

# **Stochastic Complexities of Gaussian Mixtures in Variational Bayesian Approximation**

Presented by Iulian Pruteanu  
March 9, 2007

# Outline

- **Introduction**
- **Gaussian mixture models**
- **Bayesian learning**
- **Variational Bayesian learning**
- **Results**
- **Conclusion**

# **Introduction**

---

**Gaussian mixture model**

**Bayesian learning: MCMC, Laplace approximation (not for hidden variables)**

**Variational Bayesian learning: suitable for models with hidden variables)**

**Stochastic complexity (main contribution of the paper: upper and lower bounds on the variational stochastic complexity)**

**Variational stochastic complexity**

- the accuracy of the variational Bayesian learning
- the influence of the hyperparameters on the learning process

# Gaussian mixture model

$$p(\mathbf{x} | \boldsymbol{\theta}) = \sum_{k=1}^K \frac{a_k}{\sqrt{2\pi\sigma_k^2}^M} \exp\left(-\frac{\|\mathbf{x} - \boldsymbol{\mu}_k\|^2}{2\sigma_k^2}\right)$$

$M$  - dimensional normal distribution ( $\mathbf{x} \in \mathbb{R}^M$ )

$K$  - number of components

$$\boldsymbol{\theta} = \{a_k, \boldsymbol{\mu}_k\}_{k=1}^K$$

$$\sigma_k > 0$$

# Bayesian learning

$$\mathbf{X}^n = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\} \quad p_0(\mathbf{x}) = \text{true distribution}$$

$$p(\boldsymbol{\theta} | \mathbf{X}^n) = \frac{1}{Z(\mathbf{X}^n)} \varphi(\boldsymbol{\theta}) \prod_{i=1}^n p(\mathbf{x}_i | \boldsymbol{\theta}) \quad (1)$$

$Z(\mathbf{X}^n)$  = normalization constant (marginal likelihood, evidence)

$F(\mathbf{X}^n) = -\log Z(\mathbf{X}^n) \implies$  Bayesian stochastic complexity

$$F_0(\mathbf{X}^n) = -\log Z_0(\mathbf{X}^n) = F(\mathbf{X}^n) - S(\mathbf{X}^n)$$

normalized Bayesian stochastic complexity

$$S(\mathbf{X}^n) = -\sum_{i=1}^n \log p_0(\mathbf{x}_i) \implies \text{empirical entropy}$$

$$p(\boldsymbol{\theta} | \mathbf{X}^n) = \frac{1}{Z_0(\mathbf{X}^n)} \exp(-nH_n(\boldsymbol{\theta})) \varphi(\boldsymbol{\theta}) \quad (2)$$

$$H_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n \log \frac{p_0(\mathbf{x}_i)}{p(\mathbf{x}_i | \boldsymbol{\theta})} \implies \text{empirical Kullback information}$$

## Variational Bayesian learning (1/2)

$$\begin{aligned} F(\mathbf{X}^n) &= -\log \int \sum_{\mathbf{Y}^n} \varphi(\boldsymbol{\theta}) \prod_{i=1}^n p(\mathbf{x}_i, \mathbf{y}_i | \boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= -\log \int \sum_{\mathbf{Y}^n} p(\mathbf{X}^n, \mathbf{Y}^n, \boldsymbol{\theta}) d\boldsymbol{\theta} \end{aligned}$$

$\mathbf{Y}^n = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n\} \equiv$  hidden variables

$$\begin{aligned} F(\mathbf{X}^n) &\leq \sum_{\mathbf{Y}^n} \int q(\mathbf{Y}^n, \boldsymbol{\theta} | \mathbf{X}^n) \log \frac{q(\mathbf{Y}^n, \boldsymbol{\theta} | \mathbf{X}^n)}{p(\mathbf{Y}^n, \boldsymbol{\theta} | \mathbf{X}^n)} d\boldsymbol{\theta} \\ &\equiv \bar{F}[q] \equiv \text{variational free energy} \end{aligned}$$

$$q(\mathbf{Y}^n, \boldsymbol{\theta} | \mathbf{X}^n) = Q(\mathbf{Y}^n | \mathbf{X}^n) r(\boldsymbol{\theta} | \mathbf{X}^n) \tag{3}$$

