Recursive parameter estimation and inference with incomplete data – Recursive EM & VB

Junhao Hua

2013/12/30

Reference

- Titterington, D. Michael. "**Recursive parameter estimation using incomplete data**." *Journal of the Royal Statistical Society. Series B (Methodological)* (1984): 257-267.
- Lange, Kenneth. "**A gradient algorithm locally equivalent to the EM algorithm**.“ *Journal of the Royal Statistical Society. Series B (Methodological)* (1995): 425-437.
- Neal, Radford M., and Geoffrey E. Hinton. "**A view of the EM algorithm that justifies incremental, sparse, and other variants**." *Learning in graphical models*. Springer Netherlands, 1998. 355-368.
- Sato, Masa-Aki. "**Convergence of on-line EM algorithm**." *2000*.
- Sato, Masa-Aki. "**Online model selection based on the variational Bayes**.“ *Neural Computation* 13.7 (2001): 1649-1681.
- Cappé, Olivier, and Eric Moulines. "**On-line expectation–maximization algorithm for latent data models**." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71.3 (2009): 593-613.
- Hoffman, Matt, et al. "**Stochastic variational inference**." *jmlr,2013*

Outline

- Problem statement
- A Recursive Procedure
- Recursive Expectation Maximum algorithm
  - Titterington,1984
  - Iange,1995
  - Sato,2000
  - Cappe, 2009
- A Bridge from EM to VB: Free Energy
  - Neal & Hinton, 1993,1998
- Recursive Variational Bayes
  - Sato,2000
  - Hoffman, Blei,2010,2011

Outline

- Problem statement
- A Recursive Procedure
- Recursive Expectation Maximum algorithm
  - Titterington,1984
  - Iange,1995
  - Sato,2000
  - Cappe, 2009
- A Bridge from EM to VB: Free Energy
  - Neal & Hinton, 1993,1998
- Recursive Variational Bayes
  - Sato,2000
  - Hoffman, Blei,2010,2011

Problem statement

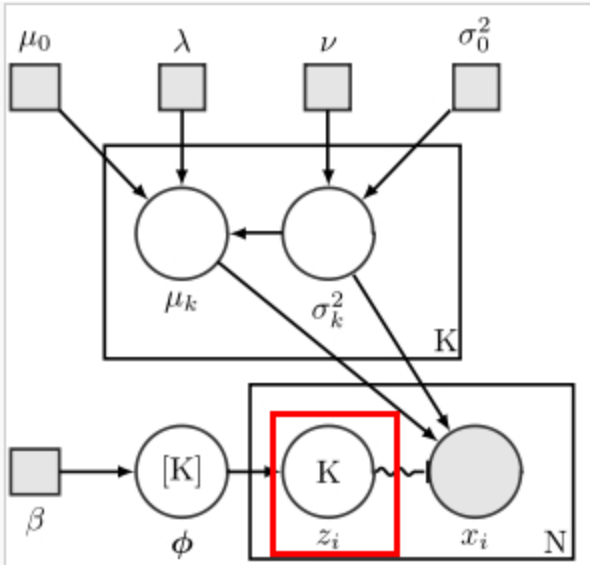
- How to compute maximum-likelihood estimates from incomplete data ?
- How when in Bayesian frameworks?
- What's the relation between EM/VB and gradient-based methods?
- How to deal with the streaming and large-scale data?

Incomplete data

The term "incomplete data" in its general form implies the existence of two sample spaces $\mathscr{Y}$ and $\mathscr{X}$ and a many–one mapping from $\mathscr{X}$ to $\mathscr{Y}$. The observed data $\mathbf{y}$ are a realization from $\mathscr{Y}$. The corresponding $\mathbf{x}$ in $\mathscr{X}$ is not observed directly, but only indirectly through $\mathbf{y}$. More specifically, we assume there is a mapping $\mathbf{x} \to \mathbf{y}(\mathbf{x})$ from $\mathscr{X}$ to $\mathscr{Y}$, and that $\mathbf{x}$ is known only to lie in $\mathscr{X}(\mathbf{y})$, the subset of $\mathscr{X}$ determined by the equation $\mathbf{y} = \mathbf{y}(\mathbf{x})$, where $\mathbf{y}$ is the observed data. We refer to $\mathbf{x}$ as the *complete data* even though in certain examples $\mathbf{x}$ includes what are traditionally called parameters.

- From: P. Dempstern, "Maximum Likelihood from Incomplete Data via the EM Algorithm", 1977

- It also refers to the latent variable.



A Recursive Procedure (Titterington, 1984)

Suppose $y_1, y_2, \ldots$ are independent observations, each with underlying probability density function (p.d.f.) $g(y \mid \boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \Theta \subset R^s$, for some s. Let $\mathbf{S}(y, \boldsymbol{\theta})$ denote the vector of scores. That is,

$$S_j(y, \boldsymbol{\theta}) = \frac{\partial}{\partial \theta_j} \log g(y \mid \boldsymbol{\theta}), \quad j = 1, \ldots, s.$$

