Bayesian Matrix Completion with Hierarchical Gaussian Prior Models

Zebediah

2024 年 7 月 21 日

论文: Bayesian Matrix Completion with Hierarchical Gaussian Prior Models, 作者为 Linxiao Yang, Jun Fang, Huiping Duan, Hongbin Li and Bing Zeng.

代码: https://www.junfang-uestc.net/codes/LRMC.rar.

目录

1 引言

低秩矩阵补全问题的规范形式可以表示为

$$\begin{aligned} \min_{\boldsymbol{X}} \quad & \operatorname{rank}(\boldsymbol{X}), \\ \text{s.t.} \quad & \boldsymbol{Y} = \boldsymbol{\Omega} * \boldsymbol{X}, \end{aligned} \tag{1.1}$$

其中 $\boldsymbol{X} \in \mathbb{R}^{M\times N}$ 是一个未知的低秩矩阵, $\Omega \in \{0,1\}^{M\times N}$ 是一个二元矩阵, 用于表示观察到 $\boldsymbol{X}$ 的哪些条目, $*$ 表示 Hadamard 乘积, $\boldsymbol{Y} \in \mathbb{R}^{M\times N}$ 是观察到的矩阵. 然而, 最小化矩阵的秩是 NP-hard 的. 最流行的替代方案是核范数 (矩阵的奇异值之和). 用核范数替换秩函数会产生以下凸优化问题:

$$\begin{aligned} \min_{\boldsymbol{X}} \quad & \|\boldsymbol{X}\|_*, \\ \text{s.t.} \quad & \boldsymbol{Y} = \boldsymbol{\Omega} * \boldsymbol{X}. \end{aligned} \tag{1.2}$$

本文开发了一种新的贝叶斯方法来解决低秩矩阵补全问题. 具体地, 假设低秩矩阵 X 的列为相互独立并服从一个具有零均值和精度矩阵的高斯分布. 精度矩阵为一个随机参数, 并在其上指定了一个 Wishart 分布作为超先验. 本文使用了 GAMP 技术, 并将其嵌入变分贝叶斯 (VB) 推断中, 从而得到了一个用于矩阵补全的高效 VB-GAMP 算法. 但由于先验分布的不可因式分解形式, GAMP 技术不能直接使用. 为了解决这个问题, 我们构建了一个替代问题, 其后验分布正是 VB 推断所需的. 同时, 这个替代问题具有可因式分解的先验和噪声分布, 因此可以直接应用 GAMP 来获得近似后验分布. 这种技巧有助于大幅降低计算复杂度.

2 贝叶斯模型

在有噪声的情况下, 矩阵补全问题的标准形式可以表述为:

$$\begin{aligned} \min_{\boldsymbol{X}} \quad & \operatorname{rank}(\boldsymbol{X}), \\ \text{s.t.} \quad & \boldsymbol{Y} = \boldsymbol{\Omega} * (\boldsymbol{X} + \boldsymbol{E}), \end{aligned} \tag{2.1}$$

其中 $\boldsymbol{E}$ 表示加性噪声, $\Omega \in \{0,1\}^{M\times N}$. 不失一般性, 我们假设 $M \leq N$. 如前所述, 该问题是 NP-hard 的.

本文考虑在贝叶斯框架内对矩阵补全问题进行建模. 我们假设 $\boldsymbol{E}$ 的条目是独立同分布 (i.i.d.) 的随机变量, 服从均值为零、方差为 γ^{-1} 的高斯分布. 为了学习 γ, 我们使用了一个 Gamma 超先验, 即

$$p(\gamma) = \operatorname{Gamma}(\gamma \mid a, b) = \Gamma(a)^{-1} b^a \gamma^{a-1} \mathrm{e}^{-b\gamma}, \tag{2.2}$$

其中 $\Gamma(a) = \int_0^\infty t^{a-1}\mathrm{e}^{-t}\mathrm{d}t$ 是 Gamma 函数. 参数 a 和 b 被设置为较小的值, 例如 10^{-8}, 这使得 Gamma 分布成为一个无信息先验.

为了促进 $\boldsymbol{X}$ 的低秩性, 我们提出了一个两层的层次高斯先验模型 (见图 1). 具体来说, 在第一层中, 假设 $\boldsymbol{X}$ 的各列是相互独立的, 并且服从一个共同的高斯分布:

$$p(\boldsymbol{X} \mid \boldsymbol{\Sigma}) = \prod_{n=1}^{N} p\left(\boldsymbol{x}_n \mid \boldsymbol{\Sigma}\right) = \prod_{n=1}^{N} \mathcal{N}\left(\boldsymbol{x}_n \mid \boldsymbol{0}, \boldsymbol{\Sigma}^{-1}\right), \tag{2.3}$$

其中 $\boldsymbol{x}_n$ 表示 $\boldsymbol{X}$ 的第 i 列, $\boldsymbol{\Sigma} \in \mathbb{R}^{M\times M}$ 是精度矩阵. 第二层指定 Wishart 分布作为精度矩阵 $\boldsymbol{\Sigma}$ 的超先验:

$$p(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{\frac{\nu-M-1}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\left(\boldsymbol{W}^{-1}\boldsymbol{\Sigma}\right)\right), \tag{2.4}$$

其中 ν 和 $\boldsymbol{W} \in \mathbb{R}^{M\times M}$ 分别表示 Wishart 分布的自由度和尺度矩阵.

