

An anime-style illustration of a character with dark hair and a red cape, holding a sword. The character is wearing a white shirt and brown mechanical gear. The background is a bright, hazy sky. A semi-transparent orange banner is overlaid across the middle of the image.

An Introduction to Variational Bayesian

Shuang Xu

Copyright © 2018 Shuang Xu

All Rights Reserved.


XU.S@OUTLOOK.COM

Version 20180406, April 6, 2018

Contents

1	变分推断	5
1.1	背景	5
1.2	变分推断	6
1.2.1	变分推断与 MCMC	6
1.2.2	信息传递算法	6
1.2.3	例子: Variational Mixture of Poisson	8
1.2.4	例子: Variational Linear regression	9
1.3	随机变分推断	11
1.3.1	指数分布族	11
1.3.2	条件共轭模型	11
1.3.3	例子: Latent Dirichlet Allocation	13
1.3.4	随机变分推断	14
1.4	变分 EM 算法	16
1.4.1	EM 算法	16
1.4.2	变分 EM 算法	17
1.4.3	例子: Latent Dirichlet Allocation (续)	18
2	非共轭模型	21
2.1	层次模型	22
2.1.1	SMN 分布族	22
2.1.2	其他层次模型	24

2.1.3	层次模型的优缺点	26
2.1.4	例子: Linear regression with Laplace noise	26
2.1.5	例子: Bayesian Lasso / Linear regression with Laplace penalty	27
2.2	构造 ELBO 下界与局部变分法	28
2.2.1	局部变分法	28
2.2.2	Logistic 函数的局部变分	29
2.2.3	例子: Variational Bayesian logistic regression	29
2.3	拉普拉斯变分推断	30
2.3.1	Laplace variational inference	31
2.3.2	LVI 的优缺点	31
2.3.3	例子: Variational Bayesian logistic regression (续)	32
2.4	黑箱变分推断	32
2.4.1	控制方差	33
2.4.2	黑箱变分推断	34
2.4.3	例子: Gaussian-Exponential model	35
3	其他变分推断方法	37
3.1	χ 变分推断	37
3.1.1	χ 散度与证据上界	37
3.1.2	优化 CUBO	38
3.2	变分提升 (Variational Boosting)	38
4	讨论与展望	41
4.1	应用	41
4.1.1	计算生物学	41
4.1.2	计算机视觉	41
4.1.3	低秩模型	42
4.1.4	变量选择	42
4.1.5	聚类分析	43
4.2	开放问题	43
4.3	总结	44



1. 变分推断

1.1 背景

在贝叶斯分析和机器学习中, 很多积分通常难以计算. 一个复杂的统计模型包括观测变量 (数据), 未知参数和局部潜变量. 由于未知参数和局部潜变量都需要推断, 在贝叶斯分析中有时统称它们为潜变量. 贝叶斯模型的一个中心任务是在给定观测数据 \mathbf{X} 的条件下, 计算潜变量 \mathbf{Z} 的后验概率分布 $p(\mathbf{Z} | \mathbf{X})$, 以及计算关于这个概率分布的期望 [1].

对于实际应用中的许多模型来说, 计算后验概率分布或者计算关于这个后验概率分布的期望是不可行的. 这可能是由于潜在空间的维度太大, 以至于无法直接计算, 或者由于后验概率分布的形式特别复杂, 从而期望无法解析地计算. 在连续变量的情形中, 需要求解的积分可能没有解析解, 而空间的维度和被积函数的复杂度可能使得数值积分变得不可行. 对于离散变量, 求边缘概率的过程涉及到对隐含变量的所有可能的情况进行求和. 这个过程虽然原则上总是可以计算的, 但是我们在实际应用中经常发现, 隐含状态的数量可能有指数多个, 从而精确的计算所需的代价过高.

贝叶斯统计学的传统推断方法是马氏链蒙特卡洛 (Markov chain Monte Carlo, MCMC) 采样. MCMC 通过抽取大量样本给出后验分布的数值近似. 越来越多的实践表明 MCMC 的计算代价非常高. 变分贝叶斯 (variational Bayesian, VB) 或变分推断 (variational inference, VI) 把原先的统计推断问题转化为了优化问题, 提供了一个分析的方法来近似潜变量的后验分布, 从而达到了统计推断的目标. 由于计算方便, 适用于大规模数据 (massive data), VI 近年来受到了越来越多的关注.

1.2 变分推断

1.2.1 变分推断与 MCMC

现代统计学的一个核心问题是如何近似一个难以计算的概率分布. 尤其在贝叶斯统计中, 后验分布的计算通常是不可行的. 统计学家提出了 MCMC, 从后验分布抽样, 并使用样本的经验分布近似代替真实分布. MCMC 的变体, 如 Metropolis-Hastings 采样 [2, 3] 和 Gibbs 采样 [4], 已经广泛地应用于贝叶斯统计 [5] 和机器学习 [6] 中. 尽管统计学家已经详细地研究了 MCMC 的理论性质, 但是许多缺点限制了它的应用. (1) 预烧期 (burn-in period): MCMC 收敛之前的样本需要被舍弃, 然而我们并不知道算法何时收敛. 预烧期过长降低了算法效率, 过短则抽取的样本与真实分布相差太大. (2) 自相关性: MCMC 抽取的相邻样本具有自相关性, 虽然有一些技巧可以使其降低, 但是同样面临抽样效率低的风险. (3) 计算速度慢: 对于一个并行系统而言, 理想的规模效应是线性的. 即, 系统增加一倍的计算资源会增加一倍的效率. 但是 MCMC 采集双倍的样本仅以 $\sqrt{2}$ 为因子降低蒙特卡洛标准误差 (Monte Carlo standard error) [7]. 随着计算资源的增加, MCMC 的规模效应会逐渐消失. 因此 MCMC 难以应用于大规模数据中.

作为 MCMC 算法的替代, 变分推断近年来受到了越来越多的关注 [1, 8, 9]. 考虑一个贝叶斯推断问题, 给定观测变量 $\mathbf{x} \in \mathbb{R}^p$ 和潜变量 $\mathbf{z} \in \mathbb{R}^d$, 其联合分布为 $p(\mathbf{z}, \mathbf{x}) = p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$, 目标是计算后验分布 $p(\mathbf{z}|\mathbf{x})$. VI 假设变分分布 $q(\mathbf{z})$ 来自于分布族 \mathcal{Q} , 通过最小化 KL 散度 (Kullback-Leibler divergence) 来近似后验分布分布 $p(\mathbf{z}|\mathbf{x})$

$$q^* = \arg \min_{q(\mathbf{z}) \in \mathcal{Q}} KL(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})) \quad (1.1)$$

VI 把贝叶斯推断问题转化为了一个优化问题. KL 散度的定义如下:

Definition 1.2.1 设概率分布 $p(x)$ 和 $q(x)$ 有相同的支撑集 \mathcal{X} , 则 $q(x)$ 对 $p(x)$ 的 KL 散度定义为

$$KL(q||p) = \int_{x \in \mathcal{X}} q(x) \log \frac{q(x)}{p(x)} = \mathbb{E}_{q(x)} \left[\log \frac{q(x)}{p(x)} \right]. \quad (1.2)$$

R KL 散度衡量了分布之间的相似性. 但是 KL 散度不是对称的, 即一般情况下, $KL(q||p) \neq KL(p||q)$. 若 $KL(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x}))$ 越小, 则变分分布 $q(\mathbf{z})$ 越接近后验分布 $p(\mathbf{z}|\mathbf{x})$.

与 MCMC 算法相比, 许多经验工作表明在大规模数据和复杂模型中 VI 的效率更高 [10]. Ormerod et al. (2017) [11] 提出了基于 VI 的稀疏线性回归, Carbonetto and Stephens (2012) [12] 针对大规模数据提出了基于 VI 的稀疏线性回归和逻辑回归. 他们分别对比了 VI 和 MCMC [13, 14] 的计算速度. 表 1 汇总了部分结果, VI 的速度明显高于 MCMC.

1.2.2 信息传递算法

通常, KL 散度的计算十分复杂. 根据定义, 我们有

$$KL(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})) = \mathbb{E}_{q(\mathbf{z})}[\log q(\mathbf{z})] - \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{x}, \mathbf{z})] + \log p(\mathbf{x}). \quad (1.3)$$

Table 1.1: VI 和 MCMC 的对比.

文献	样本数	变量数	VI 时间	MCMC 时间
[11]	80	41	2.08s	15.87s
[11]	2118	99	6.92s	40.69s
[11]	500	1000	197s	299s
[11]	600	7260	2011s	5327s
[12]	~5000	~400000	in hours	in days

令 $\text{ELBO}(q) = \mathbb{E}_{q(z)}[\log p(\mathbf{x}, z)] - \mathbb{E}_{q(z)}[\log q(z)]$, 则 $\log p(\mathbf{x}) = \text{KL}(q(z) \| p(z|\mathbf{x})) + \text{ELBO}(q)$. 由 KL 散度的非负性质可知, $\log p(\mathbf{x}) \geq \text{ELBO}(q)$, 所以函数 $\text{ELBO}(q)$ 称为 (对数) 证据下界 (Evidence Lower Bound). 当数据给定时, $\log p(\mathbf{x})$ 是一个常数, 所以原来的优化问题 Eq. (1) 等价于如下问题

$$q^* = \arg \min_{q(z) \in \mathcal{Q}} \text{ELBO}(q) = \mathbb{E}_{q(z)}[\log p(\mathbf{x}, z)/q(z)]. \quad (1.4)$$

在经典的 VI 中, 通常假设 \mathcal{Q} 是均场变分族 (mean-field variational family), 潜变量之间相互独立

$$q(z) = \prod_{j=1}^d q_j(z_j). \quad (1.5)$$

所谓的信息传递算法 (message passing algorithm) [1, 10] 通过循环坐标优化, 得到每个潜变量的更新公式. 对于第 k 个潜变量, 把与 z_k 无关的变量记为常数, 我们可以改写目标函数

$$\begin{aligned} \mathbb{E}_{q(z)}[\log(p(\mathbf{x}, z))/(q(z))] &= - \int \prod_{j=1}^d q_j \log \frac{\prod_{j=1}^d q_j}{p} dz \\ &= - \sum_{i=1}^d \int \prod_{j=1}^d q_j \log q_j dz + \int \prod_{j=1}^d q_j \log p dz \\ &= - \int q_k \{ \log q_k - \mathbb{E}_{q(z_{-k})} \log p \} dz_k + \text{constant}. \end{aligned} \quad (1.6)$$

显然, 最优的变分分布是

$$q_k^*(z_k) \propto \exp[\mathbb{E}_{q(z_{-k})} \log p(\mathbf{x}, z)]. \quad (1.7)$$

这里 $\mathbb{E}_{q(z_{-k})}$ 表示除了第 k 个潜变量, 对其他进行期望运算. 信息传递算法迭代更新每个潜变量的变分分布, 保证了 ELBO 单调不减, 直到 ELBO 收敛算法结束. 算法 1 总结了信息传递算法的工作流程.

Algorithm 1 Message passing algorithm for VI**Input:** A model $p(\mathbf{x}, \mathbf{z})$, a dataset \mathbf{X} .**Output:** $q(\mathbf{z}) = \prod_{j=1}^d q_j(z_j)$.

- 1: Initialize variational factors $q(\mathbf{z})$;
- 2: **while** the ELBO has not converged **do**
- 3: **for** $j \in 1, 2, \dots, d$ **do**
- 4: Set $q_j(z_j) \propto \exp[\mathbb{E}_{q(z_{-j})} \log p(\mathbf{x}, \mathbf{z})]$
- 5: **end for**
- 6: Compute $\text{ELBO} = \mathbb{E}[\log p(\mathbf{z}, \mathbf{x})] - \mathbb{E}[\log q(\mathbf{z})]$.
- 7: **end while**

R 由于 ELBO 通常是非凸的, 所以信息传递算法只能保证收敛到局部最小值. 当数据集很大时, 计算整个数据集上的 ELBO 十分耗费资源. 所以, 在实际中可以仅仅在原始数据的子集上计算 ELBO [10]. 但是此时不能保证 ELBO 随着算法的迭代单调不减.

R 公式 (1.7) 说明第 k 个潜变量的变分分布和全条件分布 (complete conditional distribution) $p(z_k | \mathbf{x}, \mathbf{z}_{-k})$ 的形式是一致的. 所以我们可以先求出第 k 个潜变量的全条件分布, 然后对其中无关的变量求期望, 所得结果就是变分分布 [15].

1.2.3 例子: Variational Mixture of Poisson

考虑一个混合泊松分布 (Mixture of Poisson, MoP) 的贝叶斯模型. 每个观测 x_n 加入一个 K 元的潜变量 $\mathbf{z}_n = (z_{n1}, z_{n2}, \dots, z_{nK}) \in \{0, 1\}^K$, 满足 $\sum_{k=1}^K z_{nk} = 1$, 服从多项分布 $p(\mathbf{z}_n | \boldsymbol{\pi}) = \prod_{k=1}^K \pi_k^{z_{nk}}$, 超参数 $\boldsymbol{\pi}$ 服从 Dirichlet 分布 $p(\boldsymbol{\pi}) = \text{Dirichlet}(\boldsymbol{\pi} | \boldsymbol{\alpha}_0)$. 其中 K 个泊松成分的参数为 $\boldsymbol{\lambda} = \{\lambda_1, \lambda_2, \dots, \lambda_K\}$, 参数的共轭先验为 Gamma 分布 $p(\lambda_k) = \text{Gamma}(a_0, b_0)$. 所以联合分布为

$$p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\lambda}) = p(\boldsymbol{\pi}) \prod_{k=1}^K p(\lambda_k) \prod_{n=1}^N p(x_n | \mathbf{z}_n, \boldsymbol{\pi}, \boldsymbol{\lambda}) p(\mathbf{z}_n | \boldsymbol{\pi}). \quad (1.8)$$

我们利用 Eq. (1.7) 对所有未知变量进行推断. 推断 z_{nk} :

$$\begin{aligned} \log q(z_{nk}) &= \mathbb{E}_{q(\boldsymbol{\pi}, \boldsymbol{\lambda})} [\log p(x_n | \mathbf{z}_n, \boldsymbol{\pi}, \boldsymbol{\lambda}) + \log p(\mathbf{z}_n | \boldsymbol{\pi})] + \text{constant} \\ &= \mathbb{E}_{q(\boldsymbol{\pi}, \boldsymbol{\lambda})} \{z_{nk} [\log \text{Poi}(x_n | \lambda_k) + \log \pi_k]\} + \text{constant} \\ &= z_{nk} \{x_n \mathbb{E}[\log \lambda_k] - \mathbb{E}[\lambda_k] + \log[x_n!] + \mathbb{E}[\log \pi_k]\} + \text{constant}. \end{aligned} \quad (1.9)$$

推断 $\boldsymbol{\pi}$:

$$\begin{aligned} \log q(\boldsymbol{\pi}) &= \mathbb{E}_{q(\mathbf{Z}, \boldsymbol{\lambda})} [\log p(\mathbf{Z} | \boldsymbol{\pi}) + \log p(\boldsymbol{\pi})] + \text{constant} \\ &= \mathbb{E}_{q(\mathbf{Z}, \boldsymbol{\lambda})} \left\{ \sum_{k=1}^K \log \pi_k \left[\sum_{n=1}^N z_{nk} + \alpha_0 - 1 \right] \right\} + \text{constant} \\ &= \sum_{k=1}^K \log \pi_k \left(\sum_{n=1}^N \mathbb{E}[z_{nk}] + \alpha_0 - 1 \right) + \text{constant}. \end{aligned} \quad (1.10)$$

推断 λ_k :

$$\begin{aligned}
 \log q(\lambda_k) &= \mathbb{E}_{q(\mathbf{Z}, \boldsymbol{\pi})} \left\{ \log(\lambda_k) + \sum_{n=1}^N z_{nk} \text{Poi}(x_n | \lambda_k) \right\} + \text{constant} \\
 &= (a_0 - 1) \log \lambda_k - b_0 \lambda_k + \sum_{n=1}^N \mathbb{E}_{q(\mathbf{Z})}[z_{nk}] (x_n \log \lambda_k - \lambda_k) + \text{constant} \\
 &= \left(a_0 + \sum_{n=1}^N \mathbb{E}_{q(\mathbf{Z})}[z_{nk}] x_n \right) \log \lambda_k - \left(b_0 + \sum_{n=1}^N \mathbb{E}_{q(\mathbf{Z})}[z_{nk}] \right) \lambda_k + \text{constant}.
 \end{aligned} \tag{1.11}$$

显然, $z_n, \boldsymbol{\pi}, \lambda_k$ 的变分分布与先验分布或条件分布一致, 我们记为

$$q(\mathbf{Z}) = \prod_{n=1}^N \prod_{k=1}^K \gamma_{nk}^{z_{nk}}, \quad q(\boldsymbol{\pi}) = \text{Dirichlet}(\boldsymbol{\pi} | \boldsymbol{\alpha}), \quad q(\lambda_k) = \text{Gamma}(\lambda_k | a_k, b_k), \tag{1.12}$$

其中

$$\begin{aligned}
 \gamma_{nk} &= \frac{\beta_k}{\sum_{k=1}^K \beta_k}, \\
 \log \beta_{nk} &= x_n \mathbb{E}[\log \lambda_k] - \mathbb{E}[\lambda_k] + \log[x_n!] + \mathbb{E}[\log \pi_k], \\
 &= x_n [\psi(a_k) - \log b_k] - a_k/b_k + \log[x_n!] + [\psi(\alpha_k) - \psi(\sum_k \alpha_k)], \\
 \alpha_k &= \sum_{n=1}^N \mathbb{E}[z_{nk}] + \alpha_0 = \sum_{n=1}^N \gamma_{nk} + \alpha_0, \\
 a_k &= a_0 + \sum_{n=1}^N \mathbb{E}[z_{nk}] x_n = a_0 + \sum_{n=1}^N \gamma_{nk} x_n, \\
 b_k &= b_0 + \sum_{n=1}^N \mathbb{E}[z_{nk}] = b_0 + \sum_{n=1}^N \gamma_{nk}.
 \end{aligned} \tag{1.13}$$

其中 $\psi(\cdot)$ 表示 digamma 函数, 即对数 Gamma 函数的一阶导数¹. 上式中涉及阶乘运算, 当 x_n 过大时, 在计算机中常常会出现上溢现象. 但是根据斯特林公式 $n! \sim \sqrt{2\pi n} \left(\frac{n}{e}\right)^n$. 所以 x_n 充分大时 (比如, >50), 我们可以计算近似得到阶乘的对数 $\log[x_n!] = \frac{1}{2} \log 2\pi x_n + x_n \log \frac{x_n}{e}$.