Let $\mathbf{D}^2(y, \boldsymbol{\theta})$ denote the matrix of second derivatives of $\log g(y \mid \boldsymbol{\theta})$ and let $I(\boldsymbol{\theta})$ denote the Fisher information matrix corresponding to one observation. It is assumed that all derivatives and expected values exist and that

$$\mathbb{E}_{\boldsymbol{\theta}} \mathbf{S}(y, \boldsymbol{\theta}) = \int \mathbf{S}(y, \boldsymbol{\theta})\, g(y \mid \boldsymbol{\theta})\, dy = \mathbf{0};$$

$$I(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}} \{\mathbf{S}(y, \boldsymbol{\theta})\, \mathbf{S}^{\mathrm{T}}(y, \boldsymbol{\theta})\} = -\mathbb{E}_{\boldsymbol{\theta}} \mathbf{D}^2(y, \boldsymbol{\theta}).$$

Consider the recursion

$$\boldsymbol{\theta}^*_{k+1} = \boldsymbol{\theta}^*_k + \{kI(\boldsymbol{\theta}^*_k)\}^{-1}\, \mathbf{S}(y_{k+1}, \boldsymbol{\theta}^*_k), \quad k = 0, 1, \ldots \tag{2}$$

which is recognizable as a stochastic approximation procedure. Under regularity conditions over and above those alluded to above, as $k \to \infty$,

$$\sqrt{(k)}\,(\boldsymbol{\theta}^*_k - \boldsymbol{\theta}_0) \to N(\mathbf{0},\ I(\boldsymbol{\theta}_0)^{-1}), \tag{3}$$

in distribution, where $\boldsymbol{\theta}_0$ denotes the true parameter value. This result appears in Sacks (1958), Fabian (1968), Nevel'son and Has'minskii (1973, Chapter 8) and Fabian (1978).

Alternatively…(Titterington,1984)

- Problem

As we shall see in some of the Examples in Section 3, complications may arise in applying recursions (2) and (6), in the computation and inversion, in the multiparameter case, of $I(\boldsymbol{\theta}_k^*)$. Numerical integration is often necessary and the fact that we are dealing with incomplete data will add to the complications. Suppose, with reference to (2), we write

- Modification
  - Fisher information matrix with complete observation

$$\tilde{\boldsymbol{\theta}}_{k+1} = \tilde{\boldsymbol{\theta}}_k + \{kI_c(\tilde{\boldsymbol{\theta}}_k)\}^{-1}\, \mathbf{S}(y_{k+1}, \tilde{\boldsymbol{\theta}}_k), \qquad k = 0, 1, \ldots,$$

Theorem 1

Given conditions corresponding to those above and provided $2I(\theta_0)\, I_c(\theta_0)^{-1} > 1$,

$$\sqrt{(k)}\,(\tilde{\theta}_k - \theta_0) \to N(0,\, I_c(\theta_0)^{-2}\, I(\theta_0)/\{2I(\theta_0)\, I_c(\theta_0)^{-1} - 1\})$$

in distribution as $k \to \infty$, where $\{\tilde{\theta}_k\}$ is defined by (8) or (9).

Outline

- A Recursive Procedure
- Recursive Expectation Maximum algorithm
  - Titterington,1984
  - lange,1995
  - Sato,2000
  - Cappe, 2009
- A Bridge from EM to VB: Free Energy
  - Neal & Hinton, 1993,1998
- Recursive Variational Bayes
  - Sato,2000
  - Hoffman, Blei,2010,2011

Standard EM algorithm

Suppose $x_1, \ldots, x_n$ represent n independent complete observations, corresponding to $y_1, \ldots, y_n$. Define

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}') = E_{\boldsymbol{\theta}'} \left\{ \sum_{i=1}^{n} \log f(x_i \mid \boldsymbol{\theta}) \mid y_1, \ldots, y_n \right\}.$$

The *EM* algorithm generates a sequence $\{\boldsymbol{\theta}_m\}$ of parameter estimates by repeating the following double step.

E-step: Evaluate $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_m)$.

M-step: Choose $\boldsymbol{\theta} = \boldsymbol{\theta}_{m+1}$ to maximize $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_m)$.

Connection between EM and gradient-based methods(Titterington,1984)

At stage $k+1$, with current estimate $\tilde{\theta}_k$, define

$$L_{k+1}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}_k}\{\log f(x_{k+1} \mid \boldsymbol{\theta}) \mid y_{k+1}\} + L_k(\boldsymbol{\theta}). \tag{12}$$

Choose $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}_{k+1}$ to maximize $L_{k+1}(\boldsymbol{\theta})$. Finally, estimate $\boldsymbol{\theta}_0$ by $\tilde{\boldsymbol{\theta}}_n$.

Theorem 2. Approximately, given appropriate regularity, recursion (12) can be written as

$$\tilde{\boldsymbol{\theta}}_{k+1} = \tilde{\boldsymbol{\theta}}_k + \{(k+1)\, I_c(\tilde{\boldsymbol{\theta}}_k)\}^{-1}\, \mathbf{S}(y_{k+1}, \tilde{\boldsymbol{\theta}}_k),$$

By Taylor expansion

Theorem 3. In exponential family models in which $\boldsymbol{\theta}$ is the expected value of the sufficient statistic, the recursion is exact.