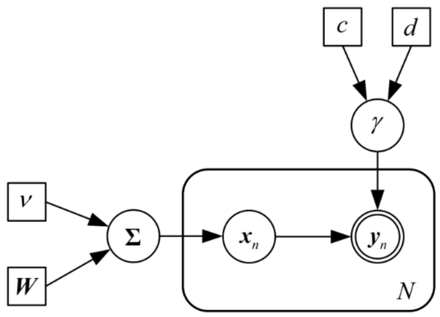

图 1: 分层高斯先验模型

我们将精度矩阵 $\boldsymbol{\Sigma}$ 积分掉得到 $\boldsymbol{X}$ 的边缘分布:

$$\begin{aligned} p(\boldsymbol{X}) &= \int \prod_{n=1}^{N} p\left(\boldsymbol{x}_n \mid \boldsymbol{\Sigma}\right) p(\boldsymbol{\Sigma}) \mathrm{d}\boldsymbol{\Sigma} \\ &\propto \left|\boldsymbol{W}^{-1} + \boldsymbol{X}\boldsymbol{X}^T\right|^{-\frac{\nu+N}{2}}. \end{aligned} \tag{2.5}$$

(2.5) 的推导细节如下:

$$\begin{aligned} p(\boldsymbol{X}) &= \int \prod_{i=1}^{N} p\left(\boldsymbol{x}_i \mid \boldsymbol{\Sigma}\right) p(\boldsymbol{\Sigma}) d\boldsymbol{\Sigma} \\ &\propto \int \left(\frac{|\boldsymbol{\Sigma}|}{(2\pi)^M}\right)^{\frac{N}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\left(\boldsymbol{X}^T\boldsymbol{\Sigma}\boldsymbol{X}\right)\right) |\boldsymbol{\Sigma}|^{\frac{\nu-M-1}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\left(\boldsymbol{W}^{-1}\boldsymbol{\Sigma}\right)\right) \mathrm{d}\boldsymbol{\Sigma} \\ &\propto 2^{\frac{\nu M}{2}} \pi^{-\frac{MN}{2}} \Gamma_M\left(\frac{\nu+N}{2}\right) \left|\boldsymbol{W}^{-1} + \boldsymbol{X}\boldsymbol{X}^T\right|^{-\frac{\nu+N}{2}} \int \frac{|\boldsymbol{\Sigma}|^{\frac{\nu+N-M-1}{2}}}{2^{\frac{(\nu+N)M}{2}} \left|\left(\boldsymbol{W}^{-1} + \boldsymbol{X}\boldsymbol{X}^T\right)^{-1}\right|^{\frac{\nu+N}{2}} \Gamma_M\left(\frac{\nu+N}{2}\right)} \mathrm{d}\boldsymbol{\Sigma}, \end{aligned}$$

其中

$$\Gamma_M(x) = \pi^{\frac{M(M-1)}{4}} \prod_{j=1}^{M} \Gamma\left(x + \frac{1-j}{2}\right).$$

注意到 $p(\boldsymbol{X})$ 右侧的积分项是一个标准 Wishart 分布, 自由度为 $\nu+N$, 尺度矩阵为 $\left(\boldsymbol{W} + \boldsymbol{X}\boldsymbol{X}^T\right)^{-1}$. 从而我们得到

$$\begin{aligned} p(\boldsymbol{X}) &\propto 2^{\frac{\nu M}{2}} \pi^{-\frac{MN}{2}} \Gamma_M\left(\frac{\nu+N}{2}\right) \left|\boldsymbol{W}^{-1} + \boldsymbol{X}\boldsymbol{X}^T\right|^{-\frac{\nu+N}{2}} \\ &\propto \left|\boldsymbol{W}^{-1} + \boldsymbol{X}\boldsymbol{X}^T\right|^{-\frac{\nu+N}{2}}. \end{aligned}$$

由 (2.5), 我们有

$$\log p(\boldsymbol{X}) \propto -\log\left|\boldsymbol{X}\boldsymbol{X}^T + \boldsymbol{W}^{-1}\right|. \tag{2.6}$$

如果我们选择 $\boldsymbol{W} = \epsilon^{-1}\boldsymbol{I}$, 并且令 ϵ 为一个小的正值, 那么

$$\begin{aligned}\log p(\boldsymbol{X}) &\propto -\log\left|\boldsymbol{X}\boldsymbol{X}^T + \epsilon\boldsymbol{I}\right| \\ &= -\sum_{m=1}^{M}\log\left(\lambda_m + \epsilon\right)\end{aligned} \tag{2.7}$$

其中 λ_m 表示 $\boldsymbol{X}\boldsymbol{X}^T$ 的第 m 个特征值. 显然, 在这种情况下, 先验 $p(\boldsymbol{X})$ 鼓励一个低秩解 $\boldsymbol{X}$. 这是因为最大化先验分布 $p(\boldsymbol{X})$ 等价于对 $\{\lambda_m\}$ 最小化 $\sum_{m=1}^{M}\log\left(\lambda_m + \epsilon\right)$. 而对数和函数 $\sum_{m=1}^{M}\log\left(\lambda_m + \epsilon\right)$ 是一种有效的促进稀疏的泛函.