**Theorem 1:** If the functional  $\bar{F}[q]$  is minimized under the constraint (3) then the variational posteriors,  $r(\boldsymbol{\theta} | \mathbf{X}^n)$  and  $Q(\mathbf{Y}^n | \mathbf{X}^n)$  satisfy

$$r(\boldsymbol{\theta} | \mathbf{X}^n) = \frac{1}{C_r} \varphi(\boldsymbol{\theta}) \exp \left\langle \log p(\mathbf{X}^n, \mathbf{Y}^n | \boldsymbol{\theta}) \right\rangle_{Q(\mathbf{Y}^n | \mathbf{X}^n)}$$

$$Q(\mathbf{Y}^n | \mathbf{X}^n) = \frac{1}{C_Q} \exp \left\langle \log p(\mathbf{X}^n, \mathbf{Y}^n | \boldsymbol{\theta}) \right\rangle_{r(\boldsymbol{\theta} | \mathbf{X}^n)}$$

## Variational Bayesian learning (2/2)

$$\bar{F}(\mathbf{X}^n) = \min_{r, Q} \bar{F}[q] \implies \text{variational stochastic complexity}$$

$$\bar{F}(\mathbf{X}^n) - F(\mathbf{X}^n) = \min_{r, Q} \text{KL}(q(\mathbf{Y}^n, \boldsymbol{\theta} | \mathbf{X}^n) \| p(\mathbf{Y}^n, \boldsymbol{\theta} | \mathbf{X}^n))$$

$$\bar{F}_0(\mathbf{X}^n) = \bar{F}(\mathbf{X}^n) - S(\mathbf{X}^n) \implies \text{normalized variational stochastic complexity}$$

### Lemma 2:

$$\bar{F}_0(\mathbf{X}^n) = \min_{r(\boldsymbol{\theta} | \mathbf{X}^n)} \{ \text{KL}(r(\boldsymbol{\theta} | \mathbf{X}^n) \| \varphi(\boldsymbol{\theta})) - (\log C_Q + S(\mathbf{X}^n)) \} \quad (4)$$

$$C_Q = \sum_{\mathbf{Y}^n} \exp \left\langle \log p(\mathbf{X}^n, \mathbf{Y}^n | \boldsymbol{\theta}) \right\rangle_{r(\boldsymbol{\theta} | \mathbf{X}^n)}$$

$$\bar{\boldsymbol{\theta}}_{vb} = \operatorname{argmin}_{\bar{\boldsymbol{\theta}}} \{ \text{KL}(r(\boldsymbol{\theta} | \bar{\boldsymbol{\theta}}) \| \varphi(\boldsymbol{\theta})) - (\log C_Q(\bar{\boldsymbol{\theta}}) + S(\mathbf{X}^n)) \} \quad (5)$$

$$\bar{\boldsymbol{\theta}} = \left\langle \boldsymbol{\theta} \right\rangle_{r(\boldsymbol{\theta} | \mathbf{X}^n)}$$

## Main results (1/3)

$p_0(\mathbf{x})$  == true distribution

$$p_0(\mathbf{x} | \boldsymbol{\theta}_0) = \sum_{k=1}^{K_0} \frac{a_k^*}{\sqrt{2\pi}^M} \exp\left(-\frac{\|\mathbf{x} - \boldsymbol{\mu}_k^*\|^2}{2}\right) \quad (6)$$

$$\boldsymbol{\theta}_0 = \{a_k^*, \boldsymbol{\mu}_k^*\}_{k=1}^{K_0}$$

Suppose that the true distribution can be realized by the model, with  $K$  components

$$p_0(\mathbf{x} | \boldsymbol{\theta}) = \sum_{k=1}^K \frac{a_k}{\sqrt{2\pi}^M} \exp\left(-\frac{\|\mathbf{x} - \boldsymbol{\mu}_k\|^2}{2}\right) \quad K \geq K_0$$