1.2.4 例子: Variational Linear regression

在这个例子中, 我们考虑变分贝叶斯线性回归 [16]. 假设 $\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon}$, 其中

$$\begin{aligned}
 p(\mathbf{y} | \mathbf{w}) &= \prod_{n=1}^N \mathcal{N}(y_n | \mathbf{x}_n^T \mathbf{w}, \beta^{-1}) \\
 p(\mathbf{w} | \mathbf{T}) &= \mathcal{N}(\mathbf{w} | \mathbf{0}, \mathbf{T}^{-1}) \\
 p(\boldsymbol{\beta}) &= \Gamma(\boldsymbol{\beta} | a_0, b_0) \\
 p(t_i) &= \Gamma(t_i | c_0, d_0)
 \end{aligned} \tag{1.14}$$

这里 \mathbf{x}_n 表示第 n 个观测值, $\mathbf{T} = \text{diag}(t_1, \dots, t_p)$. 变量的联合分布为

$$p(\mathbf{y}, \mathbf{w}, \mathbf{T}, \boldsymbol{\beta}) = p(\mathbf{y} | \mathbf{w}, \boldsymbol{\beta}) p(\mathbf{w} | \mathbf{T}) p(\boldsymbol{\beta}) \prod_{j=1}^p p(t_j). \tag{1.15}$$

¹https://en.wikipedia.org/wiki/Digamma_function

利用更新公式 Eq. (1.7) 计算变分分布. 在下面的书写中, 为了方便, 我们省略期望符号的下标. 推断 \mathbf{w} :

$$\begin{aligned}
 \log q(\mathbf{w}) &= \mathbb{E} \{ \log p(\mathbf{y}|\mathbf{w}, \beta) + \log p(\mathbf{w}|\mathbf{T}) \} \\
 &= \mathbb{E} \left\{ -\frac{\beta}{2} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 - \frac{1}{2} \mathbf{w}^\top \mathbf{T} \mathbf{w} \right\} + constant \\
 &= \mathbb{E} \left\{ -\frac{\beta}{2} (-2\mathbf{y}^\top \mathbf{X}\mathbf{w} + \mathbf{w}^\top \mathbf{X}^\top \mathbf{X} \mathbf{w}) - \frac{1}{2} \mathbf{w}^\top \mathbf{T} \mathbf{w} \right\} + constant \\
 &= \mathbb{E} \left\{ \beta \mathbf{y}^\top \mathbf{X} \mathbf{w} - \frac{1}{2} \mathbf{w}^\top (\beta \mathbf{X}^\top \mathbf{X} + \mathbf{T}) \mathbf{w} \right\} + constant \\
 &= \mathbb{E}[\beta] \mathbf{y}^\top \mathbf{X} \mathbf{w} - \frac{1}{2} \mathbf{w}^\top (\mathbb{E}[\beta] \mathbf{X}^\top \mathbf{X} + \mathbb{E}[\mathbf{T}]) \mathbf{w} + constant.
 \end{aligned} \tag{1.16}$$

推断 β :

$$\begin{aligned}
 \log q(\beta) &= \mathbb{E} \{ \log p(\mathbf{y}|\mathbf{w}, \beta) + \log p(\beta) \} \\
 &= \mathbb{E} \left\{ \frac{N}{2} \log \beta - \frac{\beta}{2} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + a_0 \log \beta - b_0 \beta \right\} + constant \\
 &= \mathbb{E} \left\{ \left(\frac{N}{2} + a_0 \right) \log \beta - \left(\frac{1}{2} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + b_0 \right) \beta \right\} + constant \\
 &= \left(\frac{N}{2} + a_0 \right) \log \beta - \left(\frac{1}{2} \mathbb{E} [\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2] + b_0 \right) \beta + constant.
 \end{aligned} \tag{1.17}$$

推断 t_j :

$$\begin{aligned}
 \log q(t_j) &= \mathbb{E} \{ \log p(w_j|t_j) + \log p(t_j) \} \\
 &= \mathbb{E} \left\{ \frac{1}{2} \log t_j - \frac{w_j^2}{2} t_j + c_0 \log t_j - d_0 t_j \right\} + constant \\
 &= \mathbb{E} \left\{ \left(\frac{1}{2} + c_0 \right) \log t_j - \left(\frac{w_j^2}{2} + d_0 \right) t_j \right\} + constant \\
 &= \left(\frac{1}{2} + c_0 \right) \log t_j - \left(\frac{\mathbb{E}[w_j^2]}{2} + d_0 \right) t_j + constant.
 \end{aligned} \tag{1.18}$$

所以 \mathbf{w}, β, t_j 的变分分布分别为高斯, Gamma, Gamma. 我们记为

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w} | \mathbf{m}, \mathbf{S}) \quad q(\beta) = \text{Gamma}(\beta | a, b) \quad q(t_j) = \text{Gamma}(t_j | c_j, d_j) \tag{1.19}$$

其中

$$\begin{aligned}
 S^{-1} &= \mathbb{E}[\beta] \mathbf{X}^\top \mathbf{X} + \mathbb{E}[\mathbf{T}] = \frac{a}{b} \mathbf{X}^\top \mathbf{X} + \text{diag} \left(\frac{c_j}{d_j} \right), \\
 \mathbf{m} &= \mathbb{E}[\beta] \mathbf{S} \mathbf{X}^\top \mathbf{y} = \frac{a}{b} \mathbf{S} \mathbf{X}^\top \mathbf{y}, \\
 a &= \frac{N}{2} + a_0, \\
 b &= \frac{1}{2} \mathbb{E} [\|\mathbf{y} - \mathbf{X} \mathbf{w}\|_2^2] + b_0, \\
 c_j &= \frac{1}{2} + c_0, \\
 d_j &= \frac{\mathbb{E}[w_j^2]}{2} + d_0.
 \end{aligned} \tag{1.20}$$

关于两个期望的求解:

1. 在求 w_j^2 的期望时, 我们知道 $\mathbb{E}[\mathbf{w} \mathbf{w}^\top] = \mathbf{m} \mathbf{m}^\top + \mathbf{S}$. 所以 $\mathbb{E}[w_j^2] = m_{jj}^2 + S_{jj}$;
2. $\mathbb{E} [\|\mathbf{y} - \mathbf{X} \mathbf{w}\|_2^2] = \mathbb{E} [\mathbf{y}^\top \mathbf{y} - 2 \mathbf{y}^\top \mathbf{X} \mathbf{w} + \mathbf{w}^\top \mathbf{X}^\top \mathbf{X} \mathbf{w}] = \mathbf{y}^\top \mathbf{y} - 2 \mathbf{y}^\top \mathbf{X} \mathbf{m} + \text{tr} [\mathbf{X}^\top \mathbf{X} (\mathbf{m} \mathbf{m}^\top + \mathbf{S})]$.

1.3 随机变分推断

1.3.1 指数分布族

指数分布族包括了一族十分常用的分布, 如: 高斯分布、指数分布、泊松分布等. 得益于指数分布族的优良性质, 如果潜变量来自其中, 则变分分布的推导会得到极大地简化 [1].

Definition 1.3.1 指数分布族有如下形式的分布函数

$$f(\mathbf{x} | \boldsymbol{\theta}) = h(\mathbf{x}) \exp [\boldsymbol{\eta}(\boldsymbol{\theta})^\top \cdot \mathbf{t}(\mathbf{x}) - a(\boldsymbol{\eta}(\boldsymbol{\theta}))], \tag{1.21}$$

其中 $h(\mathbf{x})$ 为基测函数 (base measure function), $a(\cdot)$ 为对数归一化函数 (log-normalizer), $\boldsymbol{\eta}(\boldsymbol{\theta})$ 为自然参数, $\mathbf{t}(\mathbf{x})$ 为充分统计量.

■ **Example 1.1** 泊松分布

$$f(x | \lambda) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{1}{x!} e^{-\lambda} e^{x \log \lambda}, \tag{1.22}$$

base measure 为 $h = 1/x!$, 对数归一化函数为 $a = -\lambda$, 自然参数为 $\log \lambda$, 充分统计量为 x . 若令 $\boldsymbol{\eta} = \log \lambda$, 则泊松分布可以改写为

$$f(x | \boldsymbol{\eta}) = \frac{1}{x!} e^{-\exp(\boldsymbol{\eta})} e^{x \boldsymbol{\eta}}. \tag{1.23}$$

■

1.3.2 条件共轭模型

回到贝叶斯模型, 在之前的推导中我们不加区分地把所有地未知变量都记为了潜变量. 现在规定全局潜变量 (global latent variable) 为 $\boldsymbol{\beta}$, 局部潜变量为 $\mathbf{Z} = \{\mathbf{z}_n\}_{n=1}^N$, 数据记为

$\mathbf{X} = \{\mathbf{x}_n\}_{n=1}^N$. 例如在混合泊松模型中, λ 为全局潜变量, \mathbf{z}_n 为局部潜变量. 此外我们令 β 的超参数为 α . 现在的联合分布为

$$p(\mathbf{X}, \mathbf{Z}, \beta | \alpha) = p(\beta | \alpha) \prod_{n=1}^N p(\mathbf{x}_n, \mathbf{z}_n | \beta). \quad (1.24)$$

假设潜变量和观测变量的联合分布来自指数分布族, 自然参数为 η , 则

$$p(\mathbf{X}, \mathbf{Z} | \eta) = \prod_{n=1}^N h(\mathbf{z}_n, \mathbf{x}_n) g(\eta) \exp[\eta^\top t(\mathbf{x}_n, \mathbf{z}_n)]. \quad (1.25)$$

Definition 1.3.2 条件共轭模型 (conditional conjugate model) 包括了一类广泛的概率模型. 它假设:

(1) 潜变量的后验分布在指数族中

$$p(\beta | \mathbf{X}, \mathbf{Z}, \alpha) = h(\beta) \exp \{ \eta_g(\mathbf{X}, \mathbf{Z}, \alpha)^\top t(\beta) - a_g(\eta_g(\mathbf{X}, \mathbf{Z}, \alpha)) \}, \quad (1.26)$$

$$p(z_{nj} | \mathbf{x}_n, \mathbf{z}_{n,-j}, \beta) = h(z_{nj}) \exp \{ \eta_l(\mathbf{x}_n, \mathbf{z}_{n,-j}, \beta)^\top t(z_{nj}) - a_l(\eta_l(\mathbf{x}_n, \mathbf{z}_{n,-j}, \beta)) \}, \quad (1.27)$$

这里 $\mathbf{z}_{n,-j}$ 表示删除了第 j 个元素的 \mathbf{z}_n .

(2) 给定全局变量后, 局部变量的分布在如下的指数族中

$$p(\mathbf{x}_n, \mathbf{z}_n | \beta) = h(\mathbf{x}_n, \mathbf{z}_n) \exp \{ \beta^\top t(\mathbf{x}_n, \mathbf{z}_n) - a_l(\beta) \}. \quad (1.28)$$

(3) 全局变量在指数族中

$$p(\beta) = h(\beta) \exp \{ \alpha^\top t(\beta) - a_g(\alpha) \} \quad (1.29)$$

充分统计量为 $t(\beta) = (\beta, -a_l(\beta))$, 超参数 α 可以划分成两部分 $\alpha = (\alpha_1, \alpha_2)$, 第一个和 β 的维数相同, 第二个是标量.

把 Eq. (1.28), (1.29) 带入 Eq. (1.26) 中可知

$$\eta_g(\mathbf{X}, \mathbf{Z}, \alpha) = \left(\alpha_1 + \sum_{n=1}^N t(\mathbf{z}_n, \mathbf{x}_n), \alpha_2 + N \right). \quad (1.30)$$

根据信息传递算法的更新公式, 我们可以计算潜变量的变分分布. 实际上, 由于条件共轭模型的假设, 变分分布和条件分布有相同的形式. 不妨记 β, \mathbf{z}_{nj} 的变分参数分别为 λ, ϕ_{nj} , 则

$$q(\beta | \lambda) = h(\beta) \exp \{ \lambda^\top t(\beta) - a_g(\lambda) \}, \quad (1.31)$$

$$q(z_{nj} | \phi_{nj}) = h(z_{nj}) \exp \{ \phi_{nj}^\top t(z_{nj}) - a_l(\phi_{nj}) \}, \quad (1.32)$$

其中

$$\lambda = \mathbb{E}_q[\eta_g(\mathbf{X}, \mathbf{Z}, \boldsymbol{\alpha})], \quad (1.33)$$

$$\phi_{nj} = \mathbb{E}_q[\eta_l(\mathbf{x}_n, \mathbf{z}_{n,-j}, \boldsymbol{\beta})]. \quad (1.34)$$

条件共轭模型的好处是, 我们能得到变分分布的解析解. 这是大多数非条件共轭模型做不到的.

1.3.3 例子: Latent Dirichlet Allocation

LDA 模型

文本聚类的一个常用算法是 Latent Dirichlet Allocation (LDA) [17]. 首先约定一些符号:

- 观测数据是文章中的词, 记第 d 篇文章的第 n 个词是 w_{dn} . 文章中的每个词都来自于一个固定的词汇表, 该词汇表中共有 V 个单词. 我们用 one-hot 对 w_{dn} 进行编码, 即 $w_{dn}^v = 1$ 表示第 d 篇文章的第 n 个词是词汇表的第 v 个单词. 同时, $w_{dn} = v$ 也表示 $w_{dn}^v = 1$.
- 单词比例变量 $\beta_k = (\beta_{k1}, \beta_{k2}, \dots, \beta_{kV})$ 表示定义在词汇表上的分布, 满足 $\sum_{v=1}^V \beta_{kv} = 1$.
- 话题比例变量 $\theta_d = (\theta_{d1}, \theta_{d2}, \dots, \theta_{dK})$ 表示定义在话题上的分布, 满足 $\sum_{k=1}^K \theta_{dk} = 1$.
- 假设每篇文章的每个单词来自于一个话题中. 用 z_{dn} 表示单词 w_{dn} 来自的话题. 类似地, $z_{dn}^k = 1$ 表示第 d 篇文章的第 n 个词来自第 k 个话题. $z_{dn} = k$ 表示 $z_{dn}^k = 1$.

LDA 认为一篇文章 (document) 中隐含了几个主题 (topic), 而一个主题常常伴随了一些高频词汇 (word). 所以, 为了生成一篇文章, 我们需要做的是:

1. 确定单词对话题 k 的归属度: 对单词比例进行抽样 $\beta_k \sim \text{Dirichlet}(\eta, \dots, \eta), (k = 1, 2, \dots, K)$.
2. 对每篇文章 $d = 1, 2, \dots, D$
 - (a) 确定一片文章的话题比例: 对话题比例进行抽样 $\theta_d \sim \text{Dirichlet}(\alpha, \dots, \alpha)$.
 - (b) 对每个单词 $w = 1, 2, \dots, N$
 - i. 确定这个位置的话题: $z_{dn} \sim \text{Multinomial}(\theta_d)$.
 - ii. 对该话题的单词进行抽样: $w_{dn} \sim \text{Multinomial}(\beta_{z_{dn}})$.

所以我们的概率模型为

$$\begin{aligned} W | Z, \beta &\sim \prod_{n=1}^N \prod_{d=1}^D \prod_{v=1}^V \beta_{kv}^{z_{dn}^k w_{dn}^v} \\ Z | \theta &\sim \prod_{d=1}^D \prod_{n=1}^N \prod_{k=1}^K \theta_{dk}^{z_{dn}^k} \\ \beta &\sim \prod_{k=1}^K \text{Dirichlet}(\eta_0) \\ \theta &\sim \prod_{d=1}^D \text{Dirichlet}(\alpha_0) \end{aligned} \quad (1.35)$$

这里的局部变量有话题比例 θ_d 和话题分配 z_{dn} ; 全局变量为 β_k . 下面对 LDA 模型进行变分推断 [15]. 为了方便叙述, 我们先计算待估参数的全条件分布.

(1)

$$\begin{aligned}
 p(z_{dn} = k \mid \theta_d, \beta, w_{dn}) &\propto p(z_{dn} = k, \theta_d, \beta, w_{dn}) \\
 &\propto p(w_{dn} \mid \beta, z_{dn} = k) p(z_{dn} = k \mid \theta_d) \\
 &= \exp \left(\sum_{v=1}^V w_{dn}^v \log \beta_{kv} + \log \theta_{dk} \right) \\
 &= \exp (\log \beta_{k, w_{dn}} + \log \theta_{dk})
 \end{aligned} \tag{1.36}$$

(2)

$$\begin{aligned}
 p(\theta_{dk} \mid z_d) &\propto p(z_d, \theta_{dk}) \\
 &= p(z_d \mid \theta_{dk}) p(\theta_{dk}) \\
 &= \exp \left(\sum_{n=1}^N z_{dn}^k \log \theta_{dk} + (\alpha_0 - 1) \log \theta_{dk} \right) \\
 &= \exp \left[\left(\alpha_0 + \sum_{n=1}^N z_{dn}^k - 1 \right) \log \theta_{dk} \right]
 \end{aligned} \tag{1.37}$$

(3)

$$\begin{aligned}
 p(\beta_{kv} \mid \mathbf{Z}, \mathbf{W}) &\propto p(\mathbf{Z}, \mathbf{W}, \beta_{kv}) \\
 &= p(z_d \mid \theta_{dk}) p(\beta_{kv}) \\
 &= \exp \left(\sum_{n=1}^N \sum_{d=1}^D z_{dn}^k w_{dn}^v \log \beta_{kv} + (\eta_0 - 1) \log \beta_{kv} \right) \\
 &= \exp \left[\left(\eta_0 + \sum_{n=1}^N \sum_{d=1}^D z_{dn}^k w_{dn}^v - 1 \right) \log \theta_{dk} \right]
 \end{aligned} \tag{1.38}$$

我们分别记潜变量的变分分布为 $q(z_{dn}) = \text{Multinomial}(\phi_{dn})$, $q(\theta_d) = \text{Dirichlet}(\alpha_d)$, $q(\beta_k) = \text{Dirichlet}(\eta_k)$, 其中

$$\hat{\phi}_{dn}^k = \mathbb{E}[\log \beta_{k, w_{dn}} + \log \theta_{dk}], \phi_{dn}^k = \hat{\phi}_{dn}^k / \sum_{k=1}^K \hat{\phi}_{dn}^k \tag{1.39}$$

$$\eta_k = \eta_0 + \sum_{n=1}^N \sum_{d=1}^D \phi_{dn}^k w_{dn}, \alpha_d = \alpha_0 + \sum_{n=1}^N \phi_{dn} \tag{1.40}$$

1.3.4 随机变分推断

上节中介绍了经典的变分推断. VI 在条件共轭模型中有非常优美的结果, 我们只需要根据 Eq. (9-12) 对参数进行更新就可以得到后验分布的近似. 然而, 经典 VI 的一次迭代是对整个数据集进行操作. 这个计算过程难以应对大规模数据. Hoffman et al. 结合了随机

梯度优化 (stochastic gradient optimization) 的思想提出了随机变分推断 (stochastic variational inference, StoVI) [15].

