Statistical Models & Computing Methods

Lecture 12: Variational EM

Cheng Zhang

School of Mathematical Sciences, Peking University

November 15, 2021

EM Recap

- EM algorithm finds the MLE for latent variable model

$$\mathcal{L}(\theta) = \log p(x|\theta) = \log \sum_z p(x, z|\theta)$$

- EM update formula

$$\theta^{(t+1)} = \arg\max_\theta Q^{(t)}(\theta) = \arg\max_\theta \mathbb{E}_{p(z|x,\theta^{(t)})} \log p(x, z|\theta)$$

- EM requires the posterior $p(z|x, \theta^{(t)})$ is known. What if $p(z|x, \theta^{(t)})$ is unknown?
  - If somehow we can sample from $p(z|x, \theta^{(t)})$, we can use Monte Carlo estimates, that is Monte Carlo EM.
  - However, the associated computation may be expensive.

Variational EM

- Recall EM maximizes the lower bound

$$\mathcal{F}(q, \theta^{(t)}) = \mathbb{E}_{q(z)} \log \frac{p(x, z|\theta)}{q(z)} \leq \mathcal{L}(\theta), \quad \forall q(z)$$

- When the best $q(z) = p(z|x, \theta^{(t)})$ is not available, we can use approximate $q(z)$ instead.
- A widely used approximation is the mean-field approximation

$$q(z) = \prod_{i=1}^{d} q_i(z_i)$$

Mean-Field Lower Bound

- In that case, the lower bound is

$$
\begin{aligned}
\mathcal{F}(q(z), \theta^{(t)}) &= \int \prod_{i=1}^{d} q_i(z_i) \log \frac{p(x, z|\theta^{(t)})}{\prod_{i=1}^{d} q_i(z_i)} \, dz_1 dz_2 \dots dz_d \\
&= \int \prod_{i=1}^{d} q_i(z_i) \log p(x, z|\theta^{(t)}) \, dz_1 dz_2 \dots dz_d \\
&\quad - \sum_{i=1}^{d} \int q_i(z_i) \log q_i(z_i) \, dz_i
\end{aligned}
$$

- Coordinate Ascent

$$
q_i^{(t)}(z_i) \propto \exp\left(\mathbb{E}_{-q_i} \log p(x, z|\theta^{(t)})\right), i = 1, \dots, d
$$

- E-step. Run coordinate ascent several times to obtain good mean-field approximation

$$q^{(t)}(z) = \prod_{i=1}^{d} q_i^{(t)}(z_i)$$

compute the expected complete data log-likelihood

$$Q^{(t)}(\theta) = \mathbb{E}_{q^{(t)}(z)} \log p(x, z|\theta)$$

- M-step. Update θ to maximize $Q^{(t)}(\theta)$

$$\theta^{(t+1)} = \arg\max_{\theta} Q^{(t)}(\theta)$$

Variational Bayesian EM

- Now let us consider Bayesian inference for latent variable models

$$p(z, \theta|x) \propto p(x, z|\theta)p(\theta)$$

- We can lower bound the marginal likelihood

$$\begin{aligned}
\mathcal{L}(x) = \log p(x) &= \log \int p(x, z|\theta)p(\theta)\ dzd\theta \\
&= \log \int q(z, \theta) \frac{p(x, z|\theta)p(\theta)}{q(z, \theta)}\ dzd\theta \\
&\geq \int q(z, \theta) \log \frac{p(x, z|\theta)p(\theta)}{q(z, \theta)}\ dzd\theta \\
&= \mathcal{F}(q(z, \theta))
\end{aligned}$$

- Maximizing this lower bound $\mathcal{F}$ is equivalent to minimizing $D_{\mathrm{KL}}(q(z, \theta) \| p(z, \theta|x))$

- Again, we consider a simple factorized approximation $q(z,\theta) = q_z(z)q_\theta(\theta)$

$$\mathcal{L}(x) \geq \int q_z(z)q_\theta(\theta)\log\frac{p(x,z|\theta)p(\theta)}{q_z(z)q_\theta(\theta)}\,dzd\theta$$
$$= \mathcal{F}(q_z(z), q_\theta(\theta))$$

- Maximizing this lower bound $\mathcal{F}$, leads to **EM**-like iterative updates

$$q_z^{(t+1)}(z) \propto \exp\left(\mathbb{E}_{q_\theta^{(t)}(\theta)}\log p(x,z|\theta)\right)$$
$$q_\theta^{(t+1)}(\theta) \propto p(\theta)\cdot\exp\left(\mathbb{E}_{q_z^{(t+1)}(z)}\log p(x,z|\theta)\right)$$

