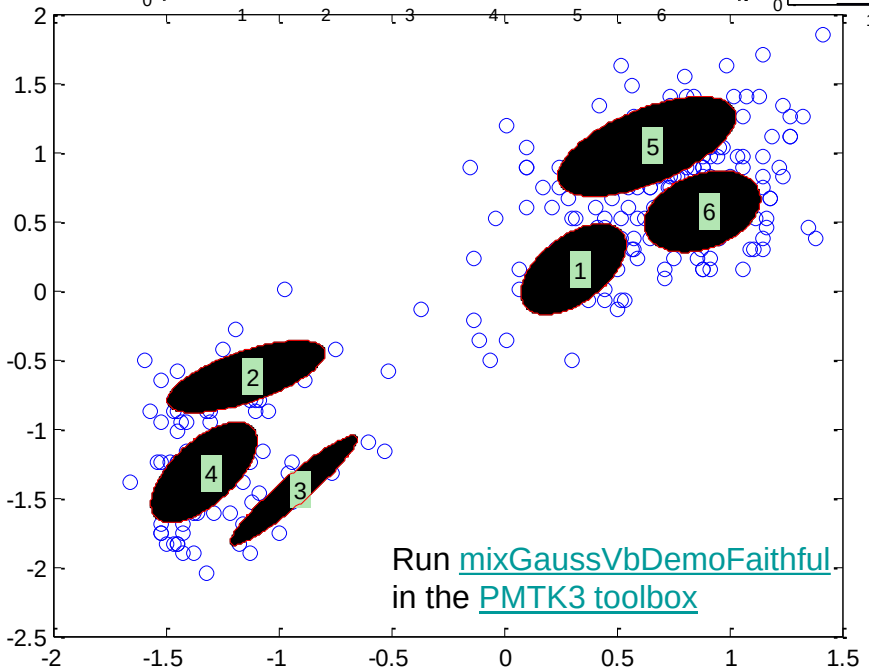


We initialize using 6 clusters with K-means.

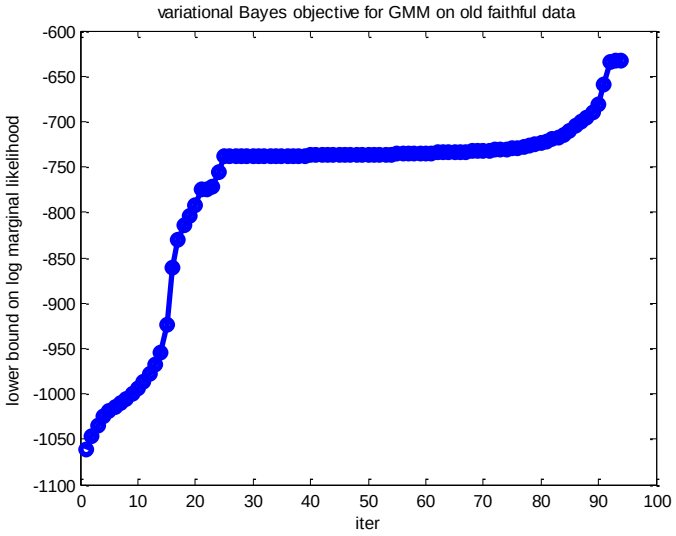
We use $\alpha_0 = 0.001$ to promote sparsity.

The posterior values of α_k are shown.

Unnecessary components are killed off.



Run mixGaussVbDemoFaithful in the PMTK3 toolbox



Automatic Pruning

To promote sparsity, can use a single model (K large) with $\alpha_0 \ll 1$.

Using $q^*(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha})$ where $\alpha_k = \alpha_0 + N_k$, we write $p(\boldsymbol{\pi}|\boldsymbol{\alpha}) = exp\left(\sum_k \alpha_k ln\pi_k - A(\boldsymbol{\alpha})\right), A(\boldsymbol{\alpha}) = \sum_k ln\Gamma(\alpha_k) - ln\Gamma\left(\sum_k \alpha_k\right)$ and using the definition of the digamma function,

$$\mathbb{E}[ln\pi_k] = \frac{\partial A(\boldsymbol{\alpha})}{\partial \alpha_k} = \psi(\alpha_k) - \psi\left(\sum_{k\prime} \alpha_{k\prime}\right)$$

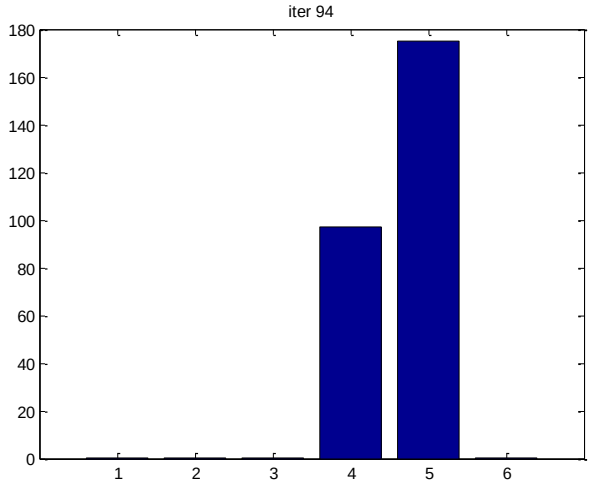
In VBEM, we use $\tilde{\pi}_k \equiv \frac{exp(\psi(\alpha_k))}{exp\left(\psi\left(\sum_{k\prime} \alpha_{k\prime}\right)\right)}$, $\alpha_k = \alpha_0 + N_k$. This is better than using the mode $\hat{\pi}_k \propto \alpha_k - 1$ that can be negative for $\alpha_0 = 0$ and $N_k = 0$.

Using the approximation $exp\left(\psi(x)\right) \approx x - 0.5, x > 1$, we simplify as:

$$\tilde{\pi}_k \propto \alpha_k - 0.5$$

i.,e. we remove 0.5 from each posterior count. So clusters with few weighted members become more empty with iterations while popular clusters get more members (the rich get richer).

This approach is more efficient than performing a discrete search over the number of clusters and comparing the marginal likelihood.



- Figueiredo, M. A. T. and A. K. Jain (2002). Unsupervised learning of finite mixture models. *IEEE Trans. On Pattern Analysis and Machine Intelligence 24*(3), 381–396. Matlab code at http://www.lx.it.pt/ mtf/mixturecode.zip.
- Liang, F., S. Mukherjee, and M. West (2007). Understanding the use of unlabelled data in predictive modelling. *Statistical Science 22*, 189–205.

Bayesian Treatment Vs. MLE

- For $N \rightarrow \infty$, the Bayesian treatment converges to the MLE EM algorithm.
- There is little computational overhead in using the Bayesian approach as compared to the traditional MLE approach.
- However, there are substantial advantages.
  - The singularities that arise in MLE when a Gaussian component 'collapses' onto a specific data point are absent in the Bayesian treatment. These singularities are removed if we simply introduce a prior and then use a MAP estimate instead of MLE.
  - There is no over-fitting for large number K of components in the mixture.
  - The variational treatment allows determining the optimal number of components in the mixture without cross validation.

Variational Lower Bound

It is useful to monitor the bound during the re-estimation in order to test for convergence.

At each step of the iterative re-estimation procedure the value of this bound should not decrease.

Use this to verify the derivation of the update equations and their implementation.

For the variational mixture of Gaussians, the lower bound is given by

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

- Svensén, M. and C. M. Bishop (2004). Robust Bayesian mixture modelling. *Neurocomputing* **64**, 235–252.

Variational Lower Bound

The various terms are given below (for proof see following Appendix):

$$\mathbb{E}[\ln p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})] =$$

$$\frac{1}{2}\sum_{k=1}^{K} N_k\left\{\ln\tilde{\boldsymbol{\Lambda}}_k - D\beta_k^{-1} - \nu_k Tr(\boldsymbol{S}_k\boldsymbol{W}_k) - \nu_k(\bar{\boldsymbol{x}}_k - \boldsymbol{m}_k)\boldsymbol{W}_k(\bar{\boldsymbol{x}}_k - \boldsymbol{m}_k)^T - D\ln(2\pi)\right\}$$

$$\mathbb{E}[\ln p(\boldsymbol{Z}|\boldsymbol{\pi})] = \sum_{\boldsymbol{n=1}}^{\boldsymbol{N}}\sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} r_{nk}\ln\tilde{\pi}_k$$

$$\mathbb{E}[\ln p(\boldsymbol{\pi})] = \ln C(\alpha_0) + (\alpha_0 - 1)\sum_{k=1}^{K}\ln\tilde{\pi}_k$$

$$\mathbb{E}[\ln p(\boldsymbol{\mu},\boldsymbol{\Lambda})] = \frac{1}{2}\sum_{k=1}^{K}\left\{D\ln(\beta_0/2\,\pi) + \ln\tilde{\boldsymbol{\Lambda}}_k - D\frac{\beta_0}{\beta_k} - \beta_0\nu_k(\boldsymbol{m}_k - \boldsymbol{m}_0)^T\boldsymbol{W}_k(\boldsymbol{m}_k - \boldsymbol{m}_0)\right\}$$

$$+ K\ln B(\boldsymbol{W}_0,\nu_0)$$

$$+\frac{\nu_0 - D - 1}{2}\sum_{k=1}^{K}\ln\tilde{\boldsymbol{\Lambda}}_k - \frac{1}{2}\sum_{k=1}^{K}\nu_k Tr\,(\boldsymbol{W}_0^{-1}\boldsymbol{W}_k)$$

Variational Lower Bound

The remaining terms are:

$$\mathbb{E}[\ln q(\boldsymbol{Z})] = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\ln r_{nk}$$

$$\mathbb{E}[\ln q(\boldsymbol{\pi})] = \sum_{k=1}^{K}(\alpha_k - 1)\ln\tilde{\pi}_k + \ln C(\alpha)$$

$$\mathbb{E}[\ln q(\boldsymbol{\mu}, \boldsymbol{\Lambda})] = \sum_{k=1}^{K}\left\{\frac{1}{2}\ln\tilde{\boldsymbol{\Lambda}}_k + \frac{D}{2}\ln\frac{\beta_k}{2\pi} - \frac{D}{2} - H[q(\boldsymbol{\Lambda}_k)]\right\}$$

D is the dimensionality of **x**, H[$q(\boldsymbol{\Lambda}_k)$] is the entropy of the Wishart, and $C(\boldsymbol{\alpha})$ and $B(\mathbf{W}, v)$ are normalization factors for the Dirichlet and Wishart distributions.

Note that the terms involving expectations of the logs of the *q* distributions simply represent the negative entropies of those distributions.

Appendix: Variational Lower Bound

Using $p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\Lambda}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k^{-1})^{z_{nk}}$ we derive:

$$\mathbb{E}\left[\ln p(\boldsymbol{X}\mid\boldsymbol{Z},\boldsymbol{\mu},\Lambda)\right]=\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\underbrace{\mathbb{E}\left[z_{nk}\right]}_{r_{nk}}\left\{\mathbb{E}\left[\ln|\Lambda_k|\right]-\mathbb{E}\left[\left(\boldsymbol{x}_n-\boldsymbol{\mu}_k\right)^T\Lambda_k\left(\boldsymbol{x}_n-\boldsymbol{\mu}_k\right)\right]-D\ln(2\pi)\right\}$$

We now use $ln\widetilde{\boldsymbol{\Lambda}}_k \equiv \mathbb{E}[\ln|\boldsymbol{\Lambda}_k|] = \sum_{i=1}^{D}\psi\left(\frac{\nu_k+1-i}{2}\right)+D\ln 2+\ln|\boldsymbol{W}_k|$

$$\mathbb{E}_{\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k}[(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^T\boldsymbol{\Lambda}_k(\boldsymbol{x}_n-\boldsymbol{\mu}_k)] = D\beta_k^{-1}+\nu_k(\boldsymbol{x}_n-\boldsymbol{m}_k)^T\boldsymbol{W}_k(\boldsymbol{x}_n-\boldsymbol{m}_k)$$

Substitution gives:

$$\mathbb{E}[\ln p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})] = \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}r_{nk}\{\ln\widetilde{\boldsymbol{\Lambda}}_k - D\beta_k^{-1}-\nu_k(\boldsymbol{x}_n-\boldsymbol{m}_k)^T\boldsymbol{W}_k(\boldsymbol{x}_n-\boldsymbol{m}_k)-D\ln(2\pi)\}$$

Make use of the following: $N_k=\sum_{n=1}^{N}r_{nk}$, $\bar{\boldsymbol{x}}_k=\frac{1}{N_k}\sum_{n=1}^{N}r_{nk}\boldsymbol{x}_n$

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

$$\mathbb{E}[\ln p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})] = \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}r_{nk}\left\{\ln\widetilde{\boldsymbol{\Lambda}}_k - D\beta_k^{-1}-\nu_k Tr\left(\boldsymbol{W}_k\left(\boldsymbol{x}_n\boldsymbol{x}_n^T-2\boldsymbol{x}_n\boldsymbol{m}_k^T+\boldsymbol{m}_k\boldsymbol{m}_k^T\right)\right)-D\ln(2\pi)\right\}=$$

$$\frac{1}{2}\sum_{k=1}^{K}N_k\left\{\ln\widetilde{\boldsymbol{\Lambda}}_k - D\beta_k^{-1}-\nu_k Tr\left(\boldsymbol{W}_k\left(\boldsymbol{S}_k+\bar{\boldsymbol{x}}_k\bar{\boldsymbol{x}}_k^T-2\bar{\boldsymbol{x}}_k\boldsymbol{m}_k^T+\boldsymbol{m}_k\boldsymbol{m}_k^T\right)\right)-D\ln(2\pi)\right\}=$$

$$\frac{1}{2}\sum_{k=1}^{K}N_k\{\ln\widetilde{\boldsymbol{\Lambda}}_k - D\beta_k^{-1}-\nu_k Tr(\boldsymbol{W}_k\boldsymbol{S}_k)-\nu_k(\bar{\boldsymbol{x}}_k-\boldsymbol{m}_k)^T\boldsymbol{W}_k(\bar{\boldsymbol{x}}_k-\boldsymbol{m}_k)-D\ln(2\pi)\}$$

