Machine Learning

Lecture 7 – Approximate inference

UPPSALA
UNIVERSITET

Thomas Schön
Division of Systems and Control
Department of Information Technology
Uppsala University.

Email: thomas.schon@it.uu.se,
www: `user.it.uu.se/~thosc112`

Contents – lecture 7

1. Summary of lecture 6
2. Finish EM and clustering from lecture 6
3. Bayesian reminder
4. Variational Bayesian inference
   - General derivation
   - Example – identification of an LGSS model
   - Example – Gaussian mixtures
5. Possibly start on expectation propagation
   - General derivation
   - Example – state estimation (smoothing) in dynamical systems

(Chapter 10)

This lecture builds on Umut Orguner's 2011 lecture.

Summary of lecture 6 (I/II)

The **Expectation Maximization (EM)** algorithm computes maximum likelihood estimates of unknown parameters in probabilistic models involving latent variables.

Expectation (E) step: Compute

$$\begin{aligned}\mathcal{Q}(\theta,\theta_i) &= \mathbf{E}_{\theta_i}\left\{\ln p_\theta(Z,X) \mid X\right\}\\ &= \int \ln p_\theta(Z,X) p_{\theta_i}(Z \mid X)\mathrm{d}Z.\end{aligned}$$

Maximization (M) step: Compute

$$\theta_{i+1} = \arg\max_{\theta} \mathcal{Q}(\theta,\theta_i).$$

Summary of lecture 6 (II/II)

We constructed a Gaussian mixture density using latent variables z (multinomial)

$$p(z) = \prod_{k=1}^{K} \pi_k^{z_k}, \qquad p(x \mid z) = \prod_{k=1}^{K} \mathcal{N}(x \mid \mu_k, \Sigma_k)^{z_k}$$

This allowed us to (start) deriving an EM algorithm for estimating a Gaussian mixture.

$$\begin{aligned}
\mathcal{Q}(\theta,\theta_i) &= \mathrm{E}_{\theta_i}\left[\ln p_\theta(Z,X) \mid X\right] \\
&= \mathrm{E}_{\theta_i}\left[\sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\left(\ln \pi_k + \ln \mathcal{N}\left(x_n \mid \mu_k, \Sigma_k\right)\right) \mid X\right] \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} \underbrace{\mathrm{E}_{\theta_i}\left[z_{nk} \mid X\right]}\left(\ln \pi_k + \ln \mathcal{N}\left(x_n \mid \mu_k, \Sigma_k\right)\right)
\end{aligned}$$

Hence, the E step amounts to finding $\mathrm{E}_{\theta_i}\left[z_{nk} \mid X\right]$, which is given by

$$\mathrm{E}_{\theta_i}\left[z_{nk} \mid X\right] = \sum_{Z} z_{nk} p_{\theta_i}(Z \mid X) = \sum_{z_{nk}} z_{nk} p_{\theta_i}(z_{nk} \mid x_n)$$

$$\begin{aligned}
\mathrm{E}_{\theta_i}\left[z_{nk} \mid X\right] &= \sum_{z_{nk}} z_{nk} \frac{p_{\theta_i}(x_n \mid z_{nk}) p_{\theta_i}(z_{nk})}{p_{\theta_i}(x_n)} \\
&= \frac{\sum_{z_{nk}} z_{nk}\left(\pi_k \mathcal{N}\left(x_n \mid \mu_k, \Sigma_k\right)\right)^{z_{nk}}}{\sum_{j=1}^{K} \pi_j \mathcal{N}\left(x_n \mid \mu_j, \Sigma_j\right)} \\
&= \frac{\pi_k \mathcal{N}\left(x_n \mid \mu_k, \Sigma_k\right)}{\sum_{j=1}^{K} \pi_j \mathcal{N}\left(x_n \mid \mu_j, \Sigma_j\right)} \triangleq \gamma(z_{nk}),
\end{aligned}$$

where the last equality follows from the fact that $z_{nk} \in \{0,1\}$.