沿用 2.4 节的记号. 易得 ELBO 关于 λ 的欧式梯度为 $\nabla_{\lambda} \text{ELBO} = a_g''(\lambda)(\mathbb{E}_{\phi}[\eta_g(\mathbf{X}, \mathbf{Z}, \alpha)] - \lambda)$. 进而可得自然梯度 (natural gradient) $g(\lambda)$,

$$g(\lambda) = \mathbb{E}_{\phi}[\eta_g(\mathbf{X}, \mathbf{Z}, \alpha)] - \lambda. \quad (1.41)$$

引入自然梯度的原因是: 自然梯度可以解释概率参数的几何结构, 即参数在自然梯度方向上移动可以解释对应了 KL 散度的变化, 但是欧式梯度没有这个性质 [18]. 随机梯度优化在一次更新中只抽取一个样本, 不妨记为第 n 个样本, 所以全局参数的随机梯度优化的更新公式为

$$\lambda_t = \lambda_{t-1} + e_t g(\lambda_t) = (1 - e_t) \lambda_{t-1} + e_t \mathbb{E}_{\phi}[\eta_g(\mathbf{x}_n^{(N)}, \mathbf{z}_n^{(N)}, \alpha)]. \quad (1.42)$$

其中 e_t 为学习率, 且满足 $\sum e_t = \infty, \sum e_t^2 < \infty$, 比如可令 $e_t = t^{-k}, k \in (0.5, 1]$. $\mathbf{x}_n^{(N)}, \mathbf{z}_n^{(N)}$ 表示 N 个 \mathbf{x}_n 组成的数据集, 故

$$\mathbb{E}_{\phi}[\eta_g(\mathbf{x}_n^{(N)}, \mathbf{z}_n^{(N)}, \alpha)] = (\alpha_1 + N \mathbb{E}[t(\mathbf{z}_n, \mathbf{x}_n)], \alpha_2 + N). \quad (1.43)$$

算法 2 为 StoVI 的工作流程.

Algorithm 2 Stochastic Variational Inference.

Input: Model $p(\mathbf{x}, \mathbf{z})$, data \mathbf{X} and step size sequence e_t .

Output: Global variational densities $q(\beta; \lambda)$.

- 1: Initialize variational parameter λ_0 ;
 - 2: **while** the ELBO has not converged **do**
 - 3: Uniformly sample a data point \mathbf{x}_n from \mathbf{X} .
 - 4: Compute its local variational parameters $\phi_n = \{\phi_{nj}\}_{j=1}^d$, where $\phi_{nj} = \mathbb{E}_q[\eta_l(\mathbf{x}_n, \mathbf{z}_{n,-j}, \beta)]$.
 - 5: Compute the update as though \mathbf{x}_n were repeated N times,

$$\mathbb{E}_{\phi}[\eta_g(\mathbf{x}_n^{(N)}, \mathbf{z}_n^{(N)}, \alpha)] = (\alpha_1 + N \mathbb{E}[t(\mathbf{z}_n, \mathbf{x}_n)], \alpha_2 + N).$$
 - 6: Update global variational parameter, $(1 - e_t) \lambda_{t-1} + e_t \mathbb{E}_{\phi}[\eta_g(\mathbf{x}_n^{(N)}, \mathbf{z}_n^{(N)}, \alpha)]$.
 - 7: **end while**
-

LDA 模型可以用 StoVI 解决. LDA 的局部潜变量为 θ_d, \mathbf{z}_d , 全局参数为 β . 所以每次更新 β 时, 先从 D 篇文章中抽取一篇, 更新它的潜变量分布, 然后更新全局参数的变分分布.

Algorithm 3 Stochastic Variational Inference for LDA.

```

1: Initialize variational parameter  $\eta_0$ ;
2: while not converged do
3:   Uniformly sample a document  $w_n$ .
4:   Initialize  $\alpha_{dk} = 1$ , for  $k \in \{1, 2, \dots, K\}$ .
5:   repeat
6:     Update  $\phi_{dn}^k$  for  $n \in \{1, 2, \dots, N\}, k \in \{1, 2, \dots, K\}$ .
7:     Update  $\alpha_d = \alpha_0 + \sum_{n=1}^N \phi_{dn}$ .
8:   until  $\phi_{dn}^k$  and  $\eta_d$  converge.
9:   Set  $\hat{\eta}_k = \eta_0 + D \sum_{n=1}^N \phi_{dn}^k w_{dn}$  for  $k \in \{1, 2, \dots, K\}$ .
10:  Set  $\eta^{(t)} = (1 - e_t) \eta^{(t-1)} + e_t \hat{\eta}$ 
11: end while

```

1.4 变分 EM 算法

EM 算法 (expectation-maximization algorithm) [19] 是统计优化的重要工具, 适合处理含有潜变量的概率模型. 这里考虑一般的概率模型, \mathbf{X}, \mathbf{Z} 分别表示数据和潜变量, θ 表示参数, 则似然函数为

$$l(\theta; \mathbf{X}) = p(\mathbf{X} | \theta) = \int p(\mathbf{X}, \mathbf{Z} | \theta) d\mathbf{Z}. \quad (1.44)$$

根据最大似然法则, 我们的目标是 $\max_{\mathbf{Z}, \theta} l(\theta; \mathbf{X})$. EM 算法通过迭代 E 步和 M 步来优化似然函数. 其中 E 步固定参数 θ , 优化潜变量分布; M 步固定潜变量分布, 优化参数. 例如在混合高斯模型中, 潜变量 z_n 表示第 n 个样本来自哪个高斯成分. E 步实际上在计算第 n 个样本属于各个高斯成分的后验概率, 而 M 步在优化每个高斯成分的均值和方差. 但是, 如果我们对潜变量分布加入约束, 或者潜变量分布难以处理时, E 步通常是无法达到最优的. 变分 EM 算法在 EM 算法的框架下加入了变分的思想, 经常被用来处理这种情况. 在介绍变分 EM 算法之前, 我们先回顾 EM 算法.

1.4.1 EM 算法

EM 算法的目标是极大化 (对数) 似然函数, 令 \mathbf{Z} 的分布是 $q(\mathbf{Z})$, 即满足 $\int q(\mathbf{Z}) d\mathbf{Z} = 1$, 则

$$\begin{aligned}
\log p(\mathbf{X} | \theta) &= \int q(\mathbf{Z}) \log p(\mathbf{X} | \theta) d\mathbf{Z} \\
&= \mathbb{E}_q \left[\log \frac{p(\mathbf{X}, \mathbf{Z} | \theta) q(\mathbf{Z})}{p(\mathbf{Z} | \mathbf{X}, \theta) q(\mathbf{Z})} \right] \\
&= \mathbb{E}_q [\log p(\mathbf{X}, \mathbf{Z} | \theta)] - \mathbb{E}_q [\log q(\mathbf{Z})] + \mathbb{E}_q \left[\log \frac{q(\mathbf{Z})}{p(\mathbf{Z} | \mathbf{X}, \theta)} \right] \\
&= \mathbb{E}_q [\log p(\mathbf{X}, \mathbf{Z} | \theta)] + \mathcal{H}(q(\mathbf{Z})) + KL(q(\mathbf{Z}) || p(\mathbf{Z} | \mathbf{X}, \theta))
\end{aligned} \quad (1.45)$$

由 KL 散度的非负性, 可得似然的下界

$$\log p(\mathbf{X} | \theta) \geq \mathbb{E}_q [\log p(\mathbf{X}, \mathbf{Z} | \theta)] + \mathcal{H}(q(\mathbf{Z})) \equiv \mathcal{F}(q, \theta) \quad (1.46)$$

我们称这个下界为自由能. 优化似然函数相当于优化自由能.

自由能包含了潜分布和参数两部分, 所以可以使用交替优化. (E 步:) 若给定第 t 步的参数估计 θ^t , 更新分布 $q^{t+1}(\mathbf{Z}) = \arg \max_q \mathcal{F}(q, \theta^t)$. 由公式 (1.45) 可知,

$$\mathcal{F}(q, \theta^t) = \log p(\mathbf{X} | \theta^t) - KL(q(\mathbf{Z}) || p(\mathbf{Z} | \mathbf{X}, \theta^t)). \quad (1.47)$$

所以 E 步相当于 $\min_q KL(q(\mathbf{Z}) || p(\mathbf{Z} | \mathbf{X}, \theta^t))$, 即 $q^{t+1} = p(\mathbf{Z} | \mathbf{X}, \theta^t)$. (M 步:) 然后固定潜分布, 更新参数

$$\begin{aligned} \theta^{t+1} &= \arg \max_{\theta} \mathcal{F}(q^{t+1}, \theta) = \arg \max_{\theta} \mathbb{E}_q [\log p(\mathbf{X}, \mathbf{Z} | \theta^{t+1})] + \mathcal{H}(q^{t+1}) \\ &= \arg \max_{\theta} \mathbb{E}_{p(\mathbf{Z} | \mathbf{X}, \theta^t)} [\log p(\mathbf{X}, \mathbf{Z} | \theta^{t+1})]. \end{aligned} \quad (1.48)$$

Algorithm 4 EM algorithm.

- 1: Initialize parameter θ^0 .
 - 2: **while** Likelihood has not converged **do**
 - 3: (E-step) Update posterior for latent variable, $q^{t+1} = p(\mathbf{Z} | \mathbf{X}, \theta^t)$.
 - 4: (M-step) Update parameter $\theta^{t+1} = \arg \max_{\theta} \mathbb{E}_{p(\mathbf{Z} | \mathbf{X}, \theta^t)} [\log p(\mathbf{X}, \mathbf{Z} | \theta^{t+1})]$
 - 5: **end while**
-

1.4.2 变分 EM 算法

在上一节中, 我们已经证明了 E 步 $\min_q KL(q(\mathbf{Z}) || p(\mathbf{Z} | \mathbf{X}, \theta^t))$ 的最优解为

$$q^{t+1} = p(\mathbf{Z} | \mathbf{X}, \theta^t). \quad (1.49)$$

然而, 我们需要注意, 这是潜变量的全后验分布 (full posterior). 在实际中, 后验分布极可能难以计算. 所以根据第 2 节中的变分思想 [20]², 我们可以在均场分布族中寻找潜变量的变分分布, 以此近似全后验分布. 在变分 EM 算法中, E 步为

$$\min_{q \in \mathcal{Q}} KL(q(\mathbf{Z}) || p(\mathbf{Z} | \mathbf{X}, \theta^t)) \quad (1.50)$$

这里 \mathcal{Q} 表示满足 Eq. (1.5) 的均场分布族. M 步与 EM 算法相同.

Algorithm 5 Variational Bayes EM algorithm.

- 1: Initialize parameter θ^0 .
 - 2: **while** Likelihood has not converged **do**
 - 3: (VBE-step) Update variational posterior for latent variable,
 $q^{t+1} = \arg \min_{q \in \mathcal{Q}} KL(q(\mathbf{Z}) || p(\mathbf{Z} | \mathbf{X}, \theta^t)).$
 - 4: (VBM-step) Update parameter $\theta^{t+1} = \arg \max_{\theta} \mathbb{E}_{p(\mathbf{Z} | \mathbf{X}, \theta^t)} [\log p(\mathbf{X}, \mathbf{Z} | \theta^{t+1})]$
 - 5: **end while**
-

²<http://www.cse.buffalo.edu/faculty/mbeal/papers/beal03.pdf>

1.4.3 例子: Latent Dirichlet Allocation (续)

在1.3.3节中, 我们给出了 LDA 的一个贝叶斯推断. 在这里, 我们讨论如何用变分 EM 算法解决 LDA 模型 [17]. 首先我们重新组织 LDA 模型. 对于一篇文章, 我们假设它的话题比例分布 $p(\boldsymbol{\theta}_d)$ 为 $\text{Dirichlet}(\boldsymbol{\alpha})$. 对于一个单词, 先抽取它的话题, 话题 $z_{dn} = t$ 的概率为 $p(z_{dn} = t | \boldsymbol{\theta}_d) = \theta_{dt}$; 然后从这个话题中抽取一个单词, 单词为 w_{dn} 的概率为 $p(w_{dn} | z_{dn}) = \beta_{z_{dn}, w_{dn}}$. 我们此时不指定局部潜变量 $\boldsymbol{\theta}$ 和 \mathbf{Z} 的先验分布. LDA 模型的对数似然函数为

$$\begin{aligned} \log p(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{Z}) &= \log \prod_{d=1}^D p(\boldsymbol{\theta}_d) \prod_{n=1}^N p(z_{dn} | \boldsymbol{\theta}_d) p(w_{dn} | z_{dn}) \\ &= \sum_{d=1}^D \log p(\boldsymbol{\theta}_d) + \sum_{n=1}^N \log p(z_{dn} | \boldsymbol{\theta}_d) + \log p(w_{dn} | z_{dn}) \\ &= \sum_{d=1}^D \sum_{t=1}^T \left[(\alpha_t - 1) \log \theta_{dt} + \sum_{n=1}^N I(z_{dn} = t) (\log \theta_{dt} + \log \beta_{k, w_{dn}}) \right]. \end{aligned} \quad (1.51)$$

在本模型中, 变分 EM 算法在 E 步和 M 步分别解决了

$$\text{VBE-step: } \min_{q(\boldsymbol{\theta}), q(\mathbf{Z})} KL(q(\boldsymbol{\theta})q(\mathbf{Z}) || p(\boldsymbol{\theta}, \mathbf{Z} | \mathbf{W}))$$

$$\text{VBM-step: } \max_{\boldsymbol{\beta}} \mathbb{E}_{q(\boldsymbol{\theta}), q(\mathbf{Z})} \log p(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{Z})$$

推断 $\boldsymbol{\theta}$:

$$\begin{aligned} \log q(\boldsymbol{\theta}) &= \mathbb{E}_{q(\mathbf{Z})} \log p(\boldsymbol{\theta}, \mathbf{Z} | \mathbf{W}) + \text{constant} \\ &= \mathbb{E}_{q(\mathbf{Z})} \log p(\boldsymbol{\theta}, \mathbf{Z}, \mathbf{W}) + \text{constant} \\ &= \mathbb{E}_{q(\mathbf{Z})} \left\{ \sum_{d=1}^D \sum_{t=1}^T \left[(\alpha_t - 1) \log \theta_{dt} + \sum_{n=1}^N I(z_{dn} = t) (\log \theta_{dt} + \log \beta_{k, w_{dn}}) \right] \right\} + \text{constant} \\ &= \sum_{d=1}^D \sum_{t=1}^T \left[(\alpha_t - 1) + \sum_{n=1}^N \mathbb{E}_{q(\mathbf{Z})} I(z_{dn} = t) \right] \log \theta_{dt} + \text{constant} \end{aligned} \quad (1.52)$$

则变分分布为 $q(\boldsymbol{\theta}) = \prod_{d=1}^D \prod_{t=1}^T \theta_{dt}^{\alpha_t + \sum_{n=1}^N \gamma_{dn}(t) - 1}$, 即 $q(\boldsymbol{\theta}_d) \sim \text{Dirichlet}(\boldsymbol{\theta}_d | \boldsymbol{\alpha} + \sum_{n=1}^N \gamma_{dn}(t))$, 其中 $\gamma_{dn}(t) = \mathbb{E}_{q(\mathbf{Z})} I(z_{dn} = t)$.

推断 \mathbf{Z} :

$$\begin{aligned} \log q(\mathbf{Z}) &= \mathbb{E}_{q(\boldsymbol{\theta})} \log p(\boldsymbol{\theta}, \mathbf{Z}, \mathbf{W}) + \text{constant} \\ &= \mathbb{E}_{q(\boldsymbol{\theta})} \sum_{d=1}^D \sum_{n=1}^N \sum_{t=1}^T I(z_{dn} = t) (\log \theta_{dt} + \log \beta_{t, w_{dn}}) + \text{constant} \\ &= \sum_{d=1}^D \sum_{n=1}^N \sum_{t=1}^T I(z_{dn} = t) (\mathbb{E} \log \theta_{dt} + \log \beta_{t, w_{dn}}) + \text{constant} \end{aligned} \quad (1.53)$$

则变分分布为


$$q(z_{dn} = t) = \frac{\beta_{t, w_{dn}} \exp(\mathbb{E} \log \theta_{dt})}{\sum_{t=1}^T \beta_{t, w_{dn}} \exp(\mathbb{E} \log \theta_{dt})} \equiv \gamma_{dn}(t). \quad (1.54)$$

推断 β :

$$\begin{aligned} \max_{\beta} \mathbb{E}_{q(\theta), q(Z)} \log p(\theta, \beta, Z) &= \sum_{d=1}^D \sum_{n=1}^N \sum_{t=1}^T I(z_{dn} = t) \log \beta_{t, w_{dn}} + \text{constant} \\ \text{s.t. } \beta_{kw} &\geq 0, \quad \sum_w \beta_{kw} = 1 \end{aligned} \quad (1.55)$$

使用拉格朗日乘子法易得,

$$\beta_{k, w_{dn}} = \frac{\sum_{d,n} \gamma_{dn}(t) I(w_{dn} = w)}{\sum_{d,n,t} \gamma_{dn}(t) I(w_{dn} = w)}. \quad (1.56)$$



2. 非共轭模型

在上面的讨论中, 我们假定了模型是理想化的. 即潜变量的先验分布和似然是共轭的. 这样导致变分分布和先验分布具有相同的形式, 极大地简化了工作量. 但是在研究工作中, 我们的模型可能不是共轭的. 例如, 在鲁棒统计中我们通常假设噪声来自 Laplace 分布, 而 Laplace 分布不存在共轭先验. 我们直接对这个模型进行变分推断是不可行的. 在过去的二十年中, 研究者对非共轭模型的变分推断进行了大量的探索. 非共轭模型的主要解决办法有:

1. 层次模型: 部分分布可以看成指数族分布的复合分布 [21, 22]. 例如: x 的条件分布为高斯分布 $p(x | \lambda) \sim \mathcal{N}(\mu, \lambda^{-1})$, 其中精度 λ 服从指数分布, 则 x 服从 Laplace 分布. 所以从 Laplace 分布中抽样相当于把精度作为潜变量, 首先从指数分布中对精度抽样 $\hat{\lambda}$, 然后在 $\mathcal{N}(\mu, \hat{\lambda}^{-1})$ 中抽样. 这个方法在理论上很优美, 通过引入潜变量, 在非指数族和指数族分布中建立了桥梁. 但是只有部分概率分布能够变为指数族分布的复合分布.
2. 局部变分法与构造 ELBO 下界: 在经典 VI 中, ELBO 是证据的下界

$$\log p(\mathbf{x}) \geq \text{ELBO}(q) = \mathbb{E}_{q(\mathbf{z})} \left[\log \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{q(\mathbf{z})} \right]. \quad (2.1)$$

我们通过极大化 ELBO 使得变分分布和后验分布的 KL 散度下降. 当似然没有共轭先验时, 一个直观的想法是构建似然的下界,

$$\log p(\mathbf{x}) \geq \mathbb{E}_{q(\mathbf{z})} \left[\log \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{q(\mathbf{z})} \right] \geq \mathbb{E}_{q(\mathbf{z})} \left[\log \frac{f(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{q(\mathbf{z})} \right] \equiv \text{ELBO}^{\text{new}}. \quad (2.2)$$

并且使得新下界与先验分布是共轭的. 而引入似然的下界, 通常的代价是增加一个局部变分参数. 所以这项技术称为局部变分法 (local variational method). 这项富有

技巧的方法最早由 Jaakkola 和 Jordan 提出 [23], 应用在 Variational Bayesian logistic regression 中. 但是构建新下界使其与先验分布共轭是十分困难的.

