第 7 讲：贝叶斯近似推断

张伟平

目录

1.1 贝叶斯大样本性质

Le Cam

Le Cam (1953) 首次证明了后验分布的渐近正态性。

1.1.1 后验相合性

对贝叶斯统计来说, 估计量的大样本性质往往由后验分布的收敛性来刻画: 考虑一列样本 $D^n, n = 1, 2, \ldots,$ 服从模型

$$D^n \sim p_n\left(d^n \mid \theta\right), \theta \in \Theta; \quad \theta \sim \Pi_n$$

以及记此时基于第 n 个样本的后验分布为 $\Pi_n\left(\cdot \mid D^n\right)$:

$$\Pi_n\left(B \mid D^n\right) = \frac{\int_B p_n\left(D^n \mid \theta\right) d\Pi_n(\theta)}{\int_\Theta p_n\left(D^n \mid \theta\right) d\Pi_n(\theta)}, \quad B \subset \Theta$$

如果随着 n 的增加, D^n 提供了 θ 越来越多的信息. 如果真实的 $\theta = \theta_0 \in \Theta$, 则我们可以期待 $\Pi_n\left(\cdot \mid D^n\right)$ 在 Θ 上弱收敛到一个 Dirac 测度 δ_{θ_0}.

记 Θ 为一个度量空间, d_W 为其上的一个测度 (Levy-Prokhorov 或 Wasserstein). 定义 $\rho_n(\theta) := d_W\left(\Pi_n\left(\cdot \mid D^n\right), \delta_\theta\right)$, 其为 D^n 和 θ 的函数.

Definition

设 $D^n \sim p_n\left(\cdot \mid \theta_0\right), n = 1, 2, \ldots$, $\theta \sim \Pi_n$, 则称后验分布 $\Pi_n\left(\cdot \mid D^n\right)$ 在 θ_0 处是 (弱) 相合的, 如果有 $\rho_n\left(\theta_0\right) \to 0$ a.s. (或 in probability)

直接验证 d_w 是比较困难的, 下述引理给出了在额外一些条件下后验相合性的等价描述:

引理 1. 设度量空间 (Θ, d) 是可分的, 则

$$\rho_n\left(\theta_0\right) \to 0 \Longleftrightarrow \Pi_n\left(U^c \mid D^n\right) \to 0 \text{对 } \theta_0 \text{ 的每个开邻域 } U$$

(此处收敛指几乎处处收敛或依概率收敛, 所有的收敛均在 $\theta = \theta_0$ 处计算).

参数空间可数场合

定理 1. 设 $X|\theta \sim f(\cdot|\theta)$, 其中 $\theta \in \Theta = \{\theta_1, \theta_2, \ldots\}$ 可数集，假设 $\theta_t \in \Theta$ 为 θ 的真值。如果先验分布 $\pi(\theta)$ 满足 $\pi(\theta_i) > 0, i = 1, 2, \ldots$ 以及

$$KL(f_{\theta_t} || f_{\theta_i}) = \int f(x|\theta_t) \log \frac{f(x|\theta_t)}{f(x|\theta_i)} dx > 0 \quad \forall i \neq t$$

则对样本 $\mathbf{x} = (x_1, \ldots, x_n)$ 有

$$\lim_{n\to\infty} \pi(\theta_t|\mathbf{x}) = 1 \text{ 以及 } \lim_{n\to\infty} \pi(\theta_i|\mathbf{x}) = 0 \quad \forall i \neq t$$

注 如果 $\theta_t \notin \Theta$, 则后验分布收敛到一个在 Kullback-Leibler 距离意义下距离真实模型最近的一个离散分布。

证明:

$$\begin{aligned}\pi\left(\theta_i|\mathbf{x}\right) &= \frac{\pi\left(\theta_i\right) f\left(\mathbf{x}|\theta_i\right)}{\sum_i \pi\left(\theta_i\right) f\left(\mathbf{x}|\theta_i\right)} \\ &= \frac{\pi\left(\theta_i\right) f\left(\mathbf{x}|\theta_i\right)/f\left(\mathbf{x}|\theta_t\right)}{\sum_i \pi\left(\theta_i\right) f\left(\mathbf{x}|\theta_i\right)/f\left(\mathbf{x}|\theta_t\right)} \\ &= \frac{\exp\left(\log \pi\left(\theta_i\right) + S_i\right)}{\sum_i \exp\left(\log \pi\left(\theta_i\right) + S_i\right)} \quad \text{其中} S_i = \sum_{j=1}^{n} \log \frac{f\left(x_j|\theta_i\right)}{f\left(x_j|\theta_t\right)}\end{aligned}$$

给定 θ_t, S_i 为 i.i.d 随机变量之和，因此由 LLN 有

$$\lim_{n\to\infty} \frac{1}{n} S_i = \int f\left(x|\theta_t\right) \log \frac{f\left(x|\theta_i\right)}{f\left(x|\theta_t\right)} dx$$

由定理假设知此量当 $i \neq t$ 时为负，当 $i = t$ 时为 0. 因此当 $n \to \infty$ 时，$S_t \to 0$, $S_i \to -\infty, i \neq t$。从而定理得证。

参数空间连续场合

上述定理可以推广到参数空间是连续的场合，此时，θ 取任意值的概率为 0. 因此，我们定义

$$\theta_t = \arg\min H(\theta) = \int f_t(x) \log \frac{f_t(x)}{f(x|\theta)} dx$$

定理 2. *设 $X|\theta \sim f(\cdot|\theta)$, 其中 $\theta \in \Theta$ 紧集，假设 $\theta_t \in \Theta$ 为 θ 的真值。A 为 θ_t 的先验概率非零的任意一个邻域 (例如以 θ_t 为中心的一个开集), 则对样本 $\mathbf{x} = (x_1, \ldots, x_n)$ 有*

$$\lim_{n\to\infty} \pi\left(\theta_t \in A|\mathbf{x}\right) = 1, \quad n \to \infty.$$

定理的证明可以通过将参数空间离散化来证明。由于 Θ 为紧集，故存在有限子覆盖，其中仅有 A 包含 θ_t. 于是由离散场合的结论可证。

此时，称后验分布 $\pi(\theta|\mathbf{x})$ 在 θ_t 处相合。

(例 6.1.1) 假设 $X \sim B(n, \theta)$, θ_t 为真值。假设 $\theta \sim Beta(\alpha, \beta)$, 则 θ 的后验分布在 θ_t 处相合。

↑Example

↓Example

(例 6.1.2) 假设 $X_1, \ldots, X_n$ $i.i.d \sim N(\theta, \sigma^2)$, θ_t 为真值。假设 $\theta \sim N(\mu, \tau^2)$, 其中 σ^2, μ, τ^2 为已知常数, 则 θ 的后验分布在 θ_t 处相合。

↑Example

↓Example

当后验相合性成立时候，我们有下述两个重要结论。

定理 3. 设 $\Theta^* \subset \Theta$ 是使得后验相合性在每个 $\theta_0 \in \Theta^*$ 都成立的子集. 则

1. 存在一个估计量 $\hat{\theta}_n = T\left(D^n\right)$ 对每个 $\theta_0 \in \Theta^*$ 都相合, *i.e.*, 对每个 $\theta_0 \in \Theta^*$, 当 $D^n \sim p_n\left(\cdot \mid \theta_0\right)$ 时有 $d\left(\hat{\theta}_n, \theta_0\right) \to 0$ 几乎处处/依概率. 如果后验分布以速度 ϵ_n 压缩到 $\theta_0 \in \Theta^*$, 则 $\epsilon_n^{-1} \cdot d\left(\hat{\theta}_n, \theta_0\right)$ 是几乎处处/依概率有界的.

2. 如果 Θ 是凸集以及 d 是有界且凸的, 则可取 $\hat{\theta}_n$ 为后验均值 $\bar{\theta}_n = \int \theta d\Pi_n\left(\theta \mid D^n\right)$.

