A Tutorial on Variational Bayes

Junhao Hua (华俊豪)
Laboratory of Machine and Biological Intelligence,
Department of Information Science & Electronic Engineering,
ZheJiang University
2014/3/27

Email: huajh7@gmail.com

For further information, see:

http://www.huajh7.com

Outline

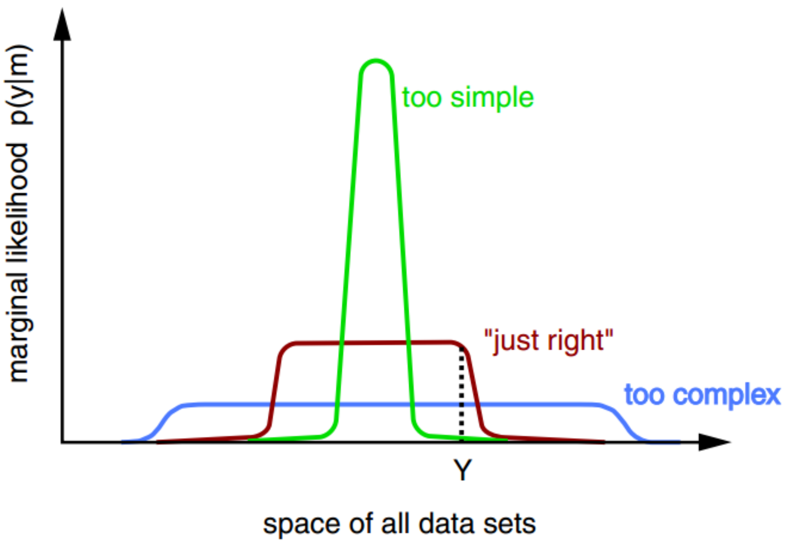
- Motivation
- The Variational Bayesian Framework
  - Variational Free Energy
  - Optimization Tech. Mean Field Approximation
  - Exponential Family
  - Bayesian Networks
- Example:
  - VB for Mixture model
- Discussion
- Application
- Reference

A Problem: How Learn From Data?

- Typical, we use a complex statistical model, but how to learn its parameters and latent variables?

- Data: X
- Model: P(X | \theta, Z)

Challenge



- Maximum Likelihood:
  - Overfits the data
  - Model Complexity
  - Computational tractability
- Bayesian Framework:
  - Arising intractable integrals:
    - partition function
    - posterior of unobserved variables
  - Approximate Inference:
    - Monte Carlo Sampling: e.g. MCMC, particle filter.
    - Variational Bayes

Outline

- Motivation
- The Variational Bayesian Framework
  - Variational Free Energy
  - Optimization Tech. Mean Field Approximation
  - Exponential Family
  - Bayesian Networks
- Example:
  - VB for Mixture model
- Discussion
- Application
- Reference

Variational Free energy

- Basic Idea:

"conditional independence is enforced as a functional constraint in the approximating distribution, and the best such approximation is found by minimization of a Kullback-Leibler divergence (KLD). "

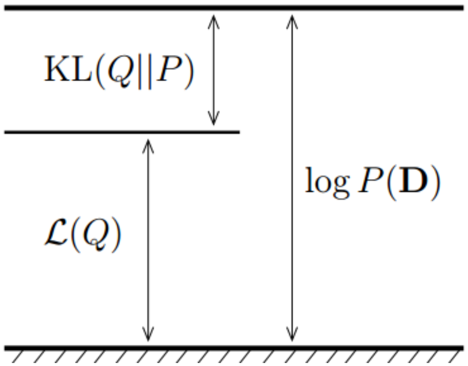
  - Use a simpler variational distribution, $Q(Z)$, to approximate the true posterior $P(Z \mid X)$

- Two alternative explanations
  - Minimize (reverse) Kullback-Leibler divergence

$$D_{KL}(Q \| P) = \sum_Z Q(Z) \log \frac{Q(Z)}{P(Z \mid D)} = \sum_Z Q(Z) \log \frac{Q(Z)}{P(Z, D)} + \log P(D)$$

  - Maximum variational free energy(lower bound)

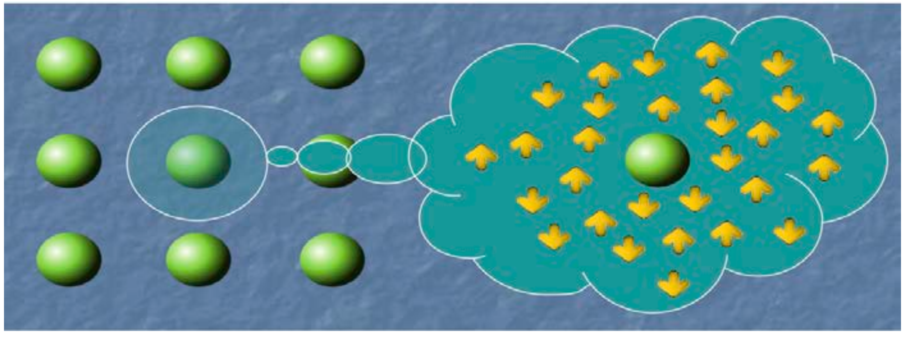
$$L(Q) = \sum_Z Q(Z) \log P(Z, D) - \sum_Z Q(Z) \log Q(Z) = E_Q[\log P(Z, D)] + H(Q)$$



Optimization Techniques: Mean Field Approximation

- Originated in the statistical physics literature
- Conditional Independence assumption
- Decoupling: intractable distribution -> a product of tractable marginal distributions (tractable subgraph)
- Factorization:

$$Q(Z) = \prod_{i=1}^{M} q(Z_i \mid D)$$



Optimization Techniques: Variational methods

- Optimization Problem:
  - Maximum the lower bound

$$L(Q(Z)) = E_{Q(Z)}[\ln P(Z,D)] + H(Q(Z))$$

  - Where $Q(Z) = \prod_i Q_i(Z_i)$
  - Subject to normalization constraints:

$$\forall i. \int Q_i(Z_i) dZ_i = 1$$

- Seek the extremum of a functional:
  - Euler – Lagrange equation

Derivation

- Consider the partition $Z = \{Z_i,\ Z_{-i}\}$, where $Z_{-i} = Z \setminus Z_i$
- Consider Energy term,

$$E_{Q(Z)}[\ln P(Z,D)] = \int (\prod_i Q_i(Z_i)) \ln(Z,D) dZ$$

$$= \int Q_i(Z_i) \int Q_{-i}(Z_{-i}) \ln(Z,D) dZ_{-i} dZ_i$$

$$= \int Q_i(Z_i) \left\langle \ln(Z,D) \right\rangle_{Q_{-i}(Z_{-i})} dZ_i$$

$$= \int Q_i(Z_i) \ln \exp \left\langle \ln(Z,D) \right\rangle_{Q_{-i}(Z_{-i})} dZ_i$$

$$= \int Q_i(Z_i) \ln Q_i^*(Z_i) dZ_i + \ln C$$

- We define $Q_i^*(Z_i) = \frac{1}{C} \exp \left\langle \ln(Z,D) \right\rangle_{Q_{-i}(Z_{-i})}$, where C is the normalization constant.