For the exponential family models considered in Theorem 3 the recursions have particularly simple forms, reminiscent of Example 1.1. Recursion (2) is

$$\boldsymbol{\theta}^*_{k+1} = \boldsymbol{\theta}^*_k + \{kI(\boldsymbol{\theta}^*_k)\}^{-1}\, I_c(\boldsymbol{\theta}^*_k)\{E(\mathbf{t}^*_{k+1} \mid y_{k+1}, \boldsymbol{\theta}^*_k) - \boldsymbol{\theta}^*_k\}.$$

Recursion (8) is

$$\boldsymbol{\theta}^*_{k+1} = \boldsymbol{\theta}^*_k + k^{-1}\{E(\mathbf{t}^*_{k+1} \mid y_{k+1}, \boldsymbol{\theta}^*_k) - \boldsymbol{\theta}^*_k\}.$$

Solve the M-step by Newton's method(lange,1995)

$$\theta^{n+1} = \theta^n - \mathrm{d}^{20}Q(\theta^n|\theta^n)^{-1}\,\mathrm{d}^{10}Q(\theta^n|\theta^n)$$
$$= \theta^n - \mathrm{d}^{20}Q(\theta^n|\theta^n)^{-1}\,\mathrm{d}L(\theta^n).$$

In equation (1) the operators d^{10} and d^{20} take first and second partial derivatives respectively with respect to the first variable of Q. The column vector $\mathrm{d}L(\theta)$ is the score of the log-likelihood $L(\theta)$. Because $L(\theta) - Q(\theta|\theta^n)$ has its minimum at $\theta = \theta^n$, the equality $\mathrm{d}^{10}Q(\theta^n|\theta^n) = \mathrm{d}L(\theta^n)$ holds whenever θ^n is an interior point of the parameter domain (Dempster *et al.*, 1977). We shall refer to algorithm (1) as the EM gradient algorithm.

Online EM: Sato,2000

- Derived for general (canonical) Exponential family model with hidden variables(EFH models)

den variables (EFH models)[1]. An EFH model for an N-dimensional vector variable $\mathbf{x} = (x_1, ..., x_N)^T$ is defined by a probability distribution,

$$P(\mathbf{x}|\boldsymbol{\theta}) = \int d\mathbf{z}\ P(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}), \quad (1)$$

$$P(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) = \exp\left[\mathbf{r}(\mathbf{x}, \mathbf{z}) \cdot \boldsymbol{\theta} + r_0(\mathbf{x}, \mathbf{z}) - \Psi(\boldsymbol{\theta})\right],$$

where $\mathbf{z} = (z_1, ..., z_M)^T$ denotes an M-dimensional vector hidden variable and $\boldsymbol{\theta} = (\theta_1, ..., \theta_K)^T$ denotes a set of model parameters called the natural parameter. A set of sufficient statistics is denoted by $\mathbf{r}(\mathbf{x}, \mathbf{z}) = (r_1(\mathbf{x}, \mathbf{z}), ..., r_K(\mathbf{x}, \mathbf{z}))^T$. An in-

Sato2000, cont.

- On-line EM with the discount factor
- From free energy aspect

up to now. A discounted free energy for $\mathbf{X}\{\tau\}$ is defined by

$$F^{\lambda}(Q\{\tau\}, \boldsymbol{\theta}|\mathbf{X}\{\tau\}) = \eta(\tau) \sum_{t=1}^{\tau} \left(\prod_{s=t+1}^{\tau} \lambda(s) \right) \quad (6)$$

$$\times \int d\mathbf{z}(t)\ Q(\mathbf{z}(t)) \log(P(\mathbf{x}(t), \mathbf{z}(t)|\boldsymbol{\theta})/Q(\mathbf{z}(t)))$$

where a time dependent discount factor $\lambda(t)$ $(0 \leq \lambda(t) \leq 1,\ t = 2, 3, ...)$ is introduced for forgetting the earlier inaccurate estimator contributions. The normalization constant $\eta(\tau)$ is given by

Sato2000,Cont.

The recursive formula for $\phi(\tau)$ is derived from the above equations:

$$\begin{aligned}\Delta\phi(\tau) &= \phi(\tau) - \phi(\tau-1) \\ &= \eta(\tau)\Big(E_{\mathbf{z}}\left[\mathbf{r}(\mathbf{x},\mathbf{z})|\boldsymbol{\theta}(\tau-1)\right] - \phi(\tau-1)\Big).\end{aligned} \tag{10}$$

The new estimator $\boldsymbol{\theta}(\tau)$ is obtained by using (2),

$$\boldsymbol{\theta}(\tau) = \boldsymbol{\theta}(\phi(\tau)) = \frac{\partial H(\phi)}{\partial \phi}|_{\phi(\tau)}. \tag{11}$$

$$\begin{aligned}\Delta\phi(\tau) &= \eta(\tau)\left(\frac{\partial L(\boldsymbol{\theta}|\mathbf{x}(\tau))}{\partial \boldsymbol{\theta}}\right)|_{\boldsymbol{\theta}(\tau-1)} \\ &= \eta(\tau)\left(\frac{\partial \phi}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial L(\boldsymbol{\theta}(\phi)|\mathbf{x}(\tau))}{\partial \phi}\right)|_{\phi(\tau-1)} \\ &= \eta(\tau)(\mathbf{V}(\phi(\tau-1)))^{-1} \\ &\quad \times \left(\frac{\partial L(\boldsymbol{\theta}(\phi)|\mathbf{x}(\tau))}{\partial \phi}\right)|_{\phi(\tau-1)},\end{aligned} \tag{14}$$

On-line expectation–maximization algorithm for latent data models, Cappe, 2009

$$\hat{Q}_{n+1}(\theta) = \hat{Q}_n(\theta) + \gamma_{n+1}(\mathbb{E}_{\hat{\theta}_n}[\log\{f(X_{n+1};\theta)\}|Y_{n+1}] - \hat{Q}_n(\theta)),$$

Assumption 1.