后验相合性可以使得从两个不同先验出发的后验是渐近相同的. 下述定理给出了 i.i.d 场合的结论：设在参数空间 Θ 上对所有 n 使用相同的先验 $\Pi_n \equiv \Pi$. Γ 是 Θ 上不同的一个先验. 记 P_Γ^∞ 为样本 $(X_1, X_2, \ldots)$ 在模型 $X_i \overset{\text{ID}}{\sim} f(\cdot \mid \theta), \theta \sim \Gamma$ 下的边际分布. 则

定理 4. *$d_W\left(\Pi\left(\cdot \mid D^n\right), \Gamma\left(\cdot \mid D^n\right)\right) \to 0$ 几乎处处 $[P_\Gamma^\infty]$ 当且仅当 $\Pi\left(\cdot \mid D^n\right)$ 在每个 $\theta \in \operatorname{supp}(\Gamma)$ 几乎处处相合, i.e., $\rho_n(\theta) \to 0$ a.s. P_θ^∞.*

The Schwartz theorem

Schwartz (1965) 对 i.i.d 样本 $D^n = (X_1, \ldots, X_n) \in \mathcal{X}^n$ 场合提供了后验相合性的一个一般且明显的充分条件, 其中 $X_i \overset{\text{IID}}{\sim} f, f \sim \Pi$, 这里先验分布 Π 是概率密度函数的集合 $\mathcal{F}$ 上的一个概率测度 (给定一个控制测度, 这里取 $\mathcal{X}$ 上的 Lebesgue 测度. 令 $d_{\text{KL}}(p, q) = \int p(x) \log\{p(x)/q(x)\}dx$ 表示 Kullback-Leibler divergence. Φ_n 表示 $\mathcal{X}^n \mapsto [0, 1]$ 的任一检验函数.

假设 $f = f_0$ 为真实的密度, 令 P_0^∞ 表示 $(X_1, X_2, \ldots)$ 在 f_0 下的联合密度. 我们称 f_0 属于 Π 的 KL 支撑, 如果对任意 $\epsilon > 0$,

$$\Pi\left(\{f : d_{\text{KL}}\left(f_0, f\right) < \epsilon\}\right) > 0$$

定理 5. 如果 f_0 属于 Π 的 KL 支撑, $U_n \subset \mathcal{F}$ 为 f_0 的邻集且使得存在检验函数 $\Phi_n, n = 1, 2, \ldots$ 满足条件

$$\mathbb{E}_{f_0} \Phi_n \leq Be^{-bn}, \quad \sup_{f \in U_n^c} \mathbb{E}_f \left(1 - \Phi_n\right) \leq Be^{-bn}$$

其中 $b, B > 0$ 为正常数, 则 $\Pi\left(U_n^c \mid D^n\right) \to 0$ 几乎处处 $[P_0^\infty]$.

注: 检验函数 Φ_n 为在样本 D^n 下假设 $H_0 : f = f_0 \leftrightarrow H_1 : f \in U_n^c$ 的检验函数.

推论 1. 若 f_0 属于 Π 的 KL 支撑, 则后验分布在 f_0 处弱相合.

证明. 定理 5 的证明. 由于 $\Phi_n\left(D^n\right) \in [0,1]$ 我们有

$$\Pi\left(U_n^c \mid D^n\right) \leq \Phi_n + \frac{\left(1-\Phi_n\right) \int_{U_n^c} \prod_{i=1}^n \frac{f(X_i)}{f_0(X_i)} d\Pi(f)}{\int_{\mathcal{F}} \prod_{i=1}^n \frac{f(X_i)}{f_0(X_i)} d\Pi(f)}$$

因为 $\mathbb{E}_{f_0} \Phi_n \leq Be^{-bn}$, 由 Borel-Canteli lemma 知对任何 $\beta < b, P_0^\infty\left(\Phi_n > e^{-n\beta} \text{ infinitely often }\right) = 0$, i.e., $\Phi_n \to 0$ a.s. $[P_0^\infty]$ 且指数快的收敛到 0.

为证右边第二项有相同结果, 我们只需证明

1. $\mathbb{E}_{f_0}\left[\left(1-\Phi_n\right) \int_{U_n^c} \prod_{i=1}^n \left\{f\left(X_i\right) / f_0\left(X_i\right)\right\} d\Pi(f)\right] \leq Be^{-bn}$,
2. 对每个 $\beta > 0, e^{n\beta} \int_{\mathcal{F}} \prod_{i=1}^n \left\{f\left(X_i\right) / f_0\left(X_i\right)\right\} d\Pi(f) \to \infty$ 几乎处处 $[P_0^\infty]$.

由 Fubini 定理知第一点成立:

$$\mathbb{E}_{f_0}\left[\left(1-\Phi_n\right) \int_{U_n^c} \prod_{i=1}^n \frac{f\left(X_i\right)}{f_0\left(X_i\right)} d\Pi(f)\right] = \int_{U_n^c} \mathbb{E}_f\left(1-\Phi_n\right) d\Pi(f) \leq Be^{-bn}$$

对第二点, 记 $K = \{f : d_{\mathrm{KL}}(f_0, f) < \beta\}$. 注意到

$$e^{n\beta} \int_{\mathcal{F}} \prod_{i=1}^{n} \frac{f(X_i)}{f_0(X_i)} d\Pi(f) \geq e^{n\beta} \int_{K} \prod_{i=1}^{n} \frac{f(X_i)}{f_0(X_i)} d\Pi(f)$$

以及如果 $d_{\mathrm{KL}}(f_0, f) < \beta$ 则由强大数律

$$e^{n\beta} \prod_{i=1}^{n} \frac{f(X_i)}{f_0(X_i)} = \exp\left\{ n \left(\beta - \frac{1}{n} \sum_{i=1}^{n} \log \frac{f_0(X_i)}{f(X_i)} \right) \right\} \to \infty \text{ a.s. } [P_0^{\infty}]$$

据此, 应用 Fubini 定理, 我们得出 $e^{n\beta} \int_K \prod_{i=1}^n \frac{f(X_i)}{f_0(X_i)} d\Pi(f) \to \infty$ 几乎处处 $[P_0^{\infty}]$.

□

Posterior contraction rates

当后验分布相合时, 可以考虑刻画收敛的速度 (或者该速度的界).

Definition

设 (Θ, d) 为度量空间, $\theta = \theta_0 \in \Theta$ 为真值. 样本 $D^n \sim p_n(\cdot \mid \theta_0)$. 后验分布 $\Pi_n(\cdot \mid D^n)$ 称为以速度 $\epsilon_n \to 0$ (或更快) 在 δ_{θ_0} 处收缩, 如果对每个 $M_n \to \infty$, $\Pi_n(\{\theta : d(\theta, \theta_0) > M_n \epsilon_n \mid D^n\}) \to 0$ 依概率成立.

序列 $M_n \to \infty$ 是为了技术方便性而需要的. 注意 $M_n \to \infty$ 非常慢. 事实上, 在许多问题中 $M_n \equiv M$ 足够了.

考虑 i.i.d 情形, 记 $D^n = (X_1, \ldots, X_n), X_i \overset{\text{IID}}{\sim} f(x_i \mid \theta)$. 令 $K(\theta_0; \theta) = d_{\text{KL}}(f(\cdot \mid \theta_0), f(\cdot \mid \theta)) = \mathbb{E}_{\theta_0} \log\{f(X_1 \mid \theta_0)/f(X_1 \mid \theta)\}$ 以及 $V(\theta_0; \theta) = \mathbb{E}_{\theta_0} \log^2\{f(X_1 \mid \theta_0)/f(X_1 \mid \theta)\}$.

记 $N(\epsilon, \mathcal{P}, d)$. 表示使用半径为 ϵ 的 d-球覆盖集合 $\mathcal{P}$ 所需要的最小个数, 称为是 $\mathcal{P}$ 的 ϵ-覆盖数.

定理 6. 令 $\epsilon_n \to 0$ 使得 $n\epsilon_n^2 \to \infty$, 存在集合 $\Theta_n \subset \Theta, n \geq 1$ 和常数 $c_1, c_2 > 0$ 满足

1. $\log N(\epsilon_n, \Theta_n, d) \leq c_1 n\epsilon_n^2$

2. $\Pi(\Theta_n^c) \leq e^{-(4+c_2)n\epsilon_n^2}$

则后验 $\Pi(\cdot \mid D^n)$ 以速度 ϵ_n 或者更快在每个满足

$$\Pi\left(\left\{\theta : K(\theta_0; \theta) < \epsilon_n^2, V(\theta_0; \theta) < \epsilon_n^2\right\}\right) \geq e^{-c_2 n\epsilon_n^2}$$

的 θ_0 处收缩.

