Variational Bayes methods and algorithms Part I

Christine KERIBIN

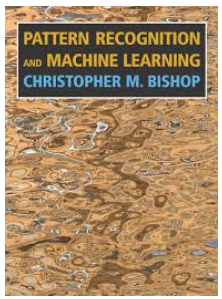
Laboratoire de Mathématique d'Orsay, Université Paris Sud and INRIA - Saclay Ile de France
Université Paris-Saclay

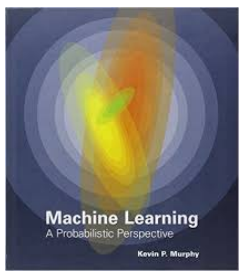
CIRM - March 2016

Outline

Main references

- Bishop, C. (2006)
  Pattern recognition and machine learning.
  Springer
- Murphy, K. (2012)
  Machine Learning: A Probabilistic Perspective.
  The MIT Press
- Keribin, C. (2010)
  Méthodes bayésiennes variationnelles : concepts et applications en neuroimagerie.
  Journal de la Société Française de Statistique, vol 151, N° 2





Attias [1999], Jordan [1999], Beal [2003], Beal and Gharamani [2004], Titterington ...

Outline

Introduction

Bayesian statistics needs to evaluate posterior distribution (and mean, mode)

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)p(\theta)}{p(\mathbf{x})}$$

- observation: $\mathbf{x} = (x_1, \ldots, x_n)$
- likelihood: $p(\mathbf{x}|\theta)$
- prior distribution: $p(\theta)$

Introduction

With latent variables $\mathbf{z}$, even more complex

- joint posterior distribution

$$p(\mathbf{z}, \theta|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{z}, \theta)p(\mathbf{z}|\theta)p(\theta)}{p(\mathbf{x})}$$

- marginal posterior distribution

$$p(\theta|\mathbf{x}) = \int p(\mathbf{z}, \theta|\mathbf{x})d\mathbf{z}; \qquad p(\mathbf{z}|\mathbf{x}) = \int p(\mathbf{z}, \theta|\mathbf{x})d\theta$$

- evidence

$$p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{z}, \theta)d\mathbf{z}d\theta$$

$\hookrightarrow$ may not have close form; may be numerically intractable.

Introduction

In these cases, approximations are needed:

- Stochastic techniques
  - generate exact results given (infinite) computational resource
  - computationally demanding
- Deterministic analytical approximations of the posterior distribution

Introduction

In these cases, approximations are needed:

- Stochastic techniques
  - generate exact results given (infinite) computational resource
  - computationally demanding
- Deterministic analytical approximations of the posterior distribution
  - Laplace approximation: local Gaussian approximation to a mode of the distribution
  - Variational inference:
    pick a family of distributions where the computation is easy, and choose the best $q^*(\theta)$ as an approximation of the posterior distribution $p(\theta|\mathbf{x})$
  - Expectation Propagation
  - ...

Outline

Basic principle

Let q be a free distribution on θ. From $p(\theta|\mathbf{x})p(\mathbf{x}) = p(\mathbf{x}, \theta)$, write:

$$\begin{aligned}
\log p(\mathbf{x}) &= \log p(\mathbf{x}, \theta) - \log p(\theta|\mathbf{x}) \\
&= \int_\theta \big(\log p(\mathbf{x}, \theta) - \log p(\theta|\mathbf{x})\big) q(\theta) d\theta \\
&= \int_\theta q(\theta) \log\left(\frac{p(\mathbf{x}, \theta)}{q(\theta)}\right) d\theta - \int_\theta q(\theta) \log\left(\frac{p(\theta|\mathbf{x})}{q(\theta)}\right) d\theta \\
&= \underbrace{\mathcal{F}(q)}_{\text{functional}} + \underbrace{KL(q, p)}_{\text{Kullback divergence}} \geq \underbrace{\mathcal{F}(q)}_{\text{lower bound}}
\end{aligned}$$

Basic principle

Let q be a free distribution on θ. From $p(\theta|\mathbf{x})p(\mathbf{x}) = p(\mathbf{x},\theta)$, write:

$$\begin{aligned}
\log p(\mathbf{x}) &= \log p(\mathbf{x},\theta) - \log p(\theta|\mathbf{x}) \\
&= \int_\theta \big(\log p(\mathbf{x},\theta) - \log p(\theta|\mathbf{x})\big) q(\theta) d\theta \\
&= \int_\theta q(\theta) \log\left(\frac{p(\mathbf{x},\theta)}{q(\theta)}\right) d\theta - \int_\theta q(\theta) \log\left(\frac{p(\theta|\mathbf{x})}{q(\theta)}\right) d\theta \\
&= \underbrace{\mathcal{F}(q)}_{\text{functional}} + \underbrace{KL(q,p)}_{\text{Kullback divergence}} \geq \underbrace{\mathcal{F}(q)}_{\text{lower bound}}
\end{aligned}$$

- considering all distributions q (no approximation)

$$p(.|\mathbf{x}) = \arg\min_q KL(q,p) = \arg\max_q \mathcal{F}(q)$$

Basic principle

Let q be a free distribution on θ. From $p(\theta|\mathbf{x})p(\mathbf{x}) = p(\mathbf{x},\theta)$, write:

$$\begin{aligned}
\log p(\mathbf{x}) &= \log p(\mathbf{x},\theta) - \log p(\theta|\mathbf{x}) \\
&= \int_\theta \big(\log p(\mathbf{x},\theta) - \log p(\theta|\mathbf{x})\big) q(\theta) d\theta \\
&= \int_\theta q(\theta) \log\left(\frac{p(\mathbf{x},\theta)}{q(\theta)}\right) d\theta - \int_\theta q(\theta) \log\left(\frac{p(\theta|\mathbf{x})}{q(\theta)}\right) d\theta \\
&= \underbrace{\mathcal{F}(q)}_{\text{functional}} + \underbrace{KL(q,p)}_{\text{Kullback divergence}} \geq \underbrace{\mathcal{F}(q)}_{\text{lower bound}}
\end{aligned}$$

- considering all distributions q (no approximation)

$$p(.|\mathbf{x}) = \arg\min_q KL(q,p) = \arg\max_q \mathcal{F}(q)$$

- considerinq only $q \in \mathcal{Q}$ where the computation is easy (approximation)

$$p(.|\mathbf{x}) \simeq q^*(.|x) = \arg\min_{q\in\mathcal{Q}} KL(q,p) = \arg\max_{q\in\mathcal{Q}} \mathcal{F}(q)$$