Derivation (cont.)

- Consider the entropy,

$$H(Q(Z)) = \sum_i \int (\prod_k Q_k(Z_k)) \ln Q_i(Z_i) dZ$$
$$= \sum_i \iint Q_i(Z_i) Q_{-i}(Z_{-i}) \ln Q_i(Z_i) dZ_i dZ_{-i}$$
$$= \sum_i \left\langle \int Q_i(Z_i) \ln Q_i(Z_i) dZ_i \right\rangle_{Q_{-i}(Z_{-i})}$$
$$= \sum_i \int Q_i(Z_i) \ln Q_i(Z_i) dZ_i$$

- Then we get the functional,

$$\mathrm{L(Q(Z))} = \int Q_i(Z_i) \ln Q_i^*(Z_i) dZ_i + \sum_i \int Q_i(Z_i) \ln Q_i(Z_i) dZ_i + \ln \mathrm{C}$$
$$= \left(\int Q_i(Z_i) \ln Q_i^*(Z_i) dZ_i - \int Q_i(Z_i) \ln Q_i(Z_i) dZ_i\right) + \sum_{k \neq i} \int Q_k(Z_k) \ln Q_k(Z_k) dZ_k + \ln \mathrm{C}$$
$$= \int Q_i(Z_i) \ln \frac{Q_i^*(Z_i)}{Q_i(Z_i)} dZ_i + \sum_{k \neq i} \int Q_k(Z_k) \ln Q_k(Z_k) dZ_k + \ln \mathrm{C}$$
$$= -\mathrm{D}_{KL}(Q_i(Z_i) \| Q_i^*(Z_i)) + H[Q_{-i}(Z_{-i})] + \ln C$$

Derivation (cont.)

- Maximizing energy functional L *w.r.t.* each Q_i could be achieved by Lagrange multipliers and functional differentiation

$$\forall i.\ \frac{\partial}{\partial Q_i(Z_i)}\{-\mathrm{D}_{KL}[Q_i(Z_i)\,\|\,Q_i^*(Z_i)] - \lambda_i(\int Q_i(Z_i)dZ_i - 1)\} := 0$$

- A long algebraic derivation would then eventually lead to a Gibbs distribution; Fortunately, L will be maximized when the KL divergence is zero,

$$Q_i(Z_i) = Q_i^*(Z_i) = \frac{1}{C}\exp\left\langle \ln P(Z_i, Z_{-i}, D)\right\rangle_{Q_{-i}(Z_{-i})}$$

  - Where C is normalization constant.

Challenge

$$Q_i(Z_i) = Q_i^*(Z_i) = \frac{1}{C}\exp\left\langle \ln P(Z_i, Z_{-i}, D)\right\rangle_{Q_{-i}(Z_{-i})}$$

- The expectation can be intractable.
- We need pick a family of distributions *Q* that allow for exact inference
- Then Find $Q' \in Q$ that maximizes the functional energy .

Challenge

- The expectation can be intractable.
- We need pick a family of distributions Q that allow for exact inference
- Then Find $Q' \in Q$ that maximizes the functional energy .

Exponential Family

Why Exponential Family ?

- Principle of maximum entropy

entropy is maximal. More formally, letting $\mathcal{P}$ be the set of all probability distributions over the random variable X, the maximum entropy solution p^* is given by the solution to the following constrained optimization problem:

$$p^* := \arg\max_{p \in \mathcal{P}} H(p) \quad \text{subject to } \mathbb{E}_p[\phi_\alpha(X)] = \widehat{\mu}_\alpha \quad \text{for all } \alpha \in \mathcal{I}. \tag{3.3}$$

- Density function:

$$p_\theta(x_1, x_2, \ldots, x_m) = \exp\left\{\langle \theta, \phi(x) \rangle \; - \; A(\theta)\right\},$$

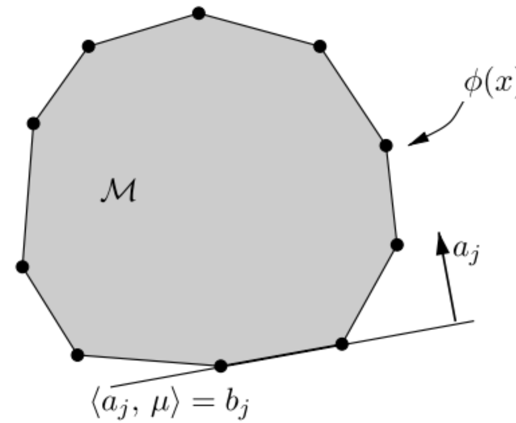
canonical parameters

Mean parameters

- Log partition function:

$$A(\theta) = \log \int_{\mathcal{X}^m} \exp\langle \theta, \phi(x) \rangle \, \nu(dx).$$

Properties of Exponential family



- Mean parameters: θ

"various statistical computations, among them marginalization and maximum likelihood estimation, can be understood as transforming from one parameterization to the other."

  - All realizable mean parameters

$$\mathcal{M} := \{ \mu \in \mathbb{R}^d \mid \exists\, p \text{ s.t. } \mathbb{E}_p[\phi_\alpha(X)] = \mu_\alpha \; \forall \alpha \in \mathcal{I} \},$$

    - Always a convex subset of $\mathbb{R}^d$

- Forward mapping
  - From canonical parameters $\phi(x)$ to the mean parameters θ
- Backward mapping
  - From mean parameters θ to the canonical parameters $\phi(x)$

- Properties of partition function A

Proposition 3.1. The cumulant function

$$A(\theta) := \log \int_{\mathcal{X}^m} \exp\langle \theta, \phi(x) \rangle \, \nu(dx) \tag{3.40}$$

associated with any regular exponential family has the following properties:

(a) It has derivatives of all orders on its domain Ω. The first two derivatives yield the cumulants of the random vector $\phi(X)$ as follows:

$$\frac{\partial A}{\partial \theta_\alpha}(\theta) = \mathbb{E}_\theta[\phi_\alpha(X)] := \int \phi_\alpha(x) p_\theta(x) \nu(dx). \tag{3.41a}$$

$$\frac{\partial^2 A}{\partial \theta_\alpha \partial \theta_\beta}(\theta) = \mathbb{E}_\theta[\phi_\alpha(X)\phi_\beta(X)] - \mathbb{E}_\theta[\phi_\alpha(X)]\mathbb{E}_\theta[\phi_\beta(X)]. \tag{3.41b}$$

(b) Moreover, A is a convex function of θ on its domain Ω, and strictly so if the representation is minimal.

Theorem 3.3. In a minimal exponential family, the gradient map ∇A is onto the interior of $\mathcal{M}$, denoted by $\mathcal{M}^\circ$. Consequently, for each $\mu \in \mathcal{M}^\circ$, there exists some $\theta = \theta(\mu) \in \Omega$ such that $\mathbb{E}_\theta[\phi(X)] = \mu$.