Algorithm 1 EM for Gaussian mixtures

1. **Initialise:** Initialize $\mu_k^1, \Sigma_k^1, \pi_k^1$ and set $i = 1$.
2. **While** not converged **do:**

(a) **Expectation (E) step:** Compute

$$\gamma(z_{nk}) = \frac{\pi_k^i \mathcal{N}(x_n \mid \mu_k^i, \Sigma_k^i)}{\sum_{j=1}^{K} \pi_j^i \mathcal{N}(x_n \mid \mu_j^i, \Sigma_j^i)}, \quad n = 1, \ldots, N, k = 1, \ldots, K.$$

(b) **Maximization (M) step:** Compute

$$\mu_k^{i+1} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})x_n, \quad \pi_k^{i+1} = \frac{N_k}{N}, \quad N_k = \sum_{n=1}^{N}\gamma(z_{nk})$$

$$\Sigma_k^{i+1} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(x_n - \mu_k^{i+1})(x_n - \mu_k^{i+1})^T$$

(c) $i \leftarrow i + 1$

Consider the same Gaussian mixture as before,

$$p(x) = \underbrace{0.3}_{\pi_1}\mathcal{N}\left(x \mid \underbrace{\begin{pmatrix}4\\4.5\end{pmatrix}}_{\mu_1}, \underbrace{\begin{pmatrix}1.2 & 0.6\\0.6 & 0.5\end{pmatrix}}_{\Sigma_1}\right) + \underbrace{0.5}_{\pi_2}\mathcal{N}\left(x \mid \underbrace{\begin{pmatrix}8\\1\end{pmatrix}}_{\mu_2}, \underbrace{\begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}}_{\Sigma_2}\right) + \underbrace{0.2}_{\pi_3}\mathcal{N}\left(x \mid \underbrace{\begin{pmatrix}9\\8\end{pmatrix}}_{\mu_3}, \underbrace{\begin{pmatrix}0.6 & 0.5\\0.5 & 1.5\end{pmatrix}}_{\Sigma_3}\right)$$

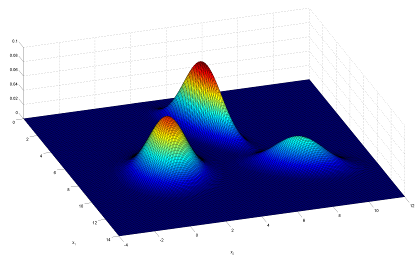


Figure: Probability density function.

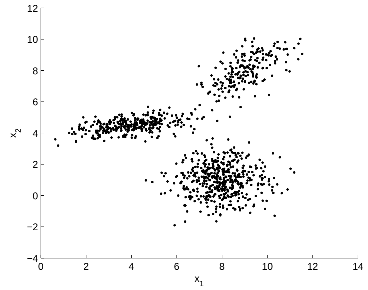


Figure: $N = 1000$ samples from the Gaussian mixture $p(x)$.

Example – EM for Gaussian mixtures (II/III)

- Apply the EM algorithm to estimate a Gaussian mixture with $K = 3$ Gaussians, i.e. use the 1000 samples to compute estimates of $\pi_1, \pi_2, \pi_3, \mu_1, \mu_2, \mu_3, \Sigma_1, \Sigma_2, \Sigma_3$.
- 200 iterations.

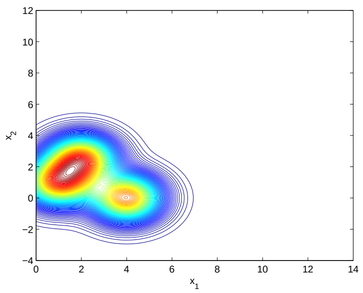


Figure: Initial guess.

Example – EM for Gaussian mixtures (III/III)

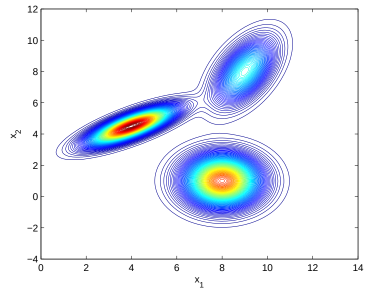


Figure: True PDF.

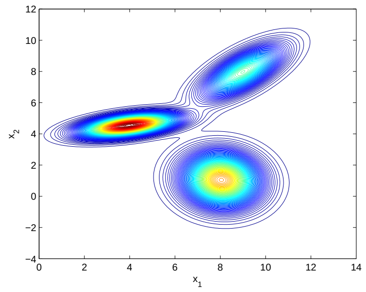


Figure: Estimate after 200 iterations of the EM algorithm.

