Variational Inference

Jiali Lin

Virginia Tech

December 4, 2016

Outline

Motivation

- **Variational inference:** deterministic approximate inference.
- Pick an approximation $q(\boldsymbol{y})$ that are tractable family, and make this approximation close to the true posterior, $p^*(\boldsymbol{y}) = p(\boldsymbol{y}|\mathcal{D})$.
- This reduces inference to an optimization problem.
- By approximating the objective, we can trade accuracy for speed.
- **Pros**:
  - For small to medium problems, it is usually faster.
  - It is deterministic.
  - Is it easy to determine when to stop.
  - It often provides a lower bound on the log likelihood.

Outline

Introduction

Variational Inference

Variational Bayes

Variational Bayes EM

Variational calculus

Variational inference is based on variational calculus.

Standard Calculus

- Functions $f : \boldsymbol{y} \to f(\boldsymbol{y})$.
- Derivatives $\frac{df}{d\boldsymbol{y}}$.

Example: maximize the likelihood expression $p(\boldsymbol{y}|\theta)$ w.r.t θ.

Variational Calculus

- Functionals $f : \boldsymbol{y} \to F(f)$.
- Derivatives $\frac{dF}{df}$.

Example: maximize the entropy $H[p]$ w.r.t a probability $p(\boldsymbol{y})$.

Variational calculus and the free energy

By appropriate choice of $q(\boldsymbol{\theta})$, $F(q, \boldsymbol{y})$ becomes tractable to compute and maximize. Hence we have both an analytical approximation $q(\boldsymbol{\theta})$ for the posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$ and a lower bound $F(q, \boldsymbol{y})$ for the evidence $\log p(\boldsymbol{y})$.

$$
\begin{aligned}
\ln p(\boldsymbol{y}) &= \ln \frac{p(\boldsymbol{y}, \boldsymbol{\theta})}{p(\boldsymbol{\theta}|\boldsymbol{y})} \\
&= \int q(\boldsymbol{\theta}) \ln \frac{p(\boldsymbol{y}, \boldsymbol{\theta})}{p(\boldsymbol{\theta}|\boldsymbol{y})} d\boldsymbol{\theta} \qquad \xleftarrow{\ln p(\boldsymbol{y}) \text{ does not depend on } \boldsymbol{\theta}} \\
&= \int q(\boldsymbol{\theta}) \ln \frac{p(\boldsymbol{y}, \boldsymbol{\theta})}{p(\boldsymbol{\theta}|\boldsymbol{y})} \frac{q(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} \\
&= \int q(\boldsymbol{\theta}) (\ln \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\boldsymbol{y})} + \ln \frac{p(\boldsymbol{y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})}) d\boldsymbol{\theta} \\
&= \int \underbrace{q(\boldsymbol{\theta}) \ln \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\boldsymbol{y})} d\boldsymbol{\theta}}_{\mathsf{KL}_{[q||p]} \text{ divergence}} + \underbrace{\int q(\boldsymbol{\theta}) \ln \frac{p(\boldsymbol{y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta}}_{F(q, \boldsymbol{y}) \text{ free energy}}
\end{aligned}
\tag{1}
$$

Computing the free energy

- $\mathsf{KL}[q||p]$ divergence is unknown and free energy $F(q, \boldsymbol{y})$ is easy to evaluate for a given q.
- Maximizing $F(q, \boldsymbol{y})$ is equivalent to minimizing $\mathsf{KL}[q||p]$ and tightening $F(q, \boldsymbol{y})$ as a lower bound to (1).

We can decompose the free energy $F(q, \boldsymbol{y})$ as follows

$$
\begin{aligned}
F(q, \boldsymbol{y}) &= \int q(\boldsymbol{\theta}) \ln \frac{p(\boldsymbol{y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} \\
&= \int q(\boldsymbol{\theta}) \ln p(\boldsymbol{y}, \boldsymbol{\theta}) d\boldsymbol{\theta} - \int q(\boldsymbol{\theta}) \ln q(\boldsymbol{\theta}) d\boldsymbol{\theta} \\
&= \underbrace{< \ln p(\boldsymbol{y}, \boldsymbol{\theta}) >_q}_{\text{expected log-joint}} + \underbrace{H[q]}_{\text{Shannon entropy}}
\end{aligned}
\tag{2}
$$

Forward or reverse KL?