Conjugate Duality: Maximum Likelihood and Maximum Entropy

- The variational representation of log partition function

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \left\{ \langle \theta, \mu \rangle - A^*(\mu) \right\}.$$

- The conjugate dual function to *A*

$$A^*(\mu) := \sup_{\theta \in \Omega} \left\{ \langle \mu, \theta \rangle - A(\theta) \right\}.$$

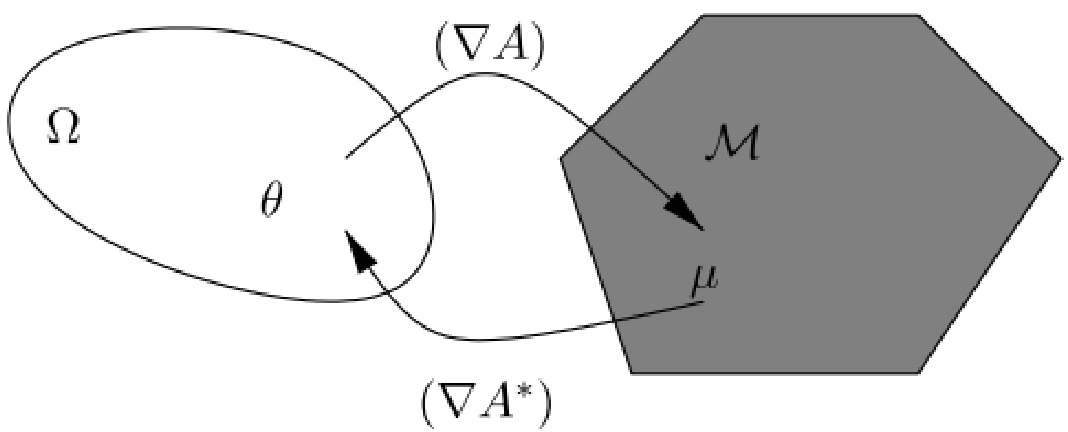


Fig. 3.8 Idealized illustration of the relation between the set Ω of valid canonical parameters, and the set $\mathcal{M}$ of valid mean parameters. The gradient mappings ∇A and ∇A^* associated with the conjugate dual pair (A, A^*) provide a bijective mapping between Ω and the interior $\mathcal{M}^\circ$.

Nonconvexity for Naïve Mean Field

- Mean field optimization is always nonconvex for any exponential family in which the state space is finite.
  - It is a strict subset of M(G)
  - Contains all of the extreme points of polytope

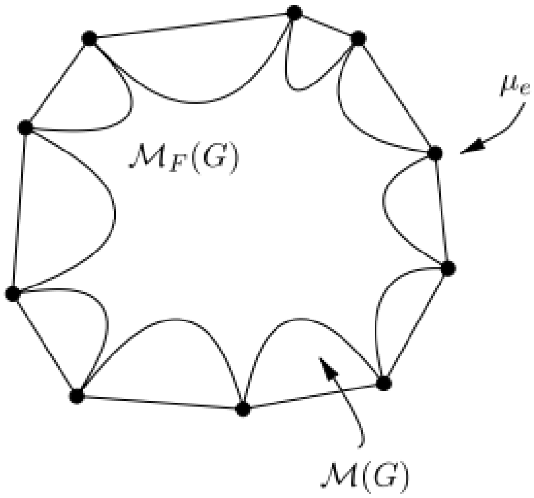


Fig. 5.3 Cartoon illustration of the set $\mathcal{M}_F(G)$ of mean parameters that arise from tractable distributions is a nonconvex inner bound on $\mathcal{M}(G)$. Illustrated here is the case of discrete random variables where $\mathcal{M}(G)$ is a polytope. The circles correspond to mean parameters that arise from delta distributions, and belong to both $\mathcal{M}(G)$ and $\mathcal{M}_F(G)$.

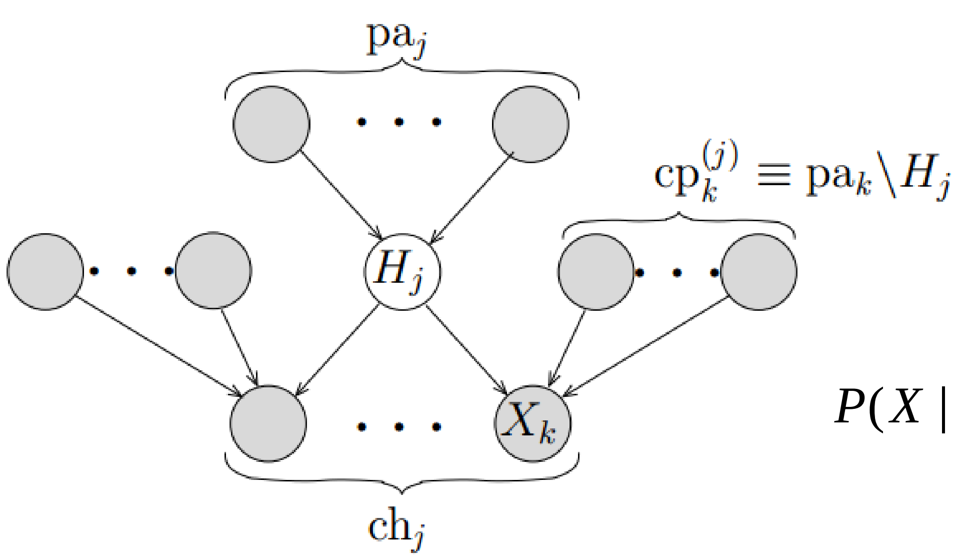
Outline

- Motivation
- The Variational Bayesian Framework
  - Variational Free Energy
  - Optimization Tech. Mean Field Approximation
  - Exponential Family
  - Bayesian Networks
- Example:
  - VB for Mixture model
- Discussion
- Application
- Reference

Inference in Bayesian Networks

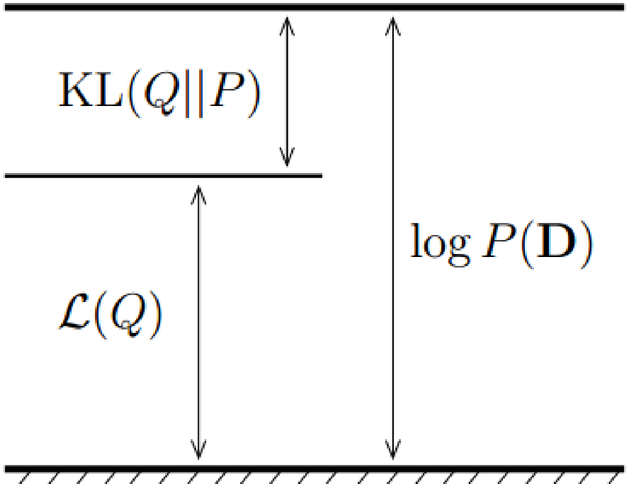
- Variational Message Passing (Winn, Bishop, 2003.)
  - Message from parents: $m_{Y \to X} = \langle u_Y \rangle$
  - Message to parents: $m_{X \to Y} = \tilde{\phi}_{XY}\left(\langle u_X \rangle, \{m_{i \to X}\}_{i \in cp_Y}\right)$
  - Update natural parameter vector :