3. 近似: 近似方法主要有两种. (1) 拉普拉斯近似 (Laplace approximation, LA) 在有些文献中也称为高斯近似, 是贝叶斯分析中对连续分布近似的方法 [24]. Mackay et al. 在研究 Bayesian logistic regression 时 [25] 使用了 LA, 用高斯分布近似回归系数的后验分布. Wang et al. 在 2013 年提出了 Laplace variational inference (LVI), 把 LA 的思想纳入了 VI 的体系中, 可以解决一类贝叶斯模型; (2) Delta 方法首次由 Braun and McAuliffe 应用在 VI 中 [26]. Delta 方法的思想是直接对 ELBO 进行泰勒展开, 近似为变分参数的二次函数, 这样会导致变分分布的形式与高斯分布一致. Wang et al. 在 2013 年把多元 Delta 方法应用在了 VI 的框架中. 虽然两种近似方法最终都使得变分分布与高斯分布的形式一致, 但是它们的思想是迥然不同的. 在计算上 Delta 方法更加复杂. Wang et al. 发现 LVI 的近似效果更好.
4. 黑箱模型: 上述方法通常只能解决部分非共轭模型. 黑箱模型的思想是直接对 ELBO 进行数值优化, 通过梯度下降获得变分分布的最优解. 所以黑箱推断可以用于任意概率模型.

2.1 层次模型

2.1.1 SMN 分布族

Andrews and Mallows 研究了一类重要的层次模型 [21], 尺度混合正态分布 (scale mixture of normal distributions, SMN).

Definition 2.1.1 假设 Z 服从标准正态分布, 若存在与 Z 独立的随机变量 V , 使得随机变量 X 和 Z/V 同分布, 则称 X 属于 SMN 分布族.

Andrews and Mallows 给出了随机变量 X 可以由 Z/V 生成的充要条件, 即

$$\left(\frac{d}{dy}\right)^k f_X(y^{1/2}) \geq 0, \forall y > 0. \quad (2.3)$$

在介绍常用 SMN 分布之前, 我们先不加证明地给出下列积分公式

Proposition 2.1.1

$$\int_0^\infty u^{a-1/2} \exp(-bu) du = \Gamma\left(a + \frac{1}{2}\right) b^{-(a+1/2)}, \quad (2.4)$$

$$\int_0^\infty u^{-1/2} \exp\left[-\frac{1}{2}(a^2u + b^2u^{-1})\right] du = \sqrt{\frac{2\pi}{a^2}} \exp(-|ab|), \quad (2.5)$$

$$\int_0^\infty u^{-3/2} \exp\left[-\frac{1}{2}(a^2u + b^2u^{-1})\right] du = \sqrt{\frac{2\pi}{b^2}} \exp(-|ab|), \quad (2.6)$$

下面我们给出两个常见的 SMN 分布.

Theorem2.1.2 t 分布属于 SMN 分布. 若 $x \sim t(x; v, \mu, \sigma^2)$, 则有如下分解

$$x | \tau \sim \mathcal{N}(x | \mu, \sigma^2 \tau^{-1}), \tau \sim \text{Gamma}(\tau | v/2, v/2). \quad (2.7)$$

Proof. 假设 $x | \tau \sim \mathcal{N}(x | \mu, \sigma^2 \tau^{-1}), \tau \sim \Gamma(\tau | a, b)$, 则

$$\begin{aligned} p(x) &= \int_0^\infty \mathcal{N}(x | \mu, \sigma^2 \tau^{-1}) \Gamma(\tau | a, b) d\tau \\ &= \int_0^\infty \sqrt{\frac{\tau}{2\pi\sigma^2}} \exp\left\{-\frac{\tau}{2\sigma^2}(x-\mu)^2\right\} \frac{b^a}{\Gamma(a)} \tau^{a-1} \exp\{-b\tau\} d\tau \\ &= \frac{b^a}{\Gamma(a)\sqrt{2\pi\sigma^2}} \int_0^\infty \tau^{a-1/2} \exp\left\{-\tau\left[b + \frac{(x-\mu)^2}{2\sigma^2}\right]\right\} d\tau \\ &= \frac{b^a}{\Gamma(a)\sqrt{2\pi\sigma^2}} \Gamma\left(a + \frac{1}{2}\right) \left(b + \frac{(x-\mu)^2}{2\sigma^2}\right)^{-(a+1/2)}. \end{aligned} \quad (2.8)$$

令 $a = b = v/2$, 则

$$p(x) = \frac{\Gamma(\frac{v+1}{2})}{\sqrt{\pi v \sigma^2} \Gamma(\frac{v}{2})} \left[1 + \frac{(x-\mu)^2}{v\sigma^2}\right]^{-\frac{v+1}{2}} = t(x; v, \mu, \sigma^2). \quad (2.9)$$

所以, $x \sim t(x; v, \mu, \sigma^2) \Leftrightarrow x | \tau \sim \mathcal{N}(x | \mu, \sigma^2 \tau^{-1}), \tau \sim \Gamma(\tau | v/2, v/2)$. ■

Theorem2.1.3 Laplace 分布属于 SMN 分布. 若 $x \sim \text{Laplace}(x; \mu, \sqrt{\lambda/2})$, 则有如下分解

$$x | z \sim \mathcal{N}(x | \mu, z), z \sim \text{Exponential}(z | \lambda). \quad (2.10)$$

Proof. 若 $x | z \sim \mathcal{N}(x | \mu, z), z \sim \text{Exponential}(z | \lambda)$, 则

$$\begin{aligned} p(x) &= \int_0^\infty \mathcal{N}(x | \mu, z) \text{Exponential}(z | \lambda) dz \\ &= \int_0^\infty \sqrt{\frac{1}{2\pi z}} \exp\left\{-\frac{1}{2z}(x-\mu)^2\right\} \frac{1}{\lambda} \exp\left\{-\frac{z}{\lambda}\right\} dz \\ &= \frac{1}{\sqrt{2\pi\lambda^2}} \int_0^\infty \sqrt{\frac{1}{z}} \exp\left\{-\frac{1}{2}\left(\frac{2}{\lambda}z + (x-\mu)^2 z^{-1}\right)\right\} dz \\ &= \frac{1}{\sqrt{2\pi\lambda^2}} \sqrt{\frac{\lambda}{2}} 2\pi \exp\left(-\sqrt{\frac{2}{\lambda}}|x-\mu|\right) = \frac{1}{2} \sqrt{\frac{2}{\lambda}} \exp\left(-\sqrt{\frac{2}{\lambda}}|x-\mu|\right) \end{aligned} \quad (2.11)$$

所以, $x \sim \text{Laplace}(x; \mu, \sqrt{\lambda/2}) \Leftrightarrow x | z \sim \mathcal{N}(x | \mu, z), z \sim \text{Exponential}(z | \lambda)$. ■

上述定理表明: 若 $X = \mu + b\sqrt{2V}Z, V \sim \text{Exponential}(1)$, 则 $X \sim \text{Laplace}(x; \mu, b)$. 由于指数分布是逆 Gamma 分布的特例, 所以, $\text{Laplace}(x; \mu, b)$ 可以分解为

$$\begin{aligned} x | z &\sim \mathcal{N}(x | \mu, b^2/z) \\ z &\sim \text{Inverse-Gamma}(1, 1/2) \end{aligned} \quad (2.12)$$

R 幂指数分布族 (power-exponential family, PE) [27] 的概率密度函数如下

$$f(x) = \frac{\beta}{2\alpha\Gamma(1/\beta)} e^{-(|x-\mu|/\alpha)^\beta}, \quad (2.13)$$

其中, α, μ 分别为尺度和位置参数, β 为形状参数. 如果 $\beta = 1, 2$, 则对应了 Laplace 和高斯分布. West 指出指数幂分布族都属于 SMN 分布族 [28].

从上面两个例子告诉我们, 对 SMN 分布族抽样由两步组成. 首先, 从方差 σ^2 的分布 $p(\sigma^2)$ 中抽样得到 $\hat{\sigma}^2$; 然后从正态分布中抽样 $\mathcal{N}(\mu, \hat{\sigma}^2)$. 这样得到的样本服从 SMN 分布.

2.1.2 其他层次模型

除了 SMN 之外, 还有其他分布可以进行层次化, 下面的例子讨论了如何将 asymmetric Laplace 分布和 skew normal 分布层次化.

Asymmetric Laplace (AL) distribution

在 Bayesian quantile regression 中, Yu and Moyeed 假设噪声服从 asymmetric Laplace 分布 [29],

$$p(x; \tau, \mu, b) = \frac{\tau(1-\tau)}{b} \exp \left\{ -\frac{x-\mu}{b} [\tau - I(x-\mu < 0)] \right\} \sim AL(x | \tau, \mu, b). \quad (2.14)$$

若令 V 为标准指数分布 $\text{Exponential}(1)$, 则

$$X = \mu + mbV + b\sqrt{\sigma^2}VZ \quad (2.15)$$

服从 $AL(x | \tau, \mu, b)$ [30, 31], 其中

$$m = \frac{1-2\tau}{\tau(1-\tau)}, \sigma^2 = \frac{2}{\tau(1-\tau)}.$$

实际上, $bV \sim \text{Exponential}(1/b)$, 我们令 $v = bV$, 则 $p(v) = \frac{1}{b} \exp\{-v/b\}$. 公式 (2.15) 可以改写为

$$X = \mu + mv + \sqrt{\sigma^2}vZ. \quad (2.16)$$

所以, $AL(x; \tau, \mu, b)$ 可以下面的层次模型构造

$$\begin{aligned} x | v &\sim \mathcal{N}(\mu + mv, \sigma^2 bv) \\ v &\sim \text{Exponential}(1/b) \end{aligned} \quad (2.17)$$

下面我们证明公式 (2.15) 成立.

Proof. 我们证明 $X = mV + \sqrt{\sigma^2}VZ$ 服从 $AL(x | \tau, 0, 1)$. 基本思想是通过特征函数证明等式两边的分布相同. $AL(x | \tau, 0, 1)$ 的特征函数为

$$\begin{aligned} f(t) &= \mathbb{E}[e^{itx}] \\ &= \int_{-\infty}^{\infty} e^{itx} p_X(x) dx \\ &= \int_{-\infty}^0 \tau(1-\tau) e^{itx+(1-\tau)x} dx + \int_0^{\infty} \tau(1-\tau) e^{itx-px} dx \\ &= \tau(1-\tau) \left\{ \frac{1}{it+(1-\tau)} + \frac{1}{\tau-it} \right\}. \end{aligned} \quad (2.18)$$

另一方面, $mV + \sqrt{\sigma^2 V}Z$ 的特征函数为

$$\begin{aligned} g(t) &= \mathbb{E}[e^{it(mV + \sqrt{\sigma^2 V}Z)}] \\ &= \int_0^\infty e^{-V} e^{itmV} \mathbb{E}[e^{it\sqrt{\sigma^2 V}Z}] dV. \end{aligned} \quad (2.19)$$

上式的期望是关于 V, Z 的, 在第二行中, 我们把关于 V 的期望展开成了积分. 由于 Z 服从标准正态分布, 所以

$$\mathbb{E}[e^{it\sqrt{\sigma^2 V}Z}] = e^{-t^2 \sigma^2 V / 2}.$$

故

$$\begin{aligned} g(t) &= \int_0^\infty e^{-V(1+t^2\sigma^2/2-itm)} dV \\ &= \left(\frac{1}{2} \sigma^2 t^2 - imt + 1 \right)^{-1} \\ &= \tau(1-\tau) \left\{ \frac{1}{it + (1-\tau)} + \frac{1}{\tau - it} \right\} = f(t). \end{aligned} \quad (2.20)$$

■

Skew normal (SN) distribution

SN 分布在数据挖掘中有重要的应用 [32, 33], 最早由 Azzalini 提出 [34, 35], 常被用于非对称分布的建模. 若 $x \sim SN(\mu, \sigma^2, \lambda)$, 则

$$p(x; \mu, \sigma^2, \lambda) = \frac{2}{\sigma} \phi\left(\frac{x-\mu}{\sigma}\right) \Phi\left(\lambda \frac{x-\mu}{\sigma}\right), \quad (2.21)$$

其中 ϕ, Φ 分别表示标准正态分布的密度函数和累计密度函数. 当 $\lambda = 0$ 时, $SN(\mu, \sigma^2, \lambda)$ 退化为 $\mathcal{N}(\mu, \sigma^2)$. 若令 $\delta = \lambda / \sqrt{1 + \lambda^2}$, U, Z 独立同分布于标准正态分布, 则

$$X = \delta|U| + \sqrt{1 - \delta^2}Z = SN(0, 1, \lambda). \quad (2.22)$$

这个结果可以通过分析的方法计算得出 [36],

$$\begin{aligned} F_X(x) &= P(X \leq x) = \mathbb{E}_U[p(X \leq x | |U|)] \\ &= \int_0^\infty p(X \leq \frac{x - \delta u}{\sqrt{1 - \delta^2}}) 2\phi(u) du = 2 \int_0^\infty \Phi\left(\frac{x - \delta u}{\sqrt{1 - \delta^2}}\right) \phi(u) du \\ f_X(x) &= \nabla_x F_X(x) = 2\phi(x)\Phi(\lambda x) = SN(0, 1, \lambda). \end{aligned} \quad (2.23)$$

根据公式 (2.22), 令 $U, Z \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$, 则有

$$X = \mu + \delta|U| + \sqrt{1 - \delta^2}Z = SN(\mu, \sigma^2, \lambda). \quad (2.24)$$

所以, $SN(\mu, \sigma^2, \lambda)$ 可以由下面的层次模型构造

$$\begin{aligned} x | u &\sim \mathcal{N}(\mu + \delta u, (1 - \delta^2)\sigma^2) \\ u &\sim HN(0, \sigma^2) \end{aligned} \quad (2.25)$$

这里的 HN 表示 half normal 分布.

最近 Ferreira [22] 把 SMN 推广到了斜尺度混合正态分布 (skew scale mixture of normals, SSMN), 并给出了 skew t , skew-slash, skew PE 分布的层次模型.

2.1.3 层次模型的优缺点

在上面的 3 个例子中, 我们证明了 t 分布, Laplace 分布, asymmetric Laplace 分布和 skew normal 分布可以层次化. 如果 x 服从上述分布, 则 x 不存在共轭分布. 但是通过引入潜变量, 我们可以把 x 转化为正态分布, 从而可以使用经典的 VI 框架进行处理. 虽然这个方法在理论上十分优雅, 但是寻找一个分布的层次分布是不容易的. 即使找到了层次分布, 也可能无法被经典的 VI 框架解决. Holmes and Held [37] 在研究贝叶斯逻辑回归时, 把似然函数分解为正态分布和渐进 Kolmogorov-Smirnov 分布的复合分布. 这虽是层次模型, 但是依然无法被经典的 VI 框架解决.