Basic principle

For inference on the parameter θ and latent variables $\mathbf{z}$

$$\begin{aligned}\log p(\mathbf{x}) &= \int_\theta \int_{\mathbf{z}} q(\theta, \mathbf{z}) \log \left(\frac{p(\mathbf{x}, \mathbf{z}, \theta)}{q(\theta, \mathbf{z})} \right) d\theta d\mathbf{z} - \int_\theta \int_{\mathbf{z}} q(\theta, \mathbf{z}) \log \left(\frac{p(\theta, \mathbf{z}|\mathbf{x})}{q(\theta, \mathbf{z})} \right) d\theta d\mathbf{z} \\ &= \underbrace{\mathcal{F}(q_{\theta z})}_{\text{Functional}} + \underbrace{KL(q_{\theta z}, p)}_{\text{Kullback divergence}} \geq \underbrace{\mathcal{F}(q_{\theta z})}_{\text{lower bound}}\end{aligned}$$

$$p(.|\mathbf{x}) \simeq q^*_{\theta z}(.|x) = \arg \min_{q_{\theta z} \in \mathcal{Q}} KL(q_{\theta z}, p) = \arg \max_{q_{\theta z} \in \mathcal{Q}} \mathcal{F}(q_{\theta z})$$

$$\log p(\mathbf{x}) \simeq_{\geq} \mathcal{F}(q^*_{\theta z})$$

$\hookrightarrow$ Replace an integration by an optimization over a set of functions where the computation is easy.

Basic principle

What kind of restrictions?

- use a given parametric distribution $q(.|\omega)$
- use factorized distributions:
  - separate latent variables and parameter

$$q_{\theta z} = q_\theta q_z$$

  - factorize variational distribution for latent variables : mean field approximation

$$q_z(\mathbf{z}) = \prod_{i=1}^{n} q_{z_i}(z_i)$$

  - $\hookrightarrow$ cycling optimization in turn along the different factorized distributions

Illustration: parametric distribution

Consider the approximation of $p(z)$ with a Gaussian $q(z; \mu, v)$[1].

$$p(z) \propto \exp(-z^2/2)(1 + e^{-(20z+4)})^{-1} = f(z)$$

Variational approximation: Minimize $KL(q, p)$ with respect to μ, v:

$$KL(p, q) = \int_z q(z; \mu, v) \log \left(\frac{q(z; \mu, v)}{p(z)} \right) dx = k(\mu, v)$$

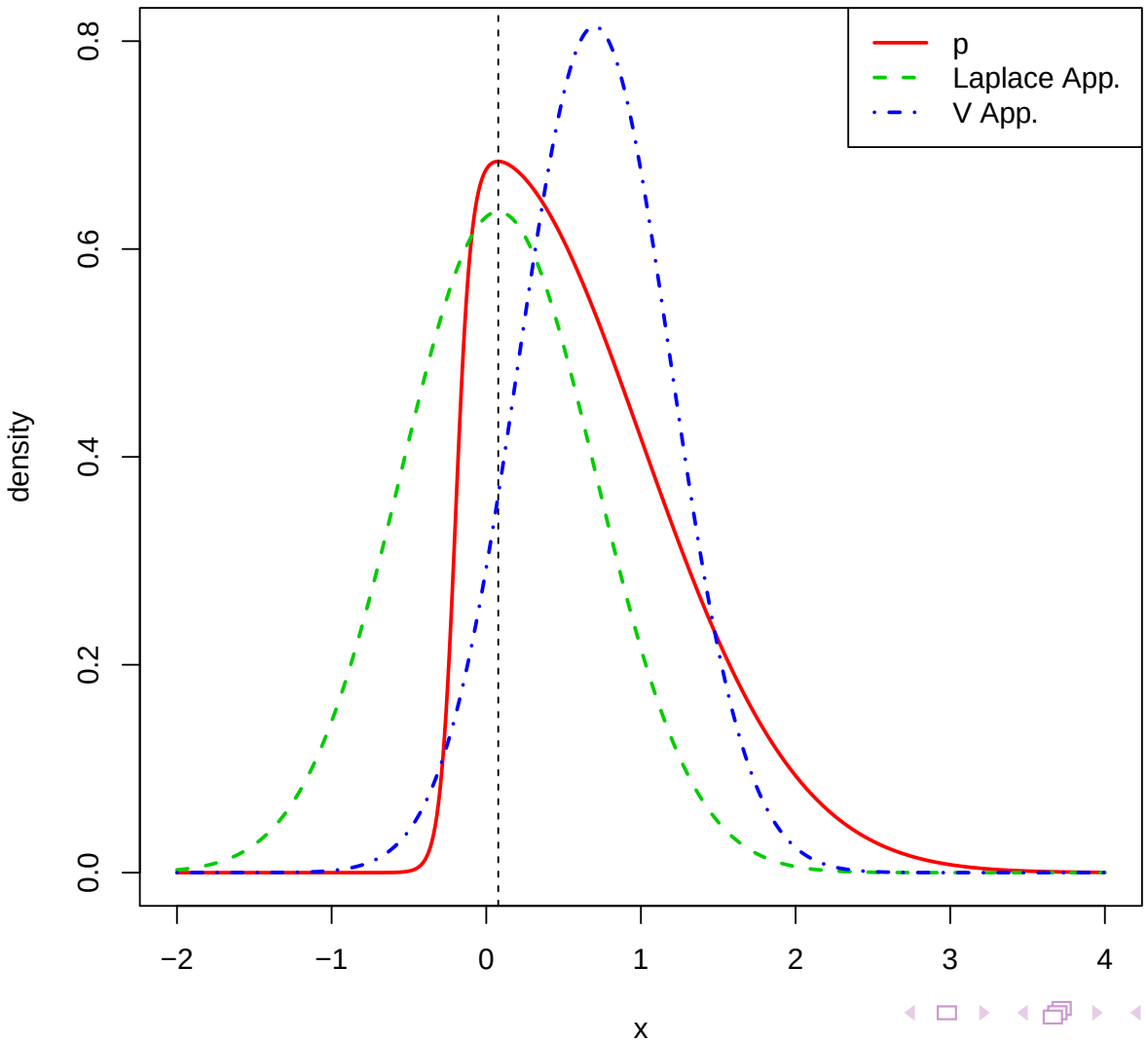
Laplace approximation

$$q_{LA}(z) = \sqrt{\frac{H}{2\pi}} \exp \left(-\frac{H}{2}(z - m_0)^2 \right)$$

with $m_0 = \frac{df(z)}{dz}|_{z=m_0}$ and $H = -\frac{d^2 \log f(z)}{dz^2}|_{z=m_0}$