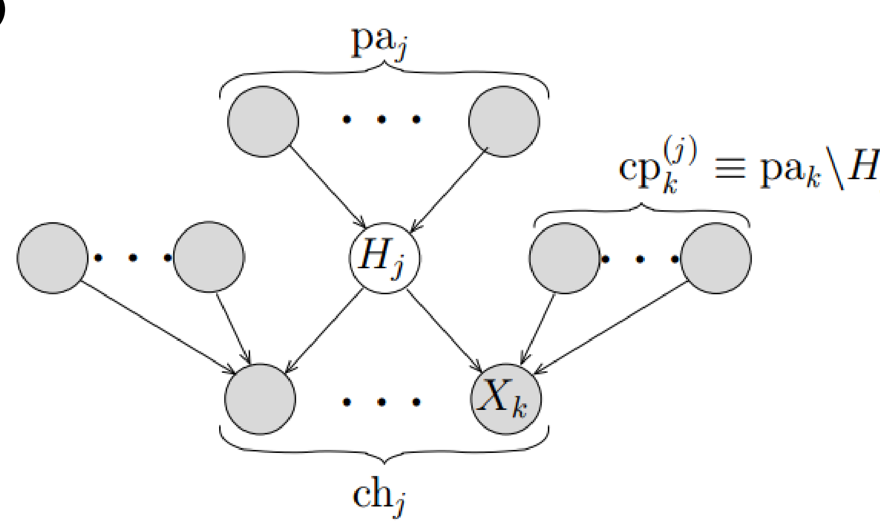
$$\phi_Y^* = \tilde{\phi}_Y\left(\{m_{i \to Y}\}_{i \in pa_Y}\right) + \sum_{j \in ch_Y} m_{j \to Y}$$



$$P(X \mid \phi) = \exp[\phi^T u(X) + f(X) + \tilde{g}(\phi)]$$

Summary of VB





Mean-field Assumption
Variational Methods

$$Q(Z_i) \propto \frac{1}{C} \exp \left\langle \ln P(Z_i, Z_{-i}, D) \right\rangle_{Q(Z_{-i}) or Q(mb(Z_i))}$$

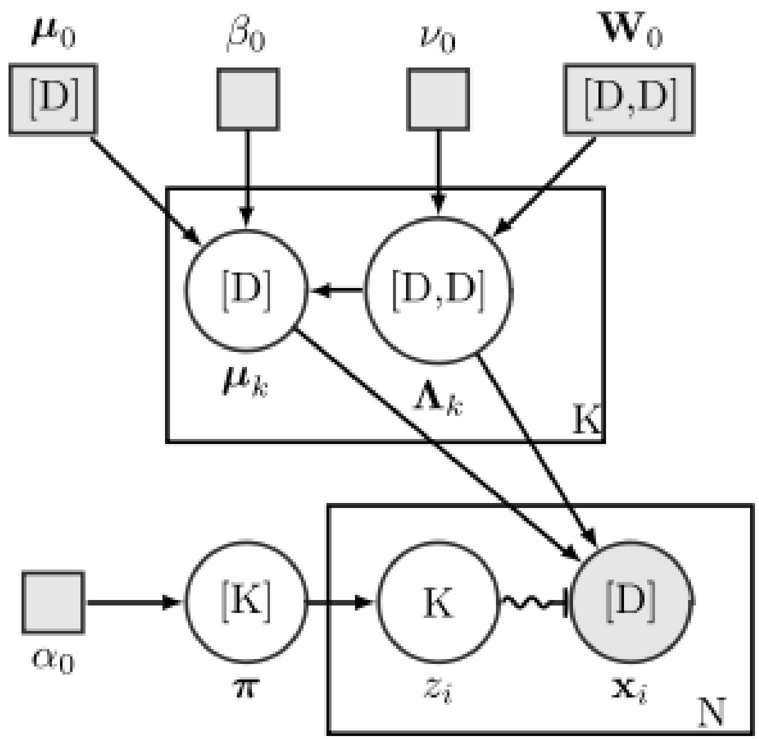
Conjugate-exponential family
Forward, backward mapping

Outline

- Motivation
- The Variational Bayesian Framework
  - Variational Free Energy
  - Optimization Tech. Mean Field Approximation
  - Exponential Family
  - Bayesian Networks
- Example
  - VB for Mixture model
- Discussion
- Application
- Reference

Mixture of Gaussian (MoG)

$$p(X \mid Z, \mu, \Lambda) = \prod_{n=1}^{N} \prod_{k=1}^{K} N(x_n \mid \mu_k, \Lambda_k^{-1})^{z_{nk}}$$



$$p(X, Z, \pi, \mu, \Lambda) = p(X \mid Z, \mu, \Lambda)\, p(Z \mid \pi)\, p(\pi)\, p(\mu \mid \Lambda)\, p(\Lambda)$$

Infinite Student's t-mixture

$DP(\alpha, G_0)$

Dirichlet Process

$G = \sum_{j=1}^{\infty} \pi_j(V)\delta_{\Theta_j} \quad \pi_j(V) = V_j \prod_{i=1}^{j-1}(1 - V_i)$

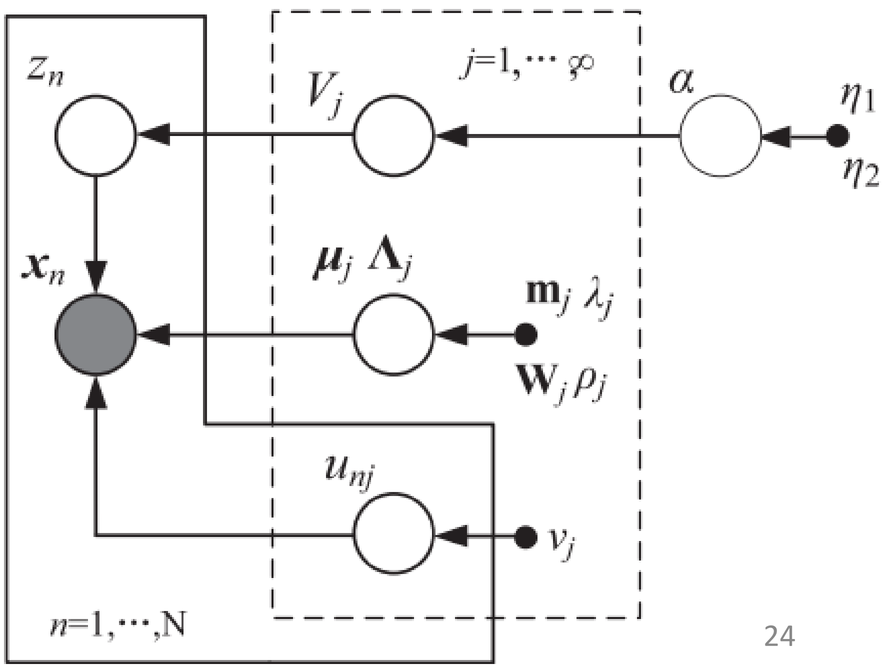
Stick-Breaking prior

$V_j \sim Beta(1, \alpha)$

$p(\alpha) = Gam(\alpha \mid \eta_1, \eta_2)$

Dirichlet Process Mixture

$$p(X) = \prod_{n=1}^{N} \sum_{j=1}^{\infty} \pi_j(V) \cdot St(x_n \mid \mu_j, \Lambda_j, v_j)$$