除了 $\boldsymbol{W} = \epsilon^{-1}\boldsymbol{I}$, 参数 $\boldsymbol{W}$ 还可以根据对 $\boldsymbol{X}$ 的额外先验知识进行设计. 例如, 在一些应用中 (如图像修复), x_n 的相邻系数之间存在空间相关性. 为了捕获相邻系数之间的平滑性, $\boldsymbol{W}$ 可以设置为

$$\boldsymbol{W} = \boldsymbol{F}^T\boldsymbol{F}, \tag{2.8}$$

其中 $\boldsymbol{F} \in \mathbb{R}^{M\times M}$ 是一个二阶差分算子, 其 (i, j) 项给出如下:

$$f_{i,j} = \begin{cases} -2, & i = j, \\ 1, & |i-j| = 1, \\ 0, & \text{else.} \end{cases} \tag{2.9}$$

另一种促进平滑解的 $\boldsymbol{W}$ 选择是拉普拉斯矩阵, i.e.

$$\boldsymbol{W} = \boldsymbol{D} - \boldsymbol{A} + \hat{\epsilon}\boldsymbol{I}, \tag{2.10}$$

其中 $\boldsymbol{A}$ 是图的邻接矩阵, 其元素为

$$a_{ij} = \exp\left(-\frac{|i-j|^2}{\theta^2}\right). \tag{2.11}$$

$\boldsymbol{D}$ 称为度矩阵, 是一个对角元素为 $d_{ii} = \sum_j a_{ij}$ 的对角矩阵, $\hat{\epsilon}$ 是一个小的正值, 用于确保 $\boldsymbol{W}$ 是满秩的. 可以证明, (2.8) 和 (2.10) 中定义的 $\boldsymbol{W}$ 促进了 $\boldsymbol{X}$ 的低秩性和平滑性. 为说明这一点, 我们首先介绍以下引理:

引理 1 对于一个正定矩阵 $\boldsymbol{W} \in \mathbb{R}^{M\times M}$, 对任意 $\boldsymbol{X} \in \mathbb{R}^{M\times N}$, 成立

$$\log\left|\boldsymbol{X}\boldsymbol{X}^T + \boldsymbol{W}^{-1}\right| = \log\left|\boldsymbol{W}^{-1}\right| + \log\left|\boldsymbol{I} + \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right|. \tag{2.12}$$

由引理 1, 我们有

$$\begin{aligned}\log p(\boldsymbol{X}) &\propto -\log\left|\boldsymbol{X}\boldsymbol{X}^T + \boldsymbol{W}^{-1}\right| \\ &\propto -\log\left|\boldsymbol{I} + \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right| \\ &= -\sum_{n=1}^{N}\log\left(\tilde{\lambda}_n + 1\right),\end{aligned} \tag{2.13}$$

其中 $\tilde{\lambda}_n$ 是 $\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}$ 的第 n 个特征值. 我们看到, 最大化先验分布等价于对 $\left\{\tilde{\lambda}_n\right\}$ 最小化 $\sum_{n=1}^{N}\log\left(\lambda_n + 1\right)$.

从而得到稀疏的 $\{\lambda_n\}$, 即 $\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}$ 是低秩的. 因为 $\boldsymbol{W}$ 是满秩的, 因此 $\boldsymbol{X}$ 是低秩的. 另一方面, 注意到 $\operatorname{tr}\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)$ 是 $\log\left|\boldsymbol{I}+\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right|$ 的一阶近似, 因此最小化 $\log\left|\boldsymbol{I}+\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right|$ 会减少 $\operatorname{tr}\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)$ 的值. 显然, 对于 (2.8) 和 (2.10) 中定义的 $\boldsymbol{W}$, 一个更平滑的解会导致 $\operatorname{tr}\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)$ 的值更小. 因此, 当 $\boldsymbol{W}$ 选取为(2.8) 或 (2.10) 时, 所得的先验分布 $p(\boldsymbol{X})$ 可以促进低秩和平滑解.

3 变分贝叶斯推断

3.1 变分贝叶斯方法回顾

我们首先简要回顾一下变分贝叶斯 (VB) 方法, 在概率模型中, 令 $\boldsymbol{y}$ 和 $\boldsymbol{\theta}$ 分别表示观测数据和隐藏变量. 很容易证明, 观测数据的边际概率可以分解为两个项:

$$\ln p(\boldsymbol{y}) = L(q) + \mathrm{KL}(q\|p), \tag{3.1}$$

其中

$$L(q) = \int q(\boldsymbol{\theta}) \ln \frac{p(\boldsymbol{y},\boldsymbol{\theta})}{q(\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta}, \quad \mathrm{KL}(q\|p) = -\int q(\boldsymbol{\theta}) \ln \frac{p(\boldsymbol{\theta} \mid \boldsymbol{y})}{q(\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta}, \tag{3.2}$$

这里 $q(\boldsymbol{\theta})$ 是任何概率密度函数, $\mathrm{KL}(q\|p)$ 是 $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ 和 $q(\boldsymbol{\theta})$ 之间的 Kullback-Leibler 散度. 由于 $\mathrm{KL}(q\|p) \geq 0$, 因此 $L(q)$ 是 $\ln p(\boldsymbol{y})$ 的严格下界. 因此最大化 $L(q)$ 等价于最小化 $\mathrm{KL}(q\|p)$, 因此后验分布 $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ 可以通过最大化 $L(q)$ 来近似为 $q(\boldsymbol{\theta})$.