[1] Bishop (2006), chap, 10

Variational and Laplace approximations



Mean field approximation

Define a factorized free distribution: $q(\mathbf{z}) = \prod_{i=1}^{n} q_i(z_i)$
First: suppose q_i fixed for $i \neq \ell$:

$$\begin{aligned}
\mathcal{F}(q) &= \int_{\mathbf{z}} [\log p(\mathbf{x}, \mathbf{z})] \prod_i q_i \, d\mathbf{z} - \sum_i \int_{z_i} q_i \log q_i \, dz_i \\
&= \int_{\mathbf{z}_\ell} q_\ell \underbrace{\left[\int_{z_{-\ell}} \log(p(\mathbf{x}, \mathbf{z}) \prod_{i \neq \ell} q_i dz_i \right]}_{\log \exp \mathbb{E}_{-\ell}(\log(p(\mathbf{x},\mathbf{z}))} dz_\ell - \int_{z_\ell} q_\ell \log q_\ell \, dz_\ell + cst
\end{aligned}$$

Mean field approximation

Define a factorized free distribution: $q(\mathbf{z}) = \prod_{i=1}^{n} q_i(z_i)$
First: suppose q_i fixed for $i \neq \ell$:

$$\begin{aligned}
\mathcal{F}(q) &= \int_{\mathbf{z}} [\log p(\mathbf{x}, \mathbf{z})] \prod_i q_i \, d\mathbf{z} - \sum_i \int_{z_i} q_i \log q_i \, dz_i \\
&= \int_{\mathbf{z}_\ell} q_\ell \underbrace{\left[\int_{z_{-\ell}} \log(p(\mathbf{x}, \mathbf{z}) \prod_{i \neq \ell} q_i dz_i \right]}_{\log \exp \mathbb{E}_{-\ell}(\log(p(\mathbf{x},\mathbf{z}))} dz_\ell - \int_{z_\ell} q_\ell \log q_\ell \, dz_\ell + cst \\
&= -KL\Big(q, A \exp \mathbb{E}_{-\ell}(\log p(\mathbf{x}, \mathbf{z}))\Big) + cst
\end{aligned}$$

Update of a given factor

$$q_\ell^* = \arg\max_{q_\ell} \mathcal{F}(q_1, \ldots, q_n) = \frac{\exp \mathbb{E}_{-\ell}(\log p(\mathbf{x}, \mathbf{z}))}{\int_{z_\ell} \exp \mathbb{E}_{-\ell}(\log p(\mathbf{x}, \mathbf{z})) \, dz_\ell} \quad (1)$$

Mean field approximation

- the set of updates $(q_i^*, i = 1, \ldots n)$ defines a set of consistency conditions for the maximum lower bound under factorized distribution.
- but each q_i^* depends on the computation of the expectation with respect to the others factors

Then: iterate...

- Initialize with some set of q_i^0, $i = 1, \ldots, n$
- Cycle through the factors and replace in turn with the corresponding update

$$\begin{aligned}
\log q_1^{t+1} &= \mathbb{E}_{q_2^t, \ldots, q_p^t} \log p(\mathbf{x}, \mathbf{z}) + cst \\
\log q_2^{t+1} &= \mathbb{E}_{q_1^{t+1}, q_3^t, \ldots, q_p^t} \log p(\mathbf{x}, \mathbf{z}) + cst \\
&\ldots \\
\log q_p^{t+1} &= \mathbb{E}_{q_1^{t+1}, \ldots, q_{p-1}^{t+1}} \log p(\mathbf{x}, \mathbf{z}) + cst
\end{aligned}$$

- Convergence is guaranted

Factorized distribution: illustration

Approximate a 2D-Gaussian distribution $p(\mathbf{z}) = \mathcal{N}(\mu, \Lambda^{-1})$ using a factorized distribution $q(\mathbf{z}) = q_1(z_1)q_2(z_2)$

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}; \quad \Lambda = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix} = V^{-1}$$

Using the update formula (1)

$$\begin{aligned}
\log q_1^*(z_1) &= \mathbb{E}_{q_2} \log p(\mathbf{z}) + cst \\
&= \mathbb{E}_{q_2} \left(-\frac{(z_1 - \mu_1)^2}{2} \Lambda_{11} - (z_1 - \mu_1)\Lambda_{12}(z_2 - \mu_2) \right) + cst \\
&= -\frac{1}{2}\Lambda_{11} z_1^2 + z_1 \mu_1 \Lambda_{11} - z_1 \Lambda_{12} \left(\mathbb{E}_{q_2}(z_2) - \mu_2 \right) + cst \\
&= -\frac{\Lambda_{11}}{2} \left[z_1 - \left(\mu_1 - \Lambda_{11}^{-1} \Lambda_{12} (\mathbb{E}_{q_2}(z_2) - \mu_2) \right) \right]^2 + cst
\end{aligned}$$

Factorized distribution: illustration

Approximate a 2D-Gaussian distribution $p(\mathbf{z}) = \mathcal{N}(\mu, \Lambda^{-1})$ using a factorized distribution $q(\mathbf{z}) = q_1(z_1)q_2(z_2)$

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}; \quad \Lambda = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix} = V^{-1}$$

Using the update formula (1)

$$\begin{aligned}
\log q_1^*(z_1) &= \mathbb{E}_{q_2} \log p(\mathbf{z}) + cst \\
&= \mathbb{E}_{q_2} \left(-\frac{(z_1 - \mu_1)^2}{2} \Lambda_{11} - (z_1 - \mu_1)\Lambda_{12}(z_2 - \mu_2) \right) + cst \\
&= -\frac{1}{2}\Lambda_{11} z_1^2 + z_1 \mu_1 \Lambda_{11} - z_1 \Lambda_{12} \left(\mathbb{E}_{q_2}(z_2) - \mu_2 \right) + cst \\
&= -\frac{\Lambda_{11}}{2} \left[z_1 - \left(\mu_1 - \Lambda_{11}^{-1} \Lambda_{12} (\mathbb{E}_{q_2}(z_2) - \mu_2) \right) \right]^2 + cst
\end{aligned}$$

q_1^* and q_2^* are Gaussian :

$$q_1^* = \mathcal{N}(m_1^*, V_1^* = \Lambda_{11}^{-1}); \quad q_2^* = \mathcal{N}(m_2^*, V_2^* = \Lambda_{22}^{-1})$$

Factorized distribution: illustration

- Approximate a 2D-Gaussian distribution $p(\mathbf{z}) = \mathcal{N}(\mu, \Lambda^{-1})$ using a factorized $q(\mathbf{z}) = q_1(z_1)q_2(z_2)$ leads to a product of Gaussian.
- The updating equations are coupled:

