Online Variational Bayesian Learning

Zoubin Ghahramani
Gatsby Computational Neuroscience Unit
University College London

December 2000

http://www.gatsby.ucl.ac.uk/

Theoretical Results

Theorem 1 Given an iid data set $\mathbf{y} = (\mathbf{y}_1, \ldots \mathbf{y}_n)$, if the model is **CE** then:

(a) $Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ is also conjugate, *i.e.*

$$Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = h(\tilde{\eta}, \tilde{\boldsymbol{\nu}}) g(\boldsymbol{\theta})^{\tilde{\eta}} \exp\left\{\boldsymbol{\phi}(\boldsymbol{\theta})^\top \tilde{\boldsymbol{\nu}}\right\}$$

(b) $Q_{\mathbf{x}}(\mathbf{x}) = \prod_{i=1}^n Q_{\mathbf{x}_i}(\mathbf{x}_i)$ is of the same form as in the E step of regular EM, but using pseudo parameters computed by averaging over $Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$

$$\begin{aligned} Q_{\mathbf{x}_i}(\mathbf{x}_i) &\propto f(\mathbf{x}_i, \mathbf{y}_i) \exp\left\{\overline{\boldsymbol{\phi}}(\boldsymbol{\theta})^\top \mathbf{u}(\mathbf{x}_i, \mathbf{y}_i)\right\} \\ &= P(\mathbf{x}_i | \mathbf{y}_i, \overline{\boldsymbol{\phi}}(\boldsymbol{\theta})) \end{aligned}$$

KEY points:

(a) the approximate parameter posterior is of the same form as the prior;

(b) the *approximate* hidden variable posterior, averaging over *all* parameters, is of the same form as the *exact* hidden variable posterior for a *single* setting of the parameters.

The Variational EM algorithm

VE Step: Compute the expected sufficient statistics $\mathbf{t}(\mathbf{y}) = \sum_i \overline{\mathbf{u}}(\mathbf{x}_i, \mathbf{y}_i)$ under the hidden variable distributions $Q_{\mathbf{x}_i}(\mathbf{x}_i)$.

VM Step: Compute expected natural parameters $\overline{\boldsymbol{\phi}}(\boldsymbol{\theta})$ under the parameter distribution given by $\tilde{\eta}$ and $\tilde{\boldsymbol{\nu}}$.

$$
\begin{aligned}
\tilde{\eta} &= \eta + n \\
\tilde{\boldsymbol{\nu}} &= \boldsymbol{\nu} + \sum_i \overline{\mathbf{u}}(\mathbf{x}_i, \mathbf{y}_i)
\end{aligned}
$$

Properties:

- VE step has same complexity as corresponding E step.
- Reduces to the EM algorithm if $Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$. M step then involves re-estimation of $\boldsymbol{\theta}^*$.
- $\mathcal{F}$ increases monotonically, and incorporates the model complexity penalty.

Online Bayesian learning for CE models

$$\tilde{\eta}_t = \tilde{\eta}_{t-1} + 1$$
$$\tilde{\boldsymbol{\nu}}_t = \boldsymbol{\nu}_{t-1} + \overline{\mathbf{u}}(\mathbf{x}_t, \mathbf{y}_t)$$

Algorithm:

1. Initialize: $\tilde{\eta}_0 = \eta$, $\tilde{\boldsymbol{\nu}}_0 = \boldsymbol{\nu}$
   Compute $\overline{\boldsymbol{\phi}}_0$ using $\tilde{\eta}_0$, $\tilde{\boldsymbol{\nu}}_0$
   Set $t = 1$

2. Get data $\mathbf{y}_t$

3. Infer $\overline{\mathbf{u}}(\mathbf{x}_t, \mathbf{y}_t)$ using $\overline{\boldsymbol{\phi}}_{t-1}$

4. Update parameters of approximating Q distributions:

$$\tilde{\eta}_t = \tilde{\eta}_{t-1} + 1$$
$$\tilde{\boldsymbol{\nu}}_t = \boldsymbol{\nu}_{t-1} + \overline{\mathbf{u}}(\mathbf{x}_t, \mathbf{y}_t)$$

5. Compute $\overline{\boldsymbol{\phi}}_t$ using $\tilde{\eta}_t$, $\tilde{\boldsymbol{\nu}}_t$.

6. $t \leftarrow t + 1$

7. Goto 2

Example: Online Mixture of Gaussians

Data: y
Hidden discrete variables: s_k

Parameters: mixing proportions π_k, means μ_k and precisions ρ_k

Hyperparameters: α_k (for Dirichlet mixing proportions)
m_k, g_k (for Gaussian means)
a_k, b_k (for Gamma precisions)

Expected natural parameters: $\langle \ln \pi_k \rangle$, $\langle \ln \rho_k \rangle$, $\langle \ln \rho_k \rangle$, $\langle \rho_k \rangle$, $\langle \rho_k \mu_k \rangle$, $\langle \rho_k \mu_k^2 \rangle$.