Conjugate-Exponential Models

Let's focus on conjugate-exponential (CE) models, which satisfy

Condition 1

The joint probability over variables is in the exponential family

$$p(x, z|\theta) = h(x, z) \exp\left(\phi(\theta) \cdot T(x, z) - A(\theta)\right)$$

Condition 2

The prior over parameters is conjugate to this joint probability

$$p(\theta|\eta, \nu) \propto \exp\left(\phi(\theta) \cdot \nu - \eta A(\theta)\right)$$

Conjugate priors are computationally convenient and have an intuitive interpretation:

- η: number of pseudo-observations
- ν: values of pseudo-observations

Conjugate-Exponential Models

Now suppose we have an iid data set $x = \{x_1, \ldots, x_n\}$

- VB E-step.

$$
\begin{aligned}
q_z^{(t+1)}(z) &\propto \exp\left(\mathbb{E}_{q_\theta^{(t)}(\theta)} \log p(x, z|\theta)\right) \\
&\propto \prod_{i=1}^{n} h(x_i, z_i) \exp\left(\bar{\phi} \cdot T(x_i, z_i)\right)
\end{aligned}
$$

  where $\bar{\phi} = \mathbb{E}_{q_\theta^{(t)}}(\phi(\theta))$

- VB M-step

$$
q_\theta^{(t+1)}(\theta) \propto \exp\left(\phi(\theta) \cdot \left(\nu + \sum_{i=1}^{n} \overline{T}(x_i, z_i)\right) - (\eta + n)A(\theta)\right)
$$

  where $\overline{T}(x_i, z_i) = \mathbb{E}_{q_z^{(t+1)}}(T(x_i, z_i))$

EM for MAP

- Goal: maximize $p(x, \theta)$
- **E-step**: compute

$$q_z^{(t+1)}(z) = p(z|x, \theta^{(t)})$$

- **M-step**:

$$\theta^{(t+1)} = \arg\max_\theta Q^{(t)}(\theta)$$
$$Q^{(t)}(\theta) = \mathbb{E}_{q_z^{(t+1)}} \log p(x, z, \theta)$$

Variational Bayesian EM

- Goal: lower bound $p(x)$
- **VB E-step**: compute

$$q_z^{(t+1)}(z) = p(z|x, \bar{\phi})$$

- **VB M-step**:

$$q_\theta^{(t+1)}(\theta) \propto \exp\left(Q^{(t)}(\theta)\right)$$
$$Q^{(t)}(\theta) = \mathbb{E}_{q_z^{(t+1)}} \log p(x, z, \theta)$$

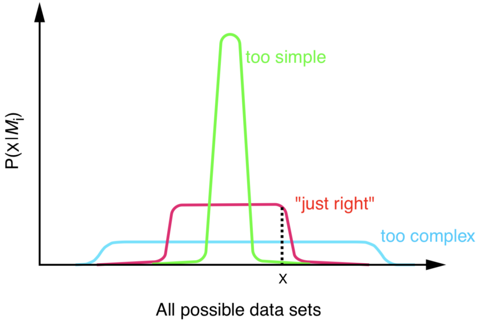
Properties of VBEM

- Reduces to the EM algorithm if $q_\theta(\theta) = \delta(\theta - \theta^*)$.
- $\mathcal{F}$ increases monotonically, and incorporates the model complexity penalty.
- Analytical parameter distributions
- VB E-step has the same complexity as corresponding E step, and is almost identical except that it uses the expected natural parameters, $\bar{\phi}$.
- The lower bound given by VBEM can be used for model selection.

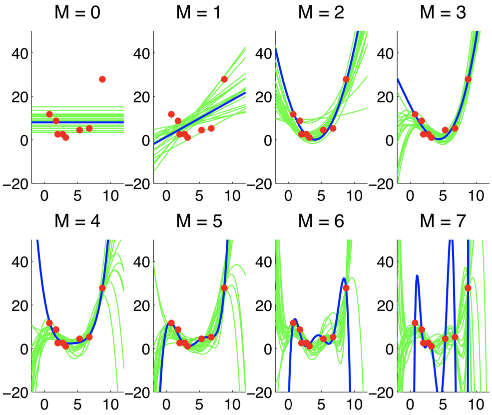
- In Bayesian model selection, we want to select the model class with the highest marginal likelihood (evidence)

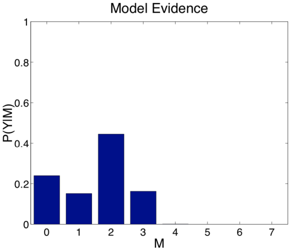
$$p(x|m) = \int p(x|\theta, m)p(\theta|m)d\theta$$