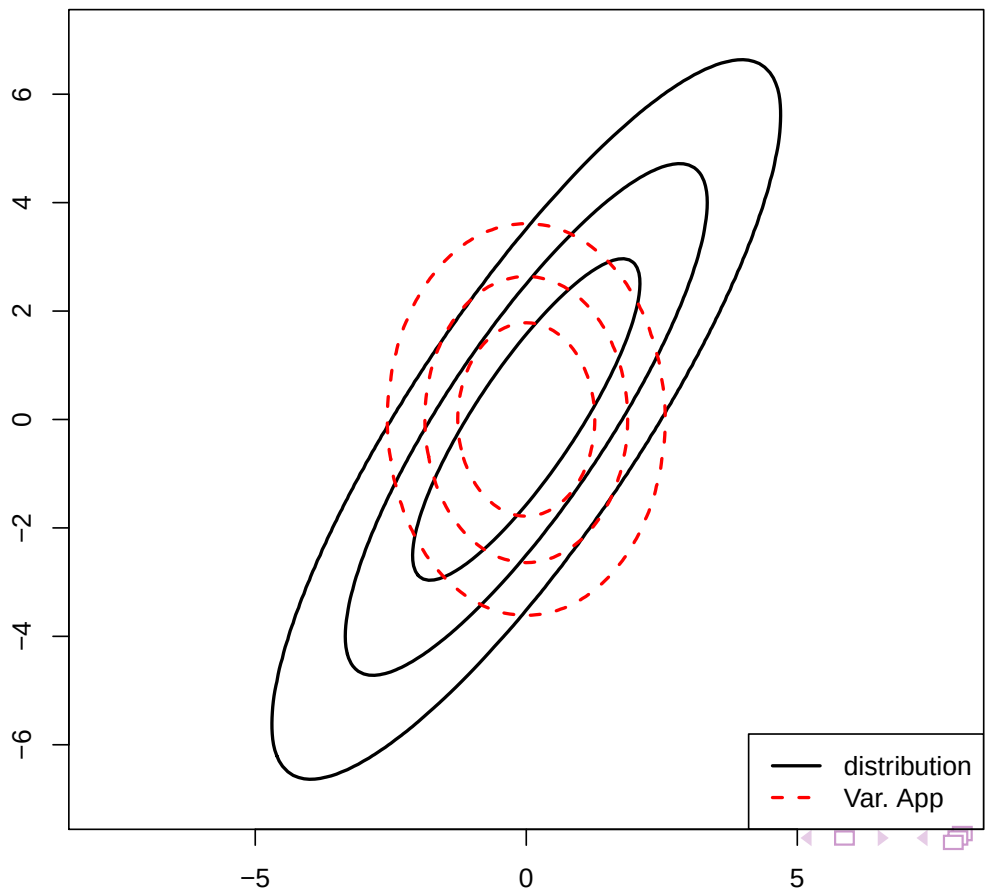
$$\begin{aligned} \mathbb{E}_{q_1}(x_1) = m_1^* &= \mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(m_2^* - \mu_2) \\ \mathbb{E}_{q_2}(x_2) = m_2^* &= \mu_2 - \Lambda_{22}^{-1}\Lambda_{12}(m_1^* - \mu_1) \end{aligned}$$

- the mean is correctly captured: $m_1^* = \mu_1$ and $m_2^* = \mu_2$
- however, the variance is underestimated and controlled by the direction of smallest variance of $p(\mathbf{z})$:

$$V_1^* = \Lambda_{11}^{-1} = V_{11} - \frac{V_{12}^2}{V_{22}} \text{ and } V_2^* = \Lambda_{22}^{-1} = V_{22} - \frac{V_{12}^2}{V_{11}}$$

Factorized distribution: illustration

$$\mu_1 = \mu_2 = 0; \quad V = \begin{pmatrix} 1 & 1.2 \\ 1.2 & 2 \end{pmatrix}$$



Reverse KL

If now we minimize the reverse KL instead: $KL(p, q)$

$$KL(p, q) = \mathbb{E}_p(\log(p/q)) = -\int p(\mathbf{z}) \sum_i \log q_i(z_i) d\mathbf{z} + cst$$

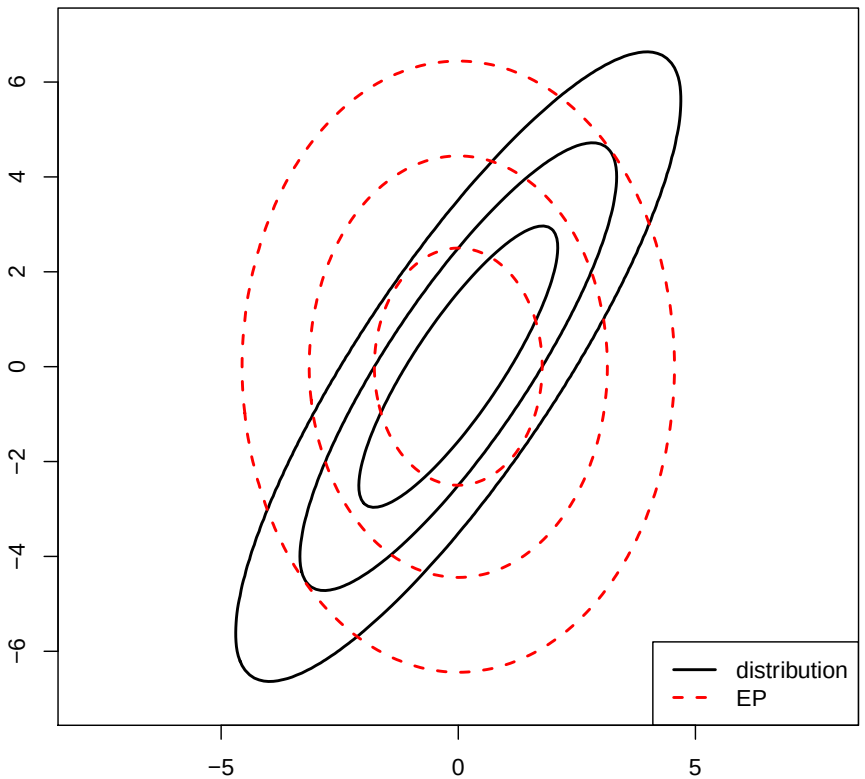
Using the factorized form of q, optimization leads to

$$q_i^*(z_i) = \int p(\mathbf{z}) \prod_{\ell \neq i} dz_i = p(z_i)$$

$\hookrightarrow$ optimal solution in closed-form as the marginal distribution

Reverse KL

Mean is successfully recovered, but approximation is too spread.



Reverse KL

- direct $KL(q, p)$ is zero forcing for q :

$$KL(q, p) = \int q(\mathbf{z}) \log \frac{q(\mathbf{z})}{p(\mathbf{z})}$$

$\hookrightarrow$ infinite if $p(\mathbf{z}) = 0$ and $q(\mathbf{z}) > 0$.