KL divergence is not symmetric in its arguments, minimizing $\mathrm{KL}[q||p]$ wrt q will give different behavior than minimizing $\mathrm{KL}[q||p]$.

- **Variational Bayes** minimize $\mathrm{KL}[q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\boldsymbol{y})]$: $q(\boldsymbol{\theta})$ will tend to be zero where $p(\boldsymbol{\theta}|\boldsymbol{y})$ is zero.
- **Expectation Propagation** minimize $\mathrm{KL}[p(\boldsymbol{\theta}|\boldsymbol{y})||q(\boldsymbol{\theta})]$: $q(\boldsymbol{\theta})$ will tend to be nonzero where $p(\boldsymbol{\theta}|\boldsymbol{y})$ is nonzero.

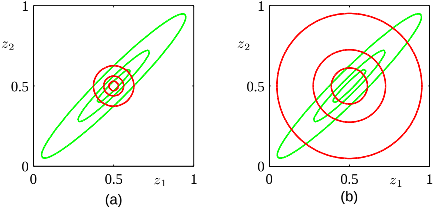
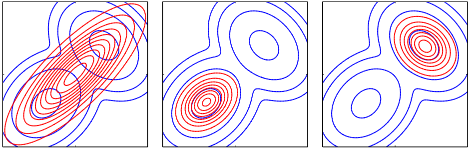


Figure: Comparison of the two alternative forms for the Kullback-Leibler divergence. (a) the Kullback-Leibler divergence $\mathrm{KL}(q||p)$, and (b) the reverse Kullback-Leibler divergence $\mathrm{KL}(p||q)$. Figure generated by `KLpqGauss`.

Forward or reverse KL? (cont'd)



(a) (b) (c)

Figure: Another comparison of the two alternative forms for the Kullback-Leibler divergence. (a) Averaging across modes may lead to poor predictive performance. (b) (c) Variational Bayes may lead to local minimum. Figure generated by `KLfwdReverseMixGauss`.

Mean field approximation

Mean field approximation assumes the posterior is a fully factorized approximation of the form

$$q(\boldsymbol{\theta}) = \prod_i q_i(\boldsymbol{\theta}_i) \tag{3}$$

Derivation of the mean field update equations

$$
\begin{aligned}
F(q, \boldsymbol{y}) &= \int q(\boldsymbol{\theta}) \ln \frac{p(\boldsymbol{y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} \\
&= \int \prod_i q_i \times (\ln p(\boldsymbol{y}, \boldsymbol{\theta}) - \sum_i \ln q_i) d\boldsymbol{\theta} \quad \xleftarrow{\textbf{mean field assumption: } q(\boldsymbol{\theta}) = \prod_i q_i(\boldsymbol{\theta}_i)} \\
&= \int q_j \prod_{\backslash j} q_i (\ln p(\boldsymbol{y}, \boldsymbol{\theta}) - \ln q_i) d\boldsymbol{\theta} - \int q_j \prod_{\backslash j} q_i \sum_{\backslash j} \ln q_i d\boldsymbol{\theta} \\
&= \int q_j (\underbrace{\int \prod_{\backslash j} q_i \ln p(\boldsymbol{y}, \boldsymbol{\theta}) d\boldsymbol{\theta}_{\backslash j}}_{< \ln p(\boldsymbol{y}, \boldsymbol{\theta}) >_{q \backslash j}} - \ln q_i) d\boldsymbol{\theta}_j - \int q_j \int \prod_{\backslash j} q_i \ln \prod_{\backslash j} q_i d\boldsymbol{\theta}_{\backslash j} d\boldsymbol{\theta}_j \\
&= \int q_j \ln \frac{\exp(< \ln p(\boldsymbol{y}, \boldsymbol{\theta}) >_{q \backslash j})}{q_j} d\boldsymbol{\theta}_j + c \quad \xleftarrow{\exp(< \ln p(\boldsymbol{y}, \boldsymbol{\theta}) >_{q \backslash j}) = E_{\backslash j}[\ln p(\boldsymbol{y}, \boldsymbol{\theta})]} \\
&= -\mathsf{KL}[q_j || \exp(< \ln p(\boldsymbol{y}, \boldsymbol{\theta}) >_{q \backslash j})] + c
\end{aligned}
\tag{4}
$$