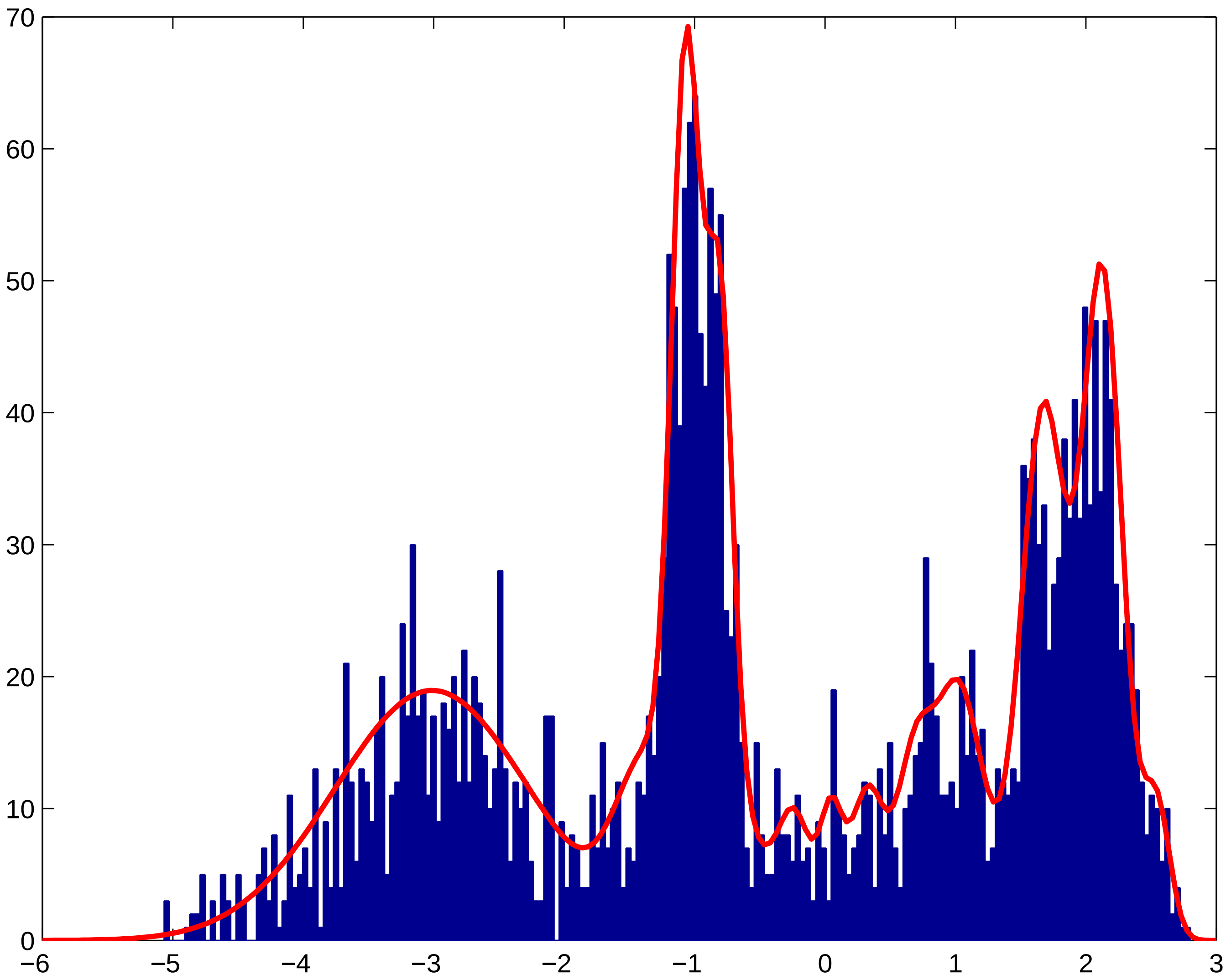
VE step:

$$Q(s_k) \propto \exp \left\{ s_k \left[\langle \ln \pi_k \rangle + \frac{1}{2} \langle \ln \rho_k \rangle - \frac{1}{2} \langle \rho_k \rangle y^2 + \langle \rho_k \mu_k \rangle y - \frac{1}{2} \langle \rho_k \mu_k^2 \rangle \right] \right\}$$

VM step:

$$\begin{aligned}
\tilde{\alpha}_k &\leftarrow \tilde{\alpha}_k + \langle s_k \rangle \\
\tilde{a}_k &\leftarrow \tilde{a}_k + \frac{1}{2} \langle s_k \rangle \\
\tilde{b}_k &\leftarrow \tilde{b}_k + \frac{1}{2} \left[\frac{\langle s_k \rangle \tilde{g}_k (y - \tilde{m}_k)^2}{\tilde{g}_k + \langle s_k \rangle} \right] \\
\tilde{m}_k &\leftarrow \frac{\tilde{g}_k \tilde{m}_k + y \langle s_k \rangle}{\tilde{g}_k + \langle s_k \rangle} \\
\tilde{g}_k &\leftarrow \tilde{g}_k + \langle s_k \rangle
\end{aligned}$$

Online Mixture of Gaussians



3000 data points generated from 5 clusters
fit using online variational Bayes with 20 clusters

Summary & Conclusions

- Tractable Bayesian learning using variational methods
- Conjugate-exponential families
- Variational EM
- Online Variational EM
- Mixture of Gaussians example