(a) The complete-data likelihood is of the form

$$f(x;\theta) = h(x)\exp\{-\psi(\theta) + \langle S(x), \phi(\theta)\rangle\}. \tag{11}$$

(b) The function

$$\bar{s}(y;\theta) \stackrel{\text{def}}{=} \mathbb{E}_\theta[S(X)|Y=y] \tag{12}$$

is well defined for all $(y,\theta) \in \mathcal{Y} \times \Theta$.

(c) There is a convex open subset $\mathcal{S} \subset \mathbb{R}^d$, which is such that

(i) for all $s \in \mathcal{S}$, $(y,\theta) \in \mathcal{Y} \times \Theta$ and $\gamma \in [0,1)$, $(1-\gamma)s + \gamma\bar{s}(y;\theta) \in \mathcal{S}$ and

(ii) for any $s \in \mathcal{S}$, the function $\theta \mapsto l(s;\theta) \stackrel{\text{def}}{=} -\psi(\theta) + \langle s, \phi(\theta)\rangle$ has a unique global maximum over Θ denoted $\bar{\theta}(s)$, i.e.

$$\bar{\theta}(s) \stackrel{\text{def}}{=} \arg\max_{\theta\in\Theta}\{l(s;\theta)\}. \tag{13}$$

Cappe2009, cont.

Assumption 1 implies that the evaluation of $\mathbb{E}_\theta[\log\{f(X;\theta)\}|Y]$, and hence the E-step of the EM algorithm reduces to the computation of the expected value $\mathbb{E}_\theta[S(X)|Y]$ of the *complete-data sufficient statistic* $S(X)$. Indeed, the EM re-estimation functional $Q_{\theta'}(Y_{1:n};\theta)$ is then defined by

$$Q_{\theta'}(Y_{1:n};\theta) = l\left\{n^{-1}\sum_{i=1}^{n}\bar{s}(Y_i;\theta');\theta\right\}.$$

The $(k+1)$th iteration of the (batch mode) EM algorithm may thus be expressed as

$$\hat{\theta}_{k+1} = \bar{\theta}\left\{n^{-1}\sum_{i=1}^{n}\bar{s}(Y_i;\hat{\theta}_k)\right\},$$

In this setting, the proposed on-line EM algorithm takes the form

$$\hat{s}_{n+1} = \hat{s}_n + \gamma_{n+1}\{\bar{s}(Y_{n+1};\hat{\theta}_n) - \hat{s}_n\},$$
$$\hat{\theta}_{n+1} = \bar{\theta}(\hat{s}_{n+1}).$$

takes care of this issue in the case of the on-line EM algorithm. As an additional comment about assumption 1, note that we do not require that ϕ be a one-to-one mapping and hence the complete-data model may also correspond to a *curved* exponential family, where typically θ is of much lower dimension than $\psi(\theta)$ (see, for instance, Chung and Böhme (2005) and Cappé

Outline

- A Recursive Procedure
- Recursive Expectation Maximum algorithm
  - Titterington,1984
  - Iange,1995
  - Sato,2000
  - Cappe, 2009
- A Bridge from EM to VB: Free Energy
  - Neal & Hinton, 1993,1998
- Recursive Variational Bayes
  - Sato,2000
  - Hoffman, Blei,2010,2011

A Bridge from EM to VB: Free Energy

- A view of the EM algorithm that justifies incremental, sparse, and other variants, Neal & Hinton, 1993,1998
- Standard EM

$$\left.\begin{array}{ll} \textbf{E Step:} & \text{Compute a distribution } \widetilde{P}^{(t)} \text{ over the range of } Y \text{ such} \\ & \text{that } \widetilde{P}^{(t)}(y) = P(y \mid z, \theta^{(t-1)}). \\ \textbf{M Step:} & \text{Set } \theta^{(t)} \text{ to the } \theta \text{ that maximizes } E_{\widetilde{P}^{(t)}}[\log P(y, z \mid \theta)]. \end{array}\right\} \quad (1)$$

- Variational free energy

$$F(\widetilde{P}, \theta) \;=\; E_{\widetilde{P}}[\log P(y, z \mid \theta)] \;+\; H(\widetilde{P}) \qquad (2)$$

where $H(\widetilde{P}) = -E_{\widetilde{P}}[\log \widetilde{P}(y)]$ is the entropy of the distribution $\widetilde{P}$. Note

Lemma 1 *For a fixed value of θ, there is a unique distribution, P_θ, that maximizes $F(\widetilde{P}, \theta)$, given by $P_\theta(y) = P(y \mid z, \theta)$. Furthermore, this P_θ varies continuously with θ.*

(Neal & Hinton, 1998) cont.

Lemma 2 *If* $\widetilde{P}(y) = P(y \mid z, \theta) = P_\theta(y)$ *then* $F(\widetilde{P}, \theta) = \log P(z \mid \theta) = L(\theta)$.