Derivation of the mean field update equations(cont'd)

Suppose the densities $q_{\setminus j} = q(\boldsymbol{\theta}_{\setminus j})$ are kept fixed. Then the approximate posterior $q(\boldsymbol{\theta}_j)$ that maximizes $F(q, \boldsymbol{y})$ is given by

$$\begin{aligned} q_j^* &= \max_{q_j} \; F(q, \boldsymbol{y}) \\ &= \frac{1}{Z} \exp(< \ln p(\boldsymbol{y}, \boldsymbol{\theta}) >_{q \setminus j}) \end{aligned} \tag{5}$$

Therefore:

$$\ln q_j^* =< \ln p(\boldsymbol{y}, \boldsymbol{\theta}) >_{q \setminus j} - \ln Z \tag{6}$$

where $Z = \int < \ln p(\boldsymbol{y}, \boldsymbol{\theta}) >_{q \setminus j} d\boldsymbol{\theta}_j$.

This implies a straightforward algorithm for variational inference:

1. Initialize all approximate posteriors $q(\boldsymbol{\theta}_i)$, e.g., by setting them to their priors.
2. Cycle over the parameters, revising each given the current estimates of the others.
3. Loop until convergence.

Outline

Introduction

- So far we are inferring latent variables $\boldsymbol{z}_i$ assuming the parameters $\boldsymbol{\theta}$ of the model are known.
- Now infer the parameters themselves.
- Mean field approximation

$$p(\boldsymbol{\theta}|\mathcal{D}) \approx \prod_i q_i(\boldsymbol{\theta}_i) \tag{7}$$

- This is **variational Bayes** or **VB**.
- If we want to infer both latent variables and parameters, then

$$p(\boldsymbol{\theta}, \boldsymbol{z}_{1:N}|\mathcal{D}) \approx q(\boldsymbol{\theta}) \prod_i q_i(\boldsymbol{z}_i) \tag{8}$$

Example: VB for a univariate Gaussian

- Assuming there are no latent variables.
- Consider applying VB to infer the posterior over the parameters for a 1d Gaussian, $p(\mu, \lambda|\mathcal{D})$, where $\lambda = 1/\sigma^2$ is the precision.
- For convenience, we will use a conjugate prior of the form.

$$\begin{aligned} p(\mu, \lambda) &= p(\mu|\lambda)p(\lambda) \\ &= N(\mu|\mu_0, (\kappa_0\lambda)^{-1})\mathsf{Ga}(\lambda|a_0, b_0) \end{aligned} \tag{9}$$

- Consider a factorized variational approximation

$$q(\mu, \lambda) = q_\mu(\mu)q_\lambda(\lambda) \tag{10}$$

Target distribution

The unnormalized log posterior has the form

$$
\begin{aligned}
\log \widetilde{p}(\mu, \lambda) &= \log p(\mu, \lambda, D) = \log p(D|\mu, \lambda) + \log p(\mu|\lambda) + \log p(\lambda) \\
&= \frac{N}{2} \log \lambda - \frac{\lambda}{2} \sum_{i=1}^{N} (x_i - \mu)^2 - \frac{1}{2} \log(\kappa_0 \lambda) \\
&+ \frac{\kappa_0 \lambda}{2} (\mu - \mu_0)^2 + (a_0 - 1) \log \lambda - b_0 \lambda + \text{const}
\end{aligned} \tag{11}
$$

Updating $q_\mu(\mu)$ (fix $q_\lambda(\lambda)$)

The optimal form for $q_\mu(\mu)$ is obtained by averaging over λ

$$
\begin{aligned}
\log q_\mu(\mu) &= E_{q_\lambda}[\log p(D|\mu,\lambda) + \log p(\mu|\lambda)] + \text{const} \\
&= -\frac{E_{q_\lambda}[\lambda]}{2}\{\kappa_0(\mu-\mu_0)^2 + \sum_{i=1}^{2}(x_i-\mu)^2\} + \text{const}
\end{aligned}
\tag{12}
$$

By completing the square one can show that $q_\mu(\mu) = N(\mu|\mu_N, \kappa_N^{-1})$, where

$$
\begin{aligned}
\mu_N &= \frac{\kappa_0\mu_0 + N\bar{x}}{\kappa_0 + N} \\
\kappa_N &= (\kappa_0 + N)E_{q_\lambda}[\lambda]
\end{aligned}
\tag{13}
$$

At this stage we don't know what $q_\lambda(\lambda)$ is, and hence we cannot compute $E[\lambda]$, but we will derive this below.