The K-means algorithm (I/II)

Algorithm 2 K-means algorithm, a.k.a. Lloyd's algorithm

1. Initialize μ_k^1 and set $i = 1$.
2. Minimize J w.r.t. r_{nk} keeping $\mu_k = \mu_k^i$ fixed.

$$r_{nk}^{i+1} = \begin{cases} 1 & \text{if } k = \arg\min_j \|x_n - \mu_j^i\|^2 \\ 0 & \text{otherwise} \end{cases}$$

3. Minimize J w.r.t. μ_k keeping $r_{nk} = r_{nk}^{i+1}$ fixed.

$$\mu_k^{i+1} = \frac{\sum_{n=1}^N r_{nk}^{i+1} x_n}{\sum_{n=1}^N r_{nk}^{i+1}}.$$

4. If not converged, update $i := i + 1$ and return to step 2.

The K-means algorithm (II/II)

The name K-means stems from the fact that in step 3 of the algorithm, u_k is give by the mean of all the data points assigned to cluster k.

Note the **similarities** between the K-means algorithm and the EM algorithm for Gaussian mixtures!

K-means is deterministic with "hard" assignment of data points to clusters (no uncertainty), whereas EM is a probabilistic method that provides a "soft" assignment.

If the Gaussian mixtures are modeled using covariance matrices

$$\Sigma_k = \epsilon I, \quad k = 1, \ldots, K,$$

it can be shown that the EM algorithm for a mixture of K Gaussian's is **equivalent** to the K-means algorithm, when $\epsilon \to 0$.

Bayesian reminder

In the Bayesian framework we are interested in the posterior density $p(Z|X)$ given by Bayes' rule as

$$p(Z|X) = \frac{p(X|Z)p(Z)}{p(X)},$$

where $X = x_1, \ldots, x_N$ denotes the measurements and $Z = z_1, \ldots, z_N$ denotes the latent variables.

Sometimes the posterior can be found exactly using the concept of **conjugate priors**.

- Gaussian case
- More generally the exponential family.

What happens when there is no exact solution?

Variational methods (I/II)

Classic calculus involves functions and defines *derivatives* to optimize them.

The so-called **calculus of variations** investigates functions of functions which are called **functionals**.

Example: Entropy $\mathcal{H}[p(\cdot)] = -\int p(x)\log(p(x))\mathrm{d}x.$

The derivatives of functionals are called **variations**.

Calculus of variations has its origins in the 18th century and the most important result is probably the so-called Euler-lagrange equation

$$C(q) \triangleq \int \underbrace{L(t, q(t), q'(t))}_{\triangleq L(t,x,v)} dt, \quad L_x(t, q_*, q'_*) + \frac{d}{dt}L_v(t, q_*, q'_*) = 0,$$

which constitutes the core of optimal control theory.

Variational methods (II/II)

In general variational methods, one generally assumes a predetermined form of the argument function, possibly parametric.

- Quadratic: $q(x) = x^T A x + b^T x + c$
- Basis functions: $q(x) = \sum_{i=1}^{N_\phi} w_i \phi(x)$

Variational inference: In the case of probabilistic inference, the variational approximation takes the form:

$$q(Z) = \prod_{i=1}^{M} q_i(Z_i)$$

where $Z = \{Z_1, \ldots, Z_M\}$ is a partitioning of the unknown variables.

Variational inference

Algorithm 3 Variational iteration

Solve the problem iteratively:

1. For $j = 1, \ldots, M$
   (a) Fix $\{q_i(Z_i)\}_{\substack{i=1\\i\neq j}}^{M}$ to their last estimated values $\{\widehat{q}_i(Z_i)\}_{\substack{i=1\\i\neq j}}^{M}$.
   (b) Find the solution of
   $$\widehat{q}_j(Z_j) = \arg\max_{q_j} \mathcal{L}(q)$$
2. Repeat 1 until convergence.

Consider the following Bayesian LGSS model

$$x_{k+1} = \theta x_k + v_k, \quad y_k = \frac{1}{2}x_k + e_k, \quad x_0 \sim \mathcal{N}(x_0; \bar{x}_0, \Sigma_0), \quad \theta \sim \mathcal{N}(\theta; 0, \sigma_\theta^2),$$

$$\begin{pmatrix} v_k \\ e_k \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_v^2 & 0 \\ 0 & \sigma_e^2 \end{pmatrix}\right).$$

Aim: Compute the posterior $p(\theta | y_{0:N})$ using the VB framework.