参考文献

1. Ghosal and van der Vaart (2017). Fundamentals of Nonparametric Bayesian Inference (Cambridge Series in Statistical and Probabilistic Mathematics). Cambridge: Cambridge University Press.

设 $(X_i, Y_i) \in \mathcal{X} \times \mathbb{R}, i = 1, \ldots, n$ 满足非参数回归

$$Y_i = f(X_i) + \epsilon_i, \epsilon_i \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$$

其中 $(f, \sigma) \in C(\mathcal{X}) \times \mathbb{R}_+$ 未知. 假设 $\mathcal{X}$ 为 $\mathbb{R}^p$ 的紧集. 考虑下述 f 的高斯过程先验 Π:

$$f \mid \psi \sim \mathrm{GP}(0, C(\cdot, \cdot \mid \psi)), \quad \psi^p \sim Ga(a, b)$$

其中 $C(s, t) = \exp\{-\psi^2 \|s - t\|^2\}$, $\psi > 0$. 假设 σ 的先验密度 H 支撑为 $\mathrm{supp}(H) \subset [c, d]$, 其中 $0 < c < d < \infty$.

↑Example

↓Example

定理 7. 设 $f_0 \in C^\alpha(\mathcal{X})$ 以及 $\sigma_0 \in \mathrm{supp}(H)$ 则 $\Pi(\{(f, \sigma) : d((f, \sigma), (f_0, \sigma_0)) > \epsilon_n\} \mid D^n) \to 0$ 依概率 P_0^∞ 和速度 $\epsilon_n = n^{-1/(2+d/\alpha)} (\log n)^t$ 成立, 其中 $t = 1 - 1/(2 + 4\alpha/d)$.

1.1.2 渐近正态性

定理 8. 在一些正则条件下，后验分布 $\sqrt{n}(\theta-\theta_t)|\mathbf{x}$ 趋于正态分布 $N(0, I(\theta_t)^{-1})$.

证明:假设 θ 为一元,$\hat{\theta}$ 为 MAP 估计,则将对数后验分布 $log\pi(\theta|\mathbf{x})$ 在 $\hat{\theta}$ 处进行 Taylor 展开得到

$$\log \pi(\theta|\mathbf{x}) = \log \pi(\hat{\theta}|\mathbf{x}) + \frac{1}{2}(\theta-\hat{\theta})^2 \frac{d^2}{d\theta^2} \log \pi(\theta|\mathbf{x})\Big|_{\theta=\hat{\theta}} + \dots$$

其中第二项满足

$$\begin{aligned}(\theta-\hat{\theta})^2 \frac{d^2}{d\theta^2} \log \pi(\theta|\mathbf{x})\Big|_{\theta=\hat{\theta}} &= (\theta-\hat{\theta})^2 \frac{d^2}{d\theta^2} \log \frac{\pi(\theta) f(\mathbf{x}|\theta)}{f(\mathbf{x})}\Big|_{\theta=\hat{\theta}} \\ &= (\theta-\hat{\theta})^2 \left(\frac{d^2}{d\theta^2} \log \pi(\hat{\theta}) + \sum_{i=1}^{n} \frac{d^2}{d\theta^2} \log f\left(x_i|\theta\right)\Big|_{\theta=\hat{\theta}} \right)\end{aligned}$$

此时括号中第一项为常数，第二项可以视为是以 x 为变量的 i.i.d 随机变量之和。

由于后验众数 $\hat{\theta}$ 为相合估计, 因此如果 $f_t(x) = f(x|\theta_t)$ 为真实分布,则上式第二项除 n 后收敛到 $-I(\theta_t)$. 否则收敛到 $E_{f_t}\left[\frac{d^2}{d\theta^2}\log f(x|\theta)\right]|_{\theta=\hat{\theta}} < 0$.

因此, Taylor 展开的第二项随 n 的阶增加。类似的, 可以证明高阶项的增加速度不超过 n. 令 $n \to \infty$, 则高阶项可忽略, 从而得证。

定理 9. 设 $X_i \sim f(\cdot|\theta)$, 先验分布为 $\pi(\theta)$。给定数据 $\mathbf{x}$, 当 $n \to \infty$ 时候

- $\theta|\mathbf{x} \approx N(E[\theta|\mathbf{x}], V[\theta|\mathbf{x}])$, 假设 θ 的均值和方差存在
- $\theta|\mathbf{x} \approx N(\hat{\theta}, i(\hat{\theta})^{-1})$, 其中 $\hat{\theta}$ 为后验众数, $i(\theta)$ 为观测信息量, 即 $i(\theta) = -\frac{d^2}{d\theta^2}\log(\pi(\theta|\mathbf{x}))$
- $\theta|\mathbf{x} \approx N\left(\tilde{\theta}, i^\star(\tilde{\theta})^{-1}\right)$, 其中 $\tilde{\theta}$ 为 θ 的 MLE, $i^*(\theta) = -\frac{d^2}{d\theta^2}\log(f(\mathbf{x}|\theta))$.

设 $\theta|x \sim Beta(\alpha+x, \beta+n-x)$，若 n 较大，试逼近该分布。

↑Example
↓Example

(1) 使用后验期望和后验方差，可以得到

$$\theta|x \approx N\left(\frac{\alpha+x}{\alpha+\beta+n}, \frac{(\alpha+x)(\beta+n-x)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)}\right)$$

(2) 注意到后验众数

$$\hat{\theta} = \frac{\alpha+x-1}{\alpha+\beta+n-2}$$

以及

$$\frac{d^2}{d\theta^2}\log(f(\theta|\mathbf{x})) = -\frac{\alpha+x-1}{\theta^2} - \frac{\beta+n-x-1}{(1-\theta)^2}$$

$$i(\hat{\theta}) = \frac{\alpha+x-1}{\hat{\theta}^2} + \frac{\beta+n-x-1}{(1-\hat{\theta})^2}$$

$$= \frac{(\alpha+\beta+n-2)^3}{(\alpha+x-1)(\beta+n-x-1)}$$

所以

$$\theta|\mathbf{x} \approx N\left(\frac{\alpha+x-1}{\alpha+\beta+n-2}, \frac{(\alpha+x-1)(\beta+n-x-1)}{(\alpha+\beta+n-2)^3}\right)$$

(3) 注意到 MLE $\tilde{\theta}=x/n$, 以及 $i^*(\tilde{\theta}=n^3/(x(n-x))$, 因此

$$\theta|x \approx N\left(\frac{x}{n}, \frac{x(n-x)}{n^3}\right)$$

比如 $\alpha=\beta=2, x=20, n=30$, 则 $P(\theta>0.5|x)=0.95993$, 而逼近值分别为 0.9660，0.9786，0.9735.

在一些情况下，上述渐近正态性可能不成立，例如

- θ_t 位于参数空间 Θ 的边界上
- 先验分布在 θ_t 处的质量为 0
- 后验分布是不正常的
- 模型是不可识别的

贝叶斯近似计算

记 $\mathbf{x}$ 为数据，使用**后验分布**对未知量 θ 进行推断时候,

$$p(\theta|\mathbf{x}) = \frac{p(\theta, \mathbf{x})}{p(\mathbf{x})}$$

绝大数情况下，分母并不能分析计算，我们需要逼近后验推断。

- 若要计算后验特征 $E[g(\theta)|\mathbf{x}]$，则使用后验渐近正态性近似该量

$$E[g(\theta)|\mathbf{x}] \approx \int g(\theta)\phi(\theta; \hat{\theta}, I(\hat{\theta})^{-1})d\theta$$

- 找一个与该后验分布“近”的容易计算的分布，即找一个可以计算的分布 $q^*(\theta) \approx p(\theta|\mathbf{x})$，使得

$$E_{q^*(\theta)}\theta \approx E_{p(\theta|x)}\theta.$$

Laplace 逼近和变分法是两种常用的近似方法。

1.2 Laplace 逼近

Laplace approximation 是逼近边际密度的一种一般化方法，它考虑如下形式的积分：

$$I(N;h)=\int e^{-Nh(x)}dx$$

- 在统计问题中 N 为样本量，我们常常感兴趣 N 趋于无穷大时候上述积分的值。
- 这里 $h(x)$ 假设是单峰的，且 $h(x)$ 在其众数 $\hat{x}$ 处有严格正的二阶导数。

Laplace 逼近本质上是 Taylor 展开，将 $h(x)$ 展开到四阶：

$$h(x)\approx h(\hat{x})+\frac{1}{2}h_2(\hat{x})(x-\hat{x})^2+\frac{1}{6}h_3(\hat{x})(x-\hat{x})^3+\frac{1}{24}h_4(\hat{x})(x-\hat{x})^4$$

其中 $h_i(\hat{x})$ 表示 $h(x)$ 在 $\hat{x}$ 处的 i 阶导数，$\hat{x}$ 满足 $h_1(\hat{x})=0$.