Updating $q_\lambda(\lambda)$ (fix $q_\mu(\mu)$)

The optimal form for $q_\lambda(\lambda)$ is given by

$$\begin{aligned} \log q_\lambda(\lambda) &= E_{q_\mu}[\log p(D|\mu,\lambda) + \log p(\mu|\lambda) + \log p(\lambda)] + \text{const} \\ &= (a_0 - 1)\log\lambda - b_0\lambda + \frac{1}{2}\log\lambda + \frac{N}{2}\log\lambda \\ &- \frac{\lambda}{2} E_{q_\mu}[\kappa_0(\mu-\mu_0)^2 + \sum_{i=1}^{N}(x_i-\mu)^2] + \text{const} \end{aligned} \tag{14}$$

We recognize this as the log of a Gamma distribution, hence $q_\lambda(\lambda) = \text{Ga}(\lambda|a_N, b_N)$, where

$$\begin{aligned} a_N &= a_0 + \frac{N+1}{2} \\ b_N &= b_0 + \frac{1}{2} E_{q_\mu}[\kappa_0(\mu-\mu_0)^2 + \sum_{i=1}^{N}(x_i-\mu)^2] \end{aligned} \tag{15}$$

Computing the expectations

Since $q(\mu) = N(\mu|\mu_N, \kappa_N^{-1})$, we have

$$\begin{aligned}
E_{q(\mu)}[\mu] &= \mu_N \\
E_{q(\mu)}[\mu^2] &= \frac{1}{\kappa_N} + \mu_N^2
\end{aligned} \tag{16}$$

Since $q(\lambda) = \mathrm{Ga}(\lambda|a_N, b_N)$, we have

$$E_q(\lambda)[\lambda] = \frac{a_N}{b_N} \tag{17}$$

Explicit forms for the update equations for $q(\mu)$ we have

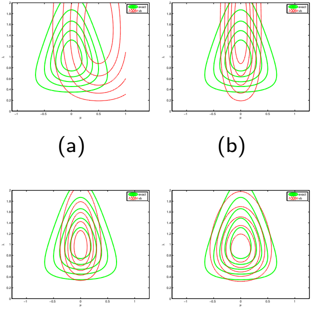
$$\begin{aligned}
\mu_N &= \frac{\kappa_0\mu_0 + N\bar{x}}{\kappa_0 + N} \xleftarrow{\text{fixed!}} \\
\kappa_N &= (\kappa_0 + N)\frac{a_N}{b_N}
\end{aligned} \tag{18}$$

and for $q(\lambda)$ we have

$$\begin{aligned}
a_N &= a_0 + \frac{N+1}{2} \xleftarrow{\text{fixed!}} \\
b_N &= b_0 + \kappa_0(E[\mu^2] + \mu_0^2 - 2E[\mu]\mu_0) + \frac{1}{2}\sum_{i=1}^{N}(x_i^2 + E[\mu^2] - 2E[\mu] - x_i)
\end{aligned} \tag{19}$$

Illustration

In the Figure, the green contours represent the exact posterior, which is Gaussian-Gamma. The dotted red contours represent the variational approximation over several iterations.



(a) (b)

(c) (d)

Figure: Factored variational approximation (red) to the Gaussian-Gamma distribution (green). (a) Initial guess. (b) After updating q_μ. (c) After updating q_λ. (d) At convergence (after 5 iterations). Based on 10.4 of (Bishop 2006b). Figure generated by `UnigaussVbDemo`.

Outline

Introduction

- Now consider latent variable models of the form $\boldsymbol{z}_i \to \boldsymbol{x}_i \leftarrow \boldsymbol{\theta}$.
- In EM, $\boldsymbol{\theta}$ are informed by all N data cases, whereas $\boldsymbol{z}_i$ is only informed by $\boldsymbol{x}_i$.
- **Variational Bayes EM** or **VBEM**: model uncertainty in $\boldsymbol{\theta}$ and $\boldsymbol{z}_i$.
- Computational cost is essentially the same as EM.
- **Same idea**

$$p(\boldsymbol{\theta}, \boldsymbol{z}_{1:N}|\mathcal{D}) \approx q(\boldsymbol{\theta}) \prod_i q_i(\boldsymbol{z}_i) \tag{20}$$

- **Pros:** marginalizing out the parameters, we can compute a lower bound on the marginal likelihood (useful for model selection).