Latent Dirichlet Allocation (LDA)

$$p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta)$$

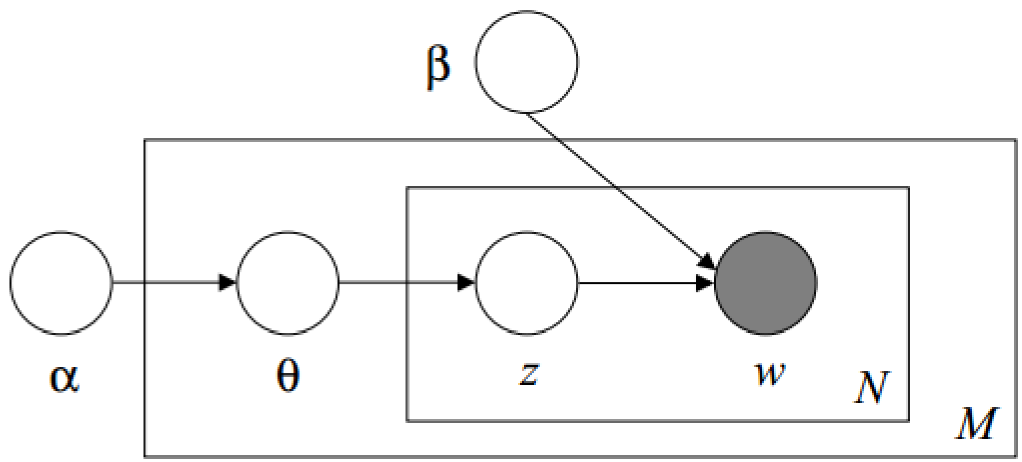


Figure 1: Graphical model representation of LDA. The boxes are "plates" representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.

Outline

- Motivation
- The Variational Bayesian Framework
  - Variational Free Energy
  - Optimization Tech. Mean Field Approximation
  - Exponential Family
  - Bayesian Networks
- Example:
  - VB for Mixture model
- Discussion
- Application
- Reference

• 步骤一：选择无信息先验分布

- 选择先验分布原则：共轭分布，Jefferys原则，最大熵原则等。
- 一般要求先验分布应取共轭分布（conjugate distribution）才合适，即先验分布h(θ)与后验分布h(θ|x)属于同一分布类型。

$$\pi_{i=1,...,k} \sim SymDir(K,\alpha_0)$$

$$\Lambda_{i=1,...,k} \sim \mathrm{W}(w_0,\upsilon_0)$$

$$\mu_{i=1,...,k} \sim N(m_0,(\beta_0\Lambda_i)^{-1})$$

$$z_{i=1,...,N} \sim Mult(1,\pi)$$

$$X_{i=1,...,N} \sim N(\mu_z)$$

说明：
- •K:单高斯分布个数，N：样本个数
- •SymDir() :K维对称 Dirichlet分布；它是分类分布（categorical）或多项式分布的共轭先验分布。
- •W() 表示Wishart分布；对多元高斯分布，它是Precision矩阵（逆协方差矩阵）的共轭先验。
- •Mult() 表示多项分布; 是二项式分布的推广，表示在一个K维向量中只有一项为1，其它都为0.
- •N() 为多元高斯分布。

$X=\{x_1,...,x_N\}$是N个训练样本，每项都是服从多元高斯分布的K维向量;

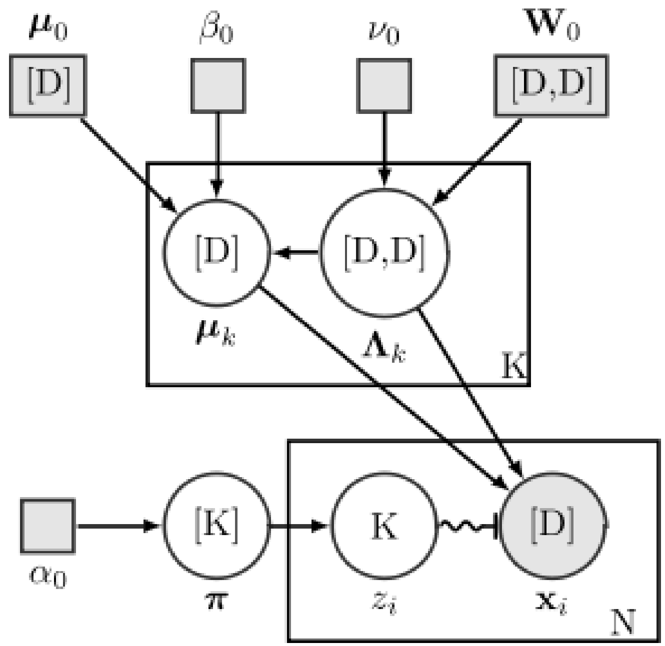
$Z=\{z_1,...,z_N\}$是一组潜在变量，每项$z_k=\{z_{1k,...},z_{nk}\}$表示对应的样本$x_k$属于哪个混合部分;

$\pi=\{\pi_1,...,\pi_K\}$表示每个单高斯分布混合比例;

$\mu_{i=1,...,k}$和$\Lambda_{i=1,...,k}$分别表示每个单高斯分布参数的均值和精度;

$K,\alpha_0,\beta_0,w_0,\upsilon_0,m_0$称为超参数（*hyperparameter*）, 都为已知量。

- 用“盘子表示法”（plate notation）表示多元高斯混合模型，如图所示。



- 小正方形表示不变的超参数，如β_0 ,ν_0 ,α_0 ,μ_0 ,W_0;
- 圆圈表示随机变量，如 $\pi, z_i, x_i, \mu_k, \Lambda_k$ ；
- 圆圈内的值为已知量，其中[K],[D]表示K、D维的向量，[D,D]表示DxD的矩阵;
- 单个K表示一个有K个值的categorical变量；
- 波浪线和开关表示变量x_i通过一个K维向量z_i来选择其他传入的变量(μ_k, Λ_k)。

- **步骤二：写出联合概率密度函数**
- 假设各参数与潜在变量条件独立，则联合概率密度函数可以表示为

$$p(X,Z,\pi,\mu,\Lambda)=p(X\mid Z,\mu,\Lambda)p(Z\mid\pi)p(\pi)p(\mu\mid\Lambda)p(\Lambda)$$

