SLC Internal tutorial

自然言語処理のための変分ベイズ法

Daichi Mochihashi

daichi.mochihashi@atr.jp

ATR SLC

2005.6.21 (Tue)

13:15–15:00@Meeting Room 1

変分ベイズ法とは?

- 確率モデルのベイズ推定を行うための近似解法
  - 最尤推定 (EM) と違い, 過学習を自動的に防ぐ
  - 最尤推定が不可能な, 複雑な確率モデル
  - 通常の EM アルゴリズムの自然な拡張
- どこで使われているか?
  - 音声認識 (實廣, 中村 2004/渡辺他 2002)
  - 混合 von Mises-Fisher 分布 (田辺他 2004)
  - HMM (MacKay 1997)
  - LDA (Blei et al. 2001)
  - PCFG (栗原他 2004)
  - ...

アウトライン

- ベイズ推定と最尤推定
- 最尤推定と EM アルゴリズム
- 変分ベイズ推定と VB-EM アルゴリズム
- 変分ベイズ推定の性質
- 自然言語処理への応用
  - LDA
  - VB-HMM
- ベイズ推定のためのその他の解法

準備: 不等式

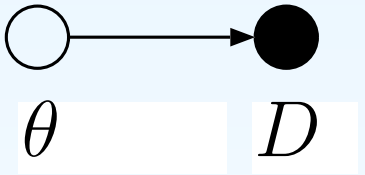
- Jensen の不等式
  - 上に凸な関数 $f(x)$ について,
$$f(E[x]) \geq E[f(x)] \tag{1}$$
  - $\log(x)$ は上に凸なので, $x = f(x)$ として
$$\log \int p(x) f(x) dx \geq \int p(x) \log f(x) dx \,. \tag{2}$$
- KL ダイバージェンス
$$D(p||q) = \int p(x) \log \frac{p(x)}{q(x)} dx \geq 0 \,. \tag{3}$$
  - $p = q$ のときに等号成立.

準備: 確率モデル

- D をデータとすると,

$$p(D) = \int p(D, \theta) d\theta \tag{4}$$

$$= \int p(D|\theta) p(\theta) d\theta. \tag{5}$$



$p(D|\theta), p(\theta)$: 確率モデル (生成モデル)

θ : 確率モデルのパラメータ

- 目標: データ D が与えられたとき, $p(\theta|D)$ を推定すること.

$$p(\theta|D) \propto p(\theta, D) = p(D|\theta) p(\theta)\,. \tag{6}$$

ベイズ推定と最尤推定

$p(\theta|D)$ がわかれば, 新しいデータ d の予測は

- ベイズ推定

$$p(d|D) = \int p(d|\theta)p(\theta|D)d\theta\,. \tag{7}$$

  - $p(d|\theta)$: 確率モデルの適用
  - $p(\theta|D)$: パラメータ θ の確率分布で期待値をとる.

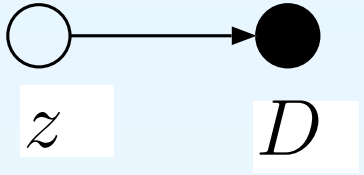
- 最尤推定

$$p(d|D) = p(d|\hat{\theta}). \tag{8}$$

  - $\theta = \hat{\theta}$ と点推定した確率モデル
  - $p(\theta|D) = \delta(\hat{\theta})$ と δ 関数で近似してしまう
  - $p(\theta|D)$ が実際はなだらかな時, 偏った推定になる

最尤推定

- データ D と隠れ変数 z があるとき,



$$p(D|\theta) = \int p(D, z|\theta) dz \quad \rightarrow \text{最大化}. \quad (9)$$

(6) を最大化する $\theta = \hat{\theta}$ と隠れ変数 z を求める.

- これを解く方法→ **EM アルゴリズム**.