The variational posterior

- The conditional distribution of $\boldsymbol{Z}$, given the mixing coefficients $\boldsymbol{\pi}$:

$$p(\boldsymbol{Z}|\boldsymbol{\pi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \tag{21}$$

- The likelihood:

$$p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{n=1}^{N} \prod_{k=1}^{K} N(\boldsymbol{X}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k^{-1})^{z_{nk}} \tag{22}$$

- Priors over the parameters $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ and $\boldsymbol{\pi}$
  - Pick a Dirichlet distribution over the mixing coefficients $\boldsymbol{\pi}$

$$p(\boldsymbol{\pi}) = \mathsf{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}_0) = C(\boldsymbol{\alpha}_0) \prod_{k=1}^{K} \boldsymbol{\pi}_k^{\alpha_0 - 1} \tag{23}$$

  - Pick an independent Gaussian-Wishart prior for the mean and precision of each Gaussian component

$$\begin{aligned} p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) &= p(\boldsymbol{\mu}|\boldsymbol{\Lambda})p(\boldsymbol{\Lambda}) \\ &= \prod_{k=1}^{K} N(\boldsymbol{\mu}_k|\boldsymbol{m}_0, (\beta_0\boldsymbol{\Lambda}_k)^{-1}) W(\boldsymbol{\Lambda}_k|\boldsymbol{W}_0, \nu_0) \end{aligned} \tag{24}$$

The variational posterior (cont'd)

- The joint distribution of all of the random variables

$$p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})p(\boldsymbol{Z}|\boldsymbol{\pi})p(\boldsymbol{\pi})p(\boldsymbol{\mu}|\boldsymbol{\Lambda})p(\boldsymbol{\Lambda}) \quad (25)$$

- Consider a variational distribution

$$q(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\boldsymbol{Z})q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \quad (26)$$

Factors $q(\boldsymbol{Z})$ and $q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ will be determined automatically by optimization of the variational distribution.

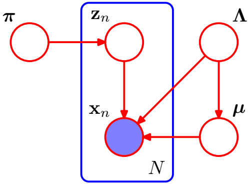


Figure: DAG of the Bayesian mixture of Gaussians model.

Derivation of $q(z)$ (variational E step)

Update for the factor $q(\boldsymbol{Z})$

$$\begin{aligned} \ln q^*(\boldsymbol{Z}) &= E_{\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda}}[\ln p(\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda})] + \text{const} \\ &= E_{\boldsymbol{\pi}}[\ln p(\boldsymbol{Z}|\boldsymbol{\pi})] + E_{\boldsymbol{\mu},\boldsymbol{\Lambda}}[\ln p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})] + \text{const} \\ &= \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\ln\rho_{nk} + \text{const} \end{aligned} \tag{27}$$

where we have defined

$$\ln\rho_{nk} = E[\ln\pi_k] + \frac{1}{2}E[\ln|\boldsymbol{\Lambda}_k|] - \frac{D}{2}\ln(2\pi) - \frac{1}{2}E_{\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T\boldsymbol{\Lambda}_k(\boldsymbol{x}_n - \boldsymbol{\mu}_k)] \tag{28}$$

Taking the exponential of both sides, we obtain

$$q^*(\boldsymbol{Z}) \propto \prod_{n=1}^{N}\prod_{k=1}^{K}\rho_{nk}^{z_{nk}} \tag{29}$$

Derivation of $q(z)$ (variational E step) (cont'd)

The factor $q(\boldsymbol{Z})$

- Requires be normalized

$$q^*(\boldsymbol{Z}) \propto \prod_{n=1}^{N} \prod_{k=1}^{K} r_{nk}^{z_{nk}} \tag{30}$$