2.1.4 例子: Linear regression with Laplace noise

在 1.2.4 节中介绍了如何使用 VI 对线性回归进行推断, 其中的一个潜在假设是噪声 ε 服从高斯分布. 实际上, 高斯分布对异常点十分敏感. 所以我们不妨假设噪声来自厚尾分布——Laplace 分布, 以此增强模型的鲁棒性. 利用公式 (2.12), 可以把高斯线性回归 (1.14) 的一个公式替换为

$$\begin{aligned} p(\mathbf{y} | \mathbf{w}, \mathbf{z}, \beta) &= \prod_{n=1}^N \mathcal{N}(y_n | \mathbf{x}_n^\top \mathbf{w}, (z_n \beta)^{-1}), \\ p(\mathbf{z}) &= \prod_{n=1}^N \text{Exponential}(z_n | \lambda). \end{aligned} \quad (2.26)$$

推断 \mathbf{w} :

$$\begin{aligned} \log q(\mathbf{w}) &= \mathbb{E} \{ \log p(\mathbf{y} | \mathbf{w}, \mathbf{z}, \beta) p(\mathbf{w} | \mathbf{T}) \} \\ &= \mathbb{E} \left\{ -\frac{\beta}{2} (\mathbf{y} - \mathbf{X}\mathbf{w})^\top \mathbf{D}_z (\mathbf{y} - \mathbf{X}\mathbf{w}) - \frac{1}{2} \mathbf{w}^\top \mathbf{T} \mathbf{w} \right\} + \text{constant} \\ &= \mathbb{E} \left\{ -\frac{\beta}{2} (-2\mathbf{y}^\top \mathbf{D}_z \mathbf{X} \mathbf{w} + \mathbf{w}^\top \mathbf{X}^\top \mathbf{D}_z \mathbf{X} \mathbf{w}) - \frac{1}{2} \mathbf{w}^\top \mathbf{T} \mathbf{w} \right\} + \text{constant} \\ &= \beta \mathbf{y}^\top \mathbb{E}[\mathbf{D}_z] \mathbf{X} \mathbf{w} - \frac{1}{2} \mathbf{w}^\top \left(\mathbb{E}[\beta] \mathbf{X}^\top \mathbb{E}[\mathbf{D}_z] \mathbf{X} + \mathbb{E}[\mathbf{T}] \right) \mathbf{w} + \text{constant} \end{aligned} \quad (2.27)$$

其中 $\mathbf{D}_z = \text{diag}(z_1, \dots, z_N)$. 记 \mathbf{w} 的后验分布为 $\mathcal{N}(\mathbf{w} | \mathbf{m}, \mathbf{S})$, 则

$$\mathbf{S} = \left(\mathbb{E}[\beta] \mathbf{X}^\top \mathbb{E}[\mathbf{D}_z] \mathbf{X} + \mathbb{E}[\mathbf{T}] \right)^{-1}, \quad \mathbf{m} = \mathbb{E}[\beta] \mathbf{S} \mathbf{X}^\top \mathbb{E}[\mathbf{D}_z] \mathbf{y}. \quad (2.28)$$

推断 z_n :

$$\begin{aligned} \log q(z_n) &= \mathbb{E} \{ \log p(y_n | \mathbf{w}, z_n, \beta) + \log p(z_n) \} \\ &= \mathbb{E} \left\{ -\frac{3}{2} \log z_n - \frac{z_n}{2} \beta (y_n - \mathbf{x}_n^\top \mathbf{w})^2 - \frac{1}{2z_n} \right\}. \end{aligned} \quad (2.29)$$

由于 z_n 不是共轭的, 所以变分分布不是逆 Gamma 分布, 而是逆高斯分布, 记为 Inverse-Gaussian($g_n, 1$)

$$g_n = \left\{ \mathbb{E} \left[\beta (y_n - \mathbf{x}_n^\top \mathbf{w})^2 \right] \right\}^{-1/2} \quad (2.30)$$

推断 β :

$$\begin{aligned}\log q(\beta) &= \mathbb{E} \{ \log \log p(\mathbf{y} | \mathbf{w}, \mathbf{z}, \beta) + \log p(\beta) \} \\ &= \mathbb{E} \left\{ \frac{N}{2} \log \beta - \frac{\beta}{2} (\mathbf{y} - \mathbf{X}\mathbf{w})^\top \mathbf{D}_z (\mathbf{y} - \mathbf{X}\mathbf{w}) + (a_0 - 1) \log \beta - b_0 \beta \right\}.\end{aligned}\quad (2.31)$$

记变分分布为 $\Gamma(a, b)$, 其中

$$a = \frac{N}{2} + a_0, \quad b = \frac{1}{2} \mathbb{E}[(\mathbf{y} - \mathbf{X}\mathbf{w})^\top \mathbf{D}_z (\mathbf{y} - \mathbf{X}\mathbf{w})] + b_0. \quad (2.32)$$

关于 \mathbf{T} 的推断没有变化. 上述更新公式的期望计算: (1) $\mathbb{E}[\mathbf{D}_z] = \text{diag}(\mathbb{E}[z_1], \dots, \mathbb{E}[z_N]) =$

$\text{diag}(g_1, \dots, g_N) = \hat{\mathbf{D}}_z$; (2) $\mathbb{E}[(\mathbf{y} - \mathbf{X}\mathbf{w})^\top \mathbf{D}_z (\mathbf{y} - \mathbf{X}\mathbf{w})] = \mathbb{E}[\mathbf{y}^\top \hat{\mathbf{D}}_z \mathbf{y} - 2\mathbf{y}^\top \hat{\mathbf{D}}_z \mathbf{X}\mathbf{w} + \text{tr}(\mathbf{X}^\top \hat{\mathbf{D}}_z \mathbf{X}(\mathbf{m}\mathbf{m}^\top + \mathbf{S}))]$;

(3) $g_n = \{ \mathbb{E}[\beta(y_n - \mathbf{x}_n^\top \mathbf{w})^2] \}^{-1/2} = \{ \frac{a}{b} [y_n^2 - 2y_n \mathbf{x}_n^\top \mathbf{m} + \text{tr}(\mathbf{x}_n \mathbf{x}_n^\top (\mathbf{m}\mathbf{m}^\top + \mathbf{S}))] \}^{-1/2}$.

2.1.5 例子: Bayesian Lasso / Linear regression with Laplace penalty

在高维统计中, 我们希望 \mathbf{w} 是稀疏的. Tibshirani 提出了 Lasso, 在损失函数上加入 L_1 范数惩罚

$$\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_1. \quad (2.33)$$

从贝叶斯观点看, 这实际上是假设 \mathbf{w} 服从 Laplace 分布. Bayesian Lasso 的模型如下,

$$\begin{aligned}p(\mathbf{y} | \mathbf{w}) &= \prod_{n=1}^N \mathcal{N}(y_n | \mathbf{x}_n^\top \mathbf{w}, \beta^{-1}) \\ p(\mathbf{w} | \mathbf{Z}, \mathbf{T}) &= \mathcal{N}(\mathbf{w} | \mathbf{0}, (\mathbf{Z}\mathbf{T})^{-1}) \\ p(\beta) &= \Gamma(\beta | a_0, b_0) \\ p(t_j) &= \Gamma(t_j | c_0, d_0) \\ p(z_j) &= \text{Inverse-Gamma}(z_j | 1, 1/2)\end{aligned}\quad (2.34)$$

虽然 \mathbf{w} 的先验分布与似然不是共轭的, 但是利用2.1.1节的结论, 引入潜变量 \mathbf{Z} 后, \mathbf{w} 的条件先验分布是共轭的.

推断 \mathbf{w} :

$$\begin{aligned}\log q(\mathbf{w}) &= \mathbb{E} \{ \log p(\mathbf{y} | \mathbf{w}, \beta) + \log p(\mathbf{w} | \mathbf{Z}, \mathbf{T}) \} \\ &= \mathbb{E} \left\{ -\frac{\beta}{2} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 - \frac{1}{2} \mathbf{w}^\top (\mathbf{Z}\mathbf{T}) \mathbf{w} \right\} + \text{constant} \\ &= \mathbb{E} \left\{ \beta \mathbf{y}^\top \mathbf{X}\mathbf{w} - \frac{1}{2} \mathbf{w}^\top (\beta \mathbf{X}^\top \mathbf{X} + \mathbf{Z}\mathbf{T}) \mathbf{w} \right\} + \text{constant}\end{aligned}\quad (2.35)$$

记 \mathbf{w} 的变分分布为 $\mathcal{N}(\mathbf{w} | \mathbf{m}, \mathbf{S})$, 其中

$$\mathbf{S} = [\mathbb{E}(\beta \mathbf{X}^\top \mathbf{X} + \mathbf{Z}\mathbf{T})]^{-1}, \quad \mathbf{m} = \mathbb{E}[\beta] \mathbf{S} \mathbf{X}^\top \mathbf{y}. \quad (2.36)$$

推断 z_j :

$$\begin{aligned}\log q(z_j) &= \mathbb{E} \{ \log p(w_j | z_j, t_j) + \log p(z_j) \} \\ &= \mathbb{E} \left\{ -\frac{3}{2} \log z_j - \frac{z_j t_j w_j^2}{2} - \frac{1}{2z_j} \right\} + \text{constant}\end{aligned}\quad (2.37)$$

记 z_j 的变分分布为 $\text{Inverse-Gaussian}(z_j | g_j, 1)$, 其中

$$g_j = \{\mathbb{E}[t_j w_j^2]\}^{-1/2} = \left\{ \frac{c}{d} (m_j^2 + s_{jj}) \right\}^{-1/2} \quad (2.38)$$

推断 t_j :

$$\begin{aligned} \log q(t_j) &= \mathbb{E} \{ \log p(w_j | z_j, t_j) + \log p(t_j) \} \\ &= \mathbb{E} \left\{ \frac{1}{2} \log t_j - \frac{z_j w_j^2 t_j}{2} + c_0 \log t_j - d_0 t_j \right\} + \text{constant} \end{aligned} \quad (2.39)$$

记 t_j 的变分分布为 $\Gamma(t_j | c, d)$, 其中

$$c = c_0 + \frac{1}{2}, \quad d = d_0 + \frac{\mathbb{E}[z_j w_j^2]}{2} = d_0 + \frac{g_j(m_j^2 + s_{jj})}{2}. \quad (2.40)$$

2.2 构造 ELBO 下界与局部变分法

VI 的基本思想是构造了对数证据 $\log p(\mathbf{x})$ 的下界 (即 ELBO), 通过优化 ELBO 使得后验分布和变分分布的 KL 散度下降. 当模型不是共轭时, 一个自然的想法是寻找似然的下界, 构造新的 ELBO, 并且使得 ELBO 中的似然和先验是共轭的.

Jaakkola 和 Jordan 提出的变分逻辑回归, 他们引入了局部变分参数, 得到了 logistic 函数的下界. 新的 ELBO 具有闭式解. Blei 和 Lafferty 在研究相关主题模型 (correlated topic model) 时 [38], 利用对数函数的下界, 把局部变分的思想推广到了变分 EM 算法的框架内.

2.2.1 局部变分法

局部变分法是寻找模型中的一个变量或变量组上定义的函数的界的 [1]. 下面以 $f(x) = \exp(-x)$ 为例. 由于 $f(x)$ 是凸函数, 则它的切线必定为其下界. 在 $x = \xi$ 处, 用一阶泰勒展开有

$$y(x, \xi) = f(\xi) + f'(\xi)(x - \xi) \leq f(x), \quad (2.41)$$

上面的等号成立, 当且仅当 $x = \xi$. 令 $\eta = -\exp(-\xi)$, 则 $y(x, \eta) = \eta x - \eta + \eta \log(-\eta)$. 为了充分逼近 $f(x)$, 我们应该选择最优的 η , 即在何处进行泰勒展开. 对于某个固定的 x , 显然

$$f(x) = \max_{\eta} \{ \eta x - g(\eta) \},$$

这里 $g(\eta) = \eta - \eta \log(-\eta)$. 对于某个固定的 η , 显然 $g(\eta)$ 需要满足

$$g(\eta) = \max_x \{ \eta x - f(x) \}.$$

类似地, 对于凹函数, 我们有

$$\begin{cases} f(x) = \min_{\eta} \{ \eta x - g(\eta) \} \\ g(x) = \min_{\eta} \{ \eta x - f(x) \} \end{cases}$$

2.2.2 Logistic 函数的局部变分

逻辑函数定义为

$$\sigma(x) = \frac{1}{1 + e^{-x}}, \quad (2.42)$$

显然这是非凹非凸函数, 我们可以先经过某些变换把它变为凹函数或凸函数, 然后进行局部变分近似, 最后通过逆变换就可以得到逻辑函数的界限 [1, 23].

- $\log \sigma(x)$ 是凹函数. 假设 $f(x) = \log \sigma(x)$ 的变分近似为 $h(x, \eta) = \eta x - g(\eta)$. 其中 $g(\eta)$ 需要满足 $g(\eta) = \min_x \{\eta x - f(x)\}$. 对 $\eta x - f(x)$ 求导, 并令导数为零, 可得 $x^* = \log \frac{1-\eta}{\eta}$. 所以 $g(\eta) = \eta x^* - f(x^*) = -\eta \log \eta - (1-\eta) \log(1-\eta)$. 则 $f(x)$ 的上界为 $\eta x - g(\eta) = \eta x + \eta \log \eta + (1-\eta) \log(1-\eta)$, 即

$$\sigma(x) \leq \exp\{\eta x + \eta \log \eta + (1-\eta) \log(1-\eta)\}.$$

- $\log \sigma(x) = -\log(1 + e^{-x}) = \frac{x}{2} - \log(e^{x/2} + e^{-x/2})$. $f(x) = -\log(e^{x/2} + e^{-x/2})$ 是 x^2 的凸函数. 假设 $f(x)$ 的变分近似为 $h(x, \eta) = \eta x - g(\eta)$. 其中 $g(\eta)$ 需要满足 $g(\eta) = \min_{x^2} \{\eta x^2 - f(\sqrt{x^2})\}$. 由驻点条件,

$$0 = \eta - \frac{dx}{dx^2} \frac{d}{dx} f(x) = \eta + \frac{1}{4x} \tanh(x/2),$$

假设上式的解为 $x^* = \xi$, 则 ξ 满足 $\eta = -\frac{1}{4\xi} \tanh(\xi/2) = -\frac{1}{2\xi} [\sigma(\xi) - 1/2] \equiv -\lambda(\xi)$. 所以 $g(\eta) = g(\lambda(\xi)) = -\lambda(\xi)\xi^2 - f(\xi) = -\lambda(\xi)\xi^2 + \log(e^{\xi/2} + e^{-\xi/2}) \leq f(x)$. 故逻辑函数的下界为

$$\sigma(x) \geq \sigma(\xi) \exp\left\{\frac{x-\xi}{2} - \lambda(\xi)(x^2 - \xi^2)\right\}, \quad \lambda(\xi) = \frac{1}{2\xi} \left[\sigma(\xi) - \frac{1}{2}\right]. \quad (2.43)$$

变分下界的对数关于 x 是二次的.

2.2.3 例子: Variational Bayesian logistic regression

逻辑回归的响应为二值变量, $y \in \{-1, 1\}$. 且假设 $P(y | \mathbf{x}, \mathbf{w}) = \sigma(\mathbf{y}\mathbf{x}^\top \mathbf{w})$. 虽然似然不存在共轭先验, 但是假设回归系数的先验为 $\mathcal{N}(\mathbf{0}, \mathbf{A}^{-1})$, $\mathbf{A} = \text{diag}(\alpha_1, \dots, \alpha_p)$, 超参数 α_j 的先验为 $\Gamma(\alpha_j | a_0, b_0)$.

根据 (2.43) 可以得出似然的下界

$$\begin{aligned} \log p(\mathbf{y} | \mathbf{X}, \mathbf{w}) &= \sum_{n=1}^N \log \sigma(y_n \mathbf{x}_n^\top \mathbf{w}) \\ &\geq \log h(\mathbf{w}, \boldsymbol{\xi}) = \log \prod_{n=1}^N \sigma(\xi_n) \exp\left\{\frac{y_n \mathbf{x}_n^\top \mathbf{w} - \xi_n}{2} - \lambda(\xi_n)[(y_n \mathbf{x}_n^\top \mathbf{w})^2 - \xi_n^2]\right\} \\ &= \frac{1}{2} \mathbf{y}^\top \mathbf{X} \mathbf{w} - \mathbf{w}^\top \mathbf{S} \mathbf{w} + \sum_{n=1}^N \left\{ \log \sigma(\xi_n) - \frac{\xi_n}{2} + \lambda(\xi_n) \xi_n^2 \right\}, \end{aligned} \quad (2.44)$$

其中 $S = \sum_{n=1}^N \lambda(\xi_n) \mathbf{x}_n \mathbf{x}_n^\top$. 进而, 我们可以得出 ELBO 的下界 [16]

$$\text{ELBO}(q) = \mathbb{E}_q \left\{ \log \frac{p(\mathbf{y} | \mathbf{X}, \mathbf{w}) p(\mathbf{w} | \boldsymbol{\alpha}) p(\boldsymbol{\alpha})}{q(\mathbf{w}) q(\boldsymbol{\alpha})} \right\} \geq \mathbb{E}_q \left\{ \log \frac{h(\mathbf{w}, \boldsymbol{\xi}) p(\mathbf{w} | \boldsymbol{\alpha}) p(\boldsymbol{\alpha})}{q(\mathbf{w}) q(\boldsymbol{\alpha})} \right\} \equiv \widetilde{\text{ELBO}}(q, \boldsymbol{\xi}). \quad (2.45)$$

推断 \mathbf{w} :

$$\begin{aligned} \log q(\mathbf{w}) &= \mathbb{E}_q \{ \log h(\mathbf{w}, \boldsymbol{\xi}) + \log p(\mathbf{w} | \boldsymbol{\alpha}) \} + \text{constant} \\ &= \mathbb{E}_q \left\{ \frac{1}{2} \mathbf{y}^\top \mathbf{X} \mathbf{w} - \mathbf{w}^\top \mathbf{S} \mathbf{w} - \frac{1}{2} \mathbf{w}^\top \mathbf{A} \mathbf{w} \right\} + \text{constant} \\ &= \frac{1}{2} \mathbf{y}^\top \mathbf{X} \mathbf{w} - \mathbf{w}^\top \left(\mathbf{S} + \frac{1}{2} \mathbb{E}_q[\mathbf{A}] \right) \mathbf{w} \end{aligned} \quad (2.46)$$

记变分分布为 $\mathcal{N}(\mathbf{w} | \boldsymbol{\mu}, \boldsymbol{\Sigma})$, 其中

$$\boldsymbol{\Sigma} = \left(\mathbf{S} + \frac{1}{2} \mathbb{E}_q[\mathbf{A}] \right)^{-1}, \quad \boldsymbol{\mu} = \frac{1}{2} \boldsymbol{\Sigma} \mathbf{X}^\top \mathbf{y}. \quad (2.47)$$

推断 α_j :

$$\begin{aligned} \log q(\alpha_j) &= \mathbb{E}_q \{ \log p(w_j | \alpha_j) + \log p(\alpha_j) \} + \text{constant} \\ &= \mathbb{E}_q \left\{ \frac{1}{2} \log \alpha_j - \frac{w_j^2}{\alpha_j} + (\alpha_0 - 1) \log \alpha_j - b_0 \alpha_j \right\} + \text{constant} \end{aligned} \quad (2.48)$$

记变分分布为 $\Gamma(\alpha_j | a_j, b_j)$, 其中

$$a_j = a_0 + \frac{1}{2}, \quad b_j = b_0 + \frac{1}{2} \mathbb{E}_q(w_j^2). \quad (2.49)$$

推断 ξ_n : 由于新的 ELBO 是局部变分参数的函数, 所以 ξ_n 在潜变量更新之后也需要更新, 即

$$\max_{\xi_n} \widetilde{\text{ELBO}}(q, \boldsymbol{\xi}).$$

通过驻点条件可得,

$$\xi_n^{\text{new}} = \sqrt{\mathbf{x}_n^\top (\boldsymbol{\Sigma} + \boldsymbol{\mu} \boldsymbol{\mu}^\top) \mathbf{x}_n}. \quad (2.50)$$

2.3 拉普拉斯变分推断

上节介绍了如何用局部变分方法对非共轭模型推断. 其基本思想是构造似然的下界, 使模型变成共轭的, 但代价是引入了局部变分参数. 在构造下界时, 一个重要的问题是下界是否紧致. 如果下界不紧致, 则变分推断的结果与后验分布可能有较大的差异.