PROOF. If $\widetilde{P}(y) = P(y \mid z, \theta)$, then

Theorem 1 *The iterations given by (1) and by (5) are equivalent.*

$$\left.\begin{array}{ll} \textbf{E Step:} & \text{Set } \widetilde{P}^{(t)} \text{ to the } \widetilde{P} \text{ that maximizes } F(\widetilde{P}, \theta^{(t-1)}). \\ \textbf{M Step:} & \text{Set } \theta^{(t)} \text{ to the } \theta \text{ that maximizes } F(\widetilde{P}^{(t)}, \theta). \end{array}\right\} \quad (5)$$

Incremental

We can then write F in the form $F(\tilde{P}, \theta) = \sum_i F_i(\tilde{P}_i, \theta)$, where

$$F_i(\tilde{P}_i, \theta) \;=\; E_{\tilde{P}_i}[\log P(y_i,\, z_i \,|\, \theta)] \;+\; H(\tilde{P}_i) \qquad (6)$$

E Step: Choose some data item, i, to be updated.
Set $P_j^{(t)} = P_j^{(t-1)}$ for $j \neq i$. (This takes no time).
Set $P_i^{(t)}$ to the $\tilde{P}_i$ that maximizes $F_i(\tilde{P}_i, \theta^{(t-1)})$, given by $\tilde{P}_i^{(t)}(y_i) = P(y_i \,|\, z_i, \theta^{(t-1)})$.

M Step: Set $\theta^{(t)}$ to the θ that maximizes $F(\tilde{P}^{(t)}, \theta)$, or, equivalently, that maximizes $E_{\tilde{P}^{(t)}}[\log P(y, z \,|\, \theta)]$.

(7)

Each E step of the above algorithm requires looking at only a single data item, but, as written, it appears that the M step requires looking at all components of $\tilde{P}$. This can be avoided in the common case where the inferential import of the complete data can be summarized by a vector of sufficient statistics that can be incrementally updated, as is the case with models in the exponential family.

Incremental

E Step: Choose some data item, i, to be updated.

Set $\tilde{s}_j^{(t)} = \tilde{s}_j^{(t-1)}$ for $j \neq i$. (This takes no time.)

Set $\tilde{s}_i^{(t)} = E_{\tilde{P}_i}[s_i(y_i, z_i)]$, for $\tilde{P}_i(y_i) = P(y_i \mid z_i, \theta^{(t-1)})$.

Set $\tilde{s}^{(t)} = \tilde{s}^{(t-1)} - \tilde{s}_i^{(t-1)} + \tilde{s}_i^{(t)}$.

M Step: Set $\theta^{(t)}$ to the θ with maximum likelihood given $s^{(t)}$. (9)

E Step: Select the next data item, i, for updating.

Set $\tilde{s}_i^{(t)} = E_{\tilde{P}_i}[s_i(y_i, z_i)]$, for $\tilde{P}_i(y_i) = P(y_i \mid z_i, \theta^{(t-1)})$.

Set $\tilde{s}^{(t)} = \gamma \tilde{s}^{(t-1)} + \tilde{s}_i^{(t)}$.

M Step: Set $\theta^{(t)}$ to the θ with maximum likelihood given $s^{(t)}$. (10)

where $0 < \gamma < 1$ is a decay constant.

Outline

- Problem statement
- A Recursive Procedure
- Recursive Expectation Maximum algorithm
  - Titterington,1984
  - Iange,1995
  - Sato,2000
  - Cappe, 2009
- A Bridge from EM to VB: Free Energy
  - Neal & Hinton, 1993,1998
- Recursive Variational Bayes
  - Sato,2000
  - Hoffman, Blei,2010,2011

ML vs. Bayesian

- There are three main problems with ML learning.
  - **Overfit :** First, it produces a model that overfits the data and subsequently have suboptimal generalization performance.
  - **Model selection**: it cannot be used to learn the structure of the graph, since more complicated graphs assign a higher likelihood to the data.
  - **Tractability** : Third, it is computationally tractable only for a small class of models.
- The Bayesian framework in principle, a solution to the first two problems.
  - one considers an ensemble of models, characterized by a probability distribution over all possible parameter values and structures
  - complex models are effectively penalized by being assigned a lower posterior probability
- Unfortunately,
  - computations in the Bayesian framework can seldom be performed exactly, due to the need to integrate over models.
  - Approximations therefore must be made, MCMC, Laplace approximation.

Online Model Selection Based on the Variational Bayes (sato2001)

- Derivation: similar to online EM(sato2000).

$$\Delta\alpha(\tau) = \alpha(\tau) - \alpha(\tau - 1)$$

$$\Delta\alpha(\tau) = \frac{1}{\gamma}\eta(\tau)\Big(TE_{\mathbf{z}}\left[\mathbf{r}(\mathbf{x}(\tau), \mathbf{z}(\tau))|\langle\theta\rangle_{\alpha(\tau-1)}\right] + \gamma_0\alpha_0 - \gamma\alpha(\tau - 1)\Big).$$

$$\bar{\theta}(\tau) = \langle\theta\rangle_{\alpha(\tau-1)} = \frac{1}{\gamma}\frac{\partial\Phi}{\partial\alpha}(\alpha(\tau - 1), \gamma).$$

$$\Delta\alpha(\tau) = \frac{1}{\gamma^2}\eta(\tau)V^{-1}_{\alpha,\alpha}(\alpha(\tau - 1), \gamma)\cdot\frac{\partial F_M}{\partial\alpha}(\mathbf{x}(\tau), \alpha(\tau - 1), T).$$

Stochastic Variational Inference (Hoffman,Blei,2011,2013)

- However, when the posterior of latent variable is intractable, the above approaches cannot be applied.