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

- Takes the same functional form as the prior $p(\boldsymbol{Z}|\boldsymbol{\pi})$.
- The discrete distribution $q(\boldsymbol{Z})$ have $E[z_{nk}] = r_{nk}$.
- The quantities r_{nk} are playing the role of responsibilities.
- Three statistics evaluated with respect to the responsibilities

$$\begin{aligned} N_k &= \sum_{n=1}^{N} r_{nk} \\ \bar{\boldsymbol{x}}_k &= \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} \boldsymbol{x}_n \\ \boldsymbol{S}_k &= \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} (\boldsymbol{x}_n - \bar{\boldsymbol{x}}_k)(\boldsymbol{x}_n - \bar{\boldsymbol{x}}_k)^T \end{aligned} \tag{31}$$

Derivation of $q(\theta)$ (variational M step)

Consider the factor $q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ in the variational posterior distribution

$$\begin{aligned}\ln q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \ln p(\boldsymbol{\pi}) + \sum_{k=1}^{K} \ln p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) + E_{\boldsymbol{Z}}[\ln p(\boldsymbol{Z}|\boldsymbol{\pi})] \\ + \sum_{n=1}^{N} \sum_{k=1}^{K} E[z_{nk}] \ln N(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}) + \text{const}\end{aligned} \tag{32}$$

This decomposes into terms involving only $\boldsymbol{\pi}$ with only $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$. Thus

$$q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\boldsymbol{\pi}) \prod_{k=1}^{K} q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) \tag{33}$$

Identifying the terms on the right-hand side of (32) that depend on $\boldsymbol{\pi}$

$$\ln q^*(\boldsymbol{\pi}) = (\alpha_0 - 1) \sum_{k=1}^{K} \ln \boldsymbol{\pi}_k + \sum_{k=1}^{K} \sum_{n=1}^{N} r_{nk} \ln \boldsymbol{\pi}_k + \text{const} \tag{34}$$

Taking the exponential of both sides, $q^*(\boldsymbol{\pi})$ is a Dirichlet distribution

$$q^*(\boldsymbol{\pi}) = \text{Dir}(\boldsymbol{\pi}|\alpha) \tag{35}$$

where α has components α_k given by $\alpha_k = \alpha_0 + N_k$.

Derivation of $q(\theta)$ (variational M step) (cont'd)

Recall: $q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) = q(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k)q(\boldsymbol{\Lambda}_k)$. Thus

$$q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) = N(\boldsymbol{\mu}_k|\boldsymbol{m}_k, (\beta_k\boldsymbol{\Lambda}_k)^{-1})W(\boldsymbol{\Lambda}_k|\boldsymbol{W}_k, \nu_k) \tag{36}$$

where we defined

$$\begin{aligned} \beta_k &= \beta_0 + N_k \\ \boldsymbol{m}_k &= \frac{1}{\beta_k}(\beta_0\boldsymbol{m}_0 + N_k\bar{\boldsymbol{x}}_k) \\ \boldsymbol{W}_k^{-1} &= \boldsymbol{W}_0^{-1} + N_kS_k + \frac{\beta_0N_k}{\beta_0 + N_k}(\bar{\boldsymbol{x}}_k - \boldsymbol{m}_0)(\bar{\boldsymbol{x}}_k - \boldsymbol{m}_0)^T \\ \nu_k &= \nu_0 + N_k \end{aligned} \tag{37}$$

Expectations of the variational distributions of the parameters

$$\begin{aligned} E_{\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k}[(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T\boldsymbol{\Lambda}_k(\boldsymbol{x}_n - \boldsymbol{\mu}_k) &= D\beta_k^{-1} + \nu_k(\boldsymbol{x}_n - \boldsymbol{m}_k)^T\boldsymbol{W}_k(\boldsymbol{x}_n - \boldsymbol{m}_k)] \\ \ln\tilde{\boldsymbol{\Lambda}}_k &= E[\ln|\boldsymbol{\Lambda}_k|] = \sum_{i=1}^{D}\psi(\frac{\nu_k + 1 - i}{2}) + D\ln 2 + \ln|\boldsymbol{W}_k| \\ \ln\tilde{\boldsymbol{\pi}}_k &= E[\ln\boldsymbol{\pi}_k] = \psi(\alpha_k) - \psi(\hat{\alpha}) \end{aligned} \tag{38}$$

where we define $\tilde{\boldsymbol{\Lambda}}_k$ and $\tilde{\boldsymbol{\pi}}_k$, $\psi(.)$ is a the diagmma function, $\hat{\alpha} = \sum_k \alpha_k$.