- We have some latent variables $x_{0:N} \triangleq \{x_0, \dots, x_N\}$.
- Different notation compared to Bishop! The observations are denoted y and the latent variables are denoted x.

With latent variables

$$p(\theta | y_{0:N}) = \int p(\theta, x_{0:N} | y_{0:N}) \mathrm{d}x_{0:N}$$

There is still no exact form for the joint density $p(\theta, x_{0:N} | y_{0:N})$.

Variational Approximation

- Approximate the posterior $p(\theta, x_{0:N} | y_{0:N})$ as

$$p(\theta, x_{0:N} | y_{0:N}) \approx q_\theta(\theta) q_x(x_{0:N})$$

- Find $q_\theta(\theta)$ and $q_x(x_{0:N})$ using

$$\begin{aligned} \log q_\theta(\theta) &= E_{q_x}\left[\log p(y_{0:N}, x_{0:N}, \theta)\right] + \text{const.} \\ \log q_x(x_{0:N}) &= E_{q_\theta}\left[\log p(y_{0:N}, x_{0:N}, \theta)\right] + \text{const.} \end{aligned}$$

Variational Bayes formulas are

$$\begin{aligned} \log q_\theta(\theta) &= E_{q_x}\left[\log p(y_{0:N}, x_{0:N}, \theta)\right] + \text{const.} \\ \log q_x(x_{0:N}) &= E_{q_\theta}\left[\log p(y_{0:N}, x_{0:N}, \theta)\right] + \text{const.} \end{aligned}$$

We have the joint density $p(y_{0:N}, x_{0:N}, \theta)$ as

$$\begin{aligned} p(y_{0:N}, x_{0:N}, \theta) &= p(y_{0:N} | x_{0:N}) p(x_{1:N} | x_{0:N-1}, \theta) p(x_0) p(\theta) \\ &= \prod_{i=0}^{N} p(y_i | x_i) \prod_{i=1}^{N} p(x_i | x_{i-1}, \theta) p(x_0) p(\theta) \end{aligned}$$

Taking the logarithm and separating the constant terms

$$\log p(y_{0:N}, x_{0:N}, \theta) = -\sum_{i=0}^{N} \frac{0.5}{\sigma_e^2}(y_i - 0.5x_i)^2 - \sum_{i=1}^{N} \frac{0.5}{\sigma_v^2}(x_i - \theta x_{i-1})^2 - 0.5/\sigma_0^2 (x_0 - \bar{x}_0)^2 - 0.5/\sigma_\theta^2 \theta^2 + \text{const.}$$

Back to the Bishop's notation: x now denotes a measurement.

- Suppose we have $x_{1:N}$ i.i.d. and distributed as

$$x_i \sim p(x | \pi_{1:K}, \mu_{1:K}, \Lambda_{1:K}) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(x; \mu_k, \Lambda_k^{-1}\right)$$

- In the Bayesian framework, all the unknowns $\{\pi_{1:K}, \mu_{1:K}, \Lambda_{1:K}\}$ are random.

$$\pi_{1:K} \sim \mathrm{Dir}(\pi_{1:K} | \alpha_0) \overset{\triangle}{\propto} \prod_{k=1}^{K} \pi_k^{\alpha_0 - 1}$$

$$\mu_{1:K}, \Lambda_{1:K} \sim p(\mu_{1:K}, \Lambda_{1:K}) \triangleq \prod_{k=1}^{K} \mathcal{N}(\mu_k; m_0, (\beta_0 \Lambda_k)^{-1}) \mathcal{W}(\Lambda_k | W_0, \nu_0)$$

VB example 2 – Gaussian mixture inference

- Define the latent variables $z_i \triangleq [z_{i1}, \cdots, z_{iK}]^T$ as in EM. Then

$$p(x_{1:N}, z_{1:N}) = \prod_{i=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{ik}} \mathcal{N}\left(x; \mu_k, \Lambda_k^{-1}\right)^{z_{ik}}$$

- The Bayesian framework then asks for the posterior density $p(z_{1:N}, \pi_{1:K}, \mu_{1:K}, \Lambda_{1:K} | x_{1:N})$.