MacKay 在研究贝叶斯模型时提出了 Laplace 近似 [25]. 如果似然是非共轭的, 参数的后验分布无法精确求解, MacKay 使用了一个高斯分布去近似真实的后验分布. 这个思想可以纳入 VI 的框架内. 对于非共轭模型, 其参数的变分分布没有解析解时, 我们可以用高斯分布去近似. 这个方法称为 Laplace variational inference (LVI) [39]. 从另外一个角度出发,

我们关于对数似然使用泰勒展开并保留 2 阶以内的项. 这个结果是随机变量的 2 次多项式, 显然这是高斯分布的核函数, 具有共轭先验. 这种近似方法叫做 Delta method variational inference (DVI) [26, 39]. Wang and Blei 在贝叶斯逻辑回归和相关话题模型中发现 LVI 的推断结果优于 DVI, 且计算效率更高. 下面主要介绍 LVI.

2.3.1 Laplace variational inference

LVI 适用于一类广泛的非共轭模型: (1) 参数 θ 为实值的, 其分布 $p(\theta)$ 二次可微; (2) $p(z | \theta) = h(z) \exp \{ \eta(\theta)^\top t(z) - a(\eta(\theta)) \}$ 在指数族中, $\eta(\theta)$ 二次可微. 注意这里不限制 $p(\theta)$ 和 $p(z | \theta)$ 为共轭对, $p(\theta | z)$ 和 $p(\theta)$ 也没必要有相同的形式; (3) $p(x | z) = h(x) \exp \{ t(z)^\top \langle t(x), 1 \rangle \}$ 在指数族中. 这条假设暗示了 z 和 x 的条件分布是共轭的, $p(z | \theta, x)$ 和 $p(z | \theta)$ 有相同的形式.

首先我们写出 θ 在信息传递算法中的更新公式,

$$q(\theta) \propto \exp \left\{ \eta(\theta)^\top \mathbb{E}_{q(z)}[t(z)] - a(\eta(\theta)) + \log p(\theta) \right\} \equiv \exp \{ f(\theta) \}. \quad (2.51)$$

对 $f(\theta)$ 在 $\hat{\theta}$ 处使用泰勒展开,

$$f(\theta) \approx f(\hat{\theta}) + \nabla f(\hat{\theta})(\theta - \hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^\top \nabla^2 f(\hat{\theta})(\theta - \hat{\theta}), \quad (2.52)$$

现在令 $\hat{\theta}$ 为随机变量 θ 的众数, 即 $\hat{\theta} = \arg \max_{\theta} f(\theta)$, 所以 $\nabla f(\hat{\theta}) = 0$. 进而

$$q(\theta) \propto \exp \{ f(\theta) \} \propto \left\{ f(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^\top \nabla^2 f(\hat{\theta})(\theta - \hat{\theta}) \right\}. \quad (2.53)$$

所以 $q(\theta) \propto \mathcal{N}(\hat{\theta}, -\nabla^2 f(\hat{\theta})^{-1})$. 注意: 在求众数时, 通常不存在解析解, 我们一般需要使用数值算法得出数值解.

根据信息传递算法, z 的更新为

$$q(z) \propto h(z) \exp \left\{ (\mathbb{E}_q[\eta(\theta)] + t(x))^\top t(z) \right\} \quad (2.54)$$

如果 $\mathbb{E}_q[\eta(\theta)]$ 不容易计算, 可以先在 μ 处进行泰勒展开 (这里要求 $\eta(\theta)$ 二次可微), 利用 $q(\theta) = \mathcal{N}(\mu, \Sigma)$, 可得

$$\mathbb{E}_q[\eta(\theta)] \approx \eta(\mu) + \frac{1}{2} \text{tr}(\nabla^2 \eta(\mu) \Sigma) \quad (2.55)$$

2.3.2 LVI 的优缺点

LVI 适用于一类比较广泛的非共轭模型, 实际上大部分可以被层次模型或局部变分方法解决的模型也可以被 LVI 解决. 但是应用 LVI 时, 我们必须注意以下几点:

1. LVI 不能处理非连续分布;
2. LVI 处理复杂分布的能力有限, 例如变分分布是多峰分布, 或者变分分布不是钟形分布 (参数小于 1 的 Beta 分布).
3. LVI 在求解众数 $\hat{\theta}$ 时, 可能没有闭式解, 计算速度缓慢.

所以使用 LVI 前, 需要判断模型是否符合 LVI 的假设, 用高斯分布近似其变分分布是否合适.

2.3.3 例子: Variational Bayesian logistic regression (续)

在2.2.3节中, 我们使用了局部变分法对贝叶斯逻辑回归进行了推断. 本节使用 LVI 进行求解. 模型中 \mathbf{w} 与 \mathbf{y} 非共轭, \mathbf{w} 的后验分布为

$$\begin{aligned}\log q(\mathbf{w}) &= \mathbb{E}_q[\log p(\mathbf{y} | \mathbf{w}) + \log p(\mathbf{w} | \alpha)] \\ &= \sum_{n=1}^N \log \sigma(y_n \mathbf{x}_n^\top \mathbf{w}) - \frac{1}{2} \mathbf{w}^\top \mathbb{E}[\mathbf{A}] \mathbf{w} + \text{constant} \\ &= - \sum_{n=1}^N \log[1 + \exp(y_n \mathbf{x}_n^\top \mathbf{w})] - \frac{1}{2} \mathbf{w}^\top \mathbb{E}[\mathbf{A}] \mathbf{w} + \text{constant}\end{aligned}\quad (2.56)$$

所以, $f(\mathbf{w}) = -\sum_{n=1}^N \log[1 + \exp(y_n \mathbf{x}_n^\top \mathbf{w})] - \frac{1}{2} \mathbf{w}^\top \mathbb{E}[\mathbf{A}] \mathbf{w}$. 为了得到 Laplace 近似, 对 f 求导,

$$\nabla f(\mathbf{w}) = \sum_{n=1}^N \sigma(-y_n \mathbf{x}_n^\top \mathbf{w}) y_n \mathbf{x}_n - \mathbb{E}[\mathbf{A}] \mathbf{w} \quad (2.57)$$

$$\nabla^2 f(\mathbf{w}) = - \sum_{n=1}^N \sigma(-y_n \mathbf{x}_n^\top \mathbf{w}) \sigma(y_n \mathbf{x}_n^\top \mathbf{w}) \mathbf{x}_n \mathbf{x}_n^\top - \mathbb{E}[\mathbf{A}] \quad (2.58)$$

所以变分分布为 $q(\mathbf{w}) = \mathcal{N}(\hat{\mathbf{w}}, -\nabla^2 f(\hat{\mathbf{w}})^{-1})$, 其中均值

$$\hat{\mathbf{w}} = \arg \max_{\mathbf{w}} \nabla f(\mathbf{w}).$$

然而, 均值的求解不存在解析解, 我们不得不使用数值解法获得均值.

2.4 黑箱变分推断

上面介绍的三种处理非共轭模型的方法各有优缺点. 层次模型和局部变分法十分优雅, 但是适用的范围仍然有限. 拉普拉斯变分推断有坚实的理论基础, 但是有很多限制. 所以这三种方法与实际模型还有一段距离. 最理想的算法是能处理任意的非共轭模型.

实际上, 非共轭模型的难点是公式 (1.7) 没有解析解. 三种方法的思想都是通过变换, 来得到变分分布的解析解, 或解析解的近似. 现在我们脱离信息传递算法的框架, 回到 VI 的目标函数, 即公式 (1.4). 一个自然的想法是通过梯度下降法 (gradient descent) 使 ELBO 最小. 记变分分布为 $q(z | \theta)$, 其中 θ 为变分参数. 证据下界的梯度为

$$\begin{aligned}\nabla_{\theta} \text{ELBO}(\theta) &= \nabla_{\theta} \int [\log p(x, z) - \log q(z | \theta)] q(z | \theta) dz \\ &= \int \nabla_{\theta} \{ [\log p(x, z) - \log q(z | \theta)] q(z | \theta) \} dz \\ &= \int q(z | \theta) \nabla_{\theta} [\log p(x, z) - \log q(z | \theta)] dz \\ &\quad + \int [\log p(x, z) - \log q(z | \theta)] \nabla_{\theta} q(z | \theta) dz \\ &= \mathbb{E}_q \{ [\log p(x, z) - \log q_{\theta}(z | \theta)] \nabla_{\theta} \log q_{\theta}(z | \theta) \}.\end{aligned}\quad (2.59)$$

由于梯度涉及了期望, 通常难以求解, 所以一般使用蒙特卡洛采样来估计梯度, 即

$$\nabla_{\theta} \text{ELBO}(\theta) \approx \frac{1}{S} \sum_{s=1}^S [\log p(x, z_s) - \log q_{\theta}(z_s | \theta)] \nabla_{\theta} \log q_{\theta}(z_s | \theta), \quad \text{where } z_s \sim q(z | \theta). \quad (2.60)$$

这种方法叫做黑箱变分推断 (black-box variational inference, BBVI) [40]. 虽然公式2.60是梯度的无偏估计, 但是直接对变分分布采样的后果是方差过大, 以至于梯度的估计和真实梯度相差太多, 影响收敛. 所以, 我们必须控制方差. 下面我们介绍两种方法.

2.4.1 控制方差

Rao-Blackwellization

Rao-Blackwellization (RB) [41] 利用了条件方差公式

$$\text{Var}(x) = \mathbb{E}[\text{Var}(x | y)] + \text{Var}(\mathbb{E}[x | y]).$$

显然, 两项都是非负的, 所以

$$\text{Var}(x) \geq \text{Var}(\mathbb{E}[x | y]). \quad (2.61)$$

假设我们的兴趣是估计期望 $\alpha = \mathbb{E}(x)$. 由于 $\mathbb{E}(\mathbb{E}[x | y]) = \mathbb{E}(x) = \alpha$, 所以 $\mathbb{E}[x | y]$ 是 $\mathbb{E}(x)$ 的无偏估计, 但是方差变小了. 实际上, RB 的思想是: 观测 y 之后, 我们可以消除部分对 x 采样的随机性, 所以能够更加精确地估计 x .

Control Variates

假设变量 y 与 x 具有强相关关系, 且已知 $\mathbb{E}(y) = \mu$. 我们的兴趣依然是估计 $\alpha = \mathbb{E}(x)$, 则我们构造它的无偏估计,

$$\mathbb{E}[x + c(y - \mu)] = \alpha,$$

其中 c 为任意常数. 该估计量的方差为

$$\text{Var}[x + c(y - \mu)] = \text{Var}(x) + c^2 \text{Var}(y) + 2c \text{Cov}(x, y),$$

为使其方差最小, 通过驻点条件可得

$$c = \frac{-\text{Cov}(x, y)}{\text{Var}(y)}. \quad (2.62)$$

变量 y 称为 x 的控制变量 (control variate) [42]. 使用这个方法的关键是控制变量与我们的兴趣变量是强相关的, 否则估计量的方差并不会明显下降. 如果 $\text{Cov}(x, y)$ 或 $\text{Var}(y)$ 未知, c 可以通过仿真估计,

$$\hat{c} = \frac{\frac{1}{S} \sum_{s=1}^S (x_s - \bar{x})(y_s - \bar{y})}{\frac{1}{S-1} \sum_{s=1}^S (y_s - \bar{y})^2}. \quad (2.63)$$

经验结果表明 \hat{c} 和 c 的效果十分类似.

■**Example 2.1—估计 π .** π 可以由蒙特卡洛进行估计. 首先, 在 $(0, 1) \times (0, 1)$ 的正方形内投点; 进而, 判断点是否在 $1/4$ 单位圆内. 即 $u_i \sim U(0, 1), i = 1, 2$, 且

$$x = \begin{cases} 1, & \text{if } u_1^2 + u_2^2 \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

则 $\mathbb{E}(x) = \pi/4, \text{Var}(x) = (\pi/4)(1 - \pi/4) = 0.1686$.

RB: 考虑 $\mathbb{E}[x | u_1] = p\{u_1^2 + u_2^2 \leq 1 | u_1\} = p\{u_2 \leq \sqrt{1 - u_1^2}\} = \sqrt{1 - u_1^2}$. 估计量 $\sqrt{1 - u_1^2}$ 的均值为 $\pi/4$, 但是方差为 $2/3 - (\pi/4)^2 = 0.0498$. 方差降低了 70.5%.

控制变量: 令 $y = I(u_1 + u_2 \leq \sqrt{2})$, 则 $\mu = 2(\sqrt{2} - 1)$. 实验结果得 $\text{Var}(x + \hat{c}(y - \mu)) = 0.0408$. 方差降低了 76%. ■

2.4.2 黑箱变分推断

我们在本节讨论如何在 BBVI 中使用 RB 和控制变量的技术. 记后验分布

$$q(z | \theta) = \prod_{j=1}^d q(z_j | \theta_j).$$

令 $z_{(j)}$ 表示 z_j 的马尔科夫毯子 (Markov blanket) 中的变量, 即与 z_j 相关的变量, 记它们的联合变分分布为 $q_{(j)}$, 与观测的联合分布为 $p_j(x, z_{(j)})$. 故关于 θ_j 的梯度为

$$\nabla_{\theta_j} \text{ELBO} = \mathbb{E}_{q_{(j)}} \{ [\log p_j(x, z_{(j)}) - \log q_j(z_j | \theta_j)] \nabla_{\theta_j} \log q_j(z_j | \theta_j) \} \quad (2.64)$$

所以, 在计算第 j 个梯度的 RB 估计量为

$$\nabla_{\theta_j} \text{ELBO}(\theta_j) \approx \frac{1}{S} \sum_{s=1}^S [\log p_j(x, z_s) - \log q_j(z_s | \theta_j)] \nabla_{\theta_j} \log q_j(z_s | \theta_j), \quad \text{where } z_s \sim q_{(i)}(z | \theta). \quad (2.65)$$

此外, 我们令控制变量为 $h_j(z) = \nabla_{\theta_j} \log q(z | \theta_j)$. 记我们感兴趣的梯度为 $f_j(z) = [\log p(x, z) - \log q(z | \theta_j)] \nabla_{\theta_j} \log q(z | \theta_j)$, 则 $\hat{c}_j = \frac{\text{Cov}(f_j, h_j)}{\text{Var}(h_j)}$. 进而可得加入控制变量的 RB 估计量

$$\nabla_{\theta_j} \text{ELBO}(\theta_j) \approx \frac{1}{S} \sum_{s=1}^S [\log p_j(x, z_s) - \log q_j(z_s | \theta_j) - \hat{c}_j] \nabla_{\theta_j} \log q_j(z_s | \theta_j), \quad \text{where } z_s \sim q_{(i)}(z | \theta). \quad (2.66)$$

算法6描述了 BBVI 的推断过程.

Algorithm 6 Black-box Variational Inference.

```

1: Initialize variational parameters  $\theta_j, j = 1, 2, \dots, d$ ;
2: repeat
3:   #Draw  $S$  samples from the variational approximation
4:   for  $s = 1$  to  $S$  do
5:      $z[s] \sim q$ 
6:   end for
7:   for  $j = 1$  do
8:     for  $s = 1$  to  $S$  do
9:        $f_j[s] = [\log p_j(x, z[s]) - \log q_j(z[s] | \theta_j)] \nabla_{\theta_j} \log q_j(z[s] | \theta_j)$ 
10:       $h_j[s] = \nabla_{\theta_j} \log q_j(z[s] | \theta_j)$ 
11:    end for
12:     $\hat{c}_j = \frac{\text{Cov}(f_j, h_j)}{\text{Var}(h_j)}$ 
13:     $\nabla_{\theta_j} \text{ELBO} = \frac{1}{S} \sum_{s=1}^S f_j[s] - \hat{c}_j h_j[s]$ 
14:  end for
15:  Compute learning rate,  $e = e_t$ .
16:   $\theta = \theta + e \nabla_{\theta} \text{ELBO}$ .
17: until change of  $\theta$  is small.

```

2.4.3 例子: Gaussian-Exponential model

在这里我们考虑一个潜变量为指数分布, 似然为高斯分布的模型,

$$x_{nk} \sim \mathcal{N}(\mu_k, 1), \mu_k \sim \text{Exponential}(\lambda_0), \text{ for } k = 1, \dots, K, n = 1, \dots, N. \quad (2.67)$$

均值使用指数分布作为先验, 好处是使得 μ_k 非负. 实际上, 很多模型都要求未知参数非负. 比如, Breiman 提出的 non-negative garotte 是一种特殊的稀疏回归, 要求回归系数要么是零, 要么是正数. 另外, 非负矩阵分解也要求两个矩阵成分的所有元素非负. 但是指数分布和高斯分布不是共轭的, 所以可以使用 BBVI 解决.

模型的对数联合分布函数为

$$\log p(x, \mu) = \sum_{n=1}^N \sum_{k=1}^K \log p(x_{nk}, \mu_k) + \log p(\mu_k), \quad (2.68)$$

我们记参数的变分分布为 $q(\mu_k | \lambda_k)$. 由于指数分布要求参数 $\lambda_k \geq 0$, 所以我们使用 soft-plus 函数, $\mathcal{S}(x) = \log(1 + \exp(x))$, 约束变分参数. 即

$$q(\mu_k | \lambda_k) = \text{EXP}(\mathcal{S}(\lambda_k)) = \mathcal{S}(\lambda_k) e^{-\mathcal{S}(\lambda_k) \mu_k}. \quad (2.69)$$

所以, 模型的 ELBO 为

$$l = \mathbb{E}_q \left[\log \frac{p(x, \mu)}{q(\mu)} \right] = \mathbb{E}_q \left[\sum_{n=1}^N \sum_{k=1}^K \log p(x_{nk} | \mu_k) + \log p(\mu_k) - \log q(\mu_k) \right]$$

第 k 个参数的梯度为

$$\nabla_{\lambda_k} l = \mathbb{E}_q[f_k] = \mathbb{E}_q \{ [\log p(x, \mu_k) - \log q(\mu_k | \lambda_k)] \nabla_{\lambda_k} \log q(\mu_k | \lambda_k) \}, \quad (2.70)$$

其中 $\log p(x, \mu_k) = \sum_{n=1}^N \log p(x_{nk} | \mu_k) + \log p(\mu_k)$. 特别地, 我们记

$$h_k = \nabla_{\lambda_k} \log q(\mu_k | \lambda_k) = \frac{\mathcal{S}'(\lambda_k)}{\mathcal{S}(\lambda_k)} - \mathcal{S}'(\lambda_k) \mu_k, \text{ where } \mathcal{S}'(x) = \frac{e^x}{1 + e^x}. \quad (2.71)$$

如果关于 μ_k 的第 s 次采样为 $\mu_k[s]$, 则高斯-指数模型的 BBVI 如下:

Algorithm 7 Black-box Variational Inference for Gaussian-Exponential model.