将其带入积分中，我们得到

$$
\begin{aligned}
I(N;h) &= \int e^{-Nh(x)}dx \\
&\approx \int \exp\left\{-N[h(\hat{x}) + \frac{1}{2}h_2(\hat{x})(x-\hat{x})^2 + \frac{1}{6}h_3(\hat{x})(x-\hat{x})^3 \right. \\
&\left. + \frac{1}{24}h_4(\hat{x})(x-\hat{x})^4]\right\}dx \\
&= e^{-Nh(\hat{x})} \int e^{-\frac{N}{2}\hat{h}_2 u^2} \exp\left\{-\frac{N}{6}\hat{h}_3 u^3 - \frac{N}{24}\hat{h}_4 u^4\right\}du
\end{aligned}
$$

最后一步我们做了变量代换，并简写了 $\hat{h}_i = h_i(\hat{x})$. 为了利用正态分布的积分，我们进一步将后面的指数项做 Taylor 展开到二阶，得到

$$I(N;h) \approx e^{-Nh(\hat{x})} \int e^{-\frac{N}{2}\hat{h}_2 u^2} \left(1 - \frac{N}{6}\hat{h}_3 u^3 - \frac{N}{24}\hat{h}_4 u^4 + \frac{1}{2}\left(\frac{N^2}{36}\hat{h}_3^2 u^6 + \frac{2N^2}{144}\hat{h}_3\hat{h}_4 u^7 + \frac{N^2}{576}\hat{h}_4^2 u^8\right)\right) du$$

$$= e^{-Nh(\hat{x})} \int e^{-\frac{N}{2}\hat{h}_2 u^2} \left(1 - \frac{N}{24}\hat{h}_4 u^4 + \frac{N^2}{72}\hat{h}_3^2 u^6 + \frac{N^2}{1052}\hat{h}_4^2 u^8\right) du$$

上述积分过程中利用奇数阶积分为 0，剩余项积分利用下述 gamma 积分

$$\int e^{-sx^2} x^{2m} dx = \frac{2m!}{m! 2^{2m}} \pi^{1/2} s^{-(m+1)/2}$$

容易得到

$$I(N;h) \approx e^{-Nh(\hat{x})} \sqrt{2\pi}\sigma N^{-1/2} \left(1 - \frac{\hat{h}_4\sigma^4}{8N} + \frac{5\hat{h}_3^2\sigma^6}{24N}\right) \tag{1.1}$$

$$= e^{-Nh(\hat{x})} \sqrt{2\pi}\sigma N^{-1/2} (1 + O(\frac{1}{N})) \tag{1.2}$$

其中 $\sigma^2 = 1/\hat{h}_2$.

当 $\hat{x}$ 无法分析得到时候，一般使用 $\tilde{x}$, 满足 $\hat{x} - \tilde{x} = O(1/N)$ 来代替。例如，在贝叶斯应用中，常有 $-Nh(\theta) = \ell(\theta) + log\pi(\theta)$，$\ell(\theta)$ 为对数似然函数，$\hat{\theta}$ 为 MAP 估计。当 N 趋于无穷时候，MAP 估计趋于 MLE $\tilde{\theta}$, 因此可以用 MLE 来近似 MAP 计算积分。

贝叶斯计算中应用 在贝叶斯分析中，可以使用 Laplace 逼近方法来逼近所要计算的后验特征. 具体的，假设

$$E^{\pi}(g(\theta)|x) = \frac{\int_{R^k} g(\theta) f(\mathbf{x}|\theta)\pi(\theta)d\theta}{\int_{R^k} f(\mathbf{x}|\theta)\pi(\theta)d\theta}$$

为感兴趣的量，其中 g，f，π 均为 θ 的光滑可微**取正值**的函数.

- 对分子，令 $h^*(\theta) = -(\log g(\theta) + \log f(\mathbf{x}|\theta) + \log \pi(\theta))/N$, $\hat{\theta} = \arg\max_{\theta\in\Theta} h^*(\theta)$, $\Sigma^* = \partial^2 h^*/\partial\theta\partial\theta'$, 则由 Laplace 逼近有

$$I(N;h^*) = \int_{R^k} g(\theta)f(\mathbf{x}|\theta)\pi(\theta)d\theta = \int e^{-Nh^*(\theta)}d\theta$$
$$= e^{-Nh^*(\hat{\theta})}(2\pi)^{k/2}|\Sigma^*|^{1/2}N^{-k/2}(1+O(\frac{1}{N}))$$

- 对分母则令 $h(\theta) = -(\log f(\mathbf{x}|\theta)+\log \pi(\theta))/N$, $\tilde{\theta} = \arg\max_{\theta\in\Theta} h(\theta)$, $\Sigma = \partial^2 h/\partial\theta\partial\theta'$ 则由 Laplace 逼近有

$$I(N;h) = \int_{R^k} f(\mathbf{x}|\theta)\pi(\theta)d\theta = \int e^{-Nh(\theta)}d\theta$$
$$= e^{-Nh(\tilde{\theta})}(2\pi)^{k/2}|\Sigma|^{-1/2}N^{-k/2}(1+O(\frac{1}{N}))$$

  因此

$$E(g(\theta)|x) \approx \frac{g(\hat{\theta})f(\mathbf{x}|\hat{\theta})\pi(\hat{\theta})|\Sigma^*|^{-1/2}}{f(\mathbf{x}|\tilde{\theta})\pi(\tilde{\theta})|\Sigma|^{-1/2}}$$

- 对不完全取正值的函数 $g(\theta)$，若给其加上一个充分大的数后可以只取正值，则上述方法仍适用。

设 $X_1, \ldots, X_n i.i.d \sim B(1, \theta)$, $\theta \sim Beta(\alpha, \beta)$ 为先验分布，求后验均值的 Laplace 近似。

↑Example

↓Example

解记 $X = \sum X_i$, 则 $\theta|X \sim Beta(\alpha + x, \beta + n - x)$. 不妨设 $0 \leq \alpha, \beta \leq 1$, 将后验密度写为

$$p(\theta|x) \propto \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1} \propto \exp(-nh(\theta))$$

其中

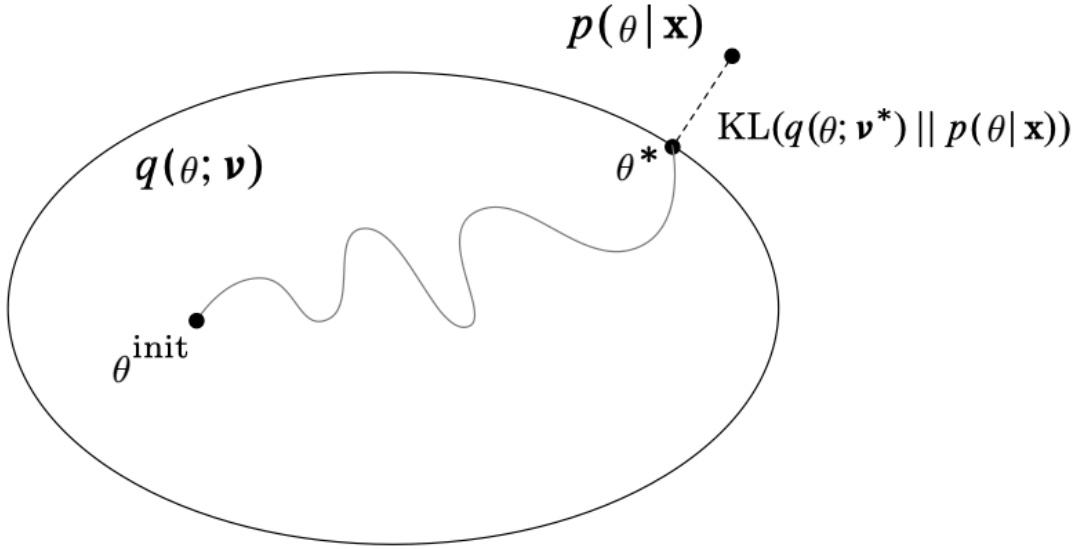
$$h(\theta) = -\frac{1}{n}[(\alpha + x - 1)\log\theta + (\beta + n - x - 1)\log(1-\theta)]$$