Appendix: Variational Lower Bound

Using $p(\boldsymbol{Z}|\boldsymbol{\pi}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\pi_k^{z_{nk}}$ we derive:

$$\mathbb{E}[\ln p(\boldsymbol{Z}|\boldsymbol{\pi})] = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\ln\tilde{\pi}_k$$

where we used: $\ln\tilde{\pi}_k \equiv \mathbb{E}[\ln\pi_k] = \psi(\alpha_k) - \psi(\hat{\alpha})$, $\hat{a}_k = \sum_k \alpha_k$

Using $p(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\alpha_0) = C(\alpha_0)\prod_{k=1}^{K}\pi_k^{\alpha_0-1}$ and the equation above we derive:

$$\mathbb{E}[\ln p(\boldsymbol{\pi})] = \ln C(\alpha_0) + (\alpha_0 - 1)\sum_{k=1}^{K}\ln\tilde{\pi}_k$$

Starting from $p(\boldsymbol{\mu}|\boldsymbol{\Lambda})p(\boldsymbol{\Lambda}) = \prod_{k=1}^{K}\mathcal{N}(\mu_k|\boldsymbol{m}_0,(\beta_0\boldsymbol{\Lambda}_k)^{-1})\,\mathcal{W}(\boldsymbol{\Lambda}_k|\boldsymbol{W}_0,\nu_0)$, and denoting with $B(\boldsymbol{W}_0,\nu_0)$ the normalization of the Wishart, we can write:

$$\begin{aligned}\mathbb{E}\left[\ln p(\boldsymbol{\mu},\Lambda)\right] = \frac{1}{2}\sum_{k=1}^{K}\Big\{ & D\ln\left(\beta_0/2\pi\right) + \mathbb{E}\left[\ln|\Lambda_k|\right] \\ & -\beta_0\mathbb{E}\left[\left(\boldsymbol{\mu}_k-\boldsymbol{m}_0\right)^T\Lambda_k\left(\boldsymbol{\mu}_k-\boldsymbol{m}_0\right)\right]\Big\} + K\ln B(\boldsymbol{W}_0,\nu_0) \\ & +\sum_{k=1}^{K}\left\{\frac{\nu_0-D-1}{2}\mathbb{E}\left[\ln|\Lambda_k|\right] - \frac{1}{2}Tr(\boldsymbol{W}_0^{-1}\mathbb{E}\left[\Lambda_k\right])\right\}\end{aligned}$$

Appendix: Variational Lower Bound

We compute $\mathbb{E}\left[(\boldsymbol{\mu}_k - \boldsymbol{m}_0)^T \boldsymbol{\Lambda}_k (\boldsymbol{\mu}_k - \boldsymbol{m}_0)\right]$ as follows:

$$\mathbb{E}_{\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k}\left[(\boldsymbol{\mu}_k - \boldsymbol{m}_0)^T \boldsymbol{\Lambda}_k (\boldsymbol{\mu}_k - \boldsymbol{m}_0)\right] = \int\int \left(Tr\left[\boldsymbol{\Lambda}_k (\boldsymbol{\mu}_k - \boldsymbol{m}_0)(\boldsymbol{\mu}_k - \boldsymbol{m}_0)^T\right] q^*(\boldsymbol{\mu}_k \mid \boldsymbol{\Lambda}_k) d\boldsymbol{\mu}_k\right) q^*(\boldsymbol{\Lambda}_k) d\boldsymbol{\Lambda}_k$$

Using $\mathbb{E}[\boldsymbol{\mu}_k] = \boldsymbol{m}_k$, $\mathbb{E}\left[\boldsymbol{\mu}_k \boldsymbol{\mu}_k^T\right] = \boldsymbol{m}_k \boldsymbol{m}_k^T + \beta_k^{-1}\boldsymbol{\Lambda}_k^{-1}$, as well as the mean of the Wishart, we simplify:

$$\mathbb{E}_{\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k}\left[(\boldsymbol{\mu}_k - \boldsymbol{m}_0)^T \boldsymbol{\Lambda}_k (\boldsymbol{\mu}_k - \boldsymbol{m}_0)\right] = \int Tr\left[\boldsymbol{\Lambda}_k \left(\boldsymbol{m}_k \boldsymbol{m}_k^T + \beta_k^{-1}\boldsymbol{\Lambda}_k^{-1} - 2\boldsymbol{m}_k \boldsymbol{m}_0^T + \boldsymbol{m}_0 \boldsymbol{m}_0^T\right)\right] q^*(\boldsymbol{\Lambda}_k) d\boldsymbol{\Lambda}_k =$$

$$D\beta_k^{-1} + \mathbb{E}_{\boldsymbol{\Lambda}_k}\left[(\boldsymbol{m}_k - \boldsymbol{m}_0)^T \boldsymbol{\Lambda}_k (\boldsymbol{m}_k - \boldsymbol{m}_0)\right] = D\beta_k^{-1} + (\boldsymbol{m}_k - \boldsymbol{m}_0)^T \mathbb{E}[\boldsymbol{\Lambda}_k](\boldsymbol{m}_k - \boldsymbol{m}_0)$$

$$= D\beta_k^{-1} + \nu_k (\boldsymbol{m}_k - \boldsymbol{m}_0)^T \boldsymbol{W}_k (\boldsymbol{m}_k - \boldsymbol{m}_0)$$

Finally using $ln\tilde{\boldsymbol{\Lambda}}_k \equiv \mathbb{E}[\ln|\boldsymbol{\Lambda}_k|]$ our expression from the last slide n for $\mathbb{E}[\ln p(\boldsymbol{\mu}, \boldsymbol{\Lambda})]$ becomes:

$$\mathbb{E}[\ln p(\boldsymbol{\mu}, \boldsymbol{\Lambda})] = \frac{1}{2}\sum_{k=1}^{K}\left\{D\ln(\beta_0/2\pi) + \ln\tilde{\boldsymbol{\Lambda}}_k - \frac{\beta_0}{\beta_k} D - \beta_0 \nu_k (\boldsymbol{m}_k - \boldsymbol{m}_0)^T \boldsymbol{W}_k (\boldsymbol{m}_k - \boldsymbol{m}_0)\right\} + K\ln B(\boldsymbol{W}_0, \nu_0)$$

$$+ \frac{\nu_0 - D - 1}{2}\sum_{k=1}^{K}\ln\tilde{\boldsymbol{\Lambda}}_k - \frac{1}{2}\sum_{k=1}^{K}\nu_k Tr\left(\boldsymbol{W}_0^{-1}\boldsymbol{W}_k\right)$$

Appendix: Variational Lower Bound

Using $q^*(\boldsymbol{Z}) = \prod_{n=1}^{N}\prod_{k=1}^{K} r_{nk}^{z_{nk}}$ we derive as before:

$$\mathbb{E}[\ln q(Z)] = \sum_{n=1}^{N}\sum_{k=1}^{K} \mathbb{E}[z_{nk}]\ln r_{nk} = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\ln r_{nk}$$

Note that $\mathbb{E}\left[\ln q(\boldsymbol{\pi})\right]$ is the negative entropy of a Dirichlet distribution. Thus:

$$\mathbb{E}[\ln q(\boldsymbol{\pi})] = \sum_{k=1}^{K} (\alpha_k - 1)\ln\tilde{\pi}_k + \ln C(\alpha)$$

c(α) is the normalization constant.

Note that here we used $\ln\tilde{\pi}_k \equiv \mathbb{E}[\ln \pi_k] = \psi(\alpha_k) - \psi(\hat{\alpha})$, $\hat{a}_k = \sum_k \alpha_k$

Finally we need to show $\mathbb{E}[\ln q(\mu, \Lambda)] = \sum_{k=1}^{K}\left\{\frac{1}{2}\ln\tilde{\Lambda}_k + \frac{D}{2}\ln\frac{\beta_k}{2\pi} - \frac{D}{2} - H[q(\Lambda_k)]\right\}$

Recall that $q(\mu,\Lambda) = \prod_{k=1}^{K} q(\mu_k,\Lambda_k) = \prod_{k=1}^{K} q(\mu_k \mid \Lambda_k)q(\Lambda_k) \Rightarrow \ln q(\mu,\Lambda) = \sum_{k=1}^{K} q(\mu_k \mid \Lambda_k) + \sum_{k=1}^{K} q(\Lambda_k)$

Use the entropy of the multivariate Gaussian $\frac{D}{2}(1 + \ln(2\pi)) - \frac{D}{2}ln\beta_k - \frac{1}{2}\ln|\boldsymbol{\Lambda}_k|$ and Wishart distributions

$$H\left[q(\Lambda_k)\right] = -\ln\left(B(\boldsymbol{W}_k, \nu_k)\right) - \frac{\nu_k - D - 1}{2}\mathbb{E}\left[\ln|\Lambda_k|\right] + \frac{\nu_k D}{2}$$

Appendix: Variational Lower Bound

Thus using $q^*(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) = \mathcal{N}(\boldsymbol{\mu}_k|\boldsymbol{m}_k, (\beta_k\boldsymbol{\Lambda}_k)^{-1})\mathcal{W}(\boldsymbol{\Lambda}_k|\boldsymbol{W}_k, \nu_k)$ we can write:

$$\mathbb{E}[\ln q(\boldsymbol{\mu}, \boldsymbol{\Lambda})] = \sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{\Lambda}_k, \boldsymbol{\mu}_k}[lnq(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k)] + \sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{\Lambda}_k}[lnq(\boldsymbol{\Lambda}_k)]$$

$$= \sum_{k=1}^{K} \left\{ \frac{1}{2}\mathbb{E}_{\boldsymbol{\Lambda}_k}[\ln\boldsymbol{\Lambda}_k] + \frac{D}{2}\ln\beta_k - \frac{D}{2}\ln 2\pi - \frac{D}{2} - H[q(\boldsymbol{\Lambda}_k)] \right\}$$

$$= \sum_{k=1}^{K} \left\{ \frac{1}{2}\ln\widetilde{\boldsymbol{\Lambda}}_k + \frac{D}{2}\ln\frac{\beta_k}{2\pi} - \frac{D}{2} - H[q(\boldsymbol{\Lambda}_k)] \right\}$$

where

$$H[q(\boldsymbol{\Lambda}_k)] = -\ln\big(B(\boldsymbol{W}_k, \nu_k)\big) - \frac{\nu_k - D - 1}{2}\ln|\widetilde{\boldsymbol{\Lambda}}_k| + \frac{\nu_k D}{2}$$

Re-Estimation Eqs. Using the Variational Lower Bound

The lower bound provides an alternative approach for deriving the re-estimation eqs using direct differentiation.

With conjugate priors, the functional form of the factors in the posterior is known: discrete for **Z**, Dirichlet for π, and Gaussian-Wishart for $(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$. By substitution, we can then derive the lower bound as a function of the parameters of these distributions. Maximizing wrt these parameters gives the re-estimation eqs.

$$\mathcal{L} = \mathbb{E}\left[\ln p(\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})\right] + \mathbb{E}\left[\ln p(\boldsymbol{Z} \mid \boldsymbol{\pi})\right] + \mathbb{E}\left[\ln p(\boldsymbol{\pi})\right] + \mathbb{E}\left[\ln p(\boldsymbol{\mu}, \boldsymbol{\Lambda})\right] - \mathbb{E}\left[\ln q(\boldsymbol{Z})\right] - \mathbb{E}\left[\ln q(\boldsymbol{\pi})\right] - \mathbb{E}\left[\ln q(\boldsymbol{\mu}, \boldsymbol{\Lambda})\right]$$

$$\mathbb{E}[\ln p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] = \frac{1}{2}\sum_{k=1}^{K} N_k \left\{\ln \tilde{\boldsymbol{\Lambda}}_k - D\beta_k^{-1} - \nu_k Tr(\boldsymbol{W}_k \boldsymbol{S}_k) - \nu_k (\bar{\boldsymbol{x}}_k - \boldsymbol{m}_k)^T \boldsymbol{W}_k (\bar{\boldsymbol{x}}_k - \boldsymbol{m}_k) - D\ln(2\pi)\right\}$$

$$\mathbb{E}[\ln p(\boldsymbol{\mu}, \boldsymbol{\Lambda})] = \frac{1}{2}\sum_{k=1}^{K} \left\{D\ln(\beta_0/2\,\pi) + \ln \tilde{\boldsymbol{\Lambda}}_k - \frac{\beta_0}{\beta_k} D - \beta_0 \nu_k (\boldsymbol{m}_k - \boldsymbol{m}_0)^T \boldsymbol{W}_k (\boldsymbol{m}_k - \boldsymbol{m}_0)\right\} + K\ln B(\boldsymbol{W}_0, \nu_0)$$

$$+ \frac{\nu_0 - D - 1}{2} \sum_{k=1}^{K} \ln \tilde{\boldsymbol{\Lambda}}_k - \frac{1}{2}\sum_{k=1}^{K} \nu_k Tr\,(\boldsymbol{W}_0^{-1} \boldsymbol{W}_k)$$