- Occam's Razor



Bayesian Model Selection





Adapted from Zoubin Ghahramani

Marginal Likelihood Estimation

- **Bayesian Information Criterion** (BIC):

$$\log p(x|m) \approx \log p(x|\hat{\theta}_{\mathrm{MAP}}, m) - \frac{d}{2}\log n$$

- **Annealed Importance Sampling** (AIS):

$$Z_k = \int p(x|\theta, m)^{\tau_k} p(\theta|m) d\theta, \quad 0 = \tau_0 < \cdots < \tau_K = 1$$

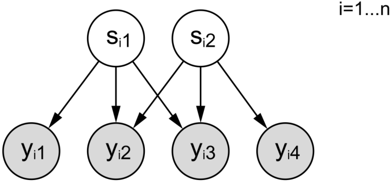
$$\log p(x|\theta) = Z_K = \prod_{k=0}^{K-1} \frac{Z_{k+1}}{Z_k}$$

  where $\frac{Z_{k+1}}{Z_k}$ can be estimated via importance sampling.

- **Variational Bayesian EM** (VB): use VBEM lower bound estimate

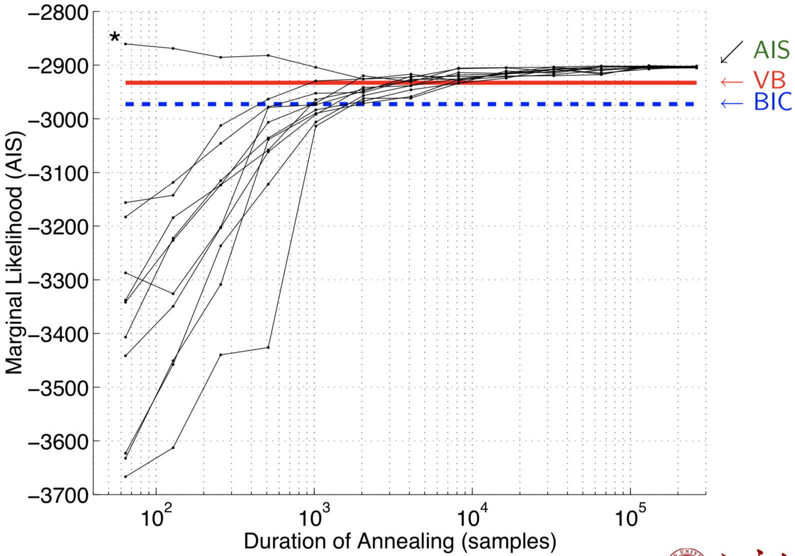
Example: A Bipartite Structured Model

- A simple bipartite graphical model: **two** binary hidden variables, and **four** five-valued discrete observed variables



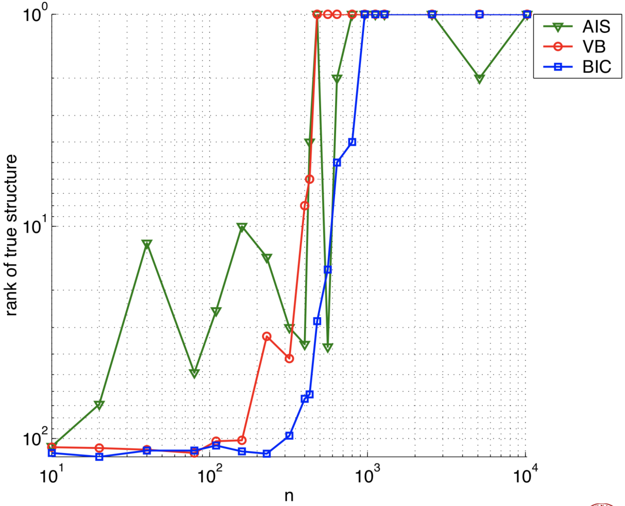
- Experiment: there are 136 distinict structures with 2 latent variables as potential parents of 4 conditionally independent observed variables
- Score each structure with 3 methods: BIC, VB and the gold standard AIS.

How Reliable is The AIS Gold Standard?



Ranking The True Structure

VB score finds correct structure earlier, and more reliably



References

- Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6:461–464.
- Neal, R. M. (2001). Annealed importance sampling. Statistics and Computing, 11:125–139.
- M. J. Beal and Z. Ghahramani, “The variational bayesian em algorithm for in- complete data: With application to scoring graphical model structures”, Bayesian statistics, vol. 7, J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D Heckerman, A. Smith, M West, et al., Eds., pp. 453–464, 2003.