最大化过程可以对每个潜在变量进行交替进行:

$$q_i\left(\theta_i\right) = \frac{e^{\langle \ln p(\boldsymbol{y},\boldsymbol{\theta})\rangle_{k\neq i}}}{\int e^{\langle \ln p(\boldsymbol{y},\boldsymbol{\theta})\rangle_{k\neq i}} \mathrm{d}\theta_i}, \tag{3.3}$$

其中 $\langle\cdot\rangle_{k\neq i}$ 表示对所有 $k \neq i$ 的分布 $q_k\left(\theta_k\right)$ 求期望. 两边取对数可得

$$\ln q_i\left(\theta_i\right) = \langle \ln p(\boldsymbol{y},\boldsymbol{\theta})\rangle_{k\neq i} + \text{ constant}. \tag{3.4}$$

3.2 提出的算法

现在我们进行对提出的层次模型的变分贝叶斯推理. 令 $\boldsymbol{\theta} \triangleq \{\boldsymbol{X},\boldsymbol{\Sigma},\gamma\}$ 表示所有隐藏变量, 我们的目标是找到后验分布 $p(\boldsymbol{\theta} \mid \boldsymbol{y})$. 由于计算 $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ 通常不可行, 我们将 $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ 近似为 $q(\boldsymbol{X},\boldsymbol{\Sigma},\gamma)$, 其因子化形式为隐藏变量 $\{\boldsymbol{X},\boldsymbol{\Sigma},\gamma\}$ 的因子化形式, i.e.

$$q(\boldsymbol{X},\boldsymbol{\Sigma},\gamma) = q_x(\boldsymbol{X}) q_{\Sigma}(\boldsymbol{\Sigma}) q_{\gamma}(\gamma). \tag{3.5}$$

如前面所述, $L(q)$ 的最大化可以对每个潜在变量交替进行, 因此我们有

$$\begin{aligned} \ln q_x(\boldsymbol{X}) &= \langle \ln p(\boldsymbol{\Sigma},\gamma)\rangle_{q_{\Sigma}(\boldsymbol{\Sigma}) q_{\gamma}(\gamma)} + \text{ constant}, \\ \ln q_{\Sigma}(\boldsymbol{\Sigma}) &= \langle \ln p(\boldsymbol{X},\gamma)\rangle_{q_x(\boldsymbol{X}) q_{\gamma}(\gamma)} + \text{ constant}, \\ \ln q_{\gamma}(\gamma) &= \langle \ln p(\boldsymbol{X},\boldsymbol{\Sigma})\rangle_{q_x(\boldsymbol{X}) q_{\Sigma}(\boldsymbol{\Sigma})} + \text{ constant}, \end{aligned} \tag{3.6}$$

其中 $\langle\cdot\rangle_{q_1(\cdot)\ldots q_K(\cdot)}$ 表示对分布 $\{q_k(\cdot)\}_{k=1}^{K}$ 的期望. 接下来将介绍算法的详细内容.

(1) $q_x(\boldsymbol{X})$ 的更新: $q_x(\boldsymbol{X})$ 的计算可以分解为一组独立任务, 每个任务计算 $\boldsymbol{X}$ 每列的后验分布近似 $q_x(\boldsymbol{x}_n)$. 我们有

$$\begin{aligned}\ln q_x(\boldsymbol{x}_n) &\propto \langle \ln [p(\boldsymbol{y}_n \mid \boldsymbol{x}_n) p(\boldsymbol{x}_n \mid \boldsymbol{\Sigma})] \rangle_{q_\Sigma(\boldsymbol{\Sigma}) q_\gamma(\gamma)} \\ &\propto \left\langle -\gamma (\boldsymbol{y}_n - \boldsymbol{x}_n)^T \boldsymbol{O}_n (\boldsymbol{y}_n - \boldsymbol{x}_n) - \boldsymbol{x}_n^T \boldsymbol{\Sigma} \boldsymbol{x}_n \right\rangle \\ &\propto -\boldsymbol{x}_n^T (\langle\gamma\rangle \boldsymbol{O}_n + \langle \boldsymbol{\Sigma} \rangle) \boldsymbol{x}_n + 2\langle\gamma\rangle \boldsymbol{x}_n^T \boldsymbol{O}_n \boldsymbol{y}_n,\end{aligned} \tag{3.7}$$

其中 $\boldsymbol{y}_n$ 是 $\boldsymbol{Y}$ 的第 n 列, $\boldsymbol{O}_n \triangleq \operatorname{diag}(\boldsymbol{o}_n)$, 这里 $\boldsymbol{o}_n$ 是 $\boldsymbol{\Omega}$ 的第 n 列. 由 (3.7), $\boldsymbol{x}_n$ 服从高斯分布

$$q_x(\boldsymbol{x}_n) = \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_n, \boldsymbol{Q}_n), \tag{3.8}$$

其中

$$\boldsymbol{\mu}_n = \langle\gamma\rangle \boldsymbol{Q}_n \boldsymbol{O}_n \boldsymbol{y}_n, \quad \boldsymbol{Q}_n = (\langle\gamma\rangle \boldsymbol{O}_n + \langle \boldsymbol{\Sigma} \rangle)^{-1}. \tag{3.9}$$

可以看出, 为了计算 $q_x(\boldsymbol{x}_n)$, 需要对一个 $M \times M$ 的矩阵求逆, 计算复杂度为 $\mathcal{O}(M^3)$.