$$\varphi(\mathbf{a}) = \frac{\Gamma(K\phi_0)}{\Gamma(\phi_0)^K} \prod_{k=1}^K a_k^{\phi_0-1}$$

$$\varphi(\boldsymbol{\mu}) = \prod_{k=1}^K \sqrt{\frac{\xi_0}{2\pi}}^M \exp\left(-\frac{\xi_0 \|\boldsymbol{\mu}_k - \mathbf{v}_0\|^2}{2}\right)$$

$$\xi_0 > 0, \mathbf{v}_0 \in \mathcal{R}^M, \phi_0 > 0 \quad (7)$$

## Main results (2/3)

**Theorem 3:** Assume the conditions (4) and (5). Then the normalized variational stochastic complexity  $\bar{F}_0(\mathbf{X}^n)$  satisfies

$$\underline{\lambda} \log n + nH_n(\bar{\theta}_{vb}) + C_1 \leq \bar{F}_0(\mathbf{X}^n) \leq \bar{\lambda} \log n + C_2$$

with probability 1 for an arbitrary natural number  $n$  where  $C_1, C_2$  are constants independent of  $n$  and the coefficients  $\underline{\lambda}, \bar{\lambda}$  are given by

$$\underline{\lambda} = \begin{cases} (K-1)\phi_0 + \frac{M}{2} & \phi_0 \leq \frac{M+1}{2} \\ \frac{MK+K-1}{2} & \phi_0 > \frac{M+1}{2} \end{cases} \quad (8)$$

$$\bar{\lambda} = \begin{cases} (K-K_0)\phi_0 + \frac{MK_0+K_0-1}{2} & \phi_0 \leq \frac{M+1}{2} \\ \frac{MK+K-1}{2} & \phi_0 > \frac{M+1}{2} \end{cases} \quad (9)$$

## Main results (3/3)

**Corollary 4:** Assume the conditions (4) and (5). Then the average of the normalized variational stochastic complexity  $\bar{F}_0(\mathbf{X}^n)$  satisfies

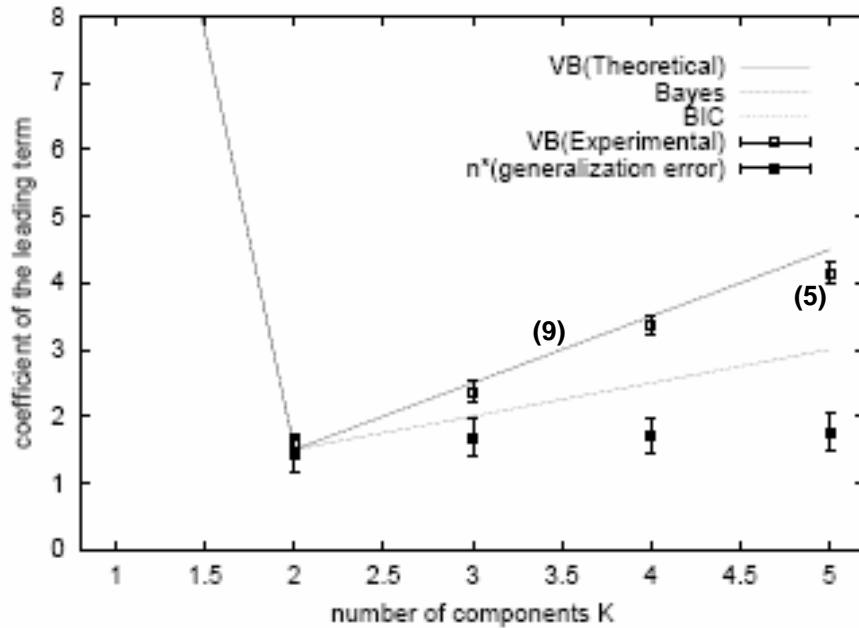
$$\underline{\lambda} \log n + E_{\mathbf{X}^n} [nH_n(\bar{\theta}_{vb})] + C_1 \leq E_{\mathbf{X}^n} [\bar{F}_0(\mathbf{X}^n)] \leq \bar{\lambda} \log n + C_2$$

where  $E_{\mathbf{X}^n}$  is the expectation over all sets of training samples.