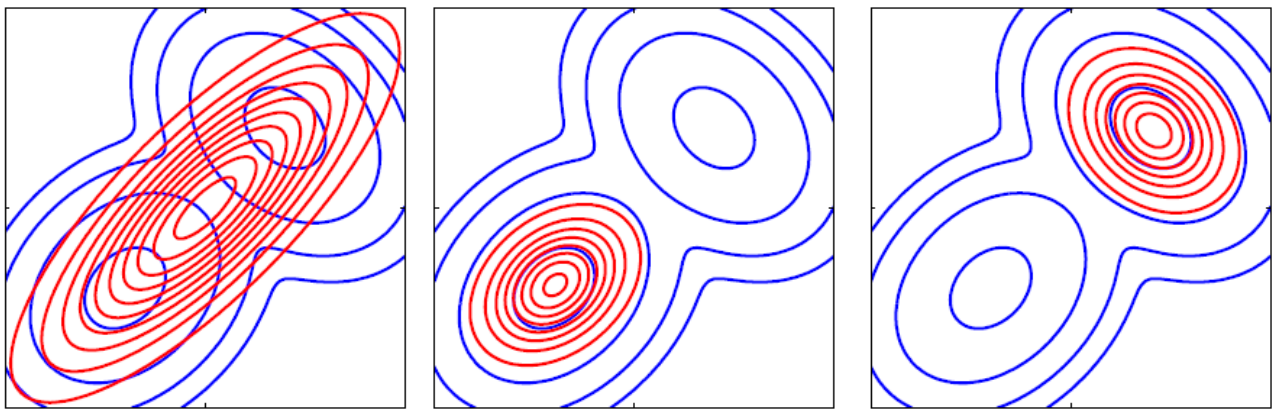
- reverse $KL(p, q)$ is zero avoiding for q :

$$KL(p, q) = \int p(\mathbf{z}) \log \frac{p(\mathbf{z})}{q(\mathbf{z})}$$

$\hookrightarrow$ infinite if $q(\mathbf{z}) = 0$ and $p(\mathbf{z}) > 0$.

Reverse KL

Approximation of a bimodal distribution (blue contours) with a unimodal distribution (red contours)[2]



Direct KL will tend to find a single mode (b, c), whereas inverse will average across all of the modes (a)
Note: use of reverse divergence : Expectation Propagation

[2] Bishop (2006)

You say variational?

- Variational inference: approximate a distribution using the optimisation of a given functional
- Variational Bayes: variational inference for the posterior distribution of a parameter θ: $p(\theta|\mathbf{x}) \simeq \prod_j q_j(\theta_j)$
- Variational Bayes EM: variational inference on both latent variables $\mathbf{z}$ and parameter θ : $p(\theta, \mathbf{z}|\mathbf{x}) \simeq q(\theta) \prod_i q_i(z_i)$
- Variational interpretation of the EM algorithm

Outline

Variational Bayes on an example

Compute a factorized variational approximation for the model of iid univariate Gaussian observations

$$p(x|\mu, \lambda = \sigma^{-2}) = \left(\frac{\lambda}{2\pi}\right)^{n/2} \exp\left(-\frac{\lambda}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right)$$

with conjugate prior Normal-Gamma

$$p(\mu, \lambda) = p(\mu|\lambda)p(\lambda) = \mathcal{N}(\mu|\mu_0, (\kappa_0\lambda)^{-1})\ \mathrm{Ga}(\lambda|a_0, b_0)$$

VB approximation with a factorized free distribution

$$q(\mu, \lambda) = q_\mu(\mu)q_\lambda(\lambda)$$

Updating q_μ

$$\begin{aligned}
\log q_\mu(\mu) &= \mathbb{E}_\lambda \log p(\mathbf{x}, \mu, \lambda) + cst \\
&= \mathbb{E}_\lambda \log p(x|\mu, \lambda) + \mathbb{E}_\lambda \log p(\mu|\lambda) + cst \\
&= \frac{n}{2}\mathbb{E}_\lambda(\log \lambda) - \frac{\mathbb{E}_\lambda(\lambda)}{2}\sum_i (x_i - \mu)^2 + \frac{1}{2}\mathbb{E}_\lambda(\log \lambda) - \frac{\mathbb{E}_\lambda(\lambda)}{2}\kappa_0(\mu - \mu_0)^2 + cst \\
&= -\frac{\mathbb{E}_\lambda(\lambda)}{2}\left(\sum_i (x_i - \mu)^2 + \kappa_0(\mu - \mu_0)^2\right) + cst \\
&= -\frac{\mathbb{E}_\lambda(\lambda)}{2}\left(\mu^2(\kappa_0 + n) - 2\mu\left(\sum_i x_i + \mu_0\kappa_0\right)\right) + cst
\end{aligned}$$

- optimal $q_\mu^* = \mathcal{N}(\mu_n, \kappa_n^{-1})$
- $\mu_n = (\mu_0\kappa_0 + n\bar{x})/(\kappa_0 + n)$
- $\kappa_n = (\kappa_0 + n)\mathbb{E}_\lambda(\lambda)$

Updating q_λ

Remind Ga distribution: $p(\lambda|a_0, b_0) = \frac{b_0^{a_0}}{\Gamma(a_0)}\lambda^{a_0-1}e^{-\lambda b_0}$. It follows:

$$
\begin{aligned}
& \log q_\lambda(\lambda) = \mathbb{E}_\mu \log p(\mathbf{x}, \mu, \lambda) + cst \\
= \;& \mathbb{E}_\mu \log p(x|\mu, \lambda) + \mathbb{E}_\mu \log p(\mu|\lambda) + \log p(\lambda) + cst \\
= \;& \frac{n}{2}\log\lambda - \frac{\lambda}{2}\mathbb{E}_\mu\Big(\sum_i (x_i - \mu)^2\Big) + \frac{1}{2}\log\lambda - \frac{\lambda\kappa_0}{2}\mathbb{E}_\mu\big((\mu - \mu_0)^2\big) \\
& + (a_0 - 1)\log\lambda - b_0\lambda + cst \\
= \;& \left(\frac{n+1}{2} + a_0 - 1\right)\log\lambda - \lambda\left[b_0 + \mathbb{E}_\mu\left(\sum_i \frac{(x_i - \mu)^2}{2} + \kappa_0\frac{(\mu - \mu_0)^2}{2}\right)\right] + cst
\end{aligned}
$$

- optimal $q_\lambda^* = \mathrm{Ga}(a_n, b_n)$
- $a_n = a_0 + \frac{n+1}{2}$
- $b_n = b_0 + \mathbb{E}_\mu\left(\sum_i (x_i - \mu)^2 + \kappa_0(\mu - \mu_0)^2\right)/2$

Optimized (q_μ^*, q_λ^*)