(2) $q_\Sigma(\boldsymbol{\Sigma})$ 的更新: 可以得到近似后验 $q_\Sigma(\boldsymbol{\Sigma})$ 为

$$\begin{aligned}\ln q_\Sigma(\boldsymbol{\Sigma}) &\propto \left\langle \ln \left[\prod_{n=1}^{N} p(\boldsymbol{x}_n \mid \boldsymbol{\Sigma}) p(\boldsymbol{\Sigma}) \right] \right\rangle_{q_x(\boldsymbol{X})} \\ &\propto \left\langle \frac{N}{2} \ln |\boldsymbol{\Sigma}| - \frac{1}{2} \operatorname{tr}(\boldsymbol{X}^T \boldsymbol{\Sigma} \boldsymbol{X}) + \frac{\nu - M - 1}{2} \ln |\boldsymbol{\Sigma}| - \frac{1}{2} \operatorname{tr}(\boldsymbol{W}^{-1} \boldsymbol{\Sigma}) \right\rangle \\ &\propto \frac{\nu + N - M - 1}{2} \ln |\boldsymbol{\Sigma}| - \frac{1}{2} \operatorname{tr}((\boldsymbol{W}^{-1} + \langle \boldsymbol{X}\boldsymbol{X}^T \rangle) \boldsymbol{\Sigma}).\end{aligned} \tag{3.10}$$

由 (3.10) 可知, $\boldsymbol{\Sigma}$ 服从 Wishart 分布, i.e.

$$q_\Sigma(\boldsymbol{\Sigma}) = \text{ Wishart } (\boldsymbol{\Sigma}; \hat{\boldsymbol{W}}, \hat{\nu}), \tag{3.11}$$

其中

$$\hat{\boldsymbol{W}} = (\boldsymbol{W}^{-1} + \langle \boldsymbol{X}\boldsymbol{X}^T \rangle)^{-1}, \quad \hat{\nu} = \nu + N. \tag{3.12}$$

(3) $q_\gamma(\gamma)$ 的更新: 通过对 $q_\gamma(\gamma)$ 的变分优化得到

$$\begin{aligned}\ln q_\gamma(\gamma) &\propto \langle \ln p(\boldsymbol{Y} \mid \boldsymbol{X}, \gamma) p(\gamma) \rangle_{q_x(\boldsymbol{X})} \\ &\propto \left\langle \ln \prod_{(m,n)\in\mathbb{S}} p(y_{mn} \mid x_{mn}, \gamma) p(\gamma) \right\rangle \\ &\propto \left\langle \frac{L}{2} \ln \gamma - \frac{\gamma}{2} \sum_{(m,n)\in\mathbb{S}} (y_{mn} - x_{mn})^2 + (c-1) \ln \gamma - d\gamma \right\rangle \\ &= \left(\frac{L}{2} + c - 1 \right) \ln \gamma - \left(\frac{1}{2} \sum_{(m,n)\in\mathbb{S}} \left\langle (y_{mn} - x_{mn})^2 \right\rangle + d \right),\end{aligned} \tag{3.13}$$

其中 x_{mn} 和 y_{mn} 表示 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的第 (m,n) 项, $\mathbb{S} \triangleq \{(m,n) \mid \Omega_{mn} = 1\}$ 是由观测值的索引组成的集合,

$L \triangleq |\mathbb{S}|$ 是集合 $\mathbb{S}$ 的基数, Ω_{mn} 表示 Ω 的第 (m,n) 项. 可以验证 $q_\gamma(\gamma)$ 服从 Gamma 分布

$$q_\gamma(\gamma) = \text{Gamma}(\gamma \mid \tilde{c}, \tilde{d}), \tag{3.14}$$

其中

$$\tilde{c} = \frac{L}{2} + c, \quad \tilde{d} = \frac{1}{2} \sum_{(m,n)\in\mathbb{S}} \left\langle (y_{mn} - x_{mn})^2 \right\rangle + d, \tag{3.15}$$

这里

$$\left\langle (y_{mn} - x_{mn})^2 \right\rangle = y_{mn}^2 - 2y_{mn} \langle x_{mn} \rangle + \langle x_{mn}^2 \rangle. \tag{3.16}$$

在更新过程中使用的一些期望和矩如下:

$$\begin{aligned} \langle \mathbf{\Sigma} \rangle &= \hat{\boldsymbol{W}} \hat{\nu}, \\ \langle \boldsymbol{X}\boldsymbol{X}^T \rangle &= \langle \boldsymbol{X} \rangle \langle \boldsymbol{X} \rangle^T + \sum_{n=1}^{N} \boldsymbol{Q}_n, \\ \langle x_{mn}^2 \rangle &= \langle x_{mn} \rangle^2 + Q_n(m,m), \end{aligned} \tag{3.17}$$

其中 $Q_n(m,m)$ 表示 $\boldsymbol{Q}_n$ 的第 m 个对角元. 上述过程总结如算法所示.