EM アルゴリズム (1)

- Jensen の不等式を用いると,

$$\log p(D|\theta) = \log \int p(D, z|\theta) dz \tag{10}$$

$$= \log \int q(z|D, \hat{\theta}) \frac{p(D, z|\theta)}{q(z|D, \hat{\theta})} \, dz \tag{11}$$

$$\geq \int q(z|D, \hat{\theta}) \log \frac{p(D, z|\theta)}{q(z|D, \hat{\theta})} \, dz = F(q(z), \theta) \tag{12}$$

- よって, 下限 $F(q(z), \theta)$ を交互に最大化すればよい.

$$\textbf{E step:} \quad q(z) = \arg\max_{q(z)} F(q(z), \theta) \,, \tag{13}$$

$$\textbf{M step:} \quad \hat{\theta} = \arg\max_{\theta} F(q(z), \theta) \,. \tag{14}$$

EM アルゴリズム (2)

- E step

$$F(q(z),\theta) = \int q(z|D,\hat{\theta}) \log \frac{p(D,z|\theta)}{q(z|D,\hat{\theta})} dz \tag{15}$$

$$= \int q(z|D,\hat{\theta}) \log \frac{p(z|D,\theta)p(D|\theta)}{q(z|D,\hat{\theta})} dz \tag{16}$$

$$= -\int q(z|D,\hat{\theta}) \log \frac{q(z|D,\hat{\theta})}{p(z|D,\theta)} dz + \log p(D|\theta) \tag{17}$$

$$= -\underline{D(q(z|D,\hat{\theta})||p(z|D,\theta))} + \log p(D|\theta) \tag{18}$$

は

$$q(z|D,\hat{\theta}) = p(z|D,\theta) \tag{19}$$

で最大 (E ステップ).

EM アルゴリズム (3)

- M step

$$F(q(z),\theta) = \int q(z|D,\hat{\theta}) \log \frac{p(D,z|\theta)}{q(z|D,\hat{\theta})} dz \tag{20}$$

$$= \left\langle \log p(D,z|\theta) \right\rangle_{q(z|D,\hat{\theta})} + H(q(z|D,\hat{\theta})) \tag{21}$$

よって, $F(q(z),\theta)$ を θ について最大化するには,

$$Q(\theta) = \left\langle \log p(D,z|\theta) \right\rangle_{q(z|D,\hat{\theta})} \quad \text{(Q 関数)} \tag{22}$$

に対して,

$$\frac{\partial Q(\theta)}{\partial \theta} = 0 \tag{23}$$

を解いた θ を新しい $\hat{\theta}$ とすればよい. (M ステップ)

EM アルゴリズム (まとめ)

$$\log p(D|\theta) \geq F(q(z), \theta) = \int q(z|D, \hat{\theta}) \log \frac{p(D, z|\theta)}{q(z|D, \hat{\theta})} dz \tag{24}$$

として, 下限 $F(q(z), \theta)$ を $q(z), \theta$ について順に最大化する (EM アルゴリズム).

ここで左辺と右辺の差は,

$$\log p(D|\theta) - F(q(z), \theta) \tag{25}$$

$$= \int q(z|D, \hat{\theta}) \log p(D|\theta) dz - \int q(z|D, \hat{\theta}) \log \frac{p(D, z|\theta)}{q(z|D, \hat{\theta})} dz \tag{26}$$

$$= \int q(z|D, \hat{\theta}) \log \frac{q(z|D, \hat{\theta})}{p(z|D, \theta)} dz \tag{27}$$

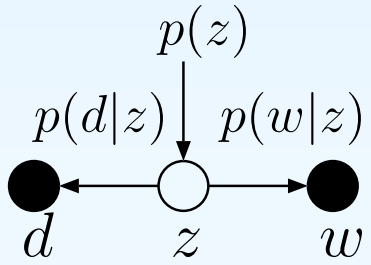
$$= D(q(z|D, \hat{\theta}) || p(z|D, \theta)) \geq 0\,. \tag{28}$$

この KL ダイバージェンスを最小化していることに相当.

Example: PLSI (1/3)

- ある単語 w が文書 d で生起したとき, 隠れ変数 z があって

$$p(d, w, z) = p(z)p(d|z)p(w|z) \tag{29}$$



と分解できたと仮定する.

- 文書 $\mathbf{w} = w_1 w_2 \cdots w_n$ の集合 $W = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_D\}$,
  文書のインデックス集合 $\mathcal{D} = \{1, 2, \ldots, D\}$ について,

$$p(\mathcal{D}, W, Z) = \prod_d p(d, \mathbf{w}_d, \mathbf{z}_d) \tag{30}$$

$$= \prod_d \prod_n p(d, w_{dn}, z_{dn}) \tag{31}$$

$$= \prod_d \prod_n p(z_{dn})p(d|z_{dn})p(w_{dn}|z_{dn}) \tag{32}$$

$$\therefore \quad \log p(\mathcal{D}, W, Z) = \sum_d \sum_n \left[\log p(z_{dn}) + \log p(d|z_{dn}) + \log p(w_{dn}|z_{dn})\right]. \tag{33}$$