- 1: Initialize variational parameters $\lambda_k, k = 1, 2, \dots, K$;
 - 2: **repeat**
 - 3: **for** $s = 1$ to S **do**
 - 4: $\mu[s] \sim q(\mu | \lambda)$
 - 5: **end for**
 - 6: **for** $k = 1$ to K **do**
 - 7: **for** $s = 1$ to S **do**
 - 8: $f_k[s] = [\log p(x, \mu_k[s]) - \log q(\mu_k[s] | \lambda_k)] \nabla_{\lambda_k} \log q(\mu_k[s] | \lambda_k)$
 - 9: $h_k[s] = \nabla_{\lambda_k} \log q(\mu_k[s] | \lambda_k)$
 - 10: **end for**
 - 11: $\hat{c}_k = \frac{\text{Cov}(f_k, h_k)}{\text{Var}(h_k)}$
 - 12: $\nabla_{\lambda_k} l = \frac{1}{S} \sum_{s=1}^S f_k[s] - \hat{c}_k h_k[s]$.
 - 13: **end for**
 - 14: Compute learning rate, $e = e_t$.
 - 15: $\lambda = \lambda + e \nabla_{\lambda} l$.
 - 16: **until** change of λ is small.
-



3. 其他变分推断方法

3.1 χ 变分推断

3.1.1 χ 散度与证据上界

χ^n 散度 (χ^n divergence) 定义为

$$D_n(p||q) = \mathbb{E}_{q(x)} \left[\left(\frac{p(x)}{q(x)} \right)^n - 1 \right]. \quad (3.1)$$

与经典的 VI 不同, χ 变分推断 [43] 使用 χ 散度量变分分布与后验分布之间的差异, 即目标函数为

$$\min_{q(z)} D_n(p(z|x)||q(z)). \quad (3.2)$$

与 VI 类似, CHIVI 可以导出证据的上界,

$$\begin{aligned} D_n &= \mathbb{E}_q \left[\left(\frac{p(z|x)}{q(z)} \right)^n \right] - 1 = p(x)^{-n} \mathbb{E}_q \left[\left(\frac{p(z,x)}{q(z)} \right)^n \right] - 1 \\ &\Leftrightarrow p(x)^n (1 + D_n) = \mathbb{E}_q \left[\left(\frac{p(z,x)}{q(z)} \right)^n \right] - 1 \end{aligned} \quad (3.3)$$

对上式两边同时取对数

$$\log p(x) = \frac{1}{n} \log \mathbb{E}_q \left[\left(\frac{p(z,x)}{q(z)} \right)^n \right] - \frac{1}{n} \log(1 + D_n) \quad (3.4)$$

χ 散度是非负的, 所以我们得到了证据的一个上界, 称其为 CUBO (χ upper bound)

$$\log p(x) \leq \frac{1}{n} \log \mathbb{E}_q \left[\left(\frac{p(z,x)}{q(z)} \right)^n \right] \equiv \text{CUBO}_n \quad (3.5)$$

由于 $\log p(x)$ 是常数, 所以最小化 χ 散度相当于最小化 CUBO_n . 下面我们不加证明地给出三明治定理

Theorem 3.1.1—Sandwich Theorem. CUBO 满足以下性质:

1. $\forall n \geq 1, \text{ELBO} \leq \log p(x) \leq \text{CUBO}_n$
2. $\forall n \geq 1, \text{CUBO}_n$ 关于 n 是非降函数
3. $\lim_{n \rightarrow 0} \text{CUBO}_n = \text{ELBO}$

这个定理表明 CUBO_n 与 ELBO 的距离随着 n 的增大而变大. 一般, 我们取 $n = 2$.

3.1.2 优化 CUBO

由于 CUBO 涉及了难以求解的期望, 所以我们使用蒙特卡洛估计. 由于对数操作, 直接使用蒙特卡洛的缺点是无法获得无偏估计. 所以我们关注指数 CUBO,

$$l = \exp(n \cdot \text{CUBO}_n) = \mathbb{E}_q \left[\left(\frac{p(z, x)}{q(z)} \right)^n \right]. \quad (3.6)$$

显然, 优化 l 和优化 CUBO_n 是等价的. 为了控制方差, 我们使用重参数化 (reparameterization) 技巧. 假设存在 g , 使得 $z = g(\lambda, \epsilon)$. 我们对 ϵ 采样, 不直接对 z 进行采样. 所以 $\nabla_\lambda l$ 的无偏蒙特卡洛估计

$$\nabla_\lambda \hat{l} = \frac{n}{S} \sum_{s=1}^S \left[\frac{p(x, g(\lambda, \epsilon[s]))}{q(g(\lambda, \epsilon[s]); \lambda)} \right]^n \nabla_\lambda \log \frac{p(x, g(\lambda, \epsilon[s]))}{q(g(\lambda, \epsilon[s]); \lambda)}. \quad (3.7)$$

这个估计需要在整个数据集上计算似然函数. 如果样本数量很大, 我们可以使用部分样本上的似然代替所有样本的似然. 算法8展示了整个推断过程. 其中 5-6 行为了防止下溢, 使用了 $\exp\text{-sum-log}$ 技巧.

Algorithm 8 χ Variational Inference.

- 1: Initialize variational parameters λ ;
 - 2: **repeat**
 - 3: Draw $\epsilon[s] \sim p(\epsilon)$ and set $z[s] = g(\lambda, \epsilon[s])$
 - 4: Draw a sub-dataset $\{x_{i_1}, \dots, x_{i_M}\}$ and set learning rate e
 - 5: Set $\log w[s] = \log p(z[s]) + \frac{N}{M} \sum_{j=1}^M p(x_{i_j} | z[s]) - \log q(z[s]; \lambda), \forall s \in \{1, \dots, S\}$
 - 6: Set $w[s] = \exp(\log w[s] - \max \log w[s]), \forall s \in \{1, \dots, S\}$
 - 7: Update $\lambda = \lambda - \frac{(1-n)e}{S} \sum_{s=1}^S [(w[s])^n \nabla_\lambda \log q(z[s]; \lambda)]$
 - 8: **until** change of λ is small.
-

3.2 变分提升 (Variational Boosting)

变分提升法 (variational boosting, VBoost) [44, 45] 是最近提出的一种处理非共轭模型的方法. VBoost 的基本思想与 LVI 相同. 如果变分分布的解析解不易获得, 那么就使用一个分布去近似变分分布. 其中 LVI 使用了一个高斯分布. LVI 的缺点是在复杂的概率模型中, 高斯分布的近似精度不够. VBoost 可以改善 LVI 的不足. VBoost 使用混合高斯分布近似变分分布. 由于混合高斯分布可以近似任意连续分布, 所以在理论上 VBoost 可以获得更好的效果.

但是 VBoost 也有缺点. 首先, 如果对每个高斯成分的协方差矩阵不加限制, 则计算量会急剧升高, 因为我们不得不对 $p(p+1)/2$ 个参数进行估计. 如果假设协方差矩阵是对角矩阵, 这么做虽然方便计算, 但是我们不得不使用额外的高斯成分去描述变分分布的相关性, 这么做还可能加剧局部最优的可能性. 作者推荐使用一个折衷的方案: 使用低秩矩阵加对角阵来限制协方差矩阵, 即 $\Sigma = \mathbf{F}\mathbf{F}^\top + \text{diag}(\exp(\mathbf{v}))$, 其中 \mathbf{F} 的秩为 $r(< p)$. 这样做一方面降低了计算量, 另一方面可能描述变分分布的相关性. 然而, r 取多少合适呢? 作者使用了监视法 (monitor), 但是这个方法主观性太强, 浪费人力. 其次, VBoost 不得不面对一个经典的问题——多少个高斯成分是合适的呢? 如果使用过少的高斯成分, VBoost 的效果可能与 LVI 类似; 但是过多的高斯成分极可能导致过拟合. 目前, 还没有合适的方案解决这些小问题.

虽然, VBoost 并不成熟, 但是它是一个强大的工具, 效果不会比 LVI 差, 而且可以处理任意的非共轭模型. 一些数值实验表明它可以有效处理许多复杂模型.



4. 讨论与展望

4.1 应用

变分推断已经应用在了很多领域. 下面我们抛砖引玉, 从 5 个角度简单地介绍变分推断地部分应用:

4.1.1 计算生物学

VI 在计算生物学上有广泛的应用, 特别是在基因数据的处理上. 例如, 基因组关联研究 (genome-wide association studies) [12, 46], 基因调控网络分析 (gene regulatory network analysis) [47], 模体识别 (motif detection) [48], 基因表达分析 (gene expression analysis) [49], 进化树隐马尔科夫模型 (phylogenetic hidden Markov models) [50], 种群遗传学 (population genetics) [51].

4.1.2 计算机视觉

在计算机视觉的初期, 变分推断是重要的工具. 计算机视觉的研究者经常分析大规模高维的图像数据集, 快速的贝叶斯推断已经成功地应用于许多问题中:

1. 去卷积 (blind image deconvolution, BID): Likas 和 Galatsanos 使用变分推断解决了图像的 BID 问题 [52]. Molina et al. 在去卷积任务中把自回归 (autoregression) 作为先验, 可以有效处理图像中的污斑 (blur) [53]. Babacan et al. 在去卷积框架内考虑了 TV (total variation) 先验, 进一步加强了模型的效果 [54]. Tzikas et al. 在模型中加入了核化技巧并考虑稀疏性, 把 t 分布作为先验, 增强了模型的鲁棒性 [55].
2. 目标追踪 (target tracking): Vermaak et al. 提出了目标追踪的变分推断框架 [56], Jin 和 Mokhtarian [57] 提出了变分粒子滤波 (variational particle filter), 实现了多目标追踪. Ba et al. [58] 通过变分 EM 算法实现了在线的多目标追踪.

3. 图像分割 (image segmentation): 近年来, 部分学者使用非参数贝叶斯模型处理图像分割问题. Sudderth 和 Jordan 使用了 Pitman-Yor 过程 [59], Kropotov et al. 在模型中加入了标签频率的限制 [60], Shyr et al. [61] 利用标签信息, 提出了有监督的 Pitman-Yor 过程, Ghosh et al. 使用了中国餐馆过程 (Chinese restaurant processes) [62], Nakamura et al. [63] 利用了层次狄利克雷马尔科夫随机场 (Hierarchical Dirichlet Process Markov Random Fields).

4.1.3 低秩模型

低秩模型在许多领域有重要的应用, 比如可以处理视频的前背景分离问题, 即背景为低秩成分, 而动态的前景是噪声. 常见的低秩模型有矩阵分解和 PCA. Salakhutdinov 和 Mnih 提出了概率低秩矩阵分解 [64], 但是没有使用贝叶斯推断. Bishop 和 Winn 使用了混合贝叶斯 PCA 对非线性图像流形进行变分推断 [65]. Watanabe et al. 研究了指数族上的 PCA 模型 [66]. Ding et al. 认为鲁棒的 PCA 应该把数据集分成 3 部分建模, 即低秩成分 + 高斯噪声 + 稀疏异常点, 并分别使用 VI 和 MCMC 进行推断 [67]. Wang et al. [68] 和 Zhao et al. [69] 分别提出了基于 L_1 损失的鲁棒矩阵分解. Meng and Torre [70] 使用混合高斯分布 (Mixture of Gaussians, MoG) 对噪声建模, 提出了鲁棒的矩阵分解, 但是推断方法为 EM 算法. Zhao et al. 使用类似的思想, 提出了基于 VI 的鲁棒 MoG-PCA. Cao et al. [71] 建议使用混合 PE 分布对噪声建模, 进一步提高了矩阵分解的鲁棒性. Han et al. 和 Chen et al. 考虑图像的局部相关性, 提出了鲁棒结构 PCA 和矩阵分解 [72, 73]. Yong et al. [74] 提出了在线鲁棒矩阵分解, 可以实时处理背景分离问题.

Tan 和 Févotte 考虑了贝叶斯非负矩阵分解 [75], 其中低秩成分的先验为半高斯分布 (half Gaussian distribution), 但是使用了最大后验进行推断. Cemgil [76] 假设似然为泊松, 先验为 Gamma, 提出了变分非负矩阵分解. Hoffman et al. 假设低秩成分的先验为 Gamma 过程, 提出了贝叶斯非参数非负矩阵分解 [77]. Paisley et al. [78] 使用 StoVI 处理了非负矩阵分解, 可以对大规模数据进行推断.

近年来, 低秩模型还应用在了张量分解 (tensor factorization) 的问题上. Zhao et al. [79] 提出了贝叶斯张量分解. Schein et al. 提出了贝叶斯泊松张量分解 [80]. Zhao et al. [81] 考虑了低秩成分 + 高斯噪声 + 稀疏异常点的鲁棒贝叶斯张量分解模型. Luo et al. [82] 考虑了 L_1 损失的贝叶斯张量分解.

4.1.4 变量选择

Lasso 模型是传统的变量选择方法, 它与最小二乘回归的区别是加入了惩罚项

$$\mathbf{w} = \arg \min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \|\mathbf{w}\|_1. \quad (4.1)$$

这个模型对回归系数的潜在假设是拉普拉斯先验. 受到 Lasso 的启发, Themelis et al. [83] 提出了 2 个贝叶斯稀疏回归, 分别通过 Laplace 分布和 t 分布诱导稀疏性. 最近, 他们把这个思想拓展到了组变量选择 (group variable selection) 的问题上 [84]. 这种方法的缺点是对回归系数进行了惩罚, 增加了回归系数的偏, 不是无偏估计.

近年来, 指示模型 (indicator model) 越来越受到研究者的关注. 指示模型把变量选择和回归系数估计分开考虑. 它对每个协变量增加了指示变量 γ , 其中 $\gamma = 1$ 表示变量重要 (回归系数非零), $\gamma = 0$ 表示变量不重要. 记 $\Gamma = \text{diag}(\gamma_1, \dots, \gamma_p)$, 则贝叶斯变量选择模型为 $y = X\Gamma w + \epsilon$. Carbonetto 和 Stephens [12] 讨论了线性回归和逻辑回归中的变量选择. 他们用采样法推断指示变量, 用变分法推断其他变量. Ormerod et al. [11] 研究了线性回归, 并考虑了两种噪声 (高斯分布和拉普拉斯分布), 但是全部变量都使用了变分推断. Ročková [85] 也考虑了线性回归, 他为了解决 EM 变量选择局部优化的缺点, 在 EM 的框架中加入了变分的思想. Subrahmanya and Shin [86] 考虑了线性回归和逻辑回归中的 group 变量选择.

除了回归模型之外, Constantinopoulos et al. 和 Valente et al. [87] 考虑了混合高斯模型的贝叶斯变量选择问题. Fan 和 Bouguila [88] 使用变分贝叶斯研究了 Dirichlet 过程中变量选择问题.

4.1.5 聚类分析

基于分布的聚类方法常假设一个类别的样本来自一个分布. 这种聚类方法相当于一个概率分布的参数估计. 所以变分推断自然可以应用于此. Attias [89] 讨论了混合高斯模型的变分推断. 我们需要指出: 这个模型可以自动确定 K (高斯成分的个数) [90]! 在初始阶段, 如果不知道 K 的大小, 我们可以设置为充分大的数, 记为 K_0 . 在变分推断中, 如果真实的 $K < K_0$, 那么多余的高斯成分会退化为具有相同参数的分布. 从而达到自动确定 K 的效果. 但是 K_0 过大, 算法收敛时可能仍然不能消除多余的高斯成分 [91]. Watanabe [92–94] 研究了混合指数族分布的变分推断.

除了基于分布的聚类方法, 最近, 研究者们逐渐如何使用贝叶斯方法描述距离学习 (distance metric learning) 和子空间聚类 (subspace clustering) [95]. Yang [96] 提出了距离学习的贝叶斯框架. Babagholami [97] 使用 Beta 过程描述了训练数据的隐空间. Ye et al. [98] 对每个样本进行建模, 提出了 Instance Specific metric Subspace 聚类方法.

4.2 开放问题

经过了 30 年的发展, 变分推断已经逐渐应用于各个领域. 但是仍然有部分问题没有被解决:

1. 虽然变分推断的算法研究已经日趋成熟, 但是相比于 MCMC, 变分推断的理论性质还没有深入研究.
2. 变分推断的目标函数是非凸的, 无论是信息传递算法还是黑箱变分推断都只能得到局部最优解.
3. 因为变分分布来自均场变分族, 所以大多数变分推断无法捕捉后验分布之间的结构. 目前已经有部分工作尝试嵌入这种结构, 但是没有很好地解决.
4. 关于非共轭模型: 黑箱变分推断和变分提升可以适用于任意模型, 但是需要对变分分布采样; 其他不需要采样的算法, 如拉普拉斯变分推断等, 不能适用于任意模型. 设计不需要采样且能适用于任意模型的算法仍有难度.

5. 目前变分推断的框架局限于 KL 散度. 然而, 其他的度量可能会产生更好的逼近, χ 变分推断是一个很好的例子. 关于其他度量的理论十分匮乏.
6. 目前模型驱动的深度学习的深度学习是一个重要的研究方向 [99]. 研究者们已经把 ADMM [100] 和 EM [101] 等优化算法与深度学习框架结合. 深度学习和变分推断结合具有十分光明的发展前景.

4.3 总结

我们介绍了变分推断的相关理论 VI 的基本思想是利用均场变分族近似后验分布, 使两者的 KL 最小. 从而把最初的贝叶斯推断问题转化为优化问题. 然后, 我们给出了经典的信息传递算法, 通过循环坐标优化是变分分布和后验分布的 KL 散度最小 (算法1). 针对大规模数据集, 我们介绍了随机变分推断. 它结合了随机优化的思想, 能够有效地获得全局参数的解 (算法2). 此外, 变分的思想可以推广至 EM 算法中. 如果 E 步难以获得局部潜变量的后验分布, 我们可以通过变分法进行优化 (算法5).