Variational Approximation

- Approximate the posterior as

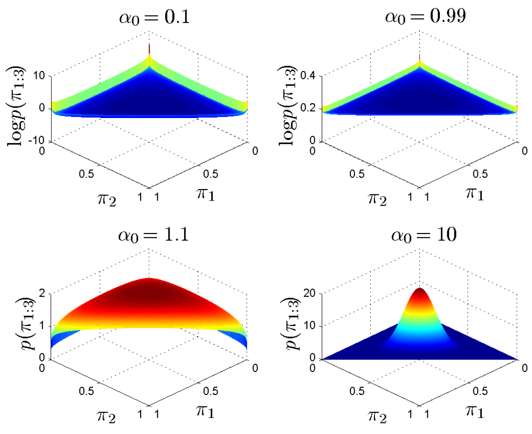
$$p(z_{1:N}, \pi_{1:K}, \mu_{1:K}, \Lambda_{1:K} | x_{1:N}) \approx q_z(z_{1:N}) q_{\pi,\mu,\Lambda}(\pi_{1:K}, \mu_{1:K}, \Lambda_{1:K})$$

- Find $q_z(z_{1:N})$ and $q_{\pi,\mu,\Lambda}(\pi_{1:K}, \mu_{1:K}, \Lambda_{1:K})$ iteratively.

VB example 2 – sparsity with Bayesian methods

Symmetric Dirichlet distribution for $K = 3$.

$$\pi_{1:3} \sim \mathrm{Dir}(\pi_{1:3} | \alpha_0) \overset{\triangle}{\propto} \prod_{k=1}^{3} \pi_k^{\alpha_0 - 1} = (\pi_1 \pi_2 (1 - \pi_1 - \pi_2))^{\alpha_0 - 1}$$



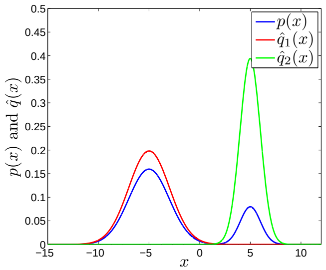
Minimization of KL-divergence (I/III)

Suppose we have

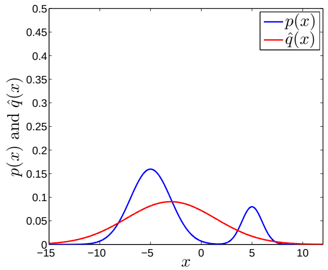
$$p(x) = 0.2\mathcal{N}(x; 5, 1) + 0.8\mathcal{N}(x, -5, 2^2)$$

Let $q_{\mu,\sigma}(x) \triangleq \mathcal{N}(x; \mu, \sigma^2)$

Find $\min_{\mu,\sigma} \mathrm{KL}(q_{\mu,\sigma} || p)$

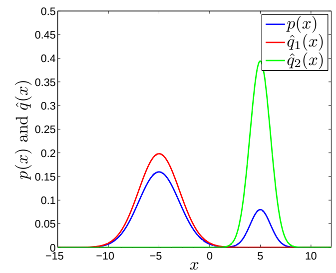


Find $\min_{\mu,\sigma} \mathrm{KL}(p || q_{\mu,\sigma})$



Minimization of KL-divergence (II/III)

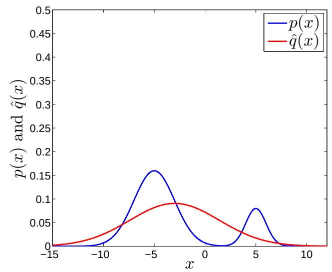
Find $\min_{\mu,\sigma} \mathrm{KL}(q_{\mu,\sigma} || p)$



$$\mathrm{KL}(q_{\mu,\sigma} || p) \triangleq \int q_{\mu,\sigma}(x) \log \frac{q_{\mu,\sigma}}{p(x)} \mathrm{d}x$$

zero–forcing

Find $\min_{\mu,\sigma} \mathrm{KL}(p || q_{\mu,\sigma})$



$$\mathrm{KL}(p || q_{\mu,\sigma}) \triangleq \int p(x) \log \frac{p(x)}{q_{\mu,\sigma}} \mathrm{d}x$$

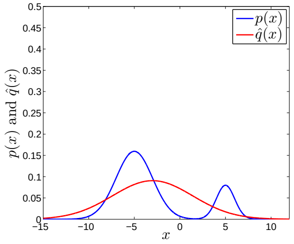
non-zero-forcing

Minimization of KL-divergence (III/III)