Example: PLSI (2/3)

- Q 関数 $\left\langle \log p(D, z|\theta) \right\rangle_{p(z|D,\theta)}$ を計算すると,

$$Q(z) = \left\langle \log p(\mathcal{D}, W, Z) \right\rangle_{p(Z|\mathcal{D},W)} \tag{34}$$

$$= \sum_d \sum_n \left[\sum_z p(z|d, w_{dn}) \log p(z_{dn}) + \sum_z p(z|d, w_{dn}) \log p(d|z_{dn}) + \sum_z p(z|d, w_{dn}) \log p(w_{dn}|z_{dn}) \right] . \tag{35}$$

- $\delta Q / \delta \theta$ を計算すると,

$$\frac{\delta Q}{\delta p(z)} = \frac{\sum_d \sum_n p(z|d, w_{dn})}{p(z)} + \lambda = 0 \tag{36}$$

$$\therefore \; p(z) \propto \sum_d \sum_n p(z|d, w_{dn}) \propto \sum_d \sum_w n(d, w) p(z|d, w) \tag{37}$$

Example: PLSI (3/3)

- 同様にして,

$$p(d|z) \propto \sum_n p(z|d, w_{dn}) \propto \sum_w n(d, w) p(z|d, w) \tag{38}$$

$$p(w|z) \propto \sum_d \sum_n p(z|d, w_{dn}) \propto \sum_d n(d, w) p(z|d, w)\,. \quad \square \tag{39}$$

- ここで

$$p(z|d, w) \propto p(z, d, w) = p(z) p(d|z) p(w|z)\,. \tag{40}$$

- この場合, 文書 d ごとにパラメータ

$$\theta(d) = p(z|d) \tag{41}$$

$$\propto p(z) p(d|z) \tag{42}$$

を点推定していることに相当する.

EM アルゴリズムの欠点

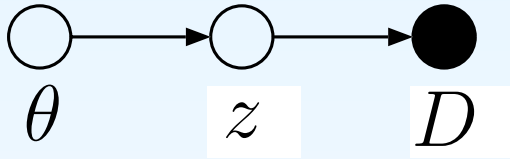
- θ は given, 点推定
- 隠れ変数 z が 1 層だけある場合にしか適用不可能
- 過学習してしまう (z はバラバラ).

  ↓

  ベイズ推定.

ベイズ推定

$$p(D) = \iint p(D, z, \theta) dz d\theta \quad \rightarrow \text{最大化.} \quad (43)$$



(26) を最大化する z, θ の確率分布 $p(z|D), p(\theta|D)$ を求めることが目標.

$$\log p(D) = \log \iint q(z, \theta|D) \frac{p(D, z, \theta)}{q(z, \theta|D)} \, dz d\theta \quad (44)$$

$$\geq \iint q(z, \theta|D) \log \frac{p(D, z, \theta)}{q(z, \theta|D)} \, dz d\theta \quad (45)$$

この下限はそのままでは z, θ のそれぞれに対して最大化できないので,

$$q(z, \theta|D) = q(z)q(\theta) \quad (46)$$

という因子分解を仮定すると,

変分ベイズ推定

$$\log p(D) \geq \iint q(z,\theta|D) \log \frac{p(D,z,\theta)}{q(z,\theta|D)} dz d\theta \tag{47}$$

$$= \iint q(z)q(\theta) \log \frac{p(D,z\theta)}{q(z)q(\theta)} dz d\theta \tag{48}$$

$$= F(q). \quad \text{(変分自由エネルギー)} \tag{49}$$

この下限 (変分下限, variational lower bound) $F(q)$ は $q(z), q(\theta)$ について逐次最大化できる.

Maximize w.r.t. $q(z)$

$$L = F(q) + \lambda \left(\int q(z) dz - 1 \right) \tag{50}$$

$$= \iint q(z) q(\theta) \log \frac{p(D, z, \theta)}{q(z) q(\theta)} \, dz d\theta + \lambda \left(\int q(z) dz - 1 \right) \quad \text{とおくと,}$$

$$\frac{\delta L}{\delta q(z)} = \iint q(\theta) \Big[\log p(D, z, \theta) - \log q(\theta) - \log q(z) - 1 \Big] dz d\theta + \lambda$$

$$= \iint q(\theta) \Big[\log p(D, z|\theta) + \log p(\theta) - \log q(\theta) - \log q(z) - 1 \Big] dz d\theta + \lambda$$

$$= \Big\langle \log p(D, z|\theta) \Big\rangle_{q(\theta)} - \log q(z) + (\text{const.}) + \lambda = 0 \tag{51}$$

$$\therefore \quad q(z) \ \propto \ \exp \Big\langle \log p(D, z|\theta) \Big\rangle_{q(\theta)}. \tag{52}$$