- 每个因子为：$p(X\mid Z,\mu,\Lambda)=\prod_{n=1}^{N}\prod_{k=1}^{K}N(x_n\mid\mu_k,\Lambda_k^{-1})^{z_{nk}}$

$$p(Z\mid\pi)=\prod_{n=1}^{N}\prod_{k=1}^{K}\pi_k^{z_{nk}}$$

$$p(\pi)=\frac{\Gamma(K\alpha_0)}{\Gamma(\alpha_0)^K}\prod_{k=1}^{K}\pi_k^{\alpha_0-1}$$

$$p(\mu\mid\Lambda)=N(\mu_k\mid m_0,(\beta_0\Lambda_k)^{-1})$$

$$p(\Lambda)=W(\Lambda_k\mid w_0,\nu_0)$$

- 其中,

$$N(x\mid\mu,\Sigma)=\frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}$$

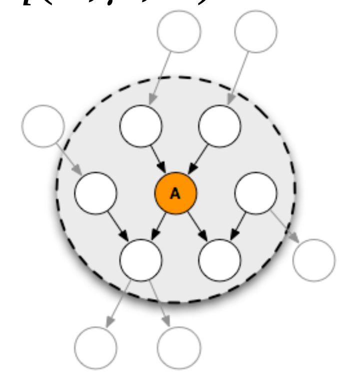
$$W(\Lambda\mid w,\nu)=B(w,\nu)|\Lambda|^{(\nu-D-1)/2}\exp(-\frac{1}{2}Tr(w^{-1}\Lambda))$$

$$B(w,\nu)=|w|^{-\nu/2}(2^{\nu D/2}\pi^{D(D-1)/4}\prod_{i=1}^{D}\Gamma(\frac{\nu+1-i}{2}))^{-1}$$

- **步骤三：计算边缘密度(VB- marginal)**

（1）计算Z的边缘密度，根据平均场假设,有$q(Z,\pi,\mu,\Lambda)=q(Z)q(\pi,\mu,\Lambda)$

$$\begin{aligned}\ln q^*(Z) &= E_{\pi,\mu,\Lambda}[\ln p(X,Z,\pi,\mu,\Lambda)]+\text{const}\\ &= E_{\pi,\mu,\Lambda}[\ln p(X\mid Z,\mu,\Lambda)p(Z\mid\pi)p(\pi)p(\mu\mid\Lambda)p(\Lambda)]+\text{const}\\ &= E_{\pi}[\ln p(Z\mid\pi)]+E_{\mu,\Lambda}[\ln p(X\mid Z,\mu,\Lambda)]+\text{const}\\ &= \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\ln\rho_{nk}+\text{const}\end{aligned}$$



- 其中 $\ln\rho_{nk}=E[\ln\pi_k]+\frac{1}{2}E[\ln|\Lambda_k|]-\frac{D}{2}\ln(2\pi)-\frac{1}{2}E_{\mu_k,\Lambda_k}[(x_n-\mu_k)^T\Lambda_k(x_n-\mu_k)]$
- 两边分别取对数可得， $q^*(Z)\propto\prod_{n=1}^{N}\prod_{k=1}^{K}\rho_{nk}^{z_{nk}}$
- 归一化，得 $q^*(Z)=\prod_{n=1}^{N}\prod_{k=1}^{K}r_{nk}^{z_{nk}}$，其中 $r_{nk}=\frac{\rho_{nk}}{\sum_{j=1}^{K}\rho_{nj}}$
- 可见 $q^*(Z)$是多个单观测多项式分布(single-observation multinomial distribution)的乘积 。
- 更进一步，根据categorical分布，有 $E[z_{nk}]=r_{nk}$

（2）计算π的概率密度, $q(\pi,\mu,\Lambda)=q(\pi)\prod_{k=1}^{K}q(\mu_k,\Lambda_k)$

$$\ln q^*(\pi)=E_{Z,\mu,\Lambda}[p(X\mid Z,\pi,\mu,\Lambda)]+const$$

$$=\ln p(\pi)+E_Z[\ln p(Z\mid\pi)]+const$$

$$=(\alpha_0\text{-}1)\sum_{k=1}^{K}\ln\pi_k+\sum_{n=1}^{N}\sum_{k=1}^{K}r_{nk}\ln\pi_k+const$$

•两边取对数 $q^*(\pi)\sim\prod_{n=1}^{K}\pi_k^{\sum_{n=1}^{N}r_{nk}+\alpha_0-1}$，可见 $q^*(\pi)$ 是Dirichlet分布，

$$q^*(\pi)\sim Dir(\alpha)$$

•其中 $\alpha=\alpha_0+N_k$，$N_k=\sum_{n=1}^{N}r_{nk}$.

- 最后同时考虑μ, Λ，对于每一个单高斯分布有，

$$\ln q^*(\mu_k, \Lambda_k) = E_{Z,\pi,\mu_{i\neq k},\Lambda_{i\neq k}}[\ln p(X \mid Z, \mu_k, \Lambda_k) p(\mu_k, \Lambda_k)]$$
$$= \ln p(\mu_k, \Lambda_k) + \sum_{n=1}^{N} E[z_{nk}] \ln N(x_n \mid \mu_k, \Lambda_k^{-1}) + const$$

- 经过一系列重组化解将得到Gaussian-Wishart分布，

$$q^*(\mu_k, \Lambda_k) = N(\mu_k \mid m_k, (\beta_k \Lambda_k)^{-1}) W(\Lambda_k \mid w_k, \nu_k)$$

其中
$$\begin{cases} \beta_{\mathrm{k}} = \beta_0 + N_k, \\ m_k = \dfrac{1}{\beta_k}(\beta_0 m_0 + N_k \overline{x}_k), \\ w_k^{-1} = w_0^{-1} + N_k S_k + \dfrac{\beta_0 N_k}{\beta_0 + N_k}(\overline{x}_k - m_0)(\overline{x}_k - m_0)^T, \\ v_k = v_0 + N_k, \\ \overline{x}_k = \dfrac{1}{N_{\mathrm{k}}} \sum_{n=1}^{N} r_{nk} x_n, \\ S_k = \dfrac{1}{N_k} \sum_{n=1}^{N} r_{nk} (\overline{x}_k - x_k)(\overline{x}_k - x_k)^T. \end{cases}$$

- **步骤四：迭代收敛**
- 最后，注意到对π,μ,Λ的边缘概率都需要且只需要r_{nk}；另一方面，r_{nk}的计算需要ρ_{nk}，而这又是基于$E[\ln\pi_k], E[\ln|\Lambda_k|], E_{\mu_k,\Lambda_k}[(x_n-\mu_k)^T\Lambda_k(x_n-\mu_k)]$，即需要知道π,μ,Λ的值。不难确定这三个期望的一般表达式为：

$$\begin{cases} \ln\tilde{\pi}_k \equiv E[\ln|\pi_k|] = \psi(\alpha_k) - \psi\left(\sum_{i=1}^{K}\alpha_i\right) \\ \ln\tilde{\Lambda}_k \equiv E[\ln|\Lambda_k|] = \sum_{i=1}^{D}\psi\left(\frac{\nu_k+1-i}{2}\right) + D\ln 2 + \ln|\Lambda_k| \\ E_{\mu_k,\Lambda_k}[(x_n-\mu_k)^T\Lambda_k(x_n-\mu_k)] = D\beta_k^{-1} + \nu_k(x_n-m_k)^T W_k(x_n-m_k) \end{cases}$$

- 这些结果能导出，

$$r_{nk} \propto \tilde{\pi}_k\tilde{\Lambda}_k^{1/2}\exp\left\{-\frac{D}{2\beta_k} - \frac{\nu_k}{2}(x_n-m_k)^T W_k(x_n-m_k)\right\}$$

且 $\sum_{k=1}^{K} r_{nk} = 1$.