上述算法往往局限于条件共轭模型. 如果似然和参数是非共轭的, 经典的变分推断可能会失效. 原因在于公式 (1.7) 没有解析解. 为了解决这个问题, 常用的做法是: (1) 把原模型转化为层次模型, 使似然和参数使共轭的; (2) 使用局部变分法, 引入局部变分参数, 构造 ELBO 的下界, 使下界和参数使共轭的; (3) 对公式 (1.7) 使用 Laplace 近似; (4) 指定变分分布的形式, 通过梯度下降获得变分参数的最优解.

最后, 我们介绍了两种新的变分推断方法. 其中, χ 变分推断使用 χ 散度度量两个分布的差异, 从而获得了证据上界. 通过最小化证据上界获得最优的变分分布. 而变分提升法使用混合高斯分布近似变分分布, 提高了 Laplace 变分推断的效果.

变分推断的工作还有很多, 本文简单介绍了它的发展, 实属冰山一角. 变分推断的潜力巨大, 在贝叶斯统计领域, 逐渐成为替代 MCMC 的重要工具.



Bibliography

- [1] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, Inc., 2006.
- [2] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6):1087–1092, 1953.
- [3] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.
- [4] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 6(6):721–741, 1984.
- [5] Alane. Gelfand and Adrianf. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990.
- [6] Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I. Jordan. An introduction to mcmc for machine learning. *Machine Learning*, 50(1-2):5–43, 2003.
- [7] Elaine Angelino, Matthew James Johnson, and Ryan P Adams. Patterns of scalable bayesian inference. *Foundations & Trends®in Machine Learning*, 9(2-3), 2016.
- [8] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. Introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

- [9] Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations & Trends® in Machine Learning*, 1(12):1–305, 2008.
- [10] David M. Blei, Alp Kucukelbir, and Jon D. Mcauliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.
- [11] John T. Ormerod, Chong You, and Samuel Müller. A variational bayes approach to variable selection. *Electronic Journal of Statistics*, 11(2):3549–3594, 2017.
- [12] Peter Carbonetto and Matthew Stephens. Scalable variational inference for bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Analysis*, 7(1):73–107, 2012.
- [13] Stefan Zeugner and Martin Feldkircher. Benchmark priors revisited: On adaptive shrinkage and the supermodel effect in bayesian model averaging. *IMF Working Papers*, 09(202):1–39, 2009.
- [14] Yongtao Guan and Matthew Stephens. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Annals of Applied Statistics*, 5(3):1780–1815, 2011.
- [15] Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347, 2013.
- [16] Jan Drugowitsch. Variational bayesian inference for linear and logistic regression. *arXiv:1310.5438*, 2013.
- [17] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
- [18] Shunichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.
- [19] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.
- [20] Matthew J. Beal. Variational algorithms for approximate bayesian inference, May 2003.
- [21] D. F. Andrews and C. L. Mallows. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society*, 36(1):99–102, 1974.
- [22] Clo Da Silva Ferreira, Heleno Bolfarine, and Vor H. Lachos. Skew scale mixtures of normal distributions: Properties and estimation. *Statistical Methodology*, 8(2):154–171, 2011.
- [23] Tommi S Jaakkola and Michael I Jordan. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37, 2000.

- [24] Luke Tierney, Robert E. Kass, and Joseph B. Kadane. Fully exponential laplace approximations to expectations and variances of nonpositive functions. *Journal of the American Statistical Association*, 84(407):710–716, 1989.
- [25] David J. C. Mackay. Information-based objective functions for active data selection. *Neural Computation*, 4(4):590–604, 1992.
- [26] Michael Braun and Jon D McAuliffe. Variational inference for large-scale models of discrete choice. *Journal of the American Statistical Association*, 105(489):324–335, 2010.
- [27] George E. P Box and George C Tiao. *Bayesian Inference in Statistical Analysis*. Wiley, 1992.
- [28] Mike West. On scale mixtures of normal distributions. *Biometrika*, 74(3):646–648, 1987.
- [29] Keming Yu and Rana A. Moyeed. Bayesian quantile regression. *Statistics & Probability Letters*, 54(4):437–447, 2001.
- [30] Hideo Kozumi and Genya Kobayashi. Gibbs sampling methods for bayesian quantile regression. *Journal of Statistical Computation & Simulation*, 81(11):1565–1578, 2011.
- [31] Qing Li, Ruibin Xi, and Nan Lin. Bayesian regularized quantile regression. *Bayesian Analysis*, 1(1):1–26, 2010.
- [32] Lin, I Tsung, Lee, C Jack, Yen, and Y. Shu. Finite mixture modelling using the skew normal distribution. *Statistica Sinica*, 17(3):909–927, 2007.
- [33] Tsung I. Lin, Jack C. Lee, and J. Hsieh Wan. Robust mixture modeling using the skew t distribution. *Statistics & Computing*, 17(2):81–92, 2007.
- [34] A. Azzalini. A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, 12(2):171–178, 1985.
- [35] Adelchi Azzalini. Further results on a class of distributions which includes the normal ones. 46(2):199–208, 1986.
- [36] Norbert Henze. A probabilistic representation of the 'skew-normal' distribution. *Scandinavian Journal of Statistics*, 13(4):271–275, 1986.
- [37] Chris C. Holmes and Leonhard Held. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Anal.*, 1(1):145–168, 03 2006.
- [38] David M. Blei and John D. Lafferty. A correlated topic model of science. *Ann. Appl. Stat.*, 1(1):17–35, 06 2007.

- [39] Chong Wang and David M Blei. Variational inference in nonconjugate models. *Journal of Machine Learning Research*, 14(4):1005–1031, 2013.
- [40] Rajesh Ranganath, Sean Gerrish, and David Blei. Black Box Variational Inference. In Samuel Kaski and Jukka Corander, editors, *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 814–822, Reykjavik, Iceland, 22–25 Apr 2014. PMLR.
- [41] Richard P. Waterman and Bruce G. Lindsay. Rao-blackwellisation of sampling schemes. *Biometrika*, 83(1):81–94, 1996.
- [42] Sheldon Ross. *Simulation, Fifth Edition*. Academic Press - Elsevier.
- [43] Adji Bousso Dieng, Dustin Tran, Rajesh Ranganath, John Paisley, and David Blei. Variational inference via χ upper bound minimization. In *Advances in Neural Information Processing Systems 30*, pages 2732–2741. Curran Associates, Inc., 2017.
- [44] Andrew C. Miller, Nicholas J. Foti, and Ryan P. Adams. Variational boosting: Iteratively refining posterior approximations. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 2420–2429, 2017.
- [45] Fangjian Guo, Xiangyu Wang, Kai Fan, Tamara Broderick, and David B. Dunson. Boosting variational inference. *arXiv:1611.05559*, 2016.
- [46] Jason G Mezey, Gabriel E Hoffman, and Benjamin A Logsdon. A variational bayes algorithm for fast and accurate multiple locus genome-wide association analysis. *Bmc Bioinformatics*, 11(1):1–13, 2010.
- [47] Guido Sanguinetti, Neil D Lawrence, and Magnus Rattray. Probabilistic inference of transcription factor concentrations and gene-specific regulatory activities. *Bioinformatics*, 22:2775–2781, 2006.
- [48] ERIC P. XING, WEI WU, MICHAEL I. JORDAN, and RICHARD M. KARP. Logos: A modular bayesian model for de novo motif detection. In *IEEE Computer Society Conference on Bioinformatics*, page 266, 2003.
- [49] O. Stegle, L. Parts, R. Durbin, and J. Winn. A bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eqtl studies. *Plos Computational Biology*, 6(5):e1000770, 2010.
- [50] Vladimir Jojic, Nebojsa Jojic, Chris Meek, Geiger Dan, Adam Siepel, David Haussler, and D. Heckerman. Efficient approximations for learning phylogenetic hmm models from data. *Bioinformatics*, 20 Suppl 1(suppl 1):i161, 2004.

- [51] Anil Raj, Matthew Stephens, and Jonathan K. Pritchard. faststructure: Variational inference of population structure in large snp data sets. *Genetics*, 197(2):573–89, 2014.
- [52] C. L Likas and N. P Galatsanos. A variational approach for bayesian blind image deconvolution. *Signal Processing IEEE Transactions on*, 52(8):2222–2233, 2004.
- [53] Rafael Molina, Javier Mateos, and Aggelos K. Katsaggelos. Blind deconvolution using a variational approach to parameter, image, and blur estimation. *IEEE Transactions on Image Processing*, 15(12):3715–27, 2006.
- [54] S. D. Babacan, R Molina, and A. K. Katsaggelos. Variational bayesian blind deconvolution using a total variation prior. *IEEE Trans Image Process*, 18(1):12–26, 2009.
- [55] D. G. Tzikas, A. C. Likas, and N. P. Galatsanos. Variational bayesian sparse kernel-based blind image deconvolution with student’s-t priors. *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, 18(4):753–64, 2009.
- [56] Jaco Vermaak, Neil D Lawrence, and Patrick Perez. Variational inference for visual tracking. In *Computer Vision and Pattern Recognition*, volume 1, pages 773–780, 2003.
- [57] Yonggang Jin and Farzin Mokhtarian. Variational particle filter for multi-object tracking. pages 1–8, 2007.
- [58] Sileye O Ba, Xavier Alamedapineda, Alessio Xompero, and Radu Horaud. An on-line variational bayesian model for multi-person tracking from cluttered scenes. *Computer Vision and Image Understanding*, 153:64–76, 2016.
- [59] Erik B. Sudderth and Michael I. Jordan. Shared segmentation of natural scenes using dependent pitman-yor processes. In *Neural Information Processing Systems, Vancouver, British Columbia, Canada, December*, pages 1585–1592, 2008.
- [60] Dmitry Kropotov, Dmitry Laptev, A Osokin, and D Vetrov. Variational segmentation algorithms with label frequency constraints. *Pattern Recognition and Image Analysis*, 20(3):324–334, 2010.
- [61] A. Shyr, T. Darrell, M. Jordan, and R. Urtasun. Supervised hierarchical pitman-yor process for natural scene segmentation. 32(14):2281–2288, 2011.
- [62] S. Ghosh, A. B. Ungureanu, E. B. Sudderth, and D. M. Blei. Spatial distance dependent chinese restaurant processes for image segmentation. *Advances in Neural Information Processing Systems*, pages 1476–1484, 2012.
- [63] T. Nakamura, T. Harada, T. Suzuki, and T. Matsumoto. Hdp-mrf: A hierarchical nonparametric model for image segmentation. In *International Conference on Pattern Recognition*, pages 2254–2257, 2012.

- [64] Ruslan Salakhutdinov and Andriy Mnih. Probabilistic matrix factorization. In *International Conference on Neural Information Processing Systems*, pages 1257–1264, 2007.
- [65] Christopher M. Bishop and John M. Winn. Non-linear bayesian image modelling. In *European Conference on Computer Vision*, pages 3–17, 2000.
- [66] Kazuho Watanabe, Shotaro Akaho, Shinichiro Omachi, and Masato Okada. Variational bayesian mixture model on a subspace of exponential family distributions. *IEEE Transactions on Neural Networks*, 20(11):1783–1796, 2009.
- [67] X. Ding, L. He, and L. Carin. Bayesian robust principal component analysis. *IEEE Transactions on Image Processing*, 20(12):3419, 2011.
- [68] Naiyan Wang, Tiansheng Yao, Jingdong Wang, and Dit Yan Yeung. A probabilistic approach to robust matrix factorization. In *European Conference on Computer Vision*, pages 126–139, 2012.
- [69] Q. Zhao, D. Meng, Z. Xu, W. Zuo, and Y. Yan. L1 -norm low-rank matrix factorization by variational bayesian method. *IEEE Transactions on Neural Networks & Learning Systems*, 26(4):825–839, 2015.
- [70] Deyu Meng and Fernando De La Torre. Robust matrix factorization with unknown noise. In *IEEE International Conference on Computer Vision*, pages 1337–1344, 2014.
- [71] Xiangyong Cao, Yang Chen, Qian Zhao, Deyu Meng, Yao Wang, Dong Wang, and Zongben Xu. Low-rank matrix factorization under general mixture noise distributions. In *IEEE International Conference on Computer Vision*, pages 1493–1501, 2016.
- [72] Ningning Han, Yumeng Song, and Zhanjie Song. Bayesian robust principal component analysis with structured sparse component. *Computational Statistics & Data Analysis*, 109:144–158, 2017.
- [73] Y. Chen, X. Cao, Q. Zhao, D. Meng, and Z. Xu. Denoising hyperspectral image with non-i.i.d. noise structure. *IEEE Trans Cybern*, PP(99):1–13, 2017.
- [74] H. Yong, D. Meng, W. Zuo, and L. Zhang. Robust online matrix factorization for dynamic background subtraction. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, PP(99):1–1, 2017.
- [75] Vincent Y. F. Tan and Cédric Févotte. Automatic relevance determination in nonnegative matrix factorization. in *SPARS, (St-Malo, 35(7):1592–1605*, 2009.
- [76] Ali Taylan Cemgil. Bayesian inference for nonnegative matrix factorisation models. *Computational Intelligence & Neuroscience*, 2009(432):785152, 2009.

- [77] Matthew D. Hoffman, David M. Blei, and Perry R. Cook. Bayesian nonparametric matrix factorization for recorded music. In *International Conference on Machine Learning*, pages 439–446, 2010.
- [78] MI Jordan JW Paisley, DM Blei. Bayesian nonnegative matrix factorization with stochastic variational inference. http://www.people.fas.harvard.edu/~airoldi/pub/books/b02.AiroldiBleiEroshevaFienberg2014HandbookMMM/Ch11_MMM2014.pdf, 2014.
- [79] Q. Zhao, L. Zhang, and A Cichocki. Bayesian cp factorization of incomplete tensors with automatic rank determination. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 37(9):1751–63, 2015.
- [80] Aaron Schein, John Paisley, David M Blei, and Hanna Wallach. Bayesian poisson tensor factorization for inferring multilateral relations from sparse dyadic event counts. pages 1045–1054, 2015.
- [81] Qibin Zhao, Guoxu Zhou, Liqing Zhang, Andrzej Cichocki, and Shunichi Amari. Bayesian robust tensor factorization for incomplete multiway data. *IEEE Transactions on Neural Networks*, 27(4):736–748, 2016.
- [82] Qiong Luo, Han Zhi, Xiai Chen, Wang Yao, Deyu Meng, Liang Dong, and Yandong Tang. Tensor rpca by bayesian cp factorization with complex noise. In *IEEE International Conference on Computer Vision*, pages 5029–5038, 2017.
- [83] Konstantinos E. Themelis, Athanasios A. Rontogiannis, and Konstantinos D. Koutroumbas. A variational bayes framework for sparse adaptive estimation. *IEEE Transactions on Signal Processing*, 62(18):4723–4736, 2014.
- [84] Konstantinos E. Themelis, Athanasios A. Rontogiannis, and Konstantinos D. Koutroumbas. Variational bayes group sparse time-adaptive parameter estimation with either known or unknown sparsity pattern. *IEEE Transactions on Signal Processing*, 64(12):3194–3206, 2016.
- [85] Veronika Ročková. Particle em for variable selection. *Journal of the American Statistical Association*, 2017.
- [86] Niranjana Subrahmanya and Yung C. Shin. A variational bayesian framework for group feature selection. *International Journal of Machine Learning & Cybernetics*, 4(6):609–619, 2013.
- [87] C Constantinopoulos, M. K. Titsias, and A Likas. Bayesian feature and model selection for gaussian mixture models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 28(6):1013–1018, 2006.

- [88] Wentao Fan and Nizar Bouguila. Variational learning of a dirichlet process of generalized dirichlet distributions for simultaneous clustering and feature selection. *Pattern Recognition*, 46(10):2754–2769, 2013.
- [89] Hagai Attias. Inferring parameters and structure of latent variable models by variational bayes. In *Proc. Conference on Uncertainty in Artificial Intelligence Proc. Conference on UAI*, pages 21–30, 1999.
- [90] Adrian Corduneanu and Christopher M Bishop. Variational bayesian model selection for mixture distribution. pages 27–34, 2001.
- [91] C. A. Mcgrory and D. M. Titterington. Variational approximations in bayesian model selection for finite mixture distributions. *Computational Statistics & Data Analysis*, 51(11):5352–5367, 2007.
- [92] Kazuho Watanabe and Sumio Watanabe. On variational bayes algorithms for exponential family mixtures. In *International Symposium on Nonlinear Theory and its Applications (NOLTA2005) Bruges, Belgium, October 18-21,, 2005*.
- [93] Kazuho Watanabe and Sumio Watanabe. Stochastic complexities of gaussian mixtures in variational bayesian approximation. *Journal of Machine Learning Research*, 7(2):625–644, 2006.
- [94] K Watanabe and S Watanabe. Stochastic complexities of general mixture models in variational bayesian learning. *Neural Networks*, 20(2):210–219, 2007.
- [95] Brian Kulis. Metric learning: A survey. *Foundations & Trends in Machine Learning*, 5(4), 2013.
- [96] Liu Yang, Rong Jin, and Rahul Sukthankar. Bayesian active distance metric learning. In *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence*, page 442449, 2007.
- [97] Behnam Babagholami-Mohamadabadi, Seyed Mahdi Roostaiyan, Ali Zarghami, and Mahdiah Soleymani Baghshah. Multi-modal distance metric learning: A bayesian non-parametric approach. In *European Conference on Computer Vision*, pages 63–77, 2014.
- [98] HJ Ye, DC Zhan, and Y Jiang. Instance specific metric subspace learning: A bayesian approach. In *AAAI*, page 22722278, 2016.
- [99] Zongben Xu and Jian Sun. Model-driven deep-learning. *National Science Review*, 5(1):22–24, 2018.

-
- [100] yan Yang, Jian Sun, Huibin Li, and Zongben Xu. Deep admm-net for compressive sensing mri. In *Advances in Neural Information Processing Systems 29*, pages 10–18. Curran Associates, Inc., 2016.
- [101] Klaus Greff, Sjoerd Van Steenkiste, and Jrgen Schmidhuber. Neural expectation maximization. In *Advances in Neural Information Processing Systems*.