可以得到后验均值的 Laplace 近似为

$$\frac{(\alpha+x)^{\alpha+x+1/2}}{(\alpha+x-1)^{\alpha+x-1/2}} \frac{(\alpha+\beta+n-2)^{\alpha+\beta+n-1/2}}{(\alpha+\beta+n-1)^{\alpha+\beta+n+1/2}}$$

1.3 贝叶斯变分推断

变分法 (Variational Inference) 将推断问题转化为优化问题



对隐变量 θ, 假设一个变分分布族 $q(\theta;\nu)$ (容易计算), 选择最优的变分参数 ν^* 使得分布 $q^*(\theta) = q(\theta;\nu^*)$ 在 KL 距离[1]下最接近精确后验分布 $p(\theta|\mathbf{x})$。

[1]也有其他准则, 与 Expectation propagation、Belief propagation 等算法有关

- 变分推断将来自统计物理的想法应用到概率推断。可以认为，Peterson and Anderson (1987) 最早使用平均场 (mean-field) 变分逼近方法来拟合神经网络。
- 在 90 年代早期，很多学者将该想法推广到许多概率模型，Jordan et al. (1999) 综述了这些工作。Hinton and Van Camp (1993) 同时发展了用于神经网络的平均场方法。Neal and Hinton (1993) 将此想法和 EM 算法联系起来，这促进变分法在专家混合系统、隐马氏链 (HMM) 等领域里的应用。
- 现在关于变分推理的新工作层出不穷，使得变分法可扩展、容易实施、快速和准确，以及应用到更多复杂模型和应用场景。
- 现代变分法已经涉及到许多重要领域：概率规划、强化学习、神经网络、凸优化、贝叶斯统计等许多领域

设 p 和 q 为两个分布，称

$$KL(q||p) = \int q(x) \log \frac{q(x)}{p(x)} dx = E_q \log \frac{q}{p}$$

Definition

为 q 到 p 的 **Kullback-Leibler distance** 或者相对熵。

- $KL(q||p) \neq KL(p||q)$
- $KL(q||p) \geq 0$, $KL(q||p) = 0 \Leftrightarrow q = p$

变分法就是在指定的分布族 Q 中求

$$q^* = \arg\min_{q \in Q} KL(q||p)$$

Evidence lower bound (ELBO) 注意到

$$
\begin{aligned}
KL(q||p) &= \int q(\theta) \log \frac{q(\theta)}{p(\theta|\mathbf{x})} d\theta \\
&= \int q(\theta) \log \frac{q(\theta)p(\mathbf{x})}{p(\mathbf{x}|\theta)p(\theta)} d\theta \\
&= \log p(\mathbf{x}) - \int q(\theta) \log \frac{p(\mathbf{x}|\theta)p(\theta)}{q(\theta)} d\theta \\
&= \underbrace{\log p(\mathbf{x})}_{\log(evidence)} - \underbrace{\int q(\theta) \log \frac{p(\mathbf{x}|\theta)p(\theta)}{q(\theta)} d\theta}_{ELBO(q)}
\end{aligned}
$$

由 $KL(q||p) \geq 0$ 知道 $\log p(\mathbf{x}) \geq ELBO(q)$, 从而等价地,

$$
q^* = \arg\max_{q \in Q} ELBO(q)
$$

我们选择一个变分分布 q 的分布族, 例如参数分布族 $Q = \{q(\theta; \nu), \nu \in \Xi\}$, 使得期望是可以计算的

平均场变分推断

平均场变分推断 (Mean-field variational inference,MFVI) 就是考虑一类满足如下条件的分布族

$$Q = \left\{ q : q(\theta) = \prod_{j=1}^{J} q_j(\theta_j) \right\}$$

即认为 $\theta = (\theta_1, \ldots, \theta_J)$ 的各个分量是相互独立的。这样做允许我们简化 $ELBO(q)$。事实上

$$\begin{aligned} ELBO(q) &:= \int q(\theta) \log \frac{p(\mathbf{x}, \theta)}{q(\theta)} d\theta \\ &= \int \left[\prod_{i=1}^{J} q_i \right] \left[\log(p(\mathbf{x}, \theta)) - \sum_{k=1}^{J} \log(q_k) \right] d\theta \end{aligned}$$

$$
\begin{aligned}
&\simeq \int q_j \int \left[\prod_{i\neq j} q_i\right] \log(p(\mathbf{x},\theta)) d\theta_{i\neq j} d\theta_j - \int q_j \log\left(q_j\right) d\theta_j \\
&\simeq \int q_j \log\left(\tilde{p}\left(\theta_j\right)\right) d\theta_j - \int q_j \ln\left(q_j\right) d\theta_j \\
&= \int q_j \log\left(\frac{\tilde{p}\left(\theta_j\right)}{q_j}\right) d\theta_j = -KL(q_j \| \tilde{p})
\end{aligned}
$$

其中 $\simeq$ 表示至多差一个可加常数。而 $\tilde{p}$ 满足

$$
\log\left(\tilde{p}\left(\theta_j\right)\right) \simeq \int \left[\prod_{i\neq j} q_i\right] \ln(p(\mathbf{x},\theta)) d\theta_{i\neq j} = \mathbb{E}_{q_{i\neq j}} \ln(p(\mathbf{x},\theta))
$$

因此，使用坐标下降算法，对 $j = 1, \ldots, J$ 有

$$
\begin{aligned}
q_j^* &= \arg\max_{q_j} ELBO(q) = \arg\min_{q_j} KL(q_j \| \tilde{p}) = \tilde{p}\left(\theta_j\right) \\
&\propto \exp\left(E_{q_{i\neq j}} \ln(p(\mathbf{x},\theta))\right)
\end{aligned} \tag{1.3}
$$

如果 $q(\theta) = \prod_{j=1}^{J} q_j(\theta_j; \nu_j)$, 以及

$$p(\mathbf{x}, \theta) = p(\mathbf{x}) \prod_{j=1}^{J} p(\theta_j | \theta_{1:(j-1)}, \mathbf{x})$$

那么

$$ELBO(q) = E_q \log \frac{p(\mathbf{x}, \theta)}{q(\theta; \nu)} = p(\mathbf{x}) + \sum_{j=1}^{J} E_q \log \frac{p(\theta_j | \theta_{1:(j-1)}, \mathbf{x})}{q_j(\theta_j; \nu_j)}$$

从而对 $\nu_1, \ldots, \nu_J$ 最大化等价于

$$\nu_j^* = \arg\max_{\nu_j} E_q \log \frac{p(\theta_j | \theta_{1:(j-1)}, \mathbf{x})}{q_j(\theta_j; \nu_j)}, j = 1, \ldots, J$$

变分推断的一般算法 给定数据 X 和联合似然 $p(X, \theta_1, \ldots, \theta_m)$，求变分近似 $q(\theta; \psi) = \prod_{i=1}^{m} q(\theta_i|\psi_i)$ 的步骤如下：

- 1. 对模型 $i = 1, \ldots, m$, 令

$$q(\theta_i|\psi_i) = \frac{e^{\mathbb{E}_{q_{j\neq i}}[\ln p(X,\theta_1,\ldots,\theta_m)]}}{\int e^{\mathbb{E}_{q_{j\neq i}}[\ln p(X,\theta_1,\ldots,\theta_m)]} d\theta_i}$$