The optimal distribution for mean and precision depends on moments evaluated with respect to the other distribution
$\hookrightarrow$ one option is to cycle through q_μ and q_λ in turn and update until convergence

- Updating $q_\mu^{t+1} = \arg\max_{q_\mu} \mathcal{F}(q_\mu, q_\lambda^t) = \mathcal{N}(\mu_n^{t+1}, 1/\kappa_n^{t+1})$

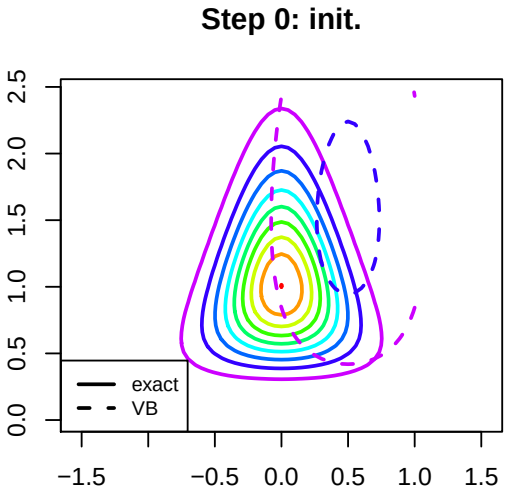
$$\mu_n^{t+1} = \frac{\mu_0\kappa_0 + n\bar{x}}{\kappa_0 + n}; \quad \kappa_n^{t+1} = (\kappa_0 + n)\underbrace{\mathbb{E}_\lambda(\lambda)}_{a_n^t/b_n^t}$$

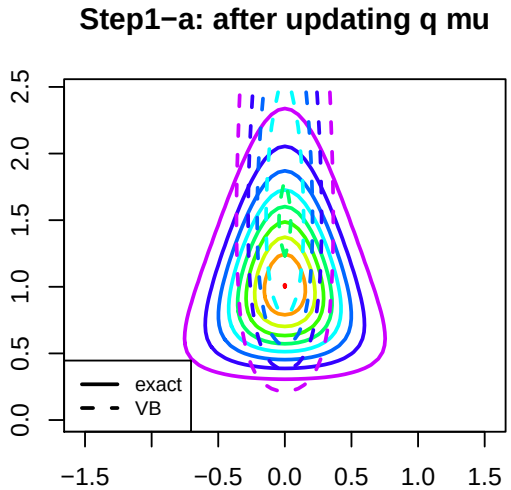
- Updating $q_\lambda^{t+1} = \arg\max_{q_\lambda} \mathcal{F}(q_\mu^{t+1}, q_\lambda) = \mathrm{Ga}(a_n, b_n)$

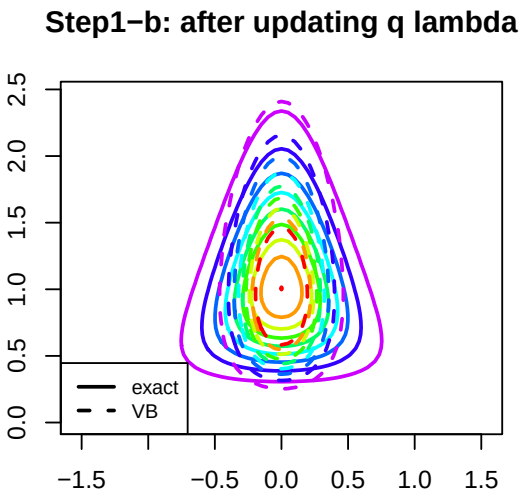
$$a_n^{t+1} = a_0 + \frac{n+1}{2}$$

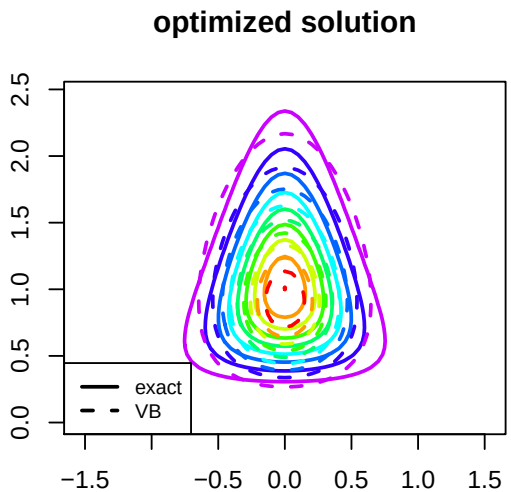
$$b_n^{t+1} = b_0 + \frac{1}{2}\left(\sum_i x_i^2 + \kappa_0\mu_0^2\right) - (n\bar{x} + \kappa_0\mu_0)\underbrace{\mathbb{E}_\mu(\mu)}_{\mu_n} + \frac{1}{2}(n+\kappa_0)\underbrace{\mathbb{E}_\mu(\mu^2)}_{\mu_n^2+\kappa_n^{-1}}$$

Convergence









Outline

Dealing with latent variables

The frequentist approach uses the EM algorithm

$$\begin{aligned}\log p(\mathbf{x};\theta) &= \log p(\mathbf{x},\mathbf{z};\theta) - \log p(\mathbf{z}|\mathbf{x};\theta) \\ &= \mathbb{E}[\log p(\mathbf{x},\mathbf{z};\theta)|\mathbf{x};\theta^{(c)}] - \mathbb{E}[\log p(\mathbf{z}|\mathbf{x};\theta)|\mathbf{x};\theta^{(c)}] \\ &= Q(\theta|\theta^{(c)}) - H(\theta|\theta^{(c)})\end{aligned}$$

Let $\tilde{\theta} \in \arg\max_\theta Q(\theta|\theta^{(c)})$

$$L(\tilde{\theta}) - L(\theta^{(c)}) = Q(\tilde{\theta}|\theta^{(c)}) - Q(\theta|\theta^{(c)}) + H(\theta^{(c)}|\theta^{(c)}) - H(\tilde{\theta}|\theta^{(c)}) \geq 0$$

EM algorithm

Repeat:

- **E Step** : compute $Q(\theta|\theta^{(c)})$, conditional expectation of the complete likelihood
  $\hookrightarrow$ need to compute $p(\mathbf{z}|\mathbf{x};\theta^{(c)})$ or its moments
- **M Step** : update θ by maximization: $\theta^{c+1} = \arg\max_\theta Q(\theta|\theta^c)$