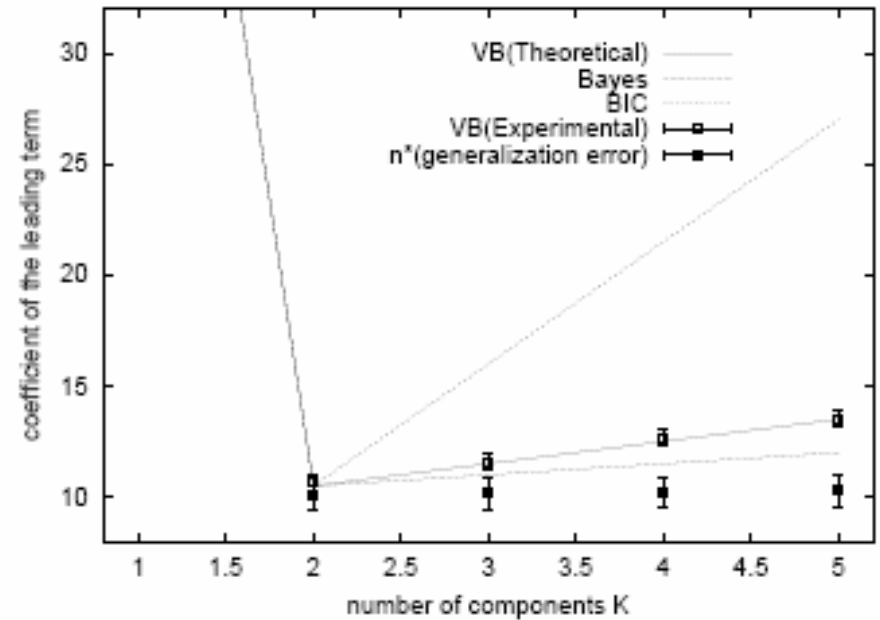
### Proof of Theorem 3:

- derivations for the variational posteriors  $r(\boldsymbol{\theta} | \mathbf{X}^n)$  and  $Q(\mathbf{Y}^n | \mathbf{X}^n)$
- evaluation of  $\text{KL}(r(\boldsymbol{\theta} | \mathbf{X}^n) || \varphi(\boldsymbol{\theta}))$  and  $\log C_Q + S(\mathbf{X}^n)$  in Lemma 2

# Experiments



(a). M=1



(b). M=10

The true distribution is a Gaussian mixture with 2 components

$$a_1^* = a_2^* = \frac{1}{2}$$

$$\lambda_{\text{VB}} = \frac{\bar{F}_0(\mathbf{X}^{1000}) - \bar{F}_0(\mathbf{X}^{100})}{\log 10} \quad \text{see (5)}$$

$$\mu_1^* = -\frac{2}{\sqrt{M}} \cdot \mathbf{1}^M; \mu_2^* = \frac{2}{\sqrt{M}} \cdot \mathbf{1}^M$$

$$\phi_0 = 1; \nu_0 = 0; \xi_0 = 1$$

## **Applications of the bounds**

---

- investigate the properties of the iterative algorithm in VB learning.
- examine whether the algorithm converges to the optimal variational posterior, instead of local minima.
- the variational stochastic complexity is used as a criterion for model selection in VB learning.

## References

---

- [1] H. Akaike. Likelihood and bayes procedure. *Bayesian Statistics*, pages 143-166, 1980.
- [2] H. Attias. Inferring parameters and structure of latent variable models by variational bayes. *Proceedings of Uncertainty in Artificial Intelligence*, 1999.
- [3] H. Alzer. On some inequalities for the Gamma and Psi functions. *Mathematics of computation*, volume 66, pages 373-389, 1997.
- [4] M. J. Beal. Variational algorithm for approximate Bayesian inference. PhD. Thesis, University College London, 2003.
- [5] Z. Ghahramani and M. J. Beal. Graphical models and variational methods. *Advanced Mean Field Methods – Theory and Practice*, MIT Press, 2000.