- 计算变分目标函数 ELBO，

$$\mathcal{L}_t = \mathbb{E}_q[\ln p(X, \theta_1, \ldots, \theta_m)] - \sum_{i=1}^{m} \mathbb{E}_{q_i}[\ln q(\theta_i|\psi_i)].$$

- 如果 $\mathcal{L}_t$ 相比于 $\mathcal{L}_{t-1}$ 增加“很小”，则停止；否则继续下一次迭代。

下面我们考虑直接计算 ELBO 的一个例子。

设 $y_i \ i.i.d \sim N(x_i'w, \alpha^{-1}), i = 1, \ldots, N, \ w \sim N(0, \lambda^{-1}I), \alpha \sim Gamma(a, b)$

↑Example

↓Example

似然为

$$p(y, w, \alpha|x) = p(\alpha)p(w)\prod_{i=1}^{N} p\left(y_i|x_i, w, \alpha\right)$$

为了逼近后验分布，我们使用如下分布

$$q(w, \alpha) = q(\alpha)q(w) = Gamma\left(\alpha|a', b'\right) N\left(w|\mu', \Sigma'\right)$$

从而 ELBO 为

$$\begin{aligned}\mathcal{L}\left(a', b', \mu', \Sigma'\right) &= \int_0^{\infty} \int_{\mathbb{R}^d} q(w, \alpha) \ln \frac{p(y, w, \alpha | X)}{q(w, \alpha)} dw d\alpha \\ &= \int q(\alpha) \ln p(\alpha) d\alpha + \int q(w) \ln p(w) dw \\ &\quad + \sum_{i=1}^{N} \iint q(\alpha) q(w) \ln p\left(y_i | x_i, w, \alpha\right) dw d\alpha \\ &\quad - \int q(\alpha) \ln q(\alpha) d\alpha - \int q(w) \ln q(w) dw\end{aligned}$$

可以看出，上述计算依赖于 $E\alpha, E\ln\alpha, Ew, Eww'$. 经过计算得到

$$\begin{aligned}
\mathcal{L}\left(a', b', \mu', \Sigma'\right) &= (a-1)\left(\psi\left(a'\right) - \ln b'\right) - b\frac{a'}{b'} + \text{ constant} \\
&- \frac{\lambda}{2}\left(\mu'^T\mu' + \operatorname{tr}\left(\Sigma'\right)\right) + \text{ constant} \\
&+ \frac{N}{2}\left(\psi\left(a'\right) - \ln b'\right) - \sum_{i=1}^{N}\frac{1}{2}\frac{a'}{b'}\left(\left(y_i - x_i^T\mu'\right)^2 + x_i^T\Sigma' x_i\right) + \text{ constant} \\
&+ a' - \ln b' + \ln\Gamma\left(a'\right) + \left(1-a'\right)\psi\left(a'\right) \\
&+ \frac{1}{2}\ln\left|\Sigma'\right| + \text{ constant}
\end{aligned}$$

其中 $\psi(x) = d\Gamma(x)/dx$.

方法一：对 ELBO 函数进行最大化，即得到 a', b', μ', Σ'.

方法二：对 $q(\alpha)$, 由一般算法

$$\begin{aligned}
q(\alpha) &\propto \exp\left\{\mathbb{E}_{q(w)}[\ln p(y|x,\alpha,w) + \ln p(\alpha) + \ln p(w)]\right\} \\
&\propto \exp\left\{\mathbb{E}_{q(w)}[\ln p(y|x,\alpha,w)] + \ln p(\alpha)\right\}
\end{aligned}$$

注意到期望只对 w 进行，舍弃无关项得到

$$q(\alpha) \propto \exp\left\{\sum_{i=1}^{N} \mathbb{E}_{q(w)}\left[\ln p\left(y_i|x_i,\alpha,w\right)\right]\right\} p(\alpha)$$

$$\propto \left[\prod_{i=1}^{N} \alpha^{\frac{1}{2}} \mathrm{e}^{-\frac{\alpha}{2}\mathbb{E}_{q(w)}\left[\left(y_i - x_i^T w\right)^2\right]}\right] \alpha^{a-1} \mathrm{e}^{-b\alpha}$$

从形式上可以看出来，

$$q(\alpha) = \mathrm{Gamma}\left(\alpha|a',b'\right), \quad a' = a + \frac{N}{2}, \quad b' = b + \frac{1}{2}\sum_{i=1}^{N} \mathbb{E}_{q(w)}\left[\left(y_i - x_i^T w\right)^2\right]$$

对 $q(w)$，类似有

$$
\begin{aligned}
q(w) &\propto \exp\left\{\mathbb{E}_{q(\alpha)}[\ln p(y|x,\alpha,w) + \ln p(\alpha) + \ln p(w)]\right\} \\
&\propto \exp\left\{\mathbb{E}_{q(\alpha)}[\ln p(y|x,\alpha,w)] + \ln p(w)\right\} \\
&\propto \exp\left\{\sum_{i=1}^{N} \mathbb{E}_{q(\alpha)}\left[\ln p\left(y_i|x_i,\alpha,w\right)\right]\right\} p(w) \\
&\propto \left[\prod_{i=1}^{N} \mathrm{e}^{\frac{1}{2}\mathbb{E}[\ln\alpha]} \mathrm{e}^{-\left(\mathbb{E}_{q(\alpha)}[\alpha]/2\right)\left(y_i - x_i^T w\right)^2}\right] \mathrm{e}^{-\frac{\lambda}{2} w^T w}
\end{aligned}
$$

从而可以看出来

$$
q(w) = N\left(w|\mu', \Sigma'\right)
$$

$$
\Sigma' = \left(\lambda I + \mathbb{E}_{q(\alpha)}[\alpha] \sum_{i=1}^{N} x_i x_i^T\right)^{-1}, \quad \mu' = \Sigma' \left(\mathbb{E}_{q(\alpha)}[\alpha] \sum_{i=1}^{N} y_i x_i\right)
$$

此时，可以计算 $q(\alpha)$ 和 $q(w)$ 中的期望

$$\mathbb{E}_{q(\alpha)}[\alpha] = a'/b'$$

$$\mathbb{E}_{q(w)}\left[\left(y_i - x_i^T w\right)^2\right] = \left(y_i - x_i^T \mu'\right)^2 + x_i^T \Sigma' x_i$$

从而，本例的变分算法可以表示为
变分分布为 $q(\alpha) = \text{Gamma}\left(\alpha|a', b'\right)$ and $q(w) = N\left(w|\mu', \Sigma'\right)$:

- 1. 初始化 $a'_0, b'_0, \mu'_0, \Sigma'_0$
- 2. 对循环 $t = 1, \ldots, T$:
  - 更新 $q(\alpha)$:

$$a'_t = a + \frac{N}{2}$$

$$b'_t = b + \frac{1}{2}\sum_{i=1}^{N}\left(y_i - x_i^T \mu'_{t-1}\right)^2 + x_i^T \Sigma'_{t-1} x_i$$

– 更新 $q(w)$:

$$\Sigma_t' = \left(\lambda I + \frac{a_t'}{b_t'} \sum_{i=1}^{N} x_i x_i^T\right)^{-1}$$

$$\mu_t' = \Sigma_t' \left(\frac{a_t'}{b_t'} \sum_{i=1}^{N} y_i x_i\right)$$

– 评估 ELBO $\mathcal{L}\left(a_t', b_t', \mu_t', \Sigma_t'\right)$ 是否达到收敛

主题模型 Latent Dirichlet Allocation [2](Blei et al., 2003) 是一种生成模型，其将文档视为一些主题的混合，其假设有 K 个主题，一个由 D 个文档组成的语料库 (corpus)，以及包含 V 个不同词的词典