Summary: Variational Inference for GMM

$q^*(\pi) \sim Dir(\alpha)$

$\alpha = \alpha_0 + N_k \qquad N_k = \sum_{n=1}^{N} r_{nk}$

Soft-count or ESS

$q^*(\mu_k, \Lambda_k) = N(\mu_k \mid m_k, (\beta_k \Lambda_k)^{-1}) W(\Lambda_k \mid w_k, \nu_k)$

$\beta_k = \beta_0 + N_k, m_k = \frac{1}{\beta_k}(\beta_0 m_0 + N_k \overline{x}_k), \nu_k = \nu_0 + N_k, \overline{x}_k = \frac{1}{N_k}\sum_{n=1}^{N} r_{nk} x_n$

$w_k^{-1} = w_0^{-1} + N_k S_k + \frac{\beta_0 N_k}{\beta_0 + N_k}(\overline{x}_k - m_0)(\overline{x}_k - m_0)^T, S_k = \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}(\overline{x}_k - x_n)(\overline{x}_k - x_n)^T$

VBM-Step

VBE-Step

Latent variable

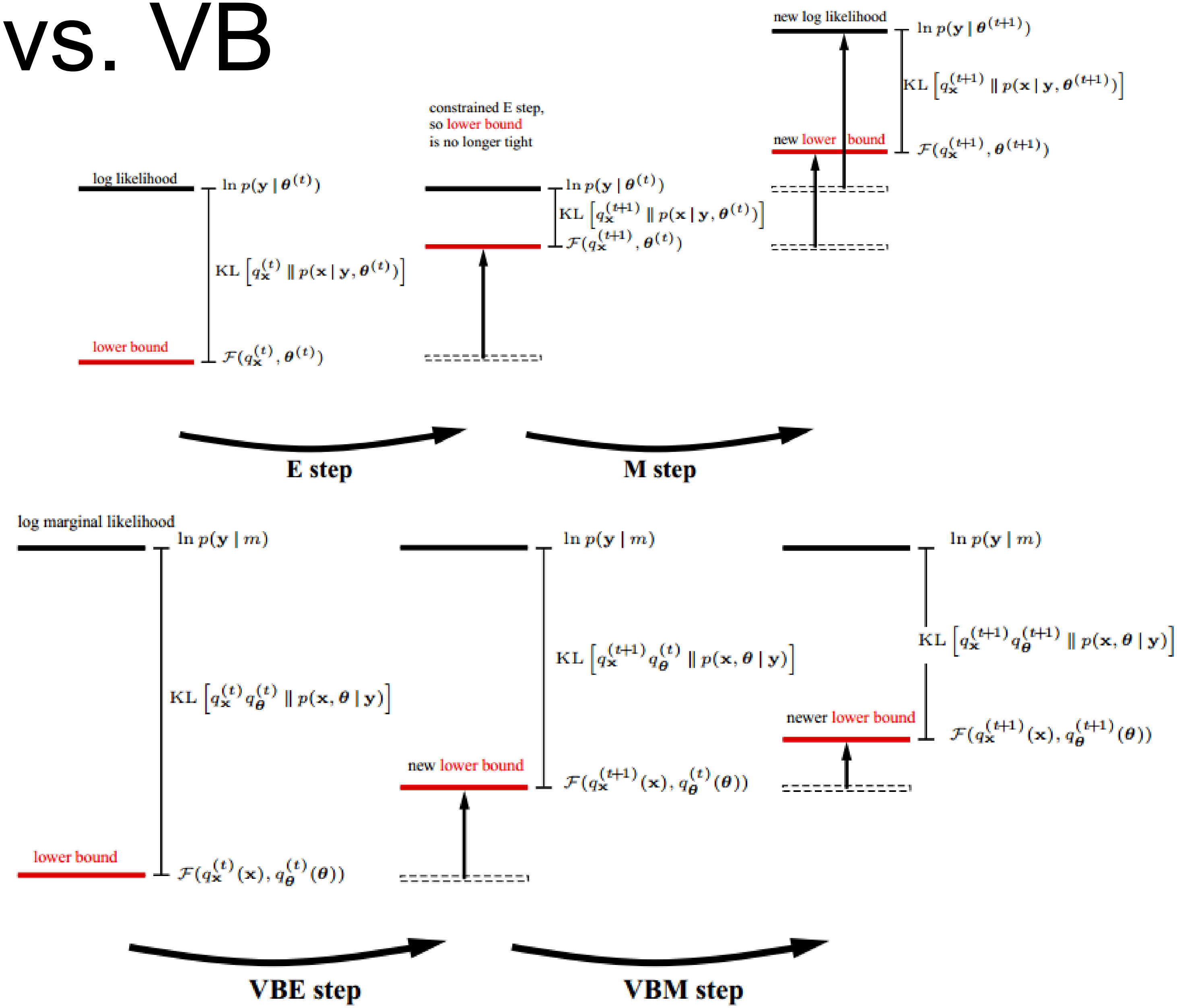
$q^*(Z) = \prod_{n=1}^{N}\prod_{k=1}^{K} r_{nk}^{z_{nk}}$

$r_{nk} \propto \tilde{\pi}_k \tilde{\Lambda}_k^{1/2} \exp\left\{-\frac{D}{2\beta_k} - \frac{\nu_k}{2}(x_n - m_k)^T W_k (x_n - m_k)\right\}$