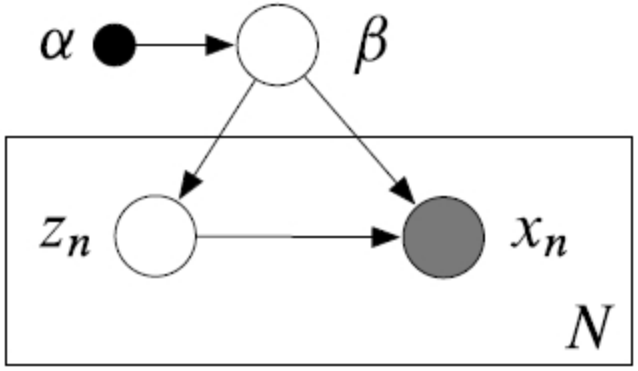


Figure 2: A graphical model with observations $x_{1:N}$, local hidden variables $z_{1:N}$ and global hidden variables β. The distribution of each observation x_n only depends on its corresponding local variable z_n and the global variables β. (Though not pictured, each hidden variable z_n, observation x_n, and global variable β may be a collection of multiple random variables.)

The joint distribution factorizes into a global term and a product of local terms,

$$p(x, z, \beta \mid \alpha) = p(\beta \mid \alpha) \prod_{n=1}^{N} p(x_n, z_n \mid \beta).$$

Complete conditionals (exponential family)

$$p(\beta \mid x, z, \alpha) = h(\beta) \exp\{\eta_g(x, z, \alpha)^\top t(\beta) - a_g(\eta_g(x, z, \alpha))\},$$
$$p(z_{nj} \mid x_n, z_{n,-j}, \beta) = h(z_{nj}) \exp\{\eta_\ell(x_n, z_{n,-j}, \beta)^\top t(z_{nj}) - a_\ell(\eta_\ell(x_n, z_{n,-j}, \beta))\}.$$

- Conjugacy relationship

$$p(x_n, z_n | \beta) = h(x_n, z_n) \exp\{\beta^\top t(x_n, z_n) - a_\ell(\beta)\}.$$

The prior distribution $p(\beta)$ must also be in an exponential family,

$$p(\beta) = h(\beta) \exp\{\alpha^\top t(\beta) - a_g(\alpha)\}. \quad (5)$$

The sufficient statistics are $t(\beta) = (\beta, -a_\ell(\beta))$ and thus the hyperparameter α has two components $\alpha = (\alpha_1, \alpha_2)$. The first component α_1 is a vector of the same dimension as β; the second component α_2 is a scalar.

$$\eta_g(x, z, \alpha) = (\alpha_1 + \textstyle\sum_{n=1}^{N} t(z_n, x_n), \alpha_2 + N).$$

- Objective function(ELBO)

$$\begin{aligned}
\log p(x) &= \log \int p(x, z, \beta) dz d\beta \\
&= \log \int p(x, z, \beta) \frac{q(z, \beta)}{q(z, \beta)} dz d\beta \\
&= \log \left(\mathbb{E}_q \left[\frac{p(x, z, \beta)}{q(z, \beta)} \right] \right) \\
&\geq \mathbb{E}_q [\log p(x, z, \beta)] - \mathbb{E}_q [\log q(z, \beta)] \\
&\triangleq \mathcal{L}(q).
\end{aligned}$$

- Mean field

$$q(z, \beta) = q(\beta \mid \lambda) \prod_{n=1}^{N} \prod_{j=1}^{J} q(z_{nj} \mid \phi_{nj}).$$

$$\begin{aligned}
q(\beta \mid \lambda) &= h(\beta) \exp\{\lambda^\top t(\beta) - a_g(\lambda)\}, \\
q(z_{nj} \mid \phi_{nj}) &= h(z_{nj}) \exp\{\phi_{nj}^\top t(z_{nj}) - a_\ell(\phi_{nj})\}.
\end{aligned}$$

Derive the coordinate update for the parameter λ

- rewrite the objective

$$\mathcal{L}(\lambda) = \mathbb{E}_q[\log p(\beta \mid x, z)] - \mathbb{E}_q[\log q(\beta)] + \text{const.}$$

$$\mathcal{L}(\lambda) = \mathbb{E}_q[\eta_g(x, z, \alpha)]^\top \nabla_\lambda a_g(\lambda) - \lambda^\top \nabla_\lambda a_g(\lambda) + a_g(\lambda) + \text{const.}$$

$$\mathbb{E}_q[t(\beta)] = \nabla_\lambda a_g(\lambda).$$

- Take gradient

$$\nabla_\lambda \mathcal{L} = \nabla^2_\lambda a_g(\lambda)(\mathbb{E}_q[\eta_g(x, z, \alpha)] - \lambda).$$

Same for local parameter: ϕ_{nj}.

$$\nabla_{\phi_{nj}} \mathcal{L} = \nabla^2_{\phi_{nj}} a_\ell(\phi_{nj})(\mathbb{E}_q[\eta_\ell(x_n, z_{n,-j}, \beta)] - \phi_{nj}).$$