Maximize w.r.t. $q(\theta)$

$$L = F(q) + \lambda \left(\int q(\theta) d\theta - 1 \right) \tag{53}$$

$$= \iint q(z) q(\theta) \log \frac{p(D, z, \theta)}{q(z) q(\theta)} dz d\theta + \lambda \left(\int q(\theta) d\theta - 1 \right) \tag{54}$$

$$\frac{\delta L}{\delta q(\theta)} = \iint q(z) \big[\log p(D, z, \theta) - \log q(\theta) - \log q(z) - 1 \big] dz d\theta + \lambda$$

$$= \iint q(z) \big[\log p(D, z|\theta) + \log p(\theta) - \log q(\theta) - \log q(z) - 1 \big] dz d\theta + \lambda$$

$$= \big\langle \log p(D, z|\theta) \big\rangle_{q(z)} + \log p(\theta) - \log q(\theta) + (\text{const.}) + \lambda = 0 \tag{55}$$

$$\therefore \quad q(\theta) \ \propto \ p(\theta) \exp \big\langle \log p(D, z|\theta) \big\rangle_{q(z)}. \tag{56}$$

変分ベイズ推定のまとめ

- 観測データ D に対して, 隠れ変数 z, パラメータ θ をすべて確率変数とみて, その確率分布を求める.

$$\log p(D) = \log \iint p(D, z, \theta) dz d\theta \tag{57}$$

$$\geq \iint q(z)q(\theta) \log \frac{p(D, z, \theta)}{q(z)q(\theta)} dz d\theta \ = F(q). \tag{58}$$

- $F(q)$ を $q(z)$, $q(\theta)$ に関して最大化すると,

$$\begin{cases} q(z) \propto \exp\left\langle \log p(D, z|\theta) \right\rangle_{q(\theta)} & \textbf{(VB-E step)} \ (59) \\ q(\theta) \propto p(\theta) \exp\left\langle \log p(D, z|\theta) \right\rangle_{q(z)} & \textbf{(VB-M step)} \ (60) \end{cases}$$

$\cdots$ VB-EM アルゴリズム.

変分ベイズ法について (1)

- VB-EM アルゴリズム:

$$\begin{cases} q(z) \propto \exp\left\langle \log p(D, z|\theta) \right\rangle_{q(\theta)} & \textbf{(VB-E step)} \ (61) \\ q(\theta) \propto p(\theta) \exp\left\langle \log p(D, z|\theta) \right\rangle_{q(z)} & \textbf{(VB-M step)} \ (62) \end{cases}$$

$q(\theta) = \delta(\hat{\theta})$ のとき, (44) 式は

$$q(z) \propto p(D, z|\hat{\theta}) \propto p(z|D, \hat{\theta}) \qquad (63)$$

$\cdots$ EM アルゴリズムの E-step と同じ.

変分ベイズ法について (2)

$$\log p(D) \geq F(q)\,. \tag{64}$$

ここで,

$$\log p(D) - F(q) \tag{65}$$

$$= \iint q(z,\theta) \log p(D) dz d\theta - \iint q(z,\theta) \log \frac{p(D,z,\theta)}{q(z,\theta)} dz d\theta \tag{66}$$

$$= \iint q(z,\theta) \Big[\log p(D) - \log p(z,\theta|D) - \log p(D) + \log q(z,\theta)\Big] dz d\theta \tag{67}$$

$$= \iint q(z,\theta) \log \frac{q(z,\theta)}{p(z,\theta|D)} dz d\theta \tag{68}$$

$$= D(q(z,\theta)||p(z,\theta|D)) \geq 0\,. \tag{69}$$

この近似誤差をできるだけ小さくするように, $q(z,\theta) = q(z)q(\theta)$ を最適化している.