This second form of optimization

$$\mathrm{KL}(p||q_{\mu,\sigma}) \triangleq \int p(x) \log \frac{p(x)}{q_{\mu,\sigma}} \mathrm{d}x$$

has the following attractive property.



$$\widehat{\mu} = E_{\widehat{q}}(x) = E_p(x)$$
$$\widehat{\sigma}^2 = E_{\widehat{q}}\left[(x - E_{\widehat{q}}(x))^2\right] = E_p\left[(x - E_p(xx^T))^2\right]$$

- Similar properties hold for the entire exponential family.
- A variational method using this type of KL-divergence minimization and hence the expectation equations above is **Expectation Propagation (EP)**.

Expectation propagation (I/II)

- Suppose we have a posterior distribution in the form of

$$p(X|Y) \propto \prod_{i=1}^{I} f_i(X)$$

which is intractable or too computationally costly to compute.

- Then EP approximates the posterior as

$$p(X|Y) \approx q(X) \triangleq \prod_{i=1}^{I} q_i(X) = \prod_{i=1}^{I} \mathcal{N}(X; \mu_i, \Sigma_i)$$

- Ideally we want to minimize the KL divergence between the true posterior and the approximation,

$$\widehat{q}(X) = \arg\min_{q} \mathrm{KL}\left(\frac{1}{Z}\prod_{i=1}^{I} f_i(X) || \prod_{i=1}^{I} q_i(X)\right)$$

Expectation propagation (II/II)

Solving this is intractable, make the approximation that we minimize the KL divergence between pairs of factors $f_i(X)$ and $q_i(X)$.

- The terms $q_j(x_j)$ are estimated iteratively as in VB by keeping the last estimates of $\{\widehat{q}_i\}_{\substack{i=1\\i\neq j}}^{I}$.

$$\widehat{q}_j(X) = \arg\min_{q_j} \mathrm{KL}\left(f_j(X)\prod_{i\neq j}\widehat{q}_i(X) \middle\| q_j(X)\prod_{i\neq j}\widehat{q}_i(X)\right)$$

- This is in the Gaussian case obtained by solving the equations

$$E_{q_j \prod_{i\neq j}\widehat{q}_i}(X) = E_{f_j \prod_{i\neq j}\widehat{q}_i}(X)$$
$$E_{q_j \prod_{i\neq j}\widehat{q}_i}(XX^T) = E_{f_j \prod_{i\neq j}\widehat{q}_i}(XX^T)$$

for the mean μ_i and the covariance Σ_i of $\widehat{q}_i(\cdot)$.

EP example – smoothing under GM noise

Consider the following LGSS model

$$x_{k+1} = x_k + v_k, \qquad x_0 = 0 \text{ is known}$$
$$y_k = x_k + e_k, \qquad v_k \sim \mathcal{N}(v_k; 0, \sigma_v^2)$$

$$e_k \sim p_e(e_k) \triangleq 0.9\mathcal{N}(e_k; 0, \sigma_e^2) + 0.1\mathcal{N}(e_k; 0, (10\sigma_e)^2)$$

Aim: Compute the posterior density $p(x_{1:N}|y_{1:N})$.

- Recall that the true posterior factorizes as

$$p(x_{1:N}|y_{1:N}) \propto \prod_{i=1}^{N} p(y_i|x_i)p(x_i|x_{i-1})$$

- The true posterior in this case is a Gaussian mixture with 2^N components which is not feasible to compute.

- Make the variational approximation

$$p(x_{1:N}|y_{1:N}) \approx q(x_{1:N}) \triangleq \prod_{i=1}^{N} \mathcal{N}(x_i; \mu_i, \sigma_i^2)$$