Re-Estimation Eqs. Using the Variational Lower Bound

$$\mathcal{L} = \mathbb{E}\left[\ln p(\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})\right] + \mathbb{E}\left[\ln p(\boldsymbol{Z} \mid \boldsymbol{\pi})\right] + \mathbb{E}\left[\ln p(\boldsymbol{\pi})\right] + \mathbb{E}\left[\ln p(\boldsymbol{\mu}, \boldsymbol{\Lambda})\right] - \mathbb{E}\left[\ln q(\boldsymbol{Z})\right] - \mathbb{E}\left[\ln q(\boldsymbol{\pi})\right] - \mathbb{E}\left[\ln q(\boldsymbol{\mu}, \boldsymbol{\Lambda})\right]$$

$$\mathbb{E}[\ln p(\boldsymbol{\pi})] = \ln C(\alpha_0) + (\alpha_0 - 1) \sum_{k=1}^{K} \ln \tilde{\pi}_k$$

$$\mathbb{E}[\ln q(\boldsymbol{\mu}, \boldsymbol{\Lambda})] = \sum_{k=1}^{K} \left\{ \frac{1}{2} \ln \tilde{\boldsymbol{\Lambda}}_k + \frac{D}{2} \ln \frac{\beta_k}{2\pi} - \frac{D}{2} - H[q(\boldsymbol{\Lambda}_k)] \right\}$$

$$\mathbb{E}[\ln p(\boldsymbol{Z} | \boldsymbol{\pi})] = \sum_{\boldsymbol{n=1}}^{\boldsymbol{N}} \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} r_{nk} \ln \tilde{\pi}_k$$

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

$$q(\boldsymbol{Z}) = \prod_{n=1}^{N} \prod_{k=1}^{K} r_{nk}^{z_{nk}}, q(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi} | \boldsymbol{\alpha}), q(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{k=1}^{K} \mathcal{N}(\mu_k | \boldsymbol{m}_k, (\beta_k \boldsymbol{\Lambda}_k)^{-1}) \mathcal{W}(\boldsymbol{\Lambda}_k | \boldsymbol{W}_k, \nu_k)$$

Re-Estimation Eqs. Using the Variational Lower Bound

The re-estimation Equations to show (in order) are as follows:

$$\beta_k = \beta_0 + N_k$$

$$\boldsymbol{m}_k = \frac{1}{\beta_k}(\beta_0 \boldsymbol{m}_0 + N_k \bar{\boldsymbol{x}}_k)$$

$$\boldsymbol{W}_k^{-1} = \boldsymbol{W}_0^{-1} + N_k \boldsymbol{S}_k + \frac{\beta_0 N_k}{\beta_0 + N_k}(\bar{\boldsymbol{x}}_k - \boldsymbol{m_0})(\bar{\boldsymbol{x}}_k - \boldsymbol{m}_0)^T$$

$$\nu_k = \nu_0 + N_k$$

$$\alpha_k = \alpha_0 + N_k$$

$$r_{nk} \propto \tilde{\pi}_k \tilde{\boldsymbol{\Lambda}}_k^{\frac{1}{2}} exp\left(-\frac{D}{2\beta_k} - \frac{\nu_k}{2}(\boldsymbol{x}_n - \boldsymbol{m}_k)^T \boldsymbol{W}_k (\boldsymbol{x}_n - \boldsymbol{m}_k)\right)$$

Re-Estimation Eqs. Using the Variational Lower Bound

Set the derivative wrt β_k^{-1} equal to zero. Then set:

$$\frac{d}{d\beta_k^{-1}}\mathcal{L} = \frac{d}{d\beta_k^{-1}}\left(-\frac{1}{2}N_k D\beta_k^{-1} - \frac{1}{2}\frac{\beta_0}{\beta_k}D - \frac{D}{2}\ln\frac{\beta_k}{2\pi}\right) = \frac{D}{2}(-N_k - \beta_0 + \beta_k) = 0 \Rightarrow \beta_k = \beta_0 + N_k$$

$$\frac{d}{d\boldsymbol{m}_k}\mathcal{L} = -N_k\nu_k(\boldsymbol{W}_k\boldsymbol{m}_k - W_k\bar{\boldsymbol{x}}_k) - \beta_0\nu_k(\boldsymbol{W}_k\boldsymbol{m}_k - \boldsymbol{W}_k\boldsymbol{m}_0) = 0 \Rightarrow \boldsymbol{m}_k = \frac{1}{\beta_k}(\beta_0\boldsymbol{m}_0 + N_k\bar{\boldsymbol{x}}_k)$$

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

$$\mathbb{E}[\ln p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})] = \frac{1}{2}\sum_{k=1}^{K} N_k\left\{\ln\tilde{\boldsymbol{\Lambda}}_k - D\beta_k^{-1} - \nu_k Tr(\boldsymbol{W}_k\boldsymbol{S}_k) - \nu_k(\bar{\boldsymbol{x}}_k - \boldsymbol{m}_k)^T\boldsymbol{W}_k(\bar{\boldsymbol{x}}_k - \boldsymbol{m}_k) - D\ln(2\pi)\right\}$$

$$\mathbb{E}[\ln p(\boldsymbol{\mu},\boldsymbol{\Lambda})] = \frac{1}{2}\sum_{k=1}^{K}\left\{D\ln(\beta_0/2\pi) + \ln\tilde{\boldsymbol{\Lambda}}_k - \frac{\beta_0}{\beta_k}D - \beta_0\nu_k(\boldsymbol{m}_k - \boldsymbol{m}_0)^T\boldsymbol{W}_k(\boldsymbol{m}_k - \boldsymbol{m}_0)\right\} + K\ln B(\boldsymbol{W}_0,\nu_0)$$

$$+\frac{\nu_0 - D - 1}{2}\sum_{k=1}^{K}\ln\tilde{\boldsymbol{\Lambda}}_k - \frac{1}{2}\sum_{k=1}^{K}\nu_k Tr\,(\boldsymbol{W}_0^{-1}\boldsymbol{W}_k)$$

$$\mathbb{E}[\ln q(\boldsymbol{\mu},\boldsymbol{\Lambda})] = \sum_{k=1}^{K}\left\{\frac{1}{2}\ln\tilde{\boldsymbol{\Lambda}}_k + \frac{D}{2}\ln\frac{\beta_k}{2\pi} - \frac{D}{2} - H[q(\boldsymbol{\Lambda}_k)]\right\}$$

Re-Estimation Eqs. Using the Variational Lower Bound

We minimize now wrt $\mathbf{W}_k, \nu_k$ (jointly). Dropping terms independent of $\mathbf{W}_k, \nu_k$:

$$\mathcal{L} = \frac{1}{2}\sum_{k=1}^{K}\Big(N_k \ln\widetilde{\boldsymbol{\Lambda}}_k - N_k\nu_k\{Tr(\boldsymbol{W}_k\boldsymbol{S}_k) + Tr(\boldsymbol{W}_k(\bar{\boldsymbol{x}}_k - \boldsymbol{m}_k)(\bar{\boldsymbol{x}}_k - \boldsymbol{m}_k)^T)\}$$

$$+\ln\widetilde{\boldsymbol{\Lambda}}_{\boldsymbol{k}} - \beta_0\nu_k(\boldsymbol{m}_k - \boldsymbol{m}_0)^T\boldsymbol{W}_k(\boldsymbol{m}_k - \boldsymbol{m}_0) + (\nu_0 - D - 1)\ln\widetilde{\boldsymbol{\Lambda}}_k - \nu_k Tr(\boldsymbol{W}_0^{-1}\boldsymbol{W}_k) - \ln\widetilde{\boldsymbol{\Lambda}}_k + 2H[q(\boldsymbol{\Lambda}_k)]\Big)$$

where $ln\widetilde{\boldsymbol{\Lambda}}_k \equiv \mathbb{E}[\ln|\boldsymbol{\Lambda}_k|] = \sum_{i=1}^{D}\psi\left(\frac{\nu_k+1-i}{2}\right) + D\ln 2 + \ln|\boldsymbol{W}_k|$ and

$$H[q(\boldsymbol{\Lambda}_k)] = -\ln\big(B(\boldsymbol{W}_k,\nu_k)\big) - \frac{\nu_k - D - 1}{2}\ln\widetilde{\boldsymbol{\Lambda}}_k + \frac{\nu_k D}{2}$$

$$\ln\big(B(\boldsymbol{W}_k,\nu_k)\big) = -\frac{\nu_k}{2}\ln|\boldsymbol{W}_k| - \frac{\nu_k D}{2}\ln 2 - \sum_{i=1}^{D}\ln\Gamma\left(\frac{\nu_k+1-i}{2}\right) - \frac{D(D-1)}{4}\ln\pi$$

$$\mathcal{L} = \mathbb{E}\big[\ln p(\boldsymbol{X}\mid\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})\big] + \mathbb{E}\big[\ln p(\boldsymbol{Z}\mid\boldsymbol{\pi})\big] + \mathbb{E}\big[\ln p(\boldsymbol{\pi})\big] + \mathbb{E}\big[\ln p(\boldsymbol{\mu},\boldsymbol{\Lambda})\big] - \mathbb{E}\big[\ln q(\boldsymbol{Z})\big] - \mathbb{E}\big[\ln q(\boldsymbol{\pi})\big] - \mathbb{E}\big[\ln q(\boldsymbol{\mu},\boldsymbol{\Lambda})\big]$$

$$\mathbb{E}[\ln p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})] = \frac{1}{2}\sum_{k=1}^{K}N_k\{\ln\widetilde{\boldsymbol{\Lambda}}_k - D\beta_k^{-1} - \nu_k Tr(\boldsymbol{W}_k\boldsymbol{S}_k) - \nu_k(\bar{\boldsymbol{x}}_k - \boldsymbol{m}_k)^T\boldsymbol{W}_k(\bar{\boldsymbol{x}}_k - \boldsymbol{m}_k) - D\ln(2\pi)\}$$

$$\mathbb{E}[\ln p(\boldsymbol{\mu},\boldsymbol{\Lambda})] = \frac{1}{2}\sum_{k=1}^{K}\left\{D\ln(\beta_0/2\pi) + \ln\widetilde{\boldsymbol{\Lambda}}_k - \frac{\beta_0}{\beta_k}D - \beta_0\nu_k(\boldsymbol{m}_k - \boldsymbol{m}_0)^T\boldsymbol{W}_k(\boldsymbol{m}_k - \boldsymbol{m}_0)\right\} + K\ln B(\boldsymbol{W}_0,\nu_0)$$

$$+\frac{\nu_0 - D - 1}{2}\sum_{k=1}^{K}\ln\widetilde{\boldsymbol{\Lambda}}_k - \frac{1}{2}\sum_{k=1}^{K}\nu_k Tr\,(\boldsymbol{W}_0^{-1}\boldsymbol{W}_k)$$

$$\mathbb{E}[\ln q(\boldsymbol{\mu},\boldsymbol{\Lambda})] = \sum_{k=1}^{K}\left\{\frac{1}{2}\ln\widetilde{\boldsymbol{\Lambda}}_k + \frac{D}{2}\ln\frac{\beta_k}{2\pi} - \frac{D}{2} - H[q(\boldsymbol{\Lambda}_k)]\right\}$$

Re-Estimation Eqs. Using the Variational Lower Bound

We minimize now wrt $\mathbf{W}_k, \nu_k$ (jointly). Dropping terms independent of $\mathbf{W}_k, \nu_k$:

$$\mathcal{L} = \frac{1}{2}\sum_{k=1}^{K}\Big(N_k \ln\widetilde{\boldsymbol{\Lambda}}_k - N_k\nu_k\{Tr(\boldsymbol{W}_k\boldsymbol{S}_k) + Tr(\boldsymbol{W}_k(\bar{\boldsymbol{x}}_k - \boldsymbol{m}_k)(\bar{\boldsymbol{x}}_k - \boldsymbol{m}_k)^T)\}$$

$$+\ln\widetilde{\boldsymbol{\Lambda}}_k - \beta_0\nu_k(\boldsymbol{m}_k - \boldsymbol{m}_0)^T\boldsymbol{W}_k(\boldsymbol{m}_k - \boldsymbol{m}_0) + (\nu_0 - D - 1)\ln\widetilde{\boldsymbol{\Lambda}}_k - \nu_k Tr(\boldsymbol{W}_0^{-1}\boldsymbol{W}_k) - \ln\widetilde{\boldsymbol{\Lambda}}_k + 2H[q(\boldsymbol{\Lambda}_k)]\Big)$$

where $ln\widetilde{\boldsymbol{\Lambda}}_k \equiv \mathbb{E}[\ln|\boldsymbol{\Lambda}_k|] = \sum_{i=1}^{D}\psi\left(\frac{\nu_k+1-i}{2}\right) + D\ln 2 + \ln|\boldsymbol{W}_k|$

and $$H[q(\boldsymbol{\Lambda}_k)] = -\ln\big(B(\boldsymbol{W}_k,\nu_k)\big) - \frac{\nu_k - D - 1}{2}\ln\widetilde{\boldsymbol{\Lambda}}_k + \frac{\nu_k D}{2}$$

$$\ln\big(B(\boldsymbol{W}_k,\nu_k)\big) = -\frac{\nu_k}{2}\ln|\boldsymbol{W}_k| - \frac{\nu_k D}{2}\ln 2 - \sum_{i=1}^{D}\ln\Gamma\left(\frac{\nu_k+1-i}{2}\right) - \frac{D(D-1)}{4}\ln\pi$$

For a single component:

$$\mathcal{L} = \frac{1}{2}(N_k + \nu_0 - \nu_k)\ln\widetilde{\boldsymbol{\Lambda}}_k - \frac{\nu_k}{2}Tr(\boldsymbol{W}_k\boldsymbol{F}_k) - \ln B(\boldsymbol{W}_k,\nu_k) + \frac{\nu_k D}{2}$$

Can prove by substituting

$$\boldsymbol{m}_k = \frac{1}{\beta_k}(\beta_0\boldsymbol{m}_0 + N_k\bar{\boldsymbol{x}}_k)$$

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

$$\frac{d}{d\nu_k}\mathcal{L} = \frac{1}{2}\left((N_k + \nu_0 - \nu_k)\frac{d\ln\widetilde{\boldsymbol{\Lambda}}_k}{d\nu_k} - \ln\widetilde{\boldsymbol{\Lambda}}_k - Tr(\boldsymbol{W}_k\boldsymbol{F}_k) + \ln|\boldsymbol{W}_k| + D\ln(2) + \sum_{i=1}^{D}\psi\left(\frac{\nu_k+1-i}{2}\right) + D\right)$$

$$= \frac{1}{2}\left((N_k + \nu_0 - \nu_k)\frac{d\ln\widetilde{\boldsymbol{\Lambda}}_k}{d\nu_k} - Tr(\boldsymbol{W}_k\boldsymbol{F}_k) + D\right)$$