変分ベイズ法について (3)

$$F(q) = \iint q(z)q(\theta) \log \frac{p(D,z,\theta)}{q(z)q(\theta)} dz d\theta \tag{70}$$

$$= \iint q(z)q(\theta) \log \frac{p(D,z|\theta)}{q(z)} \frac{p(\theta)}{q(\theta)} dz d\theta \tag{71}$$

$$= \left\langle \log \frac{p(D,z|\theta)}{q(z)} \right\rangle_{q(z)q(\theta)} - \int q(\theta) \log \frac{q(\theta)}{p(\theta)} d\theta \tag{72}$$

$$= \left\langle \log \frac{p(D,z|\theta)}{q(z)} \right\rangle_{q(z)q(\theta)} \underbrace{-D(q(\theta|D)||p(\theta))}_{\text{過学習を防ぐ (正則化項)}} \tag{73}$$

$$\left(\rightarrow \left\langle \log \frac{p(D,z|\theta)}{q(z)} \right\rangle_{q(z)q(\theta)} \underbrace{-\frac{|\hat{\theta}|}{2} \log N}_{\text{MDL, BIC}} + \underbrace{\log p(\hat{\theta})}_{\text{(const.)}} \right) \tag{74}$$

- パラメータ事前分布と事後分布の KL ダイバージェンスで, 自動的に正則化が行われる.

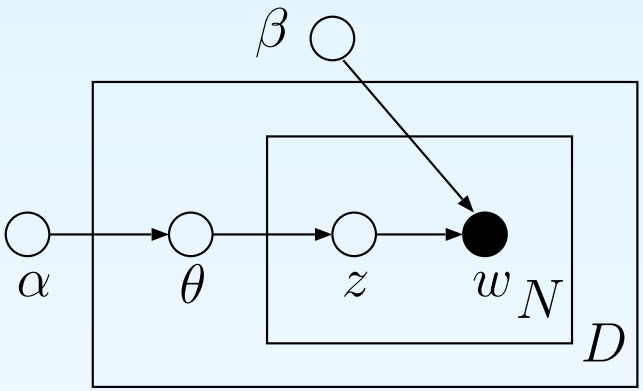
変分ベイズ法のまとめ (2)

- 学習データの尤度の下限を, 変分近似して最大化する

$$\log p(D) \geq F(q) \ \rightarrow \ \text{最大化}. \tag{75}$$

  - 左辺と右辺の差は KL ダイバージェンス ($\rightarrow$ 最小化).
- $\delta F/\delta q(z), \delta F/\delta q(\theta) \rightarrow$ VB-EM アルゴリズム.
  - パラメータの確率分布 $q(z), q(\theta)$ が求まる ($\neq$ 点推定)
  - $q(\theta)$ が δ 関数のとき, 通常の EM アルゴリズムと一致
- パラメータの過学習を防ぐ $\cdots$ パラメータの事前分布 $p(\theta)$ と事後分布 $q(\theta|D)$ との KL-ダイバージェンスで自動的に正則化
- データ数 $N \rightarrow \infty$ の極限で MDL/BIC と一致

応用: LDA



- PLSI では文書 d について対応するパラメータ $\theta = p(z|d)$ を点推定したが, これは過学習する恐れがある
- θ 自体に確率を与える (ディリクレ分布) :

$$p(\theta) \sim \mathrm{Dir}(\theta|\alpha) \tag{76}$$

- $\beta = p(w|z)$ とおくと,

$$p(\mathbf{w}|\alpha,\beta) = \int p(\theta|\alpha) \prod_n p(w_n|\theta,\beta) d\theta \tag{77}$$

$$= \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \int \prod_k \theta_k^{\alpha_k - 1} \prod_n \sum_z \prod_v (\theta_z \beta_{zv})^{w_n^v} d\theta$$

$\rightarrow$ 最大化. (78)

応用: LDA (2)

$$\log p(\mathbf{w}|\alpha,\beta) = \log \int \sum_z p(\mathbf{w},z,\theta|\alpha,\beta)d\theta \tag{79}$$

$$= \log \int \sum_z q(z,\theta|\gamma,\psi)\frac{p(\mathbf{w},z,\theta|\alpha,\beta)}{q(z,\theta|\gamma,\psi)}d\theta \tag{80}$$

$$\geq \int \sum_z q(z,\theta|\gamma,\psi)\log\frac{p(\mathbf{w},z,\theta|\alpha,\beta)}{q(z,\theta|\gamma,\psi)}d\theta \tag{81}$$