$\hookrightarrow$ convergence to a local optimum

Bayesian setting with latent variables: VBEM

Full Bayes: priors on parameter θ and latent $\mathbf{z}$; take a variational distribution that factorizes between latent variables and model parameter

$$p(\theta, \mathbf{z}|\mathbf{x}) \simeq \underbrace{q_\theta(\theta) q_z(\mathbf{z})}_{\text{crucial approximation}} = q_\theta(\theta) \underbrace{\prod_i q_i(z_i)}_{\text{follows for iid var.}}$$

$$\begin{aligned} p(\mathbf{x}) &= \mathbb{E}_q(\log(p(\mathbf{x}, \mathbf{z}, \theta)/q_\theta q_\mathbf{z})) + \mathbb{E}_q(\log(p(\mathbf{z}, \theta|\mathbf{x})/q_\theta q_\mathbf{z})) \\ &= \mathcal{F}(q_\mathbf{z}, q_\theta) + KL(q_\mathbf{z} q_\theta, p) \end{aligned}$$

Variational Bayes EM

- **VBE Step** : update $q(\mathbf{z}|\mathbf{x})$, VB posterior of the latent variables :

$$q_\mathbf{z}^{t+1} = \arg\max_{q_\mathbf{z}} \mathcal{F}(q_\mathbf{z}, q_\theta^t)$$

- **VBM Step** : update $q(\theta|\mathbf{x})$, VB posterior of the parameter

$$q_\theta^{t+1} = \arg\max_{q_\theta} \mathcal{F}(q_\mathbf{z}^{t+1}, q_\theta)$$

$\hookrightarrow$ convergence to a local optimum

Back to EM

From $\log p(\mathbf{x};\theta) = \log p(\mathbf{x},\mathbf{z};\theta) - \log p(\mathbf{z},|\mathbf{x}\theta)$:

$$\begin{aligned}\log p(\mathbf{x};\theta) &= \mathbb{E}_{q_z}\left[\log \frac{p(\mathbf{x},\mathbf{z}|\theta)}{q_z(\mathbf{z})}\right] + KL(q_z||p(\mathbf{z}|\mathbf{x};\theta)) \\ &= \mathcal{F}(q_z,\theta) + KL(q_z||p(\mathbf{z}|\mathbf{x};\theta))\end{aligned}$$

EM (revisited as a Variational algorithm)

E Update $p(\mathbf{z}|\mathbf{x};\theta)$ as

$$q_{\mathbf{z}}^{t+1} = \arg\max_{q_{\mathbf{z}}} \mathcal{F}(q_{\mathbf{z}},\theta^t); \quad Q(\theta|\theta^{t+1}) = \mathcal{F}(q_{\mathbf{z}}^{t+1},\theta)$$

M Maximize

$$\theta^{t+1} = \arg\max_{\theta} \mathcal{F}(q_{\mathbf{z}}^{t+1},\theta) = \arg\max_{\theta} = Q(\theta|\theta^{t+1})$$

Mixture of Gaussians

Remember

- iid sample of a convex combination of K distributions of a given family (here, $\mathcal{N}_D(\mu, \Lambda^{-1})$)

$$p(x_i|\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i|\mu_k, \Lambda_k^{-1}); \quad \sum_{k=1}^{K} \pi_k = 1$$

- Parameter
  $\theta = (\boldsymbol{\pi} = (\pi_k)_{k=1,\ldots,K}, \boldsymbol{\mu} = (\mu_k)_{k=1,\ldots,K}, \boldsymbol{\Lambda} = (\Lambda_k)_{k=1,\ldots,K})$
- Latent variables $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ where $\mathbf{z}_i$ is a 1-of-K binary vector with element z_{ik}: $p(z_{ik} = 1) = \pi_k$

$$p(\mathbf{z}|\boldsymbol{\pi}) = \prod_{i=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{ik}}; \quad p(\mathbf{x}|\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{i=1}^{N} \prod_{k=1}^{K} \left[\mathcal{N}(x_i|\mu_k, \Lambda_k^{-1})\right]^{z_{ik}}$$

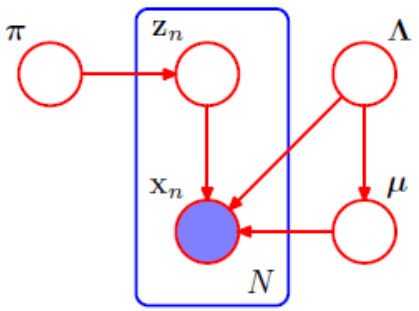
- Log-likelihood

$$p(\mathbf{x}|\theta) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \mathcal{N}(x_i|\mu_k, \Lambda_k^{-1})$$

Bayesian Mixture of Gaussians

- Introduce priors over parameter θ

$\hookrightarrow$ Dirichlet for the mixing coefficients



$$p(\pi) = \mathcal{D}(\pi|\mathbf{a}_0) \propto \prod_{k=1}^{K} \pi_k^{a_0-1}$$

$\hookrightarrow$ Independent Gaussian-Wishart prior for the modes and precisions

$$\begin{aligned} p(\mu, \Lambda) &= p(\mu|\Lambda)p(\Lambda) \\ &= \prod_{k=1}^{K} \mathcal{N}(\mu_k|\mathbf{m}_0, (b_0\Lambda)^{-1})\mathcal{W}(\Lambda_k|W_0, \nu_0) \end{aligned}$$

$$\mathcal{W}(\Lambda_k|W_k, \nu_0) \propto |\Lambda_k|^{(\nu_0-D-1)/2} \exp\left(-\frac{1}{2} Tr(W_0^{-1}\Lambda_k)\right)$$

We choose $\mathbf{m}_0 = 0$ and $W_0 = Id$

Variational distribution

- Joint distribution

$$p(\mathbf{x}, \mathbf{z}, \theta) = p(\mathbf{x}|\mathbf{z}, \mu, \mathbf{\Lambda})p(\mathbf{z}|\pi)p(\pi)p(\mu|\mathbf{\Lambda})p(\mathbf{\Lambda})$$

- Exact posterior $p(\mathbf{z}, \theta|\mathbf{x})$: K^n terms
- Variational distribution: factorization between the latent variables and model parameter

$$q(\mathbf{z}, \theta) = q_{\mathbf{z}}(\mathbf{z})q_{\theta}(\pi, \mu, \mathbf{\Lambda})$$

$\hookrightarrow$ only assumption for a tractable solution
$\hookrightarrow$ the functional form will be automatically determined by the optimization of q