Here use

$$ln\widetilde{\boldsymbol{\Lambda}}_k \equiv \sum_{i=1}^{D}\psi\left(\frac{\nu_k+1-i}{2}\right) + D\ln 2 + \ln|\boldsymbol{W}_k|$$

Re-Estimation Eqs. Using the Variational Lower Bound

$$\mathcal{L} = \frac{1}{2}(N_k + \nu_0 - \nu_k)\ln\tilde{\Lambda}_k - \frac{\nu_k}{2}Tr(\boldsymbol{W}_k\boldsymbol{F}_k) - \ln B(\boldsymbol{W}_k, \nu_k) + \frac{\nu_k D}{2}, \boldsymbol{F}_k = \boldsymbol{W}_0^{-1} + N_k\boldsymbol{S}_k + \frac{N_k\beta_0}{N_k + \beta_0}(\bar{\boldsymbol{x}}_k - \boldsymbol{m}_0)(\bar{\boldsymbol{x}}_k - \boldsymbol{m}_0)^T$$

$$\ln\left(B(\boldsymbol{W}_k, \nu_k)\right) = -\frac{\nu_k}{2}\ln|\boldsymbol{W}_k| - \frac{\nu_k D}{2}\ln 2 - \sum_{i=1}^{D}\ln\Gamma\left(\frac{\nu_k + 1 - i}{2}\right) - \frac{D(D-1)}{4}\ln\pi$$

$$ln\tilde{\Lambda}_k \equiv \mathbb{E}[\ln|\boldsymbol{\Lambda}_k|] = \sum_{i=1}^{D}\psi\left(\frac{\nu_k + 1 - i}{2}\right) + D\ln 2 + \ln|\boldsymbol{W}_k|$$

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

Similarly differentiation wrt $\mathbf{W}_k$ gives (use differentiation formulas $\frac{\partial}{\partial \boldsymbol{A}}\ln|\boldsymbol{A}| = (\boldsymbol{A}^{-1})^T$, $\frac{\partial}{\partial \boldsymbol{A}}\mathrm{tr}(\boldsymbol{A}^T\boldsymbol{B}) = \boldsymbol{B}$):

$$\frac{d}{d\boldsymbol{W}_k}\mathcal{L} = \frac{1}{2}\left((N_k + \nu_0 - \nu_k)\boldsymbol{W}_k^{-1} - \nu_k\left(\boldsymbol{F}_k - \boldsymbol{W}_k^{-1}\right)\right) = 0$$

From the above EqS. simultaneously with: $\frac{d}{d\nu_k}\mathcal{L} = \frac{1}{2}\left((N_k + \nu_0 - \nu_k)\frac{d\ln\tilde{\Lambda}_k}{d\nu_k} - Tr(\boldsymbol{W}_k\boldsymbol{F}_k) + D\right)$,
we see that the only solution is:

$$0 = N_k + \nu_0 - \nu_k \Rightarrow \nu_k = N_k + \nu_0$$

$$\boldsymbol{F}_k - \boldsymbol{W}_k^{-1} = 0 \Rightarrow \boldsymbol{W}_0^{-1} + N_k\boldsymbol{S}_k + \frac{N_k\beta_0}{N_k + \beta_0}(\bar{\boldsymbol{x}}_k - \boldsymbol{m}_0)(\bar{\boldsymbol{x}}_k - \boldsymbol{m}_0)^T - \boldsymbol{W}_k^{-1} = 0$$

from which:

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

$$\nu_k = \nu_0 + N_k$$

Re-Estimation Eqs. Using the Variational Lower Bound

We now differentiate wrt α_k. They appear in the following terms:

$$\mathbb{E}[\ln p(\boldsymbol{Z}|\boldsymbol{\pi})] = \sum_{N=1}^{n}\sum_{k=1}^{K} r_{nk}\ln\tilde{\pi}_k \qquad \mathbb{E}[\ln p(\boldsymbol{\pi})] = \ln C(\alpha_0) + (\alpha_0 - 1)\sum_{k=1}^{K}\ln\tilde{\pi}_k$$

$$\ln\tilde{\pi}_k \equiv \mathbb{E}[\ln\pi_k] = \psi(\alpha_k) - \psi(\hat{\alpha}),\ \hat{\alpha}_k = \sum_k \alpha_k$$

$$\mathbb{E}[\ln q(\boldsymbol{\pi})] = \sum_{k=1}^{K}(\alpha_k - 1)\ln\tilde{\pi}_k + \ln C(\alpha),\ C(\alpha) = \frac{\Gamma(\hat{\alpha})}{\Gamma(\alpha_1)\Gamma(\alpha_2)\ldots\Gamma(\alpha_K)}, \hat{\alpha} = \sum_{k=1}^{K}\alpha_K$$

Using the di- and tri-gamma functions $\psi(.), \psi_1(.)$ we finally have:

$$\frac{\partial}{\partial\alpha_k}\mathcal{L} = [N_k + (\alpha_0 - 1) - (\alpha_k - 1)]\frac{\partial\ln\tilde{\pi}_k}{\partial\alpha_k} - \ln\tilde{\pi}_k - \frac{\partial\ln C(\alpha)}{\partial\alpha_k}$$

$$= [N_k + (\alpha_0 - 1) - (\alpha_k - 1)]\left(\psi_1(\alpha_k) - \psi_1(\hat{\alpha})\frac{\partial\hat{\alpha}}{\partial\alpha_k}\right) + \psi(\hat{\alpha}) - \psi(\alpha_k) - \psi(\hat{\alpha})\frac{\partial\hat{\alpha}}{\partial\alpha_k} + \psi(\alpha_k)$$

$$= [N_k + (\alpha_0 - 1) - (\alpha_k - 1)]\big(\psi_1(\alpha_k) - \psi_1(\hat{\alpha})\big) = 0 \quad \underset{\alpha_0>0}{\Rightarrow} \alpha_k = N_k + \alpha_0$$

We used here that the tri-gamma function is >0 and monotonically decreasing for positive arguments: $\psi_1(z) = \sum_{n=0}^{\infty}\frac{1}{(z+n)^2}$.

Re-Estimation Eqs. Using the Variational Lower Bound

Finally we maximize $\mathcal{L}$ wrt r_{nk} subject to $1 = \sum_k r_{nk}$ for all n. r_{nk} appears in

$$\mathbb{E}[\ln p(\boldsymbol{Z}|\boldsymbol{\pi})] = \sum_{N=1}^{N}\sum_{k=1}^{K} r_{nk}\ln\tilde{\pi}_k \qquad \mathbb{E}[\ln q(\boldsymbol{Z})] = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\ln r_{nk}$$

$$\mathbb{E}[\ln p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})] = \frac{1}{2}\sum_{k=1}^{K} N_k\left\{\ln\tilde{\boldsymbol{\Lambda}}_k - D\beta_k^{-1} - \nu_k Tr(\boldsymbol{W}_k\boldsymbol{S}_k) - \nu_k(\bar{\boldsymbol{x}}_k - \boldsymbol{m}_k)^T\boldsymbol{W}_k(\bar{\boldsymbol{x}}_k - \boldsymbol{m}_k) - D\ln(2\pi)\right\}$$

In the last expression r_{nk} appears via the three terms shown

The last 2 terms in $\mathbb{E}[\ln p(\boldsymbol{X}\,|\,\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})]$ can be written as:

$$\frac{1}{2}\sum_{k=1}^{K}\nu_k Tr(\boldsymbol{W}_k\boldsymbol{Q}_k)\,,\ \boldsymbol{Q}_k = \sum_{n=1}^{N} r_{nk}(\boldsymbol{x}_n - \bar{\boldsymbol{x}}_k)(\boldsymbol{x}_n - \bar{\boldsymbol{x}}_k)^T + N_k(\bar{\boldsymbol{x}}_k - \boldsymbol{m}_k)(\bar{\boldsymbol{x}}_k - \boldsymbol{m}_k)^T$$

$$= \sum_{n=1}^{N} r_{nk}(\boldsymbol{x}_n - \boldsymbol{m}_k)(\boldsymbol{x}_n - \boldsymbol{m}_k)^T$$

Here we simply expanded the terms and used

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

Re-Estimation Eqs. Using the Variational Lower Bound

Considering all terms in $\mathcal{L}$ wrt r_{nk} subject to $1 = \sum_k r_{nk}$ for all n. r_{nk} appears in

$$\frac{1}{2}\sum_{k=1}^{K}\sum_{n=1}^{N} r_{nk}\left(\ln\tilde{\boldsymbol{\Lambda}}_k - D\beta_k^{-1}\right) - \frac{1}{2}\sum_{k=1}^{K}\sum_{n=1}^{N} r_{nk}\,\nu_k(\boldsymbol{x}_n - \boldsymbol{m}_k)^T W_k(\boldsymbol{x}_n - \boldsymbol{m}_k) + \sum_{k=1}^{K}\sum_{n=1}^{N} r_{nk}\ln\tilde{\pi}_k - \sum_{k=1}^{K}\sum_{n=1}^{N} r_{nk}\ln r_{nk} + \sum_{n=1}^{N}\lambda_n\left(1 - \sum_{k=1}^{K} r_{nk}\right)$$

Taking derivatives wrt r_{nk}:

$$0 = \frac{1}{2}\ln\tilde{\boldsymbol{\Lambda}}_k - \frac{D}{2\beta_k} - \frac{1}{2}\nu_k(\boldsymbol{x}_n - \boldsymbol{m}_k)^T\boldsymbol{W}_k(\boldsymbol{x}_n - \boldsymbol{m}_k) + \ln\tilde{\pi}_k - \ln r_{nk} - 1 - \lambda_n$$

Moving $\ln r_{nk}$ to the l.h.s. and exponentiating leads to:

$$r_{nk} \propto \tilde{\pi}_k\tilde{\boldsymbol{\Lambda}}_k^{\frac{1}{2}} exp\left(-\frac{D}{2\beta_k} - \frac{\nu_k}{2}(\boldsymbol{x}_n - \boldsymbol{m}_k)^T\boldsymbol{W}_k(\boldsymbol{x}_n - \boldsymbol{m}_k)\right)$$

This can then be normalized as usual. This step completes all the update Eqs. for the parameters of the approximations to the posterior distributions.

Predictive Distribution

In applications of the Bayesian mixture of Gaussians model we will often be interested in the predictive density for a new value $\hat{\boldsymbol{x}}$ of the observed variable. Associated with this observation will be a corresponding latent variable $\hat{\boldsymbol{z}}$, and the predictive density is then given by

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

where $p(\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda}|\boldsymbol{X})$ is the (unknown) true posterior distribution of the parameters. Using $p(\hat{\boldsymbol{z}}|\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{\hat{z}_k}$ and $p(\hat{\boldsymbol{x}}|\hat{\boldsymbol{z}}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{k=1}^{K} \mathcal{N}(\hat{\boldsymbol{x}}|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})^{\hat{z}_k}$, we can first perform the summation over $\hat{\boldsymbol{z}}$ to give

$$p(\hat{\boldsymbol{x}}|\boldsymbol{X}) = \sum_{k=1}^{K} \iiint \pi_k \mathcal{N}(\hat{\boldsymbol{x}}|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})\, p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{X}) d\boldsymbol{\pi} d\boldsymbol{\mu} d\boldsymbol{\Lambda}$$

Because the remaining integrations are intractable, we approximate the predictive density by replacing the true posterior distribution $p(\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda}|\boldsymbol{X})$ with its variational approximation $q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\boldsymbol{\pi}) q(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\boldsymbol{\pi}) \prod_{k=1}^{K} q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$ to give (and in each term we have implicitly integrated out all variables $\{\boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j\}$ for $j \neq k$)

$$p(\hat{\boldsymbol{x}}|\boldsymbol{X}) \approx \sum_{k=1}^{K} \iiint \pi_k \mathcal{N}(\hat{\boldsymbol{x}}|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}) q(\boldsymbol{\pi}) q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) d\boldsymbol{\pi} d\boldsymbol{\mu}_k d\boldsymbol{\Lambda}_k$$

Predictive Distribution

$$p(\hat{\boldsymbol{x}}|\boldsymbol{X}) = \sum_{k=1}^{K} \iiint \pi_k \mathcal{N}(\hat{\boldsymbol{x}}|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}) q(\boldsymbol{\pi}) q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) d\boldsymbol{\pi} d\boldsymbol{\mu}_k d\boldsymbol{\Lambda}_k$$

Performing the integration in $\boldsymbol{\pi}$ and using $\mathbb{E}[\pi_k] = \frac{\alpha_k}{\hat{\alpha}}$, we can simplify as

$$p(\hat{\boldsymbol{x}}|\boldsymbol{X}) = \sum_{k=1}^{K} \frac{\alpha_k}{\hat{\alpha}} \iint \mathcal{N}(\hat{\boldsymbol{x}}|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}) q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) d\boldsymbol{\mu}_k d\boldsymbol{\Lambda}_k$$

We perform next the integration in $\boldsymbol{\mu}_k$ exactly using $\mathrm{q}^*(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) = \mathcal{N}(\boldsymbol{\mu}_k|\boldsymbol{m}_k, (\beta_k \boldsymbol{\Lambda}_k)^{-1}) \mathcal{W}(\boldsymbol{\Lambda}_k|\boldsymbol{W}_k, \nu_k)$ and convolution of Gaussian linear models:

$$p(\hat{\boldsymbol{x}}|\boldsymbol{X}) = \sum_{k=1}^{K} \frac{\alpha_k}{\hat{\alpha}} \iint \mathcal{N}(\hat{\boldsymbol{x}}|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}) \mathcal{N}(\boldsymbol{\mu}_k|\boldsymbol{m}_k, (\beta_k \boldsymbol{\Lambda}_k)^{-1}) \mathcal{W}(\boldsymbol{\Lambda}_k|\boldsymbol{W}_k, \nu_k) d\boldsymbol{\mu}_k d\boldsymbol{\Lambda}_k$$

$$= \sum_{k=1}^{K} \frac{\alpha_k}{\hat{\alpha}} \int \mathcal{N}\left(\hat{\boldsymbol{x}} \middle| \boldsymbol{m}_k, \left(1 + \beta_k^{-1}\right) \boldsymbol{\Lambda}_k^{-1}\right) \mathcal{W}(\boldsymbol{\Lambda}_k|\boldsymbol{W}_k, \nu_k) d\boldsymbol{\Lambda}_k$$

This is the convolution over $\boldsymbol{\Lambda}_k$ of a Wishart with a Gaussian.