- $q(z,\theta|\mathbf{w},\gamma,\psi) = q(\theta|\gamma)\prod_n q(z_n|w_n,\psi)$ と近似すると,

$$\begin{aligned}\log p(\mathbf{w}|\alpha,\beta) \geq & \left\langle \log p(\theta|\alpha)\right\rangle_{q(\theta|\gamma)} + \sum_n \left\langle \log p(z_n|\theta)\right\rangle_{q(\theta|\gamma),q(z_n|w_n,\psi)} \\ & + \sum_n \left\langle \log p(w_n|z_n,\beta)\right\rangle_{q(z_n|w_n,\psi)} \\ & - \left\langle \log q(\theta|\gamma)\right\rangle_{q(\theta|\gamma)} - \sum_n \left\langle \log q(z_n|w_n,\psi)\right\rangle_{q(z_n|w_n,\psi)} .\end{aligned} \tag{82}$$

VB-HMM

- 観測系列 $\mathbf{y} = y_1 y_2 \cdots y_T$ に対して, 隠れた真の状態系列 $\mathbf{s} = s_1 s_2 \cdots s_T$ があって,
- HMM のパラメータ
  - 初期状態確率 $\pi \quad (1 \times K)$
  - 状態遷移行列 $C \quad (K \times K)$
  - 出力確率行列 $A \quad (K \times W)$ について

$$\log p(\mathbf{y})$$

$$= \log \int d\pi \int dA \int dC \sum_{\mathbf{s}} p(\pi, A, C) p(\mathbf{y}, \mathbf{s} | \pi, A, C) \tag{83}$$

$$\geq \int d\pi \int dA \int dC \sum_{\mathbf{s}} q(\pi, A, C, \mathbf{s}) \log \frac{p(\pi, A, C) p(\mathbf{y}, \mathbf{s} | \pi, A, C)}{q(\pi, A, C, \mathbf{s})} \,. \tag{84}$$

VB-HMM (2)

- π,C の各列, A の各列にそれぞれディリクレ事前分布 $\mathrm{Dir}(\alpha), \mathrm{Dir}(\beta), \mathrm{Dir}(\gamma)$ を考えると,
- VB-Estep:

$$\langle \pi_k \rangle \propto \exp\bigl(\Psi(\alpha_k^*) - \Psi(\sum_k \alpha_k^*)\bigr) \tag{85}$$

$$\langle A_{ij} \rangle \propto \exp\bigl(\Psi(\beta_{ij}^*) - \Psi(\sum_j \beta_{ij}^*)\bigr) \tag{86}$$

$$\langle C_{ij} \rangle \propto \exp\bigl(\Psi(\gamma_{ij}^*) - \Psi(\sum_j \gamma_{ij}^*)\bigr) \tag{87}$$

- VB-Mstep:
  - パラメータ α, β, γ の事後分布 $\alpha^*, \beta^*, \gamma^*$ を Forward-Backward から更新.
- 詳細は Beal, M.J. (2003) Variational Algorithms for Approximate Bayesian Inference. PhD thesis, Gatsby UCL. http://www.cse.buffalo.edu/faculty/mbeal/thesis/ Chapter.3.

ベイズ推定のための解法

- 変分ベイズ法は, グラフィカルモデル (or, ベイジアンネットワーク) を解くための方法の一つ
- Gibbs sampling, MCMC
  - モデルから実際にサンプリングして, 平均を取る
  - 近似のない, 正確な推定が可能
  - 複雑なモデルでも, 多くの場合適用できる
  - 計算時間が長い (LDA の場合, 3 倍くらい (らしい))
- EP (Expectation Propagation) (Minka 2001), Power EP (Minka 2004)
  - VB とは別の解析的近似
  - EP … KL-ダイバージェンス最小
  - Power EP … α-ダイバージェンス最小

Readings

- Hagai Attias. A Variational Bayesian Framework for Graphical Models. In *NIPS 1999*, 1999.
- Thomas Minka. Expectation-Maximization as lower bound maximization, 1998. http://research.microsoft.com/˜minka/papers/em.html.
- Radford M. Neal and Geoffrey E. Hinton. *A View of the EM Algorithm that Justifies Incremental, Sparse, and other Variants*. in *Learning in Graphical Models*, pages 355–368. Dordrecht: Kluwer Academic Publishers, 1998.
- Zoubin Ghahramani. *Unsupervised Learning*. in *Advanced Lectures on Machine Learning LNAI 3176*. Springer-Verlag, Berlin, 2004. http://www.gatsby.ucl.ac.uk/˜zoubin/course04/ul.pdf.