- Consider the density for x_j given as

$$\bar{p}(x_j) \propto \int \int p(y_j|x_j)p(x_{j+1}|x_j)p(x_j|x_{j-1}) \\ \times \mathcal{N}(x_{j+1}; \mu_{j+1}, \sigma_{j+1}^2)\mathcal{N}(x_{j-1}; \mu_{j-1}, \sigma_{j-1}^2)\, \mathrm{d}x_{j+1}\mathrm{d}x_{j-1}$$

which can be calculated as

$$\begin{aligned}\bar{p}(x_j) =& w_1(\mu_{j\pm1}, \sigma_{j\pm1})\mathcal{N}\left(x_j; \eta_1(\mu_{j\pm1}, \sigma_{j\pm1}), \rho_1^2(\mu_{j\pm1}, \sigma_{j\pm1})\right) \\ &+ w_2(\mu_{j\pm1}, \sigma_{j\pm1})\mathcal{N}\left(x_j; \eta_2(\mu_{j\pm1}, \sigma_{j\pm1}), \rho_2^2(\mu_{j\pm1}, \sigma_{j\pm1})\right)\end{aligned}$$

$$\begin{aligned}\bar{p}(x_j) =& w_1(\mu_{j\pm1}, \sigma_{j\pm1})\mathcal{N}\left(x_j; \eta_1(\mu_{j\pm1}, \sigma_{j\pm1}), \rho_1^2(\mu_{j\pm1}, \sigma_{j\pm1})\right) \\ &+ w_2(\mu_{j\pm1}, \sigma_{j\pm1})\mathcal{N}\left(x_j; \eta_2(\mu_{j\pm1}, \sigma_{j\pm1}), \rho_2^2(\mu_{j\pm1}, \sigma_{j\pm1})\right)\end{aligned}$$

where the parameters $w_{1,2}, \eta_{1,2}$ and $\rho_{1,2}$ are

$$\eta_1 = \rho_1^2\left(\frac{\bar{\eta}}{\bar{\rho}^2} + \frac{y_j}{\sigma_e^2}\right) \qquad \eta_2 = \rho_2^2\left(\frac{\bar{\eta}}{\bar{\rho}^2} + \frac{y_j}{(10\sigma_e)^2}\right)$$

$$\rho_1^2 = \left(\frac{1}{\bar{\rho}^2} + \frac{1}{\sigma_e^2}\right)^{-1} \qquad \rho_2^2 = \left(\frac{1}{\bar{\rho}^2} + \frac{1}{(10\sigma_e)^2}\right)^{-1}$$

$$w_1 \propto 0.9\mathcal{N}\left(y_j; \bar{\eta}, \bar{\rho}^2 + \sigma_e^2\right) \qquad w_2 \propto 0.1\mathcal{N}\left(y_j; \bar{\eta}, \bar{\rho}^2 + (10\sigma_e)^2\right)$$

$$\bar{\eta} = \bar{\rho}^2\left(\frac{\mu_{j-1}}{\sigma_{j-1}^2 + \sigma_v^2} + \frac{\mu_{j+1}}{\sigma_{j+1}^2 + \sigma_v^2}\right) \qquad \bar{\rho}^2 = \left(\frac{1}{\sigma_{j-1}^2 + \sigma_v^2} + \frac{1}{\sigma_{j+1}^2 + \sigma_v^2}\right)^{-1}$$

$$\begin{aligned}\bar{p}(x_j) =& w_1(\mu_{j\pm1}, \sigma_{j\pm1})\mathcal{N}\left(x_j; \eta_1(\mu_{j\pm1}, \sigma_{j\pm1}), \rho_1^2(\mu_{j\pm1}, \sigma_{j\pm1})\right) \\ &+ w_2(\mu_{j\pm1}, \sigma_{j\pm1})\mathcal{N}\left(x_j; \eta_2(\mu_{j\pm1}, \sigma_{j\pm1}), \rho_2^2(\mu_{j\pm1}, \sigma_{j\pm1})\right)\end{aligned}$$

The EP solution for $q_j(x_j) = \mathcal{N}(x_j; \mu_j, \sigma_j^2)$ is obtained by matching (propagating) expectations between $q_j(\cdot)$ and $\bar{p}(x_j)$.

$$\begin{aligned}\mu_j =& w_1\eta_1 + w_2\eta_2 \\ \sigma_j^2 =& w_1\left(\rho_1^2 + (\eta_1 - \mu_j)^2\right) + w_2\left(\rho_2^2 + (\eta_2 - \mu_j)^2\right)\end{aligned}$$

Variational Inference: Approximate Bayesian inference where factorial approximations are made on the form of the posteriors.

Kullback-Leibler (KL) Divergence: A cost function to find optimal approximations for the posteriors in two different forms.

Variational Bayes: A form of variational inference where $\mathrm{KL}(q||p)$ is used for the optimization.

Expectation Propagation: A form of variational inference where $\mathrm{KL}(p||q)$ is used for the optimization.