Predictive Distribution

$$p(\hat{\boldsymbol{x}}|\boldsymbol{X}) = \sum_{k=1}^{K} \frac{\alpha_k}{\hat{\alpha}} \int \mathcal{N}\left(\hat{\boldsymbol{x}} \middle| \boldsymbol{m}_k, \left(1+\beta_k^{-1}\right) \boldsymbol{\Lambda}_k^{-1}\right) \mathcal{W}(\boldsymbol{\Lambda}_k|\boldsymbol{W}_k, \nu_k) d\boldsymbol{\Lambda}_k$$

$$\propto \sum_{k=1}^{K} \frac{\alpha_k}{\hat{\alpha}} \int |\boldsymbol{\Lambda}_k|^{1/2+(\nu_k-D-1)/2} exp\left(-\frac{1}{2(1+\beta_k^{-1})} Tr\left[\boldsymbol{\Lambda}_k (\hat{\boldsymbol{x}}-\boldsymbol{m}_k)(\hat{\boldsymbol{x}}-\boldsymbol{m}_k)^T\right]\right.$$

Predictive Distribution

$$p(\hat{\boldsymbol{x}}|\boldsymbol{X}) \approx \sum_{k=1}^{K} \frac{\alpha_k}{\hat{\alpha}} \left(1 + \frac{1}{1+\beta_k^{-1}} (\hat{\boldsymbol{x}} - \boldsymbol{m}_k)^T \boldsymbol{W}_k (\hat{\boldsymbol{x}} - \boldsymbol{m}_k)\right)^{-(\nu_k+1)/2}$$

We can recognize in the parenthesis above the Student's t distribution:

$$St(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) = \frac{\Gamma\left(\frac{\nu}{2}+\frac{D}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{|\boldsymbol{\Lambda}|^{1/2}}{(\nu\pi)^{D/2}} \left(1+\frac{\Delta^2}{\nu}\right)^{-\frac{\nu}{2}-\frac{D}{2}}, \Delta^2=(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Lambda}(\boldsymbol{x}-\boldsymbol{\mu})$$

Finally:

$$p(\hat{\boldsymbol{x}}|\boldsymbol{X}) \cong \sum_{k=1}^{K} \frac{\alpha_k}{\hat{\alpha}} St(\hat{\boldsymbol{x}}|\boldsymbol{m}_k, \boldsymbol{L}_k, \nu_k + 1 - D)$$

in which the *k*th component has mean $\mathbf{m}_k$, and the precision is given by

$$\boldsymbol{L}_k = \frac{(\nu_k + 1 - D)}{1 + \beta_k^{-1}} \boldsymbol{W}_k$$

with $\nu_k = \nu_0 + N_k$. We will show next that when the size *N* of the data set is large the predictive distribution reduces to a mixture of Gaussians.

Predictive Distribution for Large Data Set

We will show that the variational Bayes solution $p(\hat{\boldsymbol{x}}|\boldsymbol{X}) \cong \sum_{k=1}^{K} \frac{\alpha_k}{\hat{\alpha}} St(\hat{\boldsymbol{x}}|\boldsymbol{m}_k, \boldsymbol{L}_k, \nu_k + 1 - D)$ for the mixture of Gaussians model when the size *N* of the data set is large reduces (as we would expect) to the MLE solution based on the EM.

$$p(\hat{\boldsymbol{x}}|\boldsymbol{X}) \cong \sum_{k=1}^{K} \frac{N_k}{N} \mathcal{N}(\hat{\boldsymbol{x}}|\bar{\boldsymbol{x}}_k, \boldsymbol{S}_k)$$

The derivation is based on the following steps:

1. We first show that the posterior distribution $q^*(\boldsymbol{\Lambda}_k)$ of the precisions becomes sharply peaked around the maximum likelihood solution.

2. We do the same for the posterior distribution of the means $q^*(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k)$.

3. Next consider the posterior distribution $q(\boldsymbol{\pi})$ for the mixing coefficients and we show that this too becomes sharply peaked around the maximum likelihood solution.

4. Similarly, we show that the responsibilities become equal to the corresponding maximum likelihood values for large *N*, by making use of the following asymptotic result for the digamma function for large *x*, $\psi(x) = \ln x + O(1/x)$

Finally, we show that for large *N*, the predictive distribution becomes a mixture of Gaussians.

Predictive Distribution for Large Data Set

1. We first show that the posterior distribution $q^*(\boldsymbol{\Lambda}_k)$ of the precisions becomes sharply peaked around the maximum likelihood solution.

Consider first the posterior distribution over the precision of component k given by

$$q^*(\boldsymbol{\Lambda}_k) = \mathcal{W}(\boldsymbol{\Lambda}_k | \boldsymbol{W}_k, \nu_k)$$

From $\nu_k = \nu_0 + N_k$ we see that for large N we have $\nu_k \to N_k$, and similarly from $\boldsymbol{W}_k^{-1} = \boldsymbol{W}_0^{-1} + N_k \boldsymbol{S}_k + \frac{\beta_0 N_k}{\beta_0 + N_k} (\bar{\boldsymbol{x}}_k - \boldsymbol{m}_0)(\bar{\boldsymbol{x}}_k - \boldsymbol{m}_0)^T$ we see that $\boldsymbol{W}_k \to N_k^{-1} \boldsymbol{S}_k^{-1}$

The mean over $\boldsymbol{\Lambda}_k$ is $\mathbb{E}[\boldsymbol{\Lambda}_k] = \nu_k \boldsymbol{W}_k \to \boldsymbol{S}_k^{-1}$, $\boldsymbol{S}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} (\boldsymbol{x}_n - \bar{\boldsymbol{x}}_k)(\boldsymbol{x}_n - \bar{\boldsymbol{x}}_k)^T$

which is the maximum likelihood value (this assumes that the quantities r_{nk} reduce to the corresponding EM values, which is indeed the case as we shall show shortly).

In order to show that this posterior is also sharply peaked, we consider the differential entropy, $H[\boldsymbol{\Lambda}_k]$ given by

$$H[\Lambda] = -\ln B(\boldsymbol{W}, \nu) - \frac{\nu - D - 1}{2} \mathbb{E}[\ln |\Lambda|] + \frac{\nu D}{2}$$

$$\mathbb{E}[\ln |\Lambda|] = \sum_{i=1}^{D} \psi\left(\frac{\nu + 1 - i}{2}\right) + D \ln 2 + \ln |\boldsymbol{W}|$$

and show that, as

$$N_k \to \infty \Rightarrow H[\boldsymbol{\Lambda}_k] \to 0$$

Predictive Distribution for Large Data Set

$$H[\Lambda] = -\ln B(\boldsymbol{W},\nu) - \frac{\nu - D - 1}{2}\mathbb{E}[\ln|\Lambda|] + \frac{\nu D}{2}, \quad \mathbb{E}[\ln|\Lambda|] = \sum_{i=1}^{D}\psi\left(\frac{\nu+1-i}{2}\right) + D\ln 2 + \ln|\boldsymbol{W}|$$

To show $N_k \to \infty \Rightarrow H[\Lambda_k] \to 0$, consider first the normalizing factor. Since $\boldsymbol{W}_k \to N_k^{-1}\boldsymbol{S}_k^{-1}$ and $\nu_k \to N_k$:

$$B(\boldsymbol{W},\nu) = |\boldsymbol{W}|^{-\nu/2}\left(2^{\nu D/2}\pi^{D(D-1)/4}\prod_{i=}^{D}\Gamma\left(\frac{\nu+1-i}{2}\right)\right)^{-1}$$

$$-\ln B(\boldsymbol{W}_k,\nu_k) \to -\frac{N_k}{2}(D\ln N_k + \ln|\boldsymbol{S}_k| - D\ln 2) + \sum_{i=1}^{D}\ln\Gamma\left(\frac{N_k+1-i}{2}\right)$$

$$Stirling's\ approximation:\ \ln\Gamma\left(\frac{N_k+1-i}{2}\right) \simeq \frac{N_k}{2}(\ln N_k - \ln 2 - 1)$$

$$For\ x >> 1:\ \Gamma(x+1) \simeq (2\pi)^{1/2}e^{-x}x^{x+1/2}$$

This leads to:

$$-\ln B(\boldsymbol{W}_k,\nu_k) \to -\frac{N_k D}{2}(\ln N_k - \ln 2 - \ln N_k + \ln 2 + 1) - \frac{N_k}{2}\ln|\boldsymbol{S}_k| = -\frac{N_k}{2}(\ln|\boldsymbol{S}_k| + D)$$

Now using $\mathbb{E}[\ln|\Lambda_k|] = \sum_{i=1}^{D}\psi\left(\frac{\nu+1-i}{2}\right) + D\ln 2 + \ln|\boldsymbol{W}|$, $\boldsymbol{W}_k \to N_k^{-1}\boldsymbol{S}_k^{-1}$ and $\psi(x) = \ln x + O(1/x)$:

$$\mathbb{E}[\ln|\Lambda_k|] \to D\ln\frac{N_k}{2} + D\ln 2 - D\ln N_k - \ln|\boldsymbol{S}_k| = -\ln|\boldsymbol{S}_k|$$

Finally with $\nu_k \to N_k$ and $N_k \to \infty$

$$H[\Lambda_k] = -\ln B(\boldsymbol{W}_k,\nu_k) - \frac{\nu_k - D - 1}{2}\mathbb{E}[\ln|\Lambda_k|] + \frac{\nu_k D}{2} \to -\frac{N_k}{2}(\ln|\boldsymbol{S}_k| + D) + \frac{N_k - D - 1}{2}\ln|\boldsymbol{S}_k| + \frac{N_k D}{2} \to 0 \Rightarrow \quad \Lambda_k \to \delta\left(\Lambda_k - \boldsymbol{S}_k^{-1}\right)$$

Predictive Distribution for Large Data Set

We do the same for the posterior distribution of the means $q^*(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k)$.

$$\mathrm{q}^*(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k) = \mathcal{N}(\boldsymbol{\mu}_k|\boldsymbol{m}_k, (\beta_k\boldsymbol{\Lambda}_k)^{-1})$$

From $\boldsymbol{m}_k = \frac{1}{\beta_0+N_k}(\beta_0\boldsymbol{m}_0 + N_k\bar{\boldsymbol{x}}_k)$ we see that for large N, the mean of this distribution reduces to $\bar{\boldsymbol{x}}_k$ which is the corresponding MLE value.

From $\beta_k = \beta_0 + N_k$, we see that $\beta_k \to N_k$ and thus $\beta_k\boldsymbol{\Lambda}_k \to N_k\boldsymbol{S}_k^{-1}$ which is large for large N and hence this distribution is sharply peaked around its mean. Thus $\mathrm{q}^*(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k) \to \delta(\boldsymbol{\mu}_k - \boldsymbol{m}_k)$

Now consider the posterior distribution $q^*(\boldsymbol{\pi})$ given by $q(\boldsymbol{\pi}) = \mathcal{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha})$. For large N we have $\alpha_k = \alpha_0 + N_k \to N_k$ and so from

$$\mathbb{E}[\pi_k] = \frac{\alpha_k}{\hat{\alpha}} \to \frac{N_k}{N}, \mathrm{var}[\pi_i] = \frac{\alpha_i(\hat{\alpha} - \alpha_i)}{\hat{\alpha}^2(\hat{\alpha}+1)} \to 0, \hat{\alpha} = \sum_i \alpha_i$$

we see that the posterior distribution becomes sharply peaked around its mean

$$\mathbb{E}[\pi_k] = \frac{\alpha_k}{\hat{\alpha}} \to \frac{N_k}{N}$$

which is the maximum likelihood solution.

Predictive Distribution for Large Data Set

For the distribution q*(**z**) we consider the responsibilities given by

$$r_{nk} \propto \tilde{\pi}_k \tilde{\boldsymbol{\Lambda}}_k^{\frac{1}{2}} exp\left(-\frac{D}{2\beta_k} - \frac{\nu_k}{2}(\boldsymbol{x}_n - \boldsymbol{m}_k)^T \boldsymbol{W}_k (\boldsymbol{x}_n - \boldsymbol{m}_k)\right).$$

Using

$$\ln \tilde{\pi}_k \equiv \mathbb{E}[\ln \pi_k] = \psi(\alpha_k) - \psi(\hat{\alpha}) \rightarrow lnN_k - lnN, \Longrightarrow \tilde{\pi}_k \longrightarrow \frac{N_k}{N}$$

$ln\tilde{\boldsymbol{\Lambda}}_k \equiv \mathbb{E}[\ln|\boldsymbol{\Lambda}_k|] = \sum_{i=1}^{D} \psi\left(\frac{\nu_k+1-i}{2}\right) + D\ln 2 + \ln|\boldsymbol{W}_k| \rightarrow -ln|\boldsymbol{S}_k|$ (shown earlier),

$\boldsymbol{m}_k \rightarrow \overline{\boldsymbol{x}}_k$, $\frac{D}{2\beta_k} = \frac{D}{2(\beta_0+N_k)} \rightarrow 0$ and $\nu_k \boldsymbol{W}_k \rightarrow N_k N_k^{-1} \boldsymbol{S}_k^{-1} = \boldsymbol{S}_k^{-1}$. We again obtain the MLE expression for the responsibilities for large N: $r_{nk} \propto \pi_k \mathcal{N}(\boldsymbol{x}_n|\overline{\boldsymbol{x}}_k, \boldsymbol{S}_k)$.