↑Example

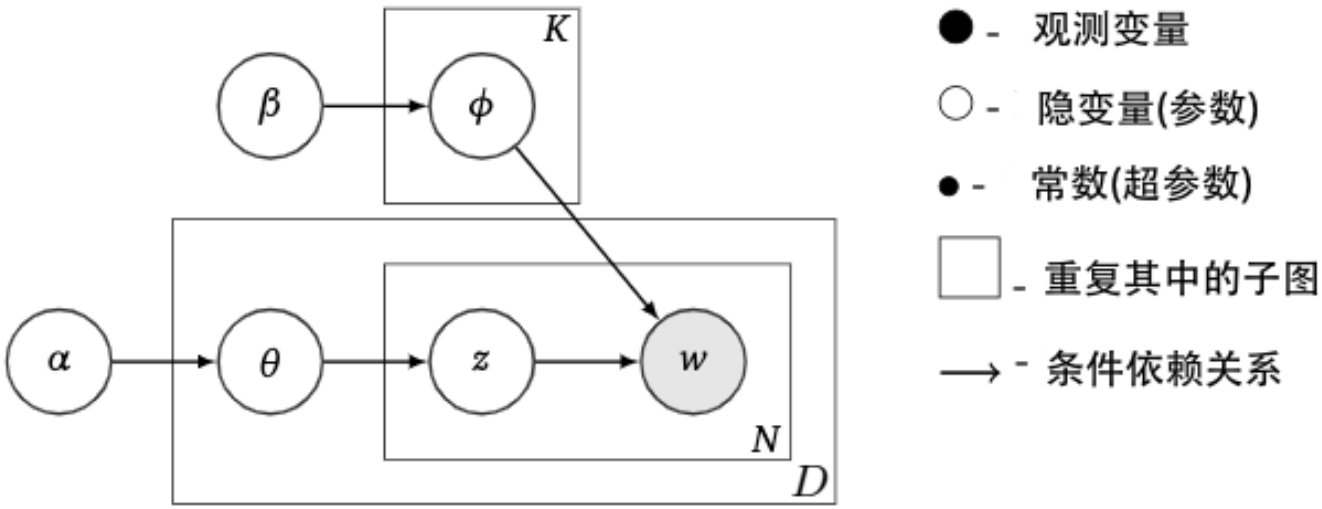
↓Example

1. for $k = 1, \ldots, K$
   (a) Draw $\phi_k \sim \text{Dir}(\beta)$
2. for $d = 1, \ldots, D$
   (a) Draw $\theta_d \sim \text{Dir}(\alpha)$
   (b) for $t = 1, \ldots, N_d$
      - Draw $z_{d,t} \sim Mult(\theta_d)$
      - Draw $w_{d,t} \sim Mult(\phi_{z_{d,t}})$

- V: 词典中词个数
- D: 文档个数
- N_d: 文档 d 中词个数
- $w_{d,t}$: 文档 d 中的第 t 个词.
- $z_{d,t}$: 文档 d 中第 t^{th} 个主题
- K: 主题的个数

[2] Latent dirichlet allocation. J. Mach. Learn. Res., 3:993–1022, March 2003.

上述模型的图模型表示如下



LDA 假设 Unigram model, 即文档之间相互独立, 文档中的词汇相互独立。一篇文档, 可以看成是一组有序的词的序列 $d = (w_1, \ldots, w_{N_d})$. 因此, 我们得到主题混合参数 $\Theta_{D\times K}$, 主题标签 $\mathbf{z}$, 语料库中的词 $\mathbf{w}$ 和主题 $\Phi_{K\times V}$ 的联合分布

$$P(\mathbf{w}, \mathbf{z}, \Theta, \Phi|\alpha, \beta) = \prod_{d=1}^{D} p(\mathbf{w}_d, \mathbf{z}_d, \theta_d, \Phi|\alpha, \beta)$$

$$= \prod_{d=1}^{D} \left\{ P\left(\theta_d|\alpha\right) \prod_{t=1}^{N_d} P\left(z_{d,t}|\theta_d\right) P\left(w_{d,t}|\phi_{z_{d,t}}\right) \times \prod_{i=1}^{K} P\left(\phi_k|\beta\right) \right\} \tag{1.4}$$

我们感兴趣的是观测到词 $\mathbf{w}$ 后的后验分布

$$P(\mathbf{z}, \Theta, \Phi|\mathbf{w}, \alpha, \beta) = \frac{P(\mathbf{w}, \mathbf{z}, \Theta, \Phi|\alpha, \beta)}{P(\mathbf{w}|\alpha, \beta)}$$

其中

$$\begin{aligned} P(\mathbf{w}|\alpha, \beta) &= \int_{\Phi} \int_{\Theta} \sum_{\mathbf{z}} P(\mathbf{w}, \mathbf{z}, \Theta, \Phi|\alpha, \beta) d\Theta d\Phi \\ &= \int_{\Phi} p(\Phi|\beta) \int_{\Theta} p(\Theta|\alpha) \sum_{\mathbf{z}} p(\mathbf{z}|\Theta) p(\mathbf{w}|\mathbf{z}, \Phi) d\Theta d\Phi \end{aligned}$$

显然是不能得出的。因此需要考虑近似计算。

Blei et al.(2003) 使用变分贝叶斯推断方法，考虑如下变分分布

$$q(\Phi, \mathbf{z}, \Theta|\lambda, \pi, \gamma) = \prod_{d=1}^{D} \{q_d(\theta_d, \mathbf{z}_d, \Phi|\gamma_d, \pi_d, \lambda)\}$$
$$= \prod_{d=1}^{D} \left\{ q_d(\theta_d|\gamma_d) \prod_{t=1}^{N_d} q_d(z_{d,t}|\pi_{d,t}) \prod_{i=1}^{K} Dir(\phi_i|\lambda_i) \right\} \tag{1.5}$$

其中 $\gamma_d, \pi_d = (\pi_{d,1}, \ldots, \pi_{d,N_d})$ 为文档 d 的变分分布 $q_d(\cdot)$ 的参数

我们考虑直接计算和利用 (1.3) 两种方法求解。

（一）直接最大化 ELBO：

注意到对 LDA 模型 (1.3) 和变分分布 (1.5) 有

$$ELBO(q) = E_q \log \frac{p(\mathbf{w}, \mathbf{z}, \Theta, \Phi|\alpha, \beta)}{q(\mathbf{z}, \Theta, \Phi|\lambda, \pi, \gamma)} = \sum_{d=1}^{D} E_q \log \frac{p(\mathbf{w}_d, \mathbf{z}_d, \theta_d, \Phi|\alpha, \beta)}{q(\theta_d, \mathbf{z}_d, \Phi|\gamma_d, \pi_d, \lambda)}$$

以及

$$E_q \log \frac{p(\mathbf{w}_d, \mathbf{z}_d, \theta_d, \Phi|\alpha, \beta)}{q(\theta_d, \mathbf{z}_d, \Phi|\gamma_d, \pi_d, \lambda)} = E_d \log \frac{p_d(\mathbf{w}_d, \mathbf{z}_d, \theta_d|\Phi, \alpha)}{q_d(\theta_d, \mathbf{z}_d|\gamma_d, \pi_d)} + E_q \log \frac{p(\Phi|\beta)}{q(\Phi|\lambda)}$$

其中 E_d 表示在 $q_d = q_d(\theta_d, \mathbf{z}_d|\gamma_d, \pi_d)$ 分布下计算期望。因此，我们可以关注于第 d 个文档，注意到

$$\begin{aligned} E_d \log \frac{p_d(\mathbf{w}_d, \mathbf{z}_d, \theta_d|\Phi, \alpha)}{q_d(\theta_d, \mathbf{z}_d|\Phi, \gamma_d, \pi_d)} &= E_d \log p_d(\mathbf{w}_d, \mathbf{z}_d, \theta_d|\Phi, \alpha) \\ &\quad - \log q_d(\theta_d, \mathbf{z}_d|\Phi, \gamma_d, \pi_d) \\ &= E_d[\log p_d(\theta_d|\alpha) + p_d(\mathbf{z}_d|\theta_d) + p(\mathbf{w}_d|\mathbf{z}_d, \Phi) \\ &\quad - \log q_d(\theta_d|\gamma_d) - \log q(\mathbf{z}_d|\pi_d)] := l(\gamma_d, \pi_d|\Phi, \alpha) \end{aligned}$$

我们分别计算各项。

(1) 首先 $\theta_d|\alpha \sim Dir(\alpha)$, 在 q_d 分布下 $\theta|\gamma \sim Dir(\gamma)$, 因此

$$
\begin{aligned}
E_d \log p_d(\theta_d|\alpha) &\propto \sum_{i=1}^{K}(\alpha_i - 1)E_d[\log \theta_{d,i}] \\
&= \sum_{i=1}^{K}(\alpha_i - 1)\left[\Psi(\gamma_{d,i}) - \Psi\left(\sum_{l=1}^{K}\gamma_{d,l}\right)\right]
\end{aligned}
$$

其中 $\Psi(x) = d\log\Gamma(x)/dx$.