Algorithm 1 VB Algorithm for Matrix Completion

```
Input: Y, Ω, ν and W.
Initialization: ⟨Σ⟩ and ⟨γ⟩.
while not converge do
  for n = 1 to N do
    Update q_x(x_n) via (3.8), with q_Σ(Σ) and q_γ(γ) fixed;
  end for
  Update q_Σ(Σ) via (3.11), with q_x(X) and q_γ(γ) fixed;
  Update q_γ(γ) via (3.14);
end while
Output: q_x(X), q_Σ(Σ), q_γ(γ).
```

4 VB-GAMP

GAMP 算法要求先验分布和噪声分布都具有因子化形式. 然而在我们的模型中, 先验分布 $p(\boldsymbol{x}_n \mid \mathbf{\Sigma})$ 具有非因子化形式. 为了解决这个问题, 我们构建了一个替代问题, 该问题旨在从 $\boldsymbol{b} \in \mathbb{R}^M$ 中恢复 $\boldsymbol{x} \in \mathbb{R}^M$:

$$\boldsymbol{b} = \boldsymbol{U}^T \boldsymbol{x} + \boldsymbol{e}, \tag{4.1}$$

其中 $\boldsymbol{U} \in \mathbb{C}^{M\times M}$ 是通过对 $\langle \mathbf{\Sigma} \rangle = \boldsymbol{U}\boldsymbol{S}\boldsymbol{U}^T$ 进行奇异值分解得到的, $\boldsymbol{e}$ 表示加性高斯噪声, 均值为零, 协方差矩阵为 $\boldsymbol{S}^{-1}$. 我们假设 $\boldsymbol{x}$ 的各个条目是相互独立的, 并服从以下分布:

$$p(x_m) = \begin{cases} \mathcal{N}(\kappa_m, \xi^{-1}), & \text{if } \pi_m = 1, \\ C, & \text{if } \pi_m = 0, \end{cases} \tag{4.2}$$

其中 π_m, x_m, κ_m 分别表示 $\boldsymbol{\pi}, \boldsymbol{x}, \boldsymbol{\kappa}$ 的第 m 项, C 是一个常数, $\boldsymbol{\pi}, \boldsymbol{\kappa} \in \mathbb{R}^{M\times 1}, \xi$ 是已知参数. 考虑问题 (4.1), $\boldsymbol{x}$ 的后验分布可以计算为:

$$\begin{aligned} p(\boldsymbol{x} \mid \boldsymbol{b}) &\propto p(\boldsymbol{b} \mid \boldsymbol{x}) p(\boldsymbol{x}) \\ &\propto p(\boldsymbol{b} \mid \boldsymbol{x}) \prod_{m\in S} p\left(x_m\right) \\ &= \mathcal{N}\left(\boldsymbol{U}^T\boldsymbol{x}, \boldsymbol{S}^{-1}\right) \prod_{m\in S} \mathcal{N}\left(\kappa_m, \xi^{-1}\right), \end{aligned} \tag{4.3}$$

其中 $S \triangleq \{m \mid \pi_m = 1\}$.

对 $p(\boldsymbol{x} \mid \boldsymbol{b})$ 取对数有

$$\begin{aligned} \ln p(\boldsymbol{x} \mid \boldsymbol{b}) &\propto -\frac{1}{2}\left(\boldsymbol{b}-\boldsymbol{U}^T\boldsymbol{x}\right)^T \boldsymbol{S}\left(\boldsymbol{b}-\boldsymbol{U}^T\boldsymbol{x}\right) - \frac{1}{2}\xi \sum_{m\in S}\left(x_m-\kappa_m\right)^2 \\ &= -\frac{1}{2}\left(\boldsymbol{b}-\boldsymbol{U}^T\boldsymbol{x}\right)^T \boldsymbol{S}\left(\boldsymbol{b}-\boldsymbol{U}^T\boldsymbol{x}\right) - \frac{1}{2}\xi(\boldsymbol{x}-\boldsymbol{\kappa})^T\boldsymbol{\Pi}(\boldsymbol{x}-\boldsymbol{\kappa}) \\ &\propto -\frac{1}{2}\boldsymbol{x}^T\left(\boldsymbol{U}\boldsymbol{S}\boldsymbol{U}^T+\xi\boldsymbol{\Pi}\right)\boldsymbol{x}^T + \left(\boldsymbol{b}^T\boldsymbol{S}\boldsymbol{U}^T+\xi\boldsymbol{\kappa}^T\boldsymbol{\Pi}\right)\boldsymbol{x}, \end{aligned} \tag{4.4}$$

其中 $\boldsymbol{\Pi}$ 是一个对角矩阵, 其第 m 个对角元素等于 π_m. 显然, $p(\boldsymbol{x} \mid \boldsymbol{b})$ 服从高斯分布, 其均值 $\boldsymbol{\mu}$ 和协方差矩阵 $\boldsymbol{Q}$ 为

$$\boldsymbol{\mu} = \boldsymbol{Q}(\boldsymbol{U}\boldsymbol{S}\boldsymbol{b}+\xi\boldsymbol{\Pi}\boldsymbol{\kappa})\boldsymbol{Q} = \left(\boldsymbol{U}\boldsymbol{S}\boldsymbol{U}^T+\xi\boldsymbol{\Pi}\right)^{-1} = (\langle\boldsymbol{\Sigma}\rangle+\xi\boldsymbol{\Pi})^{-1}. \tag{4.5}$$

比较 (3.9) 和 (4.5), 我们可以验证, 当 $\boldsymbol{b}=\boldsymbol{0}, \boldsymbol{\kappa}=\boldsymbol{y}_n, \boldsymbol{\pi}=\boldsymbol{o}_n$(i.e. $\boldsymbol{\Pi}=\boldsymbol{O}_n$), 且 $\xi=\langle\gamma\rangle$ 时, $p(\boldsymbol{x} \mid \boldsymbol{b})$ 正好是所需的后验分布 $q_x\left(\boldsymbol{x}_n\right)$. 同时对于替代问题 (4.1), 先验分布和噪声分布都是可因子化的. 因此, 可以直接应用 GAMP 算法到 (4.1) 以找到后验分布 $p(\boldsymbol{x} \mid \boldsymbol{b})$ 的近似. 通过设置 $\boldsymbol{b}=\boldsymbol{0}, \boldsymbol{\kappa}=\boldsymbol{y}_n, \boldsymbol{\pi}=\boldsymbol{o}_n, \xi=\langle\gamma\rangle$ 可以有效地获得 (3.8) 中 $q_x\left(x_n\right)$ 的近似.