Finally for the predictive distribution after the integration over π, we have:

$$p(\hat{\boldsymbol{x}}|\boldsymbol{X}) \approx \sum_{k=1}^{K} \frac{\alpha_k}{\hat{\alpha}} \iint \mathcal{N}(\hat{\boldsymbol{x}}|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}) q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) d\boldsymbol{\mu}_k d\boldsymbol{\Lambda}_k$$

$\boldsymbol{\Lambda}_k \rightarrow \delta\left(\boldsymbol{\Lambda}_k - \boldsymbol{S}_k^{-1}\right)$

$\boldsymbol{m}_k \rightarrow \overline{\boldsymbol{x}}_k$

$q^*(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k) \rightarrow \delta(\boldsymbol{\mu}_k - \boldsymbol{m}_k)$

The integration is simple as the arguments are now delta functions. We obtain:

$$p(\hat{\boldsymbol{x}}|\boldsymbol{X}) \cong \sum_{k=1}^{K} \frac{N_k}{N} \mathcal{N}(\hat{\boldsymbol{x}}|\overline{\boldsymbol{x}}_k, \boldsymbol{S}_k)$$

MAP Estimate vs MLE

The singularities arising in the MLE treatment of Gaussian mixture models do not arise if the Bayesian model were solved using maximum posterior (MAP) estimation.

Recall that the singularities that may arise in maximum likelihood estimation are caused by a mixture component, k, collapsing on a data point, $\mathbf{x}_n$, i.e., $r_{kn} = 1$, $\boldsymbol{\mu}_k = \mathbf{x}_n$ and $|\Lambda_k| \rightarrow \infty$.

However, the prior distribution $p(\boldsymbol{\mu}, \Lambda) = \prod_{k=1}^{K} \mathcal{N}(\mu_k | \boldsymbol{m}_0, (\beta_0 \boldsymbol{\Lambda}_k)^{-1}) \mathcal{W}(\boldsymbol{\Lambda}_k | \boldsymbol{W}_0, \nu_0)$ will prevent this from happening, also in the case of MAP estimation.

Let us compute the MAP $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Lambda}})$ by maximizing the expected log of the product of the complete-likelihood and prior $p(\boldsymbol{\mu}, \Lambda)$ as a function of $\boldsymbol{\Lambda}_k$:

$$\mathbb{E}_{q(\mathbf{Z})}\left[\ln p(\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{\mu}, \Lambda) p(\boldsymbol{\mu}, \Lambda)\right] = \frac{1}{2} \sum_{n=1}^{N} r_{kn} \left(\ln |\Lambda_k| - \left(\boldsymbol{x}_n - \boldsymbol{\mu}_k\right)^T \Lambda_k \left(\boldsymbol{x}_n - \boldsymbol{\mu}_k\right) \right)$$

$$+ \frac{1}{2} \left(\ln |\Lambda_k| - \beta_0 \left(\boldsymbol{\mu}_k - \boldsymbol{m}_0\right)^T \Lambda_k \left(\boldsymbol{\mu}_k - \boldsymbol{m}_0\right) + \left(\nu_o - D - 1\right) \ln |\Lambda_k| - Tr\left|\boldsymbol{W}_0^{-1} \Lambda_k\right| \right) + const$$

where we have used $p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})^{z_{nk}}$, $p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{k=1}^{K} \mathcal{N}(\mu_k | \boldsymbol{m}_0, (\beta_0 \boldsymbol{\Lambda}_k)^{-1}) \mathcal{W}(\boldsymbol{\Lambda}_k | \boldsymbol{W}_0, \nu_0)$ and $\mathbb{E}[z_{nk}] = r_{nk}$ together with the definitions for the Gaussian and Wishart distributions; the last term summarizes terms independent of Λ_k. Using $N_k = \sum_{n=1}^{N} r_{nk}$, $\bar{\boldsymbol{x}}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} \boldsymbol{x}_n$, $\boldsymbol{S}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} (\boldsymbol{x}_n - \bar{\boldsymbol{x}}_k)(\boldsymbol{x}_n - \bar{\boldsymbol{x}}_k)^T$, we can rewrite this as shown next.

MAP Estimate Vs MLE

$$\mathbb{E}_{q(\boldsymbol{Z})}[\ln p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})p(\boldsymbol{\mu},\boldsymbol{\Lambda})] = \frac{1}{2}\sum_{n=1}^{N} r_{kn}\left(\ln|\boldsymbol{\Lambda}_k| - (\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T\boldsymbol{\Lambda}_k(\boldsymbol{x}_n - \boldsymbol{\mu}_k)\right)$$

$$+\frac{1}{2}(\ln|\boldsymbol{\Lambda}_k| - \beta_0(\boldsymbol{\mu}_k - m_0)^T\boldsymbol{\Lambda}_k(\boldsymbol{\mu}_k - \boldsymbol{m}_0) + (v_o - D - 1)\ln|\boldsymbol{\Lambda}_k| - Tr|\boldsymbol{W}_0^{-1}\boldsymbol{\Lambda}_k|) + const$$

The terms that involve only. $\boldsymbol{\Lambda}_k$ are as follows:

$$(v_o + N_k - D)\ln|\boldsymbol{\Lambda}_k| - Tr[(\boldsymbol{W}_0^{-1} + \beta_0(\boldsymbol{\mu}_k - \boldsymbol{m}_0)(\boldsymbol{\mu}_k - \boldsymbol{m}_0)^T + N_k(\bar{\boldsymbol{x}}_k - \boldsymbol{\mu}_k)(\bar{\boldsymbol{x}}_k - \boldsymbol{\mu}_k)^T + N_k\boldsymbol{S}_k)\boldsymbol{\Lambda}_k] + const$$

Here we used $N_k\boldsymbol{S}_k = \sum_{n=1}^{N} r_{nk}\boldsymbol{x}_n\boldsymbol{x}_n^T - N_k\bar{\boldsymbol{x}}_k\bar{\boldsymbol{x}}_k^T$ Using $\frac{\partial}{\partial \boldsymbol{A}}Tr(\boldsymbol{AB}) = \boldsymbol{B}^T$, $\frac{\partial}{\partial \boldsymbol{A}}\ln|\boldsymbol{A}| = \left(\boldsymbol{A}^{-1}\right)^T$ we can compute the derivative of this w.r.t. $\boldsymbol{\Lambda}_k$ and setting the result equal to zero, we find the MAP estimate for $\boldsymbol{\Lambda}_k$ to be

$$\widehat{\boldsymbol{\Lambda}}_k^{-1} = \frac{1}{(v_o + N_k - D)}(\boldsymbol{W}_0^{-1} + \beta_0(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{m}_0)(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{m}_0)^T + N_k(\bar{\boldsymbol{x}}_k - \widehat{\boldsymbol{\mu}}_k)(\bar{\boldsymbol{x}}_k - \widehat{\boldsymbol{\mu}}_k)^T + N_k\boldsymbol{S}_k)$$

Here we use for the MLE estimate $\widehat{\boldsymbol{\mu}}_k = \boldsymbol{m}_k$ and $\boldsymbol{W}_k^{-1} = \boldsymbol{W}_0^{-1} + N_k\boldsymbol{S}_k + \frac{\beta_0 N_k}{\beta_0 + N_k}(\bar{\boldsymbol{x}}_k - \boldsymbol{m}_0)(\bar{\boldsymbol{x}}_k - \boldsymbol{m}_0)^T$

From this we see that $|\widehat{\boldsymbol{\Lambda}}_k^{-1}|$ can never become 0, because of the presence of $\boldsymbol{W}_0^{-1}$ (which we must chose to be positive definite) in the expression on the r.h.s. Note this result is not the same as the mode of

$$mode\left[q^*(\boldsymbol{\Lambda}_k)\right] = (\nu_k - D - 1)\boldsymbol{W}_k$$

Selecting the Number of Mixture Components

Determining the Number of Components

Consider a mixture of two Gaussians and a single observed variable x.

Case A: the parameters have the values $\pi_1 = a$, $\pi_2 = b$, $\mu_1 = c$, $\mu_2 = d$, $\sigma_1 = e$, $\sigma_2 = f$.

Case B: the parameter values $\pi_1 = b$, $\pi_2 = a$, $\mu_1 = d$, $\mu_2 = c$, $\sigma_1 = f$, $\sigma_2 = e$, in which the two components have been exchanged,

Both cases by symmetry give rise to the same value of $p(x)$.

If we have a mixture model comprising K components, then each parameter setting will be a member of a family of $K!$ equivalent settings.

The lower bound needs to be modified somewhat to take into account the lack of identifiability of the parameters.

Although VB will approximate the volume occupied by the parameter posterior, it will only do so around one of the local modes.

With K components, there are $K!$ equivalent modes, which differ merely by permuting the labels.

Therefore we should use $\log p(\boldsymbol{X}|K) \approx \mathcal{L}(K) + \log(K!)$.

Determining the Number of Components

In the context of MLE this redundancy is irrelevant because the parameter optimization algorithm (e.g. EM) will, depending on the initialization of the parameters, find one specific solution, and the other equivalent solutions play no role.

In a Bayesian setting, however, we marginalize over all possible parameter values. Variational inference based on the minimization of KL($q||p$) approximates the distribution in the neighbourhood of one of the modes and ignores the others.

Again, because equivalent modes have equivalent predictive densities, this is of no concern provided we are considering a model having a specific number K of components. 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.

$$\log p(\mathbf{X}|K) \approx \mathcal{L}(K) + \log(K!)$$

Determining the Number of Components

We have seen that the variational lower bound can be used to determine a posterior distribution over the number K of components in the mixture model.

The simplest way to select K when using VB is to fit several models, and then to use the variational lower bound to the log marginal likelihood, $\mathcal{L}(K) \leq \log p(\boldsymbol{X}|K)$, to approximate

$$q(k) \propto p(k)\exp\left(\mathcal{L}(k)\right)$$

$$where: \mathcal{L}(k) = \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}|k)\ln\frac{p(\boldsymbol{Z},\boldsymbol{X}|k)}{q(\boldsymbol{Z}|k)}$$

or without the contribution from the prior (all models with the same prior probability):

$$p(K|\boldsymbol{X}) \approx \frac{e^{\mathcal{L}(K)}}{\sum_{K'} e^{\mathcal{L}(K')}}$$

The lower bound should be modified $\mathcal{L}(K) + \log(K!)$ to take into account the lack of identifiability of the parameters.

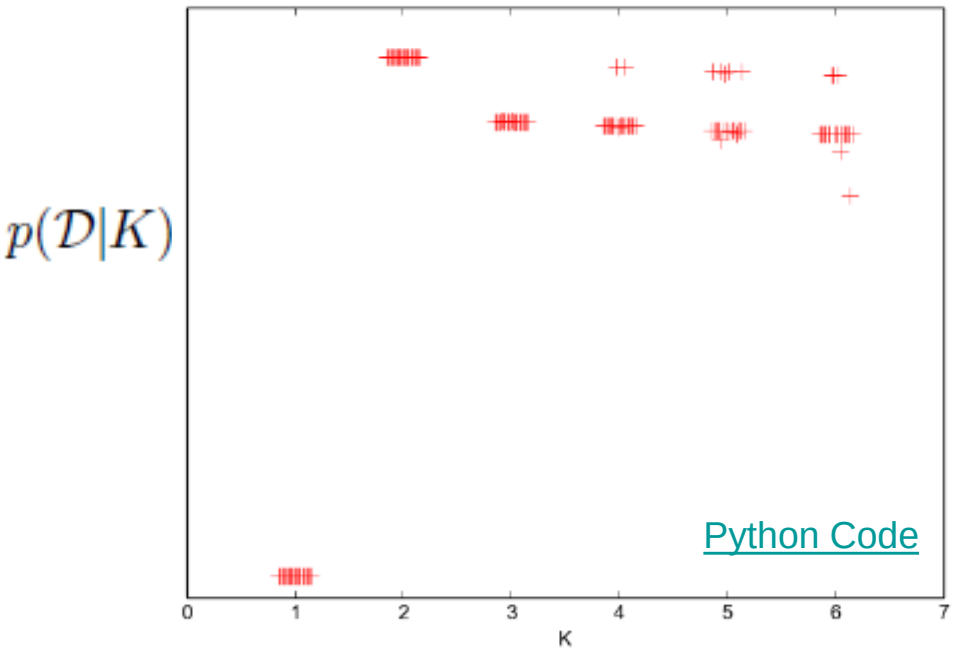
.

Determining the Number of Components

The Fig. shows a plot of the lower bound, including the multimodality factor, versus the number *K* of components for the Old Faithful data set. K=2 is the maximum of the lower bound.

MLE leads to values of the likelihood function that increase monotonically with *K* (assuming the singular solutions have been avoided, and discounting the effects of local maxima) and so cannot be used to determine an appropriate model complexity.

By contrast, Bayesian inference automatically makes the trade-off between model complexity and fitting the data.



For each *K*, the model is trained from 100 different random starts (sample from $p(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\alpha_0)$), and the results are shown as '+' with small random horizontal perturbations so that they can be distinguished.

Some solutions find suboptimal local maxima.

Determining the Number of Components

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 as in the fully Bayesian approach.