Outline

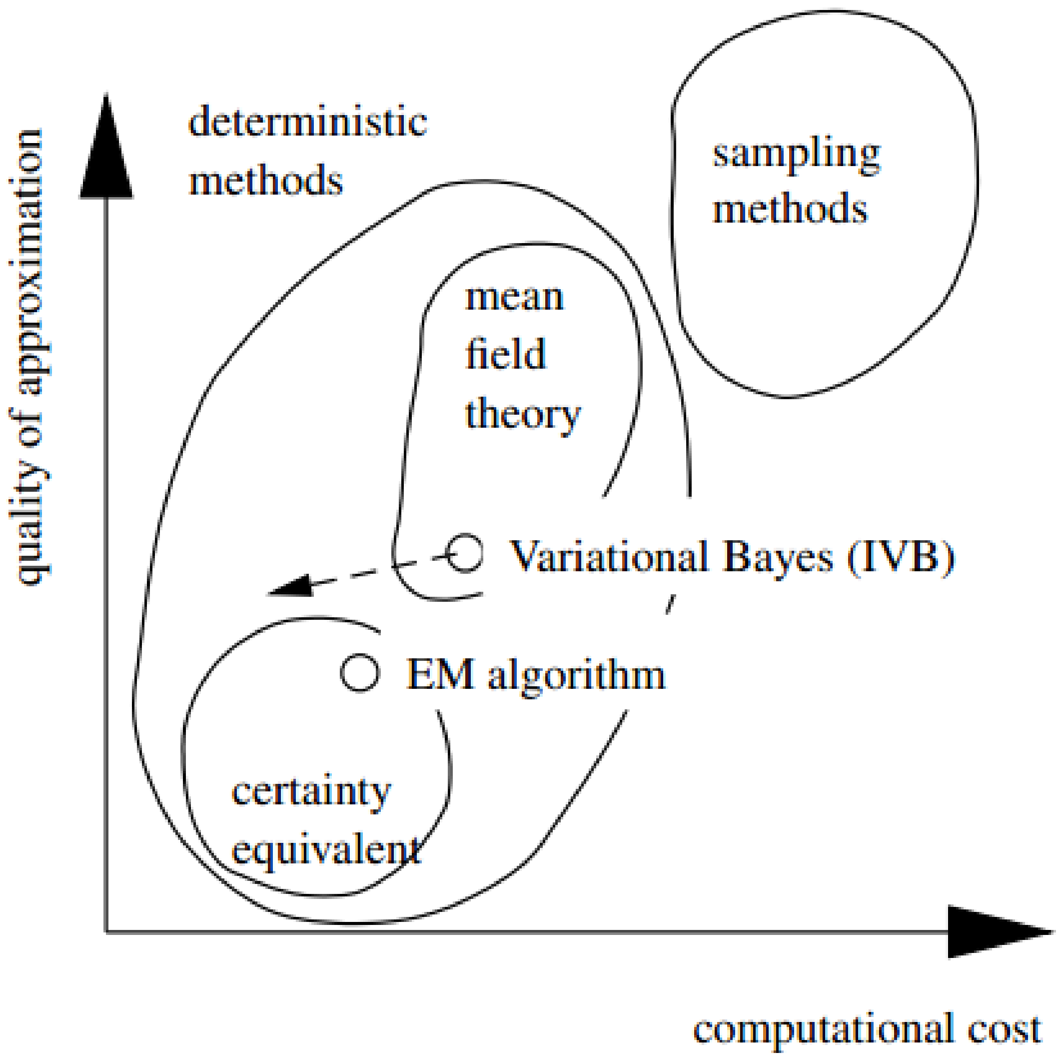
- Motivation
- The Variational Bayesian Framework
  - Variational Free Energy
  - Optimization Tech. Mean Field Approximation
  - Exponential Family
  - Bayesian Networks
- Example:
  - VB for Mixture model
- Discussion
- Application
- Reference

EM vs. VB



constrained E step,
so lower bound
is no longer tight
new log likelihood
$\ln p(\mathbf{y} \mid \boldsymbol{\theta}^{(t+1)})$
$\mathrm{KL}\left[q_{\mathbf{x}}^{(t+1)} \parallel p(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\theta}^{(t+1)})\right]$
new lower bound
$\mathcal{F}(q_{\mathbf{x}}^{(t+1)}, \boldsymbol{\theta}^{(t+1)})$
log likelihood
$\ln p(\mathbf{y} \mid \boldsymbol{\theta}^{(t)})$
$\mathrm{KL}\left[q_{\mathbf{x}}^{(t)} \parallel p(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\theta}^{(t)})\right]$
lower bound
$\mathcal{F}(q_{\mathbf{x}}^{(t)}, \boldsymbol{\theta}^{(t)})$
$\ln p(\mathbf{y} \mid \boldsymbol{\theta}^{(t)})$
$\mathrm{KL}\left[q_{\mathbf{x}}^{(t+1)} \parallel p(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\theta}^{(t)})\right]$
$\mathcal{F}(q_{\mathbf{x}}^{(t+1)}, \boldsymbol{\theta}^{(t)})$
E step
M step
log marginal likelihood
$\ln p(\mathbf{y} \mid m)$
$\mathrm{KL}\left[q_{\mathbf{x}}^{(t)} q_{\boldsymbol{\theta}}^{(t)} \parallel p(\mathbf{x}, \boldsymbol{\theta} \mid \mathbf{y})\right]$
lower bound
$\mathcal{F}(q_{\mathbf{x}}^{(t)}(\mathbf{x}), q_{\boldsymbol{\theta}}^{(t)}(\boldsymbol{\theta}))$
$\ln p(\mathbf{y} \mid m)$
$\mathrm{KL}\left[q_{\mathbf{x}}^{(t+1)} q_{\boldsymbol{\theta}}^{(t)} \parallel p(\mathbf{x}, \boldsymbol{\theta} \mid \mathbf{y})\right]$
new lower bound
$\mathcal{F}(q_{\mathbf{x}}^{(t+1)}(\mathbf{x}), q_{\boldsymbol{\theta}}^{(t)}(\boldsymbol{\theta}))$
$\ln p(\mathbf{y} \mid m)$
$\mathrm{KL}\left[q_{\mathbf{x}}^{(t+1)} q_{\boldsymbol{\theta}}^{(t+1)} \parallel p(\mathbf{x}, \boldsymbol{\theta} \mid \mathbf{y})\right]$
newer lower bound
$\mathcal{F}(q_{\mathbf{x}}^{(t+1)}(\mathbf{x}), q_{\boldsymbol{\theta}}^{(t+1)}(\boldsymbol{\theta}))$
VBE step
VBM step

The Accuracy-vs-Complexity trade-off



Application

- Matrix Factorization: Probabilistic PCA, Mixtures of PPCA, Independent Factor Analysis(IFA), nonlinear ICA/IFA/SSM, Mixture of Bayesian ICA, Bayesian Mixture of Factor Analyzers, etc.
- Time Series: Bayesian HMMs, variational Kalman filtering, Switching State-space models, etc.
- Topic model: Latent Dirichlet Allocation(LDA), (Hierarchical) Dirichlet Process (Mixture) Model, Bayesian Nonparametrical Models, etc.
- Variational Gaussian Process Classifiers
- Sparse Bayesian Learning
- Variational Bayesian Filtering, etc.

Reference

- 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.
- Attias, Hagai. "Inferring parameters and structure of latent variable models by variational Bayes." *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1999.
- Winn, John, Christopher M. Bishop, and Tommi Jaakkola. "Variational Message Passing." *Journal of Machine Learning Research* 6.4 (2005).
- Wainwright, Martin J., and Michael I. Jordan. "Graphical models, exponential families, and variational inference." *Foundations and Trends® in Machine Learning* 1.1-2 (2008): 1-305.
- Šmídl, Václav, and Anthony Quinn. *The variational Bayes method in signal processing*. Springer, 2006.
- Wikipedia, Variational Bayesian methods, http://en.wikipedia.org/wiki/Variational_Bayes

Any Question ?