Derivation of $q(\theta)$ (variational M step) (cont'd)

If we substitute (38) into $r_{nk} = \frac{\rho_{nk}}{\sum_{j=1}^{K} \rho_{nj}}$

$$r_{nk} \propto \tilde{\boldsymbol{\pi}}_k \tilde{\boldsymbol{\Lambda}}_k^{1/2} \exp\{-\frac{D}{2\beta_k} - \frac{\nu_k}{2}(\boldsymbol{x}_n - \boldsymbol{m}_k)^T \boldsymbol{W}_k (\boldsymbol{x}_n - \boldsymbol{m}_k)\} \tag{39}$$

Notice the similarity to the responsibilities in maximum likelihood EM

$$r_{nk} \propto \boldsymbol{\pi}_k |\boldsymbol{\Lambda}_k|^{1/2} \exp\{-\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Lambda}_k (\boldsymbol{x}_n - \boldsymbol{\mu}_k)\} \tag{40}$$

Lower bound on the marginal likelihood

In VB, we are maximizing a lower bound on the log marginal likelihood. Why?

- To assess convergence of the algorithm.
- To assess the correctness of one's code: as with EM, if the bound does not increase monotonically, there must be a bug.
- To approximateto the marginal likelihood, which can be used for **Bayesian model selection**.

The algorithm is trying to maximize the following lower bound (i.e. $F(q, y)$ free energy)

$$L = \sum_{\boldsymbol{Z}} \int \int \int q(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \ln\{\frac{p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})}{(Z, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})}\} d\boldsymbol{\pi}\, d\boldsymbol{\mu}\, d\boldsymbol{\Lambda} \leq \ln p(\boldsymbol{X}) \tag{41}$$

This quantity increases monotonically with each iteration, in Figure. (Exercise)

Posterior predictive distribution

The predictive density is then given by

$$p(\boldsymbol{x}^*|\boldsymbol{X}) = \sum_{\boldsymbol{z}^*} \int\int\int p(\boldsymbol{x}^*|\boldsymbol{z}^*, \boldsymbol{\mu}, \boldsymbol{\Lambda}) p(\boldsymbol{z}^*|\boldsymbol{\pi}) p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{X}) d\boldsymbol{\pi}\ d\boldsymbol{\mu}\ d\boldsymbol{\Lambda} \tag{42}$$

Using (21) and (22) we can first perform the summation over $\boldsymbol{z}^*$

$$p(\boldsymbol{x}^*|\boldsymbol{X}) = \sum_{k=1}^{K} \int\int\int \pi_k N(\boldsymbol{x}^*|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}) p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{X}) d\boldsymbol{\pi}\ d\boldsymbol{\mu}\ d\boldsymbol{\Lambda} \tag{43}$$

Because the remaining integrations are intractable, we approximate the predictive density with $q(\boldsymbol{\pi})q(\boldsymbol{\mu}, \boldsymbol{\Lambda})$

$$p(\boldsymbol{x}^*|\boldsymbol{X}) = \sum_{k=1}^{K} \int\int\int \pi_k N(\boldsymbol{x}^*|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}) q(\boldsymbol{\pi}) q(\boldsymbol{\mu}, \boldsymbol{\Lambda}) d\boldsymbol{\pi}\ d\boldsymbol{\mu}\ d\Lambda \tag{44}$$

where we have made use of the factorization (33).

Posterior predictive distribution (cont'd)

The remaining integrations can now be evaluated analytically giving a mixture of Student's t-distributions

$$p(\boldsymbol{x}^*|\boldsymbol{X}) = \frac{1}{\hat{\alpha}} \sum_{k=1}^{K} \alpha_k \mathsf{St}(\boldsymbol{x}^*|\boldsymbol{m}_k, \boldsymbol{L}_k, \nu + 1 - D) \tag{45}$$

in which the k^{th} component has mean $\boldsymbol{m}_k$, and the precision is

$$\boldsymbol{L}_k = \frac{(\nu_k + 1 - D)\beta_k}{1 + \beta_k} \boldsymbol{W}_k \tag{46}$$

in which ν_k is given by (37). When the size N of the data set is large the predictive distribution (45) reduces to a mixture of Gaussians.