When we are treating $\boldsymbol{\pi}$ as a parameter, there is neither a prior, nor a variational posterior distribution, over . Therefore, the only term remaining from the lower bound

$$\mathcal{L} = \mathbb{E}\left[\ln p(\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})\right] + \mathbb{E}\left[\ln p(\boldsymbol{Z} \mid \boldsymbol{\pi})\right] + \mathbb{E}\left[\ln p(\boldsymbol{\pi})\right] + \mathbb{E}\left[\ln p(\boldsymbol{\mu}, \boldsymbol{\Lambda})\right] - \mathbb{E}\left[\ln q(\boldsymbol{Z})\right] - \mathbb{E}\left[\ln q(\boldsymbol{\pi})\right] - \mathbb{E}\left[\ln q(\boldsymbol{\mu}, \boldsymbol{\Lambda})\right]$$

that involves $\boldsymbol{\pi}$ is the second term.

Note however, that $\mathbb{E}[\ln p(\boldsymbol{Z}|\boldsymbol{\pi})] = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk} \ln \tilde{\pi}_k$ involves the of expectation of $\ln\pi_k$ under $q(\boldsymbol{\pi})$,

whereas here, we operate directly with π_k, yielding

$$\mathbb{E}_{q(\boldsymbol{Z})}\left[\ln p(\boldsymbol{Z} \mid \boldsymbol{\pi})\right] = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk} \ln \pi_k$$

Using a Lagrange multiplier to enforce the constraint $\sum_{k=1}^{K} \pi_k = 1$ leads to $N_k/\pi_k + \lambda = 0$ or

$N_k + \lambda\pi_k = 0$ and summing over k to $-\lambda = N$ and thus: $\pi_k = N_k/N$.

- Corduneanu, A. and C. M. Bishop (2001). Variational Bayesian model selection for mixture distributions. In T. Richardson and T. Jaakkola (Eds.), *Proceedings Eighth International Conference on Artificial Intelligence and Statistics*, pp. 27–34. Morgan Kaufmann.

Determining the Number of Components

This leads to the re-estimation equation

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

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 through *automatic relevance determination*.

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.

- Corduneanu, A. and C. M. Bishop (2001). Variational Bayesian model selection for mixture distributions. In T. Richardson and T. Jaakkola (Eds.), *Proceedings Eighth International Conference on Artificial Intelligence and Statistics*, pp. 27–34. Morgan Kaufmann.

Selecting the Number of Mixture Components

The factorized assumption causes the variance of the posterior distribution to be under-estimated.

As the number of mixture components K grows, so does the number of variables that may be correlated (which are treated as independent under a variational approximation)

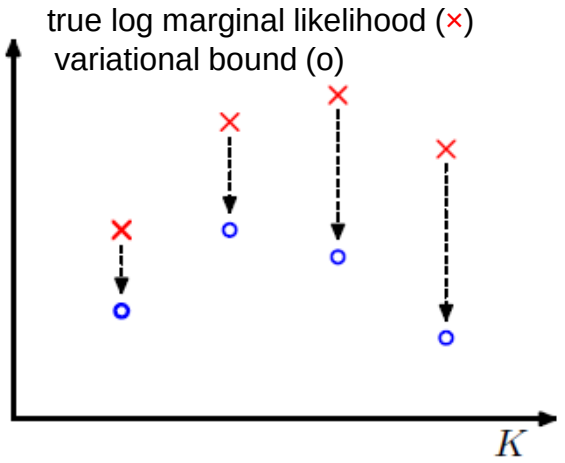
Thus, the proportion of probability mass under the true distribution, $p(\mathbf{Z},\pi,\boldsymbol{\mu},\Sigma|\mathbf{X})$, that the variational approximation $q(\mathbf{Z},\pi,\boldsymbol{\mu},\Sigma)$ does not capture, will grow.

The consequence will be that the KL divergence between $q(\mathbf{Z},\pi,\boldsymbol{\mu},\Sigma)$ and $p(\mathbf{Z},\pi,\boldsymbol{\mu},\Sigma|\mathbf{X})$ will increase.

Thus the lower bound must decrease compared to the true log marginal. Thus choosing the number of components based on the lower bound will tend to underestimate the optimal number of components.

The dashed arrows emphasize the typical increase in the difference between the true log marginal likelihood and the bound.

As a consequence, the bound tends to have its peak at a lower value of K than the true log marginal likelihood.



Exponential Family Distributions

Exponential Family Distributions

For many models the complete-data likelihood is drawn from the exponential family. However, in general this will not be the case for the marginal likelihood function for the observed data.

E.g. in a mixture of Gaussians, the joint distribution of observations $\mathbf{x}_n$ and hidden variables $\mathbf{z}_n$ is a member of the exponential family but the marginal of $\mathbf{x}_n$ is a Gaussian mixture.

Here we make a distinction between latent variables, $\mathbf{Z}$, and parameters, $\boldsymbol{\theta}$, where parameters are *intensive* (fixed in number independent of the size of the data set), whereas latent variables are *extensive* (scale in number with the size of the data set).

E.g., in a Gaussian mixture model, the indicator variables z_{kn} (which specify which component k is responsible for generating data point $\mathbf{x}_n$) represent the latent variables, whereas the means $\boldsymbol{\mu}_k$, precisions $\boldsymbol{\Lambda}_k$ and mixing proportions π_k represent the parameters.

Consider i.i.d. data $\mathbf{X} = \{\mathbf{x}_n\}$, $n = 1, \ldots N$, with latent variables $\mathbf{Z} = \{\mathbf{z}_n\}$. Let the joint distribution of observed and latent variables be a member of the exponential family, parameterized by natural parameters $\boldsymbol{\eta}$. Using a conjugate prior $p(\boldsymbol{\eta}|\boldsymbol{\chi}_0, \mathrm{v}_0)$, we write:

$$p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\eta}) = \prod_{n=1}^{N} h(\boldsymbol{x}_n, \boldsymbol{z}_n) g(\boldsymbol{\eta}) \exp\{\boldsymbol{\eta}^T \boldsymbol{u}(\boldsymbol{x}_n, \boldsymbol{z}_n)\}$$

$$p(\boldsymbol{\eta}|\boldsymbol{\chi}_0, \mathrm{v}_0) = f(\boldsymbol{\chi}_0, \mathrm{v}_0) g(\boldsymbol{\eta})^{\mathrm{v}0} \exp\{v_0 \boldsymbol{\eta}^T \boldsymbol{\chi}_0\}$$

Exponential Family Distributions

Consider a variational distribution $q(\mathbf{Z}, \boldsymbol{\eta}) = q(\mathbf{Z})q(\boldsymbol{\eta})$. We solve for the two factors as follows

$$\ln q^*(\mathbf{Z}) = \mathbb{E}_{\boldsymbol{\eta}}[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\eta})] + \text{const} = \sum_{n=1}^{N} \{\ln h(\mathbf{x}_n, \mathbf{z}_n) + \mathbb{E}[\boldsymbol{\eta}^T] u(\mathbf{x}_n, \mathbf{z}_n)\} + \text{const}$$

So it factorizes as : $q^*(\mathbf{Z}) = \prod_{n=1}^{N} q^*(\mathbf{z}_n)$. This is an induced factorization leading (after normalization) to:

$$q^*(\mathbf{z}_n) = h(\mathbf{x}_n, \mathbf{z}_n) g(\mathbb{E}[\boldsymbol{\eta}]) \exp\{\mathbb{E}[\boldsymbol{\eta}^T] u(\boldsymbol{x}_n, \boldsymbol{z}_n)\}$$

Similarly:

$$\ln q^*(\boldsymbol{\eta}) = \ln p(\boldsymbol{\eta}|\boldsymbol{\chi}_0, \nu_0) + \mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\eta})] + \text{const}$$
$$= \nu_0 \ln g(\boldsymbol{\eta}) + \nu_0 \boldsymbol{\eta}^T \boldsymbol{\chi}_0 + \sum_{n=1}^{N} \{\ln g(\boldsymbol{\eta}) + \boldsymbol{\eta}^T \mathbb{E}_{\mathbf{z}_n}[u(\mathbf{x}_n, \mathbf{z}_n)]\} + \text{const}$$

From this after normalization we compute:

$$q^*(\boldsymbol{\eta}) = f(\boldsymbol{\chi}_N, \nu_N) g(\boldsymbol{\eta})^{\nu_N} \exp\{\nu_N \boldsymbol{\eta}^T \boldsymbol{\chi}_N\}$$
$$\nu_N = \nu_0 + N, \; \nu_N \boldsymbol{\chi}_N = \nu_0 \boldsymbol{\chi}_0 + \sum_{n=1}^{N} \mathbb{E}_{\mathbf{z}_n}[u(\mathbf{x}_n, \mathbf{z}_n)]$$

$q^*(\mathbf{z}_n)$ and $q^*(\boldsymbol{\eta})$ are coupled and we solve them iteratively.

- E step: compute $\mathbb{E}_{\mathbf{z}_n}[u(\mathbf{x}_n, \mathbf{z}_n)]$ using $q(\mathbf{z}_n)$ and use this to compute a revised $q(\boldsymbol{\eta})$.
- M step: use $q(\boldsymbol{\eta})$ to find $\mathbb{E}[\boldsymbol{\eta}^T]$, which gives rise to a revised $q^*(\mathbf{z}_n)$.

Exponential Family and the Mixture of Gaussians

Rewrite the model for the Bayesian mixture of Gaussians as a conjugate model from the exponential family.

Hence use the general results

$$\ln q^*(\mathbf{Z}) = \mathbb{E}_{\boldsymbol{\eta}}[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\eta})] + \text{const} = \sum_{n=1}^{N} \{\ln h(\mathbf{x}_n, \mathbf{z}_n) + \mathbb{E}[\boldsymbol{\eta}^T] u(\mathbf{x}_n, \mathbf{z}_n)\} + \text{const}$$

$$q^*(\boldsymbol{\eta}) = f(\boldsymbol{\chi}_N, \nu_N) g(\boldsymbol{\eta})^{\nu_N} \exp\{\nu_N \boldsymbol{\eta}^T \boldsymbol{\chi}_N\}, \quad \nu_N = \nu_0 + N, \quad \nu_N \boldsymbol{\chi}_N = \nu_0 \boldsymbol{\chi}_0 + \sum_{n=1}^{N} \mathbb{E}_{\mathbf{z}_n}[u(\mathbf{x}_n, \mathbf{z}_n)]$$

to derive the specific results

$$q^*(\boldsymbol{Z}) = \prod_{n=1}^{N} \prod_{k=1}^{K} r_{nk}^{z_{nk}}$$

$$q^*(\boldsymbol{\pi}) = \text{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) \text{ where } \alpha_k = \alpha_0 + N_k$$

$$q^*(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) = \mathcal{N}(\boldsymbol{\mu}_k|\boldsymbol{m}_k, (\beta_k \boldsymbol{\Lambda}_k)^{-1}) \mathcal{W}(\boldsymbol{\Lambda}_k|\boldsymbol{W}_k, \nu_k)$$

where we have defined:

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

$$\nu_k = \nu_0 + N_k$$

Exponential Family and the Mixture of Gaussians

We start with the complete data log-likelihood using $p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\Lambda}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k^{-1})^{z_{nk}}$ and $p(\boldsymbol{Z}|\boldsymbol{\pi}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\pi_k^{z_{nk}}$:

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

$$= \prod_{n=1}^{N} exp\left\{\sum_{k=1}^{K} z_{nk}\left(ln\pi_k + \frac{1}{2}ln|\boldsymbol{\Lambda}_k| - \frac{D}{2}\ln(2\pi) - \frac{1}{2}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^T\boldsymbol{\Lambda}_k(\boldsymbol{x}_n-\boldsymbol{\mu}_k)\right)\right\}$$

Exploring the similarities with $p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\eta}) = \prod_{n=1}^{N} h(\boldsymbol{x}_n,\boldsymbol{z}_n)g(\boldsymbol{\eta})\exp\{\boldsymbol{\eta}^T\boldsymbol{u}(\boldsymbol{x}_n,\boldsymbol{z}_n)\}$, we can write by inspection:

$$\eta = \begin{pmatrix} \boldsymbol{\Lambda}_k\boldsymbol{\mu}_k \\ \overrightarrow{\boldsymbol{\Lambda}_k} \\ \boldsymbol{\mu}_k^T\boldsymbol{\Lambda}_k\boldsymbol{\mu}_k \\ ln|\boldsymbol{\Lambda}_k| \\ ln\pi_k \end{pmatrix}_{k=1,\dots,K}, \boldsymbol{u}(\boldsymbol{x}_n,\boldsymbol{z}_n) = \begin{pmatrix} z_{nk}\begin{pmatrix} \boldsymbol{x_n} \\ \frac{1}{2}\overrightarrow{\boldsymbol{x_n}\boldsymbol{x_n^T}} \\ -\frac{1}{2} \\ \frac{1}{2} \\ 1 \end{pmatrix} \end{pmatrix}_{k=1,\dots,K}, \boldsymbol{h}(\boldsymbol{x}_n,\boldsymbol{z}_n) = \prod_{k=1}^{K}\left((2\pi)^{-D/2}\right)^{z_{nk}}, g(\boldsymbol{\eta}) = 1$$

Arrows above matrices return a vector formed by stacking the columns of the matrix on top of each other.