4.1 使用 GAMP 求解 (4.1)

首先, GAMP 通过以下公式近似真实的边际后验分布 $p\left(x_m \mid \boldsymbol{b}\right)$:

$$\hat{p}\left(x_m \mid \boldsymbol{b}, \hat{r}_m, \tau_m^r\right) = \frac{p\left(x_m\right)\mathcal{N}\left(x_m \mid \hat{r}_m, \tau_m^r\right)}{\int_x p\left(x_m\right)\mathcal{N}\left(x_m \mid \hat{r}_m, \tau_m^r\right)\mathrm{d}x}, \tag{4.6}$$

其中 $\hat{r}_m$ 和 τ_m^r 是在 GAMP 算法的迭代过程中迭代更新的量. 这里为了简化, 省略了对迭代次数 k 的显式依赖 (下同). 在 $\pi_m=1$ 的情况下, 将先验分布 (4.2) 代入 (4.6) 中, 可以验证近似后验分布 $\hat{p}\left(x_m \mid \boldsymbol{b}, \hat{r}_m, \tau_m^r\right)$ 服从高斯分布, 其均值和方差分别为:

$$\mu_m^x = \phi_m^x\left(\xi\kappa_m + \hat{r}_m/\tau_m^r\right), \quad \phi_m^x = \frac{\tau_m^r}{1+\xi\tau_m^r}. \tag{4.7}$$

同样, 在 $\pi_m=0$ 的情况下, 近似后验分布 $\hat{p}\left(x_m \mid \boldsymbol{b}, \hat{r}_m, \tau_m^r\right)$ 也服从高斯分布, 其均值和方差分别为:

$$\mu_m^x = \hat{r}_m, \quad \phi_m^x = \tau_m^r. \tag{4.8}$$

另一种近似是对无噪声输出 $z_i \triangleq \boldsymbol{u}_i^T\boldsymbol{x}$ 进行, 其中 $\boldsymbol{u}_i^T$ 表示 $\boldsymbol{U}^T$ 的第 i 行. GAMP 通过以下公式近似真

实的边际后验分布 $p\left(z_i \mid \boldsymbol{b}\right)$:

$$\hat{p}\left(z_i \mid \boldsymbol{b}, \hat{p}_i, \tau_i^p\right)=\frac{p\left(b_i \mid z_i\right) \mathcal{N}\left(z_i \mid \hat{p}_i, \tau_i^p\right)}{\int_z p\left(b_i \mid z_i\right) \mathcal{N}\left(z_i \mid \hat{p}_i, \tau_i^p\right) \mathrm{d} z}, \tag{4.9}$$

其中 $\hat{p}_i$ 和 τ_i^p 是在 GAMP 算法的迭代过程中迭代更新的量. 在加性白高斯噪声假设下, 我们有 $p\left(b_i \mid z_i\right)=\mathcal{N}\left(b_i \mid z_i, s_i^{-1}\right)$, 其中 s_i 为 S 的第 i 个对角元素. 故 $\hat{p}\left(z_i \mid \boldsymbol{b}, \hat{p}_i, \tau_i^p\right)$ 也服从高斯分布, 其均值和方差分别为

$$\mu_i^z=\frac{\tau_i^p s_i b_i+\hat{p}_i}{1+s_i \tau_i^p}, \quad \phi_i^z=\frac{\tau_i^p}{1+s_i \tau_i^p}. \tag{4.10}$$

有了上述近似, 我们现在可以定义 GAMP 算法中使用的两个标量函数: $g_{\text{in}}(\cdot)$ 和 $g_{\text{out}}(\cdot)$. 输入标量函数 $g_{\text{in}}(\cdot)$ 简单地定义为后验均值 μ_m^x, i.e.

$$g_{\text{in}}\left(\hat{r}_m, \tau_m^r\right)=\mu_m^x=\begin{cases}\phi_m^x\left(\xi \kappa_m+\hat{r}_m / \tau_m^r\right), & \text{if } \pi_m=1, \\ \hat{r}_m, & \text{if } \pi_m=0.\end{cases} \tag{4.11}$$

$\tau_m^r g_{\text{in}}\left(\hat{r}_m, \tau_m^r\right)$ 对 $\hat{r}_m$ 的偏导是后验方差 ϕ_m^x, i.e.

$$\tau_m^r \frac{\partial}{\partial \hat{r}_m} g_{\text{in}}\left(\hat{r}_m, \tau_m^r\right)=\phi_m^x=\begin{cases}\frac{\tau_m^r}{1+\xi \tau_m^r}, & \text{if } \pi_m=1, \\ \tau_m^r, & \text{if } \pi_m=0.\end{cases} \tag{4.12}$$