Standard VB: Coordinate ascent mean-field variational inference

1: Initialize $\lambda^{(0)}$ randomly.
2: **repeat**
3: **for** each local variational parameter ϕ_{nj} **do**
4: Update ϕ_{nj}, $\phi_{nj}^{(t)} = \mathbb{E}_{q^{(t-1)}}[\eta_{\ell,j}(x_n, z_{n,-j}, \beta)]$.
5: **end for**
6: Update the global variational parameters, $\lambda^{(t)} = \mathbb{E}_{q^{(t)}}[\eta_g(z_{1:N}, x_{1:N})]$.
7: **until** the ELBO converges

VB update based on Natural Gradient

- the classical gradient(based on Euclidean distance metric)

$$\lambda^{(t+1)} = \lambda^{(t)} + \rho \nabla_\lambda f(\lambda^{(t)}).$$

$$\arg\max_{d\lambda} f(\lambda + d\lambda) \quad \text{subject to } ||d\lambda||^2 < \varepsilon^2$$

- Natural gradients (measure of dissimilarity: symmetrized KL divergence)

$$\hat{\nabla}_\lambda f(\lambda) \triangleq G(\lambda)^{-1} \nabla_\lambda f(\lambda),$$

$$D_{KL}^{\text{sym}}(\lambda, \lambda') = \mathbb{E}_\lambda \left[\log \frac{q(\beta|\lambda)}{q(\beta|\lambda')} \right] + \mathbb{E}_{\lambda'} \left[\log \frac{q(\beta|\lambda')}{q(\beta|\lambda)} \right].$$

$$\arg\max_{d\lambda} f(\lambda + d\lambda) \quad \text{subject to } D_{KL}^{\text{sym}}(\lambda, \lambda + d\lambda) < \varepsilon.$$

$$G(\lambda) = \mathbb{E}_\lambda \left[(\nabla_\lambda \log q(\beta|\lambda))(\nabla_\lambda \log q(\beta|\lambda))^\top \right].$$

$$d\lambda^T G(\lambda) d\lambda = D_{KL}^{\text{sym}}(\lambda, \lambda + d\lambda),$$

$$\begin{aligned} G(\lambda) &= \mathbb{E}_\lambda \left[(\nabla_\lambda \log p(\beta|\lambda))(\nabla_\lambda \log p(\beta|\lambda))^\top \right] \\ &= \mathbb{E}_\lambda \left[(t(\beta) - \mathbb{E}_\lambda[t(\beta)])(t(\beta) - \mathbb{E}_\lambda[t(\beta)])^\top \right] \\ &= \nabla_\lambda^2 a_g(\lambda). \end{aligned}$$

$$\hat{\nabla}_\lambda \mathcal{L} = \mathbb{E}_\phi[\eta_g(x, z, \alpha)] - \lambda.$$

$$\hat{\nabla}_{\phi_{nj}} \mathcal{L} = \mathbb{E}_{\lambda, \phi_{n,-j}}[\eta_\ell(x_n, z_{n,-j}, \beta)] - \phi_{nj}.$$

- The classical coordinate ascent algorithm can thus be interpreted as a projected natural gradient algorithm.

$$\mathcal{L}(\lambda) \triangleq \mathcal{L}(\lambda, \phi(\lambda)).$$

$$\mathcal{L}(\lambda) = \mathbb{E}_q[\log p(\beta)] - \mathbb{E}_q[\log q(\beta)] + \sum_{\cdot}^{N} \max_{\phi_{\cdot}} (\mathbb{E}_q[\log p(x_n, z_n \mid \beta)] - \mathbb{E}_q[\log q(z_n)]).$$

Now consider a variable that chooses an index of the data uniformly at random, $I \sim \text{Unif}(1, \ldots, N)$. Define $\mathcal{L}_I(\lambda)$ to be the following random function of the variational parameters,

$$\mathcal{L}_I(\lambda) \triangleq \mathbb{E}_q[\log p(\beta)] - \mathbb{E}_q[\log q(\beta)] + N \max_{\phi_I} (\mathbb{E}_q[\log p(x_I, z_I \mid \beta) - \mathbb{E}_q[\log q(z_I)]). \tag{25}$$

$$\hat{\nabla} \mathcal{L}_i = \mathbb{E}_q \left[\eta_g \left(x_i^{(N)}, z_i^{(N)}, \alpha \right) \right] - \lambda,$$

where $\left\{ x_i^{(N)}, z_i^{(N)} \right\}$ are a data set formed by N replicates of observation x_n and hidden variables z_n.

$$\eta_g \left(x_i^{(N)}, z_i^{(N)}, \alpha \right) = \alpha + N \cdot (t(x_n, z_n), 1).$$

$$\hat{\nabla}_\lambda \mathcal{L}_i = \alpha + N \cdot \left(\mathbb{E}_{\phi_i(\lambda)}[t(x_i, z_i)], 1 \right) - \lambda,$$

$$\hat{\lambda}_t \triangleq \alpha + N \mathbb{E}_{\phi_i(\lambda)}[(t(x_i, z_i), 1)].$$

$$\begin{aligned} \lambda^{(t)} &= \lambda^{(t-1)} + \rho_t \left(\hat{\lambda}_t - \lambda^{(t-1)} \right) \\ &= (1 - \rho_t) \lambda^{(t-1)} + \rho_t \hat{\lambda}_t. \end{aligned}$$