Also $(\boldsymbol{v}_k)_{k=1,\dots,K} = \begin{pmatrix} \boldsymbol{v}_1 \\ \boldsymbol{v}_2 \\ \vdots \\ \vdots \\ \boldsymbol{v}_k \end{pmatrix}$

Exponential Family and the Mixture of Gaussians

Similarly, the prior over the parameters can be written as:

$$p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = Dir(\boldsymbol{\pi}|a_0) \prod_{k=1}^{K} \mathcal{N}\left(\boldsymbol{\mu}_k | \boldsymbol{m}_0, \left(\beta_0 \boldsymbol{\Lambda}_k\right)^{-1}\right) W\left(\boldsymbol{\Lambda}_k | \boldsymbol{W}_0, \nu_0\right)$$

$$= \mathrm{C}(a_0) \left(\frac{\beta_0}{2\pi}\right)^{KD/2} B(\boldsymbol{W}_0, \nu_0)^K exp\left\{\sum_{k=1}^{K} (\alpha_0 - 1) ln\pi_k + \frac{\nu_0 - D}{2} ln\left|\boldsymbol{\Lambda}_k\right| - \frac{1}{2} Tr\left(\boldsymbol{\Lambda}_k \left[\beta_0 (\boldsymbol{\mu}_k\right.\right.\right.$$

Exponential Family and the Mixture of Gaussians

By exponentiating both sides of

$$\ln q^*(\mathbf{Z}) = \sum_{n=1}^{N} \{\ln h(\mathbf{x}_n, \mathbf{z}_n) + \mathbb{E}[\boldsymbol{\eta}^T] u(\mathbf{x}_n, \mathbf{z}_n)\} + \text{const},$$

and using

$$\eta = \begin{pmatrix} \boldsymbol{\Lambda}_k \boldsymbol{\mu}_k \\ \overrightarrow{\boldsymbol{\Lambda}_k} \\ \boldsymbol{\mu}_k^T \boldsymbol{\Lambda}_k \boldsymbol{\mu}_k \\ ln|\boldsymbol{\Lambda}_k| \\ ln\pi_k \end{pmatrix}_{k=1,\dots,K} , \boldsymbol{u}(\boldsymbol{x}_n, \boldsymbol{z}_n) = \begin{pmatrix} z_{nk} \begin{pmatrix} \boldsymbol{x_n} \\ \frac{1}{2} \overrightarrow{\boldsymbol{x_n}\, \boldsymbol{x_n^T}} \\ -\frac{1}{2} \\ \frac{1}{2} \\ 1 \end{pmatrix} \end{pmatrix}_{k=1,\dots,K} , \boldsymbol{h}(\boldsymbol{x}_n, \boldsymbol{z}_n) = \prod_{k=1}^{K} \left((2\pi)^{-D/2}\right)^{z_{nk}}, g(\boldsymbol{\eta}) = 1$$

we obtain

$$q^*(\boldsymbol{Z}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \rho_{nk}^{z_{nk}} \text{ with } \ln \rho_{nk} = \mathbb{E}[\ln \pi_k] + \frac{1}{2} \mathbb{E}[\ln |\boldsymbol{\Lambda}_k|] - \frac{D}{2} \ln(2\pi) - \frac{1}{2} \mathbb{E}_{\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k}[(\boldsymbol{x}_n -$$

Exponential Family and the Mixture of Gaussians

Next we can use $\mathbb{E}[z_{nk}] = r_{nk}$ to take the expectation wrt **Z** in $\mathrm{v}_N \boldsymbol{\chi}_N = \mathrm{v}_0 \boldsymbol{\chi}_0 + \sum_{n=1}^{N} \mathbb{E}_{\mathbf{z}_n}[\mathrm{u}(\mathbf{x}_n, \mathbf{z}_n)]$, substituting r_{nk} for $\mathbb{E}[z_{nk}]$ in

$$\boldsymbol{u}(\boldsymbol{x}_n, \boldsymbol{z}_n) = \left(z_{nk} \begin{pmatrix} \boldsymbol{x_n} \\ \frac{1}{2} \overrightarrow{\boldsymbol{x_n}\, \boldsymbol{x_n^T}} \\ -\frac{1}{2} \\ \frac{1}{2} \\ 1 \end{pmatrix} \right)_{k=1,\ldots,K}$$

Combining this with $\boldsymbol{\chi}_0 = \begin{pmatrix} \beta_0 \boldsymbol{m}_0 \\ -\frac{1}{2}\left(\overrightarrow{\beta_0 \boldsymbol{m}_0\, \boldsymbol{m}_0^T} + \overrightarrow{\boldsymbol{W}_0^{-1}}\right) \\ -\frac{\beta_0}{2} \\ \frac{\nu_0 - D}{2} \\ \alpha_0 - 1 \end{pmatrix}_{k=1,\ldots,K}$, Eqs.

$\mathrm{v}_N = \mathrm{v}_0 + \mathrm{N}$,
$\mathrm{v}_N \boldsymbol{\chi}_N = \mathrm{v}_0 \boldsymbol{\chi}_0 + \sum_{n=1}^{N} \mathbb{E}_{\mathbf{z}_n}[\mathrm{u}(\mathbf{x}_n, \mathbf{z}_n)]$ become as follows:

Exponential Family and the Mixture of Gaussians

$v_N = v_0 + N = 1 + N$ and ,

$$v_N \boldsymbol{\chi}_N = \begin{pmatrix} \beta_0 \boldsymbol{m}_0 \\ -\frac{1}{2}\left(\overrightarrow{\beta_0 \boldsymbol{m}_0\, \boldsymbol{m}_0^T} + \overrightarrow{\boldsymbol{W}_0^{-1}}\right) \\ -\frac{\beta_0}{2} \\ \frac{\nu_0 - D}{2} \\ \alpha_0 - 1 \end{pmatrix}_{k=1,\ldots,K} + \sum_{n=1}^{N} \begin{pmatrix} r_{nk} \begin{pmatrix} \boldsymbol{x_n} \\ \frac{1}{2}\overrightarrow{\boldsymbol{x_n}\, \boldsymbol{x_n^T}} \\ -\frac{1}{2} \\ \frac{1}{2} \\ 1 \end{pmatrix} \end{pmatrix}_{k=1,\ldots,K} =$$

$$\begin{pmatrix} \beta_0 \boldsymbol{m}_0 + N_k \bar{x}_k \\ -\frac{1}{2}\left(\overrightarrow{\beta_0 \boldsymbol{m}_0\, \boldsymbol{m}_0^T} + \overrightarrow{\boldsymbol{W}_0^{-1}} + \boldsymbol{N_k}\overrightarrow{(\boldsymbol{S_k} + \bar{x}_k \bar{x}_k^T)}\right) \\ -\frac{\beta_0 + N_k}{2} \\ \frac{\nu_0 - D + N_k}{2} \\ \alpha_0 - 1 + N_k \end{pmatrix}_{k=1,\ldots,K}$$

From the bottom row of $\boldsymbol{u}(\boldsymbol{x}_n, \boldsymbol{z}_n)$ and the Eq. above, we see that the inner product of η and $v_N \boldsymbol{\chi}_N$ gives us the r.h.s. of $\ln q^*(\boldsymbol{\pi}) = (\alpha_0 - 1)\sum_{k=1}^{K} \ln \pi_k + \sum_{k=1}^{K}\sum_{n=1}^{N} r_{nk} \ln \pi_k + const$ from which $\ln q^*(\boldsymbol{\pi}) = (\alpha_0 - 1)\sum_{k=1}^{K} \ln \pi_k + N_k \sum_{k=1}^{K} \ln \pi_k + const$ i.e. $q^*(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha})$ where $\alpha_k = \alpha_0 + N_k$

Exponential Family and the Mixture of Gaussians

The remaining terms of the inner product are:

$$\sum_{k=1}^{K}\left\{\boldsymbol{\mu}_k^T\boldsymbol{\Lambda}_k\left(\beta_0\boldsymbol{m}_0 + N_k\bar{\boldsymbol{x}}_k\right) - \frac{1}{2}Tr\left(\boldsymbol{\Lambda}_k\left[\overrightarrow{\beta_0\boldsymbol{m}_0\,\boldsymbol{m}_0^{\boldsymbol{T}}} + \overrightarrow{\boldsymbol{W}_0^{-\mathbf{1}}} + \boldsymbol{N_k}\overrightarrow{(\boldsymbol{S_k} + \bar{x}_k\bar{x}_k^T)}\right]\right)\right.$$

Exponential Family and the Mixture of Gaussians

The remaining terms of the inner product are:

$$-\frac{1}{2}\beta_k \boldsymbol{\mu}_k^T \boldsymbol{\Lambda}_k \boldsymbol{\mu}_k + \beta_k \boldsymbol{\mu}_k^T \boldsymbol{\Lambda}_k \boldsymbol{m}_k - \frac{1}{2}\beta_k \boldsymbol{m}_k^T \boldsymbol{\Lambda}_k \boldsymbol{m}_k + \frac{1}{2} ln|\boldsymbol{\Lambda}_k| + \frac{1}{2}\beta_k \boldsymbol{m}_k^T \boldsymbol{\Lambda}_k \boldsymbol{m}_k$$
$$-\frac{1}{2} Tr\left(\boldsymbol{\Lambda}_k \left[\overrightarrow{\beta_0 \boldsymbol{m}_0 \boldsymbol{m}_0^T} + \overrightarrow{\boldsymbol{W}_0^{-1}} + \overrightarrow{\boldsymbol{N_k}(\boldsymbol{S_k} + \bar{x}_k \bar{x}_k^T)}\right]\right) + \frac{1}{2}(\nu_k - D - 1) ln|\boldsymbol{\Lambda}_k|$$

To make the remaining terms match the logarithm of $\boldsymbol{W}(\boldsymbol{\Lambda}_k | \boldsymbol{W}_k, \nu_k)$ we need to show that $\beta_0 \boldsymbol{m}_0 \boldsymbol{m}_0^T + N_k \bar{x}_k \bar{x}_k^T - \beta_k \boldsymbol{m}_k \boldsymbol{m}_k^T$ equals the last term on the r.h.s. of

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

Using $\beta_k = \beta_0 + N_k$, $\boldsymbol{m}_k = \frac{1}{\beta_k}(\beta_0 \boldsymbol{m}_0 + N_k \bar{\boldsymbol{x}}_k)$, and

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

we get:

$$\beta_0 \boldsymbol{m}_0 \boldsymbol{m}_0^T + N_k \bar{x}_k \bar{x}_k^T - \beta_k \boldsymbol{m}_k \boldsymbol{m}_k^T = \frac{\beta_0 N_k}{\beta_0 + N_k}(\bar{\boldsymbol{x}}_k - \boldsymbol{m}_0)(\bar{\boldsymbol{x}}_k - \boldsymbol{m}_0)^T.$$

Thus we recovered $\boldsymbol{W}(\boldsymbol{\Lambda}_k | \boldsymbol{W}_k, \nu_k)$ with missing terms accounted by in $f(\nu_N, \boldsymbol{\chi}_N)$.

Variational Message Passing

Variational Message Passing

Here we consider more generally the use of variational methods for models described by directed graphs and derive a number of widely applicable results.

The joint distribution corresponding to a directed graph can be written using the decomposition

$$p(\boldsymbol{x}) = \prod_i \; p(\boldsymbol{x}_i | pa_i)$$

where $\boldsymbol{x}_i$ denotes the variable(s) associated with node *i*, and pa_i denotes the parent set corresponding to node *i*. Note that $\boldsymbol{x}_i$ may be a latent variable or it may belong to the set of observed variables.

Now consider a variational approximation in which the distribution $q(\mathbf{x})$ is assumed to factorize with respect to the $\boldsymbol{x}_i$ so that

$$q(\boldsymbol{x}) = \prod_i \; q_i(\boldsymbol{x}_i)$$

Note that for observed nodes, there is no factor $q_i(\boldsymbol{x}_i)$ in the variational distribution. Using our general result to give

$$lnq_j^*(\boldsymbol{x}_j) = \mathbb{E}_{i \neq j}\left[\sum_i \; lnp(\boldsymbol{x}_i | pa_i)\right] + const.$$

Variational Message Passing

$$\ln q_j^*(\boldsymbol{x}_j) = \mathbb{E}_{i \neq j}\left[\sum_i \ln p(\boldsymbol{x}_i \mid pa_i)\right] + const.$$

The only terms that do depend on $\boldsymbol{x}_j$ are $p(\boldsymbol{x}_j \mid \text{pa}_j)$ together with the conditional distributions corresponding to the children of node *j*, and they therefore also depend on the *co-parents* of the child nodes, i.e., the other parents of the child nodes besides node $\boldsymbol{x}_j$ itself.

The set of all nodes on which $q(\boldsymbol{x}_j)$ depends correspond to the Markov blanket of node $\boldsymbol{x}_j$.

Thus the update of the factors in the variational posterior distribution represents a local calculation on the graph.

This makes possible the construction of general purpose software for variational inference in which the form of the model does not need to be specified in advance (Bishop *et al.*, 2003).

If we now specialize to the case of a model in which all of the conditional distributions have a conjugate-exponential structure, then the variational update procedure can be cast in terms of a local message passing algorithm (Winn and Bishop, 2005).

- Winn, J. and C. M. Bishop (2005). Variational message passing. *Journal of Machine Learning Research* **6**, 661–694 (http://vibes.sourceforge.net/tutorial/)
- Bishop, C. M., D. Spiegelhalter, and J. Winn (2003). VIBES: A variational inference engine for Bayesian networks. In S. Becker, S. Thrun, and K. Obermeyer (Eds.), *Advances in Neural Information Processing Systems*, Volume 15, pp. 793–800. MIT Press.

Variational Message Passing

$$\ln q_j^*(\boldsymbol{x}_j) = \mathbb{E}_{i \neq j}\left[\sum_i \ln p(\boldsymbol{x}_i \mid pa_i)\right] + const.$$

The distribution associated with a particular node can be updated once that node has received messages from all of its parents and all of its children.

This in turn requires that the children have already received messages from their coparents.

The evaluation of the lower bound can also be simplified because many of the required quantities are already evaluated as part of the message passing scheme.

This distributed message passing formulation has good scaling properties and is well suited to large networks.

- Winn, J. and C. M. Bishop (2005). Variational message passing. *Journal of Machine Learning Research* **6**, 661–694 (http://vibes.sourceforge.net/tutorial/)
- Bishop, C. M., D. Spiegelhalter, and J. Winn (2003). VIBES: A variational inference engine for Bayesian networks. In S. Becker, S. Thrun, and K. Obermeyer (Eds.), *Advances in Neural Information Processing Systems*, Volume 15, pp. 793–800. MIT Press.