输出标量函数 $g_{\text{out}}(\cdot)$ 与后验均值 μ_i^z 的关系如下:

$$g_{\text{out}}\left(\hat{p}_i, \tau_{i,n}^p\right)=\frac{1}{\tau_i^p}\left(\mu_i^z-\hat{p}_i\right)=\frac{s_i\left(b_i-\hat{p}_i\right)}{1+s_i \tau_i^p}. \tag{4.13}$$

标量函数 $g_{\text{out}}\left(\hat{p}_i, \tau_i^p\right)$ 的偏导数与后验方差 $\phi_{i,n}^z$ 之间的关系如下:

$$\frac{\partial}{\partial \hat{p}_i} g_{\text{out}}\left(\hat{p}_i, \tau_i^p\right)=\frac{\phi_i^z-\tau_i^p}{\left(\tau_i^p\right)^2}=\frac{-s_i}{\left(1+s_i \tau_i^p\right)} \tag{4.14}$$

综上所述, 可以总结适用于问题 (4.1) 的 GAMP 算法 (见算法 2), 其中 $u_{i,m}$ 表示 $\boldsymbol{U}^T$ 的第 (i, m) 项.

4.2 VB-GAMP 算法

我们已经推导出一个有效的算法来获得问题 (4.1) 中 $\boldsymbol{x}$ 的近似后验分布. 具体来说, x_m 的真实边际后验分布近似为一个高斯分布 $\hat{p}\left(x_m \mid \boldsymbol{b}, \hat{r}_m, \tau_m^r\right)$, 其均值和方差分别由 (4.7) 或 (4.8) 给出. 联合后验分布 $p(\boldsymbol{x} \mid \boldsymbol{b})$ 可以近似为近似边际后验分布的乘积:

$$p(\boldsymbol{x} \mid \boldsymbol{b}) \approx \hat{p}(\boldsymbol{x} \mid \boldsymbol{b})=\prod_{m=1}^M \hat{p}\left(x_m \mid \boldsymbol{b}, \hat{r}_m, \tau_m^r\right). \tag{4.15}$$

如前所述, 通过设置 $b=0, \kappa=y_n, \pi=o_n, \xi=\langle\gamma\rangle$ 可以使用 GAMP 算法得到的后验分布 $\hat{p}(x \mid b)$ 来近似 (3.8) 中的 $q_x\left(\boldsymbol{x}_n\right)$. VB-GAMP 算法如算法 3 所示.

Algorithm 2 GAMP Algorithm

1: **Input:** κ, π, b, and ξ.
2: **Initialization:** Set $\hat{\psi}_i = 0, \forall i \in \{1, \ldots, M\}$; $\{\mu_m^x\}_{m=1}^M$ are initialized as the mean variance of the prior distribution, and $\{\phi_m^x\}_{m=1}^M$ are set to small values, say 10^{-5}.
3: **while** not converge **do**
4: **Step 1.** $\forall i \in \{1, \ldots, M\}$:

$$\hat{z}_i = \sum_m u_{i,m}\mu_m^x, \quad \tau_i^p = \sum_m u_{i,m}^2 \phi_m^x \hat{p}_i = \hat{z}_i - \tau_i^p \hat{\psi}_i.$$

5: **Step 2.** $\forall i \in \{1, \ldots, M\}$:

$$\hat{\psi}_i = g_{\text{out}}\left(\hat{p}_i, \tau_i^p\right), \quad \tau_i^s = -\frac{\partial}{\partial \hat{p}_i} g_{\text{out}}\left(\hat{p}_i, \tau_i^p\right).$$

6: **Step 3.** $\forall m \in \{1, \ldots, M\}$:

$$\tau_m^r = \left(\sum_i u_{i,m}^2 \tau_i^s\right)^{-1}, \quad \hat{r}_m = \mu_m^x + \tau_m^r \sum_i u_{i,m}\hat{\psi}_i.$$

7: **Step 4.** $\forall m \in \{1, \ldots, M\}$:

$$\mu_m^x = g_{\text{in}}\left(\hat{r}_m, \tau_m^r\right), \quad \phi_m^x = \tau_m^r \frac{\partial}{\partial \hat{r}_m} g_{\text{in}}\left(\hat{r}_m, \tau_m^r\right).$$

8: **end while**
9: **Output:** $\{\hat{r}_m, \tau_m^r\}, \{\hat{p}_i, \tau_i^p\}$, and $\{\mu_m^x, \phi_m^x\}$.

Algorithm 3 VB-GAMP Algorithm for Matrix Completion

1: **Input:** $\boldsymbol{Y}, \boldsymbol{\Omega}, \nu$ and $\boldsymbol{W}$.
2: **Initialization:** $\langle \boldsymbol{X} \rangle, \langle \boldsymbol{\Sigma} \rangle$.
3: **while** not converge **do**
4: Calculate singular value decomposition of $\langle \boldsymbol{\Sigma} \rangle$;
5: **for** $n = 1$ to N **do**
6: Obtain an approximation of $q_x(\boldsymbol{x}_n)$ via Algorithm 2;
7: **end for**Update $q_\Sigma(\boldsymbol{\Sigma})$ via (3.11); Update $q_\gamma(\gamma)$ via (3.14);
8: **end while**
9: **Output:** $q_x(\boldsymbol{X}), q_\Sigma(\boldsymbol{\Sigma})$, and $q_\gamma(\gamma)$.