1: Initialize $\lambda^{(0)}$ randomly.
2: Set the step-size schedule ρ_t appropriately.
3: **repeat**
4: Sample a data point x_i uniformly from the data set.
5: Compute its local variational parameter,

$$\phi = \mathbb{E}_{\lambda^{(t-1)}}[\eta_g(x_i^{(N)}, z_i^{(N)})].$$

6: Compute intermediate global parameters as though x_i is replicated N times,

$$\hat{\lambda} = \mathbb{E}_{\phi}[\eta_g(x_i^{(N)}, z_i^{(N)})].$$

7: Update the current estimate of the global variational parameters,

$$\lambda^{(t)} = (1-\rho_t)\lambda^{(t-1)} + \rho_t \hat{\lambda}.$$

8: **until** forever

Figure 4: Stochastic variational inference.

We set the step-size at iteration t as follows,

$$\rho_t = (t+\tau)^{-\kappa}.$$

Extensions

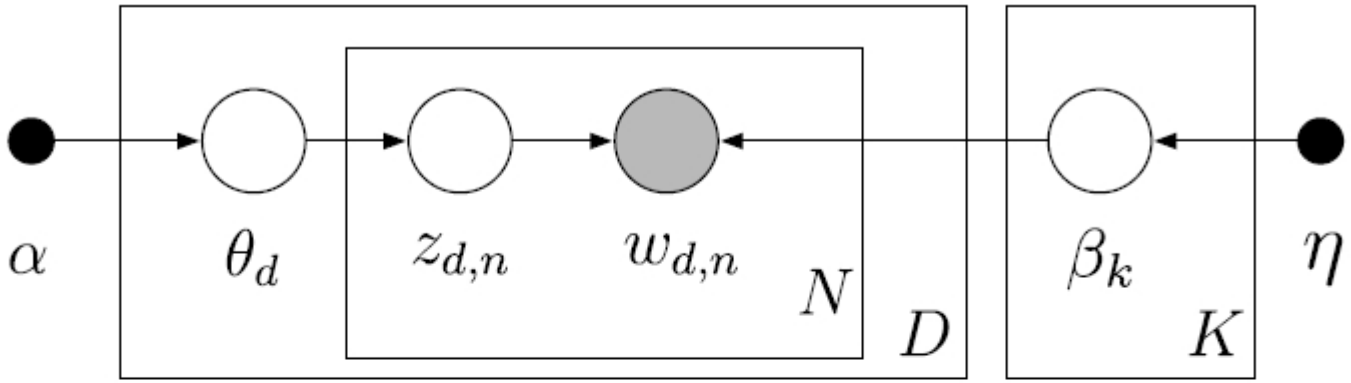
- Minibatches.

$$\lambda^{(t)} = (1-\rho_t)\lambda^{(t-1)} + \frac{\rho_t}{S}\sum_s \hat{\lambda}_s.$$

- Empirical Bayes estimation of hyperparameters.

$$\alpha^{(t)} = \alpha^{(t-1)} + \rho_t \nabla_\alpha \mathcal{L}_t(\lambda^{(t-1)}, \phi, \alpha^{(t-1)}).$$

Example: Topic Models



Var	Type	Conditional	Param	Relevant Expectations
z_{dn}	Multinomial	$\log \theta_{dk} + \log \beta_{k,w_{dn}}$	ϕ_{dn}	$\mathbb{E}[Z^k_{dn}] = \phi^k_{dn}$
θ_d	Dirichlet	$\alpha + \sum_{n=1}^{N} z_{dn}$	γ_d	$\mathbb{E}[\log \theta_{dk}] = \Psi(\gamma_{dk}) - \sum_{j=1}^{K} \Psi(\gamma_{dj})$
β_k	Dirichlet	$\eta + \sum_{d=1}^{D} \sum_{n=1}^{N} z^k_{dn} w_{dn}$	λ_k	$\mathbb{E}[\log \beta_{kv}] = \Psi(\lambda_{kv}) - \sum_{y=1}^{V} \Psi(\lambda_{ky})$

Initialize $\lambda^{(0)}$ randomly.
Set the step-size schedule ρ_t appropriately.
repeat
  Sample a document w_d uniformly from the data set.
  Initialize $\gamma_{dk} = 1$, for $k \in \{1, \ldots, K\}$.
  repeat
    For $n \in \{1, \ldots, N\}$ set

$$\phi_{dn}^k \propto \exp\left\{\mathbb{E}[\log \theta_{dk}] + \mathbb{E}[\log \beta_{k,w_{dn}}]\right\}, k \in \{1, \ldots, K\}.$$

    Set $\gamma_d = \alpha + \sum_n \phi_{dn}$.
  until local parameters ϕ_{dn} and γ_d converge.
  For $k \in \{1, \ldots, K\}$ set intermediate topics

$$\hat{\lambda}_k = \eta + D \sum_{n=1}^{N} \phi_{dn}^k w_{dn}.$$

  Set $\lambda^{(t)} = (1 - \rho_t)\lambda^{(t-1)} + \rho_t \hat{\lambda}$.
until forever

Figure 6: Stochastic variational inference for LDA. The relevant expectations for each update are found in Figure 5.

Thanks!