Sequential update for $q_{\mathbf{z}}$

- Using the general update scheme (1)

$$
\begin{aligned}
\log q^*_{\mathbf{z}}(\mathbf{z}) &= \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda}} \log p(\mathbf{x},\mathbf{z},\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda}) + cst \\
&= \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda}} (\log (p(\mathbf{x}|\mathbf{z},\boldsymbol{\mu},\boldsymbol{\Lambda}) p(\mathbf{z}|\boldsymbol{\pi}) p(\boldsymbol{\pi}) p(\boldsymbol{\mu}|\boldsymbol{\Lambda}) p(\boldsymbol{\Lambda})) + cst \\
&= \mathbb{E}_{\boldsymbol{\pi}} (\log p(\mathbf{z}|\boldsymbol{\pi})) + \mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\Lambda}} (\log p(\mathbf{x}|\mathbf{z},\boldsymbol{\mu},\boldsymbol{\Lambda})) + cst \\
&= \sum_{i=1}^{n} \sum_{k=1}^{n} \left\{ z_{ik} \mathbb{E}_{\boldsymbol{\pi}} (\log \pi_k) + z_{ik} \mathbb{E}_{\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k} \log(\mathcal{N}(x_i|\mu_k,\Lambda_k^{-1})) \right\} + cst \\
&= \sum_{i=1}^{n} \sum_{k=1}^{n} z_{ik} \log(t_{ik}) + cst
\end{aligned}
$$

with

$$
\log t_{ik} = \mathbb{E}_{\boldsymbol{\pi}} (\log \pi_k) + \frac{1}{2} \mathbb{E} (\log |\Lambda_k|) - \frac{D}{2} \log(2\pi) - \frac{1}{2} \mathbb{E}_{\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k} \left((x_i - \mu_k)' \Lambda_k (x_i - \mu_k) \right)
$$

Updating $q_{\mathbf{z}}$

- From
$$\log q^*_{\mathbf{z}}(\mathbf{z}) = \sum_{i=1}^{n}\sum_{k=1}^{n} z_{ik}\log(t_{ik}) + cst$$
- Exponentiating
$$q^*_{\mathbf{z}}(\mathbf{z}) \propto \prod_{i=1}^{n}\prod_{k=1}^{K} t_{ik}^{z_{ik}}$$
- Normalizing
$$q^*_{\mathbf{z}}(\mathbf{z}) = \prod_{i=1}^{n}\prod_{k=1}^{K} \tau_{ik}^{z_{ik}}, \text{ with } \tau_{ik} = \frac{t_{ik}^{z_{ik}}}{\sum_{\ell=1}^{K} t_{i\ell}^{z_{i\ell}}}$$
- $q_{\mathbf{z}}$ is factorized, has the same function form as the prior $p(\mathbf{z}|\pi)$ and depends on moments evaluated with respect to the other distribution $\hookrightarrow$ cycling
- $\mathbb{E}_{z_i}(z_{ik}) = \tau_{ik}$

Updating q_θ

- Using the general update scheme (1)

$$\begin{aligned}
\log q_\theta^*(\mathbf{z}) &= \mathbb{E}_{\pi,\mu,\mathbf{\Lambda}} \log p(\mathbf{x}, \mathbf{z}, \pi, \mu, \mathbf{\Lambda}) + cst \\
&= \mathbb{E}_{\mathbf{z}} \log \left(p(\mathbf{x}|\mathbf{z}, \mu, \mathbf{\Lambda}) p(\mathbf{z}|\pi) p(\pi) p(\mu|\mathbf{\Lambda}) p(\mathbf{\Lambda}) \right) + cst \\
&= \sum_{i,k} \mathbb{E}_{\mathbf{z}}(z_{ik}) \log \left(\mathcal{N}(x_i|\mu_k, \Lambda_k^{-1}) \right) + \sum_k \log p(\mu_k, \Lambda_k) \\
&\quad + \sum_{i,k} \log(\pi_k) \mathbb{E}_{z_i}(z_{ik}) + \log(p(\pi)) + cst
\end{aligned}$$

$\hookrightarrow$ factorization:

$$q_\theta^*(\pi, \mu, \mathbf{\Lambda}) = q^*(\pi) \prod_{k=1}^K q^*(\mu_k, \Lambda_k) = q^*(\pi) \prod_{k=1}^K q^*(\mu_k|\Lambda_k) q^*(\Lambda_k)$$

Updating q_π

- Remind $p(\boldsymbol{\pi}) \propto \prod_k \pi_k^{a_0 - 1}$ and $\mathbb{E}_{z_i}(z_{ik}) = \tau_{ik}$

$$\begin{aligned} \log q^*(\boldsymbol{\pi}) &= \sum_k \sum_i \log(\pi_k) \mathbb{E}_{z_i}(z_{ik}) + (a_0 - 1) \sum_k \log(\pi_k) + cst \\ &= \sum_k \left(a_0 - 1 + \sum_i \tau_{ik} \right) \log \pi_k + cst \end{aligned}$$

- the optimized variational posterior q_π is also a Dirichlet distribution $\mathcal{D}(a_0 + n_k)$, with $n_k = \sum_i \tau_{ik}$
- need to compute τ_{ik}

Updating $q_{\mu,\Lambda}$

- derive some useful statistics:
  - $n_k = \sum_i \tau_{ik}$: taille du composant
  - $\bar{x}_k = \sum_i \tau_{ik} x_i / n_k$: moyenne du composant
  - $S_k = \sum_i \tau_{ik} (x_i - \bar{x}_k)^2 / n_k$: variance du composant
- the optimized variational posterior $q_{\boldsymbol{\mu},\boldsymbol{\Lambda}}$ is also a Gaussian-Wishart distribution

$$q^*(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \mathcal{N}(\mu_k | m_k, (b_k \Lambda_k)^{-1}) \mathcal{W}(\Lambda_k | W_k, \nu_k)$$

where

- $b_k = b_0 + n_k$
- $m_k = (b_0 m_0 + n_k \bar{x}_k)/b_k$
- $\nu_k = \nu_0 + n_k$
- $W_k^{-1} = W_0^{-1} + n_k S_k + (\bar{x}_k - m_0)(\bar{x}_k - m_0)' b_0 n_k / b_k$

- need to compute τ_{ik} where moments of q_θ are needed

$$\tau_{ik} \propto \tilde{\pi}_k \tilde{\Lambda}_k^{1/2} \exp\left(-\frac{D}{2b_k} - \frac{\nu_k}{2}(x_i - m_k)' W_k (x_i - m_k) \right)$$

where $\log \tilde{\pi}_k \triangleq \mathbb{E}(\log \pi_k) = \psi(\alpha_k) - \psi(\sum'_k \alpha_{k'})$ and $\log \tilde{\Lambda}_k \triangleq \mathbb{E}(\log |\Lambda_k|)$

Some more on Bayesian Gaussian Mixtures

- The variational posterior distribution has the same functional form as the corresponding factor in the joint distribution : choice of conjugate distribution
- Computation of variational lower bound $\mathcal{F}$ is straightfoward
  - direct maximization is an alternative approach to the re-estimation equations
  - can be used for model selection. But as is tends to approximate the distribution on the neighborhood of a mode, add a log $K!$ term.
- Model selection can also be performed with a spiky prior on the corners of π (small α_0)
- Predictive density is straightforward using the VB distribution

Thank you for your attention!