(2) 其次,

$$
\begin{aligned}
E_d p_d(\mathbf{z}_d|\theta_d) &= E_d\left[\sum_{t=1}^{N_d}\sum_{i=1}^{K} 1\left(z_{d,t} = i\right)\log\theta_{d,i}\right] \\
&= \sum_{t=1}^{N_d}\sum_{i=1}^{K} E_d\left[1\left(z_{j,t} = i\right)\right] E_q\left[\log\theta_{d,i}\right] \\
&= \sum_{t=1}^{N_d}\sum_{i=1}^{K} \pi_{d,t,i}\left(\Psi\left(\gamma_{d,i}\right) - \Psi\left(\sum_{l=1}^{K}\gamma_{d,l}\right)\right)
\end{aligned}
$$

(3) 对第三项有

$$
\begin{aligned}
E_d\left[\log p\left(w_i|z_j,\Phi\right)\right] &= E_d\left[\sum_{t=1}^{N_d}\sum_{i=1}^{K}\sum_{r=1}^{V}1\left(z_{d,t}=i\right)1\left(w_{d,t}=r\right)\log\phi_{i,r}\right] \\
&= \sum_{t=1}^{N_d}\sum_{i=1}^{K}\sum_{r=1}^{V}E_q\left[1\left(z_{d,t}=i\right)1\left(w_{d,t}=r\right)\log\phi_{i,r}\right] \\
&= \sum_{t=1}^{N_d}\sum_{i=1}^{K}\sum_{r=1}^{V}\pi_{d,t,i}1\left(w_{d,t}=r\right)\log\phi_{i,r}
\end{aligned}
$$

(4) 第四项类似可得

$$E_d\left[\log q\left(\theta_d|\gamma_d\right)\right] = \log\Gamma\left(\sum_{i=1}^{K}\gamma_{d,i}\right) - \sum_{i=1}^{K}\log\Gamma\left(\gamma_{d,i}\right) + \sum_{i=1}^{K}\left(\gamma_{d,i}-1\right)\left(\Psi\left(\gamma_{d,i}\right) - \Psi\left(\sum_{l=1}^{K}\gamma_{d,l}\right)\right)$$

(5) 第五项可以得到

$$
\begin{aligned}
E_d\left[\log q\left(z_d|\pi_d\right)\right] &= E_d\left[\sum_{t=1}^{N_d}\sum_{i=1}^{K} 1\left(z_{d,t}=i\right)\log \pi_{d,t,i}\right] \\
&= \sum_{t=1}^{N_d}\sum_{i=1}^{K} E_d\left[1\left(z_{d,t}=i\right)\right]\log \pi_{d,t,i} \\
&= \sum_{t=1}^{N_d}\sum_{i=1}^{K} \pi_{d,t,i}\log \pi_{d,t,i}
\end{aligned}
$$

因此

$$
\begin{aligned}
l(\gamma_d, \pi_d | \alpha, \Phi) \propto & \sum_{i=1}^{K} (\alpha_i - 1) \left(\Psi(\gamma_{d,i}) - \Psi\left(\sum_{l=1}^{K} \gamma_{d,l} \right) \right) \\
& + \sum_{t=1}^{N_d} \sum_{i=1}^{K} \pi_{d,t,i} \left(\Psi(\gamma_{d,i}) - \Psi\left(\sum_{l=1}^{K} \gamma_{d,l} \right) \right) \\
& + \sum_{t=1}^{N_d} \sum_{i=1}^{K} \sum_{r=1}^{V} \pi_{d,t,i} 1(w_{d,t} = r) \log \phi_{i,r} - \log \Gamma\left(\sum_{i=1}^{K} \gamma_{d,i} \right) \\
& + \sum_{i=1}^{K} \log \Gamma(\gamma_{d,i}) - \sum_{i=1}^{K} (\gamma_{d,i} - 1) \left(\Psi(\gamma_{d,i}) - \Psi\left(\sum_{l=1}^{K} \gamma_{d,l} \right) \right) \\
& - \sum_{t=1}^{N_d} \sum_{i=1}^{K} \pi_{d,t,i} \log \pi_{d,t,i}
\end{aligned}
$$

注意约束条件 $\sum_{l=1}^{K} \pi_{d,t,l} = 1$, 因此使用 Lagrange 乘子法我们可以得到

$$\pi_{d,t,i} \propto \phi_{i,w_{d,t}} \exp\left\{\Psi\left(\gamma_{d,i}\right) - \Psi\left(\sum_{l=1}^{K} \gamma_{d,i}\right)\right\}$$

以及

$$\gamma_{d,i} = \alpha_i + \sum_{t=1}^{N_d} \pi_{d,t,i}$$

类似对 Φ 分布中的参数，注意可以得到

$$\lambda_{i,v} = \beta_v + \sum_{d=1}^{D} \sum_{t=1}^{N_d} \pi_{d,t,v} 1\left(w_{d,t} = v\right)$$

而超参数 α, β 可以通过变分 EM 算法估计（称为经验 Bayes 估计，见 Blei et al. 2003)

（二）利用 (1.3) 式：

注意对 LDA 模型，有 (省略对 α, β 的依赖)：

$$
\begin{aligned}
\log(p(\theta, \mathbf{z}, \Phi, \mathbf{w})) = & \sum_{d,t} \log p(w_{d,t}|z_{d,t}, \Phi) + \sum_{d,t} \log p(z_{d,t}|\theta_d) \\
& + \sum_{d} \log p(\theta_d) + \sum_{k} \log p(\phi_k) \\
\simeq & \sum_{d,t} \sum_{k} 1\{z_{d,t} = k\} \log \phi_{k,w_{d,t}} \\
& + \sum_{d,t} \sum_{k} 1\{z_{d,t} = k\} \log (\theta_{d,k}) \\
& + \sum_{d,k} (\alpha_k - 1) \log (\theta_{d,k}) \\
& + \sum_{k,v} (\beta_k - 1) \log (\phi_{k,v})
\end{aligned}
$$

(1) $z_{d,t}$

$$\begin{aligned}\log\left(q^*\left(z_{d,t}\right)\right) &= E_{-z_{d,t}}\log(p(\theta,\mathbf{z},\Phi,\mathbf{w})) \\ &\simeq \sum_k 1\left\{z_{d,t}=k\right\} E_{q^*}\left[\log\left(\phi_{k,w_{d,t}}\right)+\log\left(\theta_{d,k}\right)\right]\end{aligned}$$

其中 $\simeq$ 表示至多差一个可加常数。因此

$$\begin{aligned}\pi_{d,t,k} &:= q^*\left(z_{d,t}=k\right) \\ &\propto \exp\left(E_{q^*}\left[\log\left(\phi_{k,w_{d,t}}\right)+\log\left(\theta_{d,k}\right)\right]\right)\end{aligned}$$

(2) θ_d

$$\begin{aligned}
\log(q^*(\theta_d)) &= E_{-\theta_d} \log(p(\theta, z, \beta, w)) \\
&\simeq \sum_{k,t} [E_{q^*} 1\{z_{d,t} = k\}] \log(\theta_{d,k}) + \sum_k (\alpha_k - 1) \log(\theta_{d,k}) \\
&= \sum_{k,n} \pi_{d,t,k} \log(\theta_{d,k}) + \sum_k (\alpha_k - 1) \ln(\theta_{d,k}) \\
&= \sum_k (\alpha_k - 1 + \sum_t \phi_{d,t,k}) \log(\theta_{d,k})
\end{aligned}$$

因此

$$\begin{aligned}
q^*(\theta_d) &= Dir(\gamma_d) \\
\gamma_{d,k} &= \alpha_k + \sum_t \pi_{d,t,k}
\end{aligned}$$

(3) ϕ_k

$$q^*(\phi_k) = Dir(\lambda_k)$$

$$\lambda_{k,v} = \beta_v + \sum_{d,t} 1\{w_{d,t} = v\}\pi_{d,t,k}$$