arXiv:2601.06671v1 [stat.ME] 10 Jan 2026

删失图形马蹄先验：基于删失和缺失数据的贝叶斯稀疏精度矩阵估计

The Tien Mai [1] 和 Sayantan Banerjee[2]

[1]挪威公共卫生研究所，奥斯陆，0456，挪威。[2]运筹学与定量技术领域，印度印多尔管理学院，中央邦 453556，印度。

电子邮箱：the.tien.mai@fhi.no,
sayantanb@iimidr.ac.in

摘要

高斯图模型为研究多元数据中的条件依赖提供了强有力的框架，在生物医学、环境科学及其他数据丰富的科学领域有着广泛应用。图形马蹄先验（GHS）方法已成为稀疏精度矩阵估计领域的先进贝叶斯方法，但现有方法假设数据均被完整观测，因此在现实研究中普遍存在的删失或缺失情况下便无法适用。本文提出删失图形马蹄先验（CGHS），这是一种新颖的贝叶斯框架，将 GHS 扩展至存在删失和任意缺失的高斯数据。通过引入潜变量表示，CGHS 能够处理不完整观测，同时保留马蹄先验的自适应全局-局部收缩特性。我们推导了用于后验计算的高效吉布斯采样器，并建立了关于删失和缺失条件下后验行为的新理论结果，填补了基于套索的频率学派方法尚未解决的空白。通过大量模拟，我们证明，与惩罚似然方法相比，CGHS 能够持续提高估计精度。我们的方法已在软件包 `GHScenmis` 中实现，可在 GitHub 上获取：https://github.com/tienmt/ghscenmis。

关键词：图模型、删失数据、缺失数据、收缩先验。

1 引言

高斯图模型（GGMs）是用于建模多变量随机变量之间条件依赖关系的核心工具，应用领域涵盖基因组学、神经影像学和金融学等[35, 37, 15, 11]。在高斯图模型中，精度矩阵 $\Omega = \Sigma^{-1} \in \mathbb{R}^{p\times p}$ 编码了条件独立结构：非对角线元素为零表示相应节点之间条件独立 [20]。在高维情形下，估计稀疏精度矩阵尤为重要，此时变量数量 p 可能与样本量相当或大于样本量 n，并且可以利用稀疏性恢复底层图结构。常见的频率学派方法包括图形套索、 [16], 图形SCAD [14], 和 CLIME [9], ，它们在各种稀疏模式下可为高维逆协方差矩阵提供一致估计量 [10, 19]。

用于精度矩阵估计的贝叶斯方法能够实现自适应正则化并量化不确定性，但这方面的研究相对有限。[41]通过分块吉布斯抽样提出了贝叶斯图形套索，而 [6]引入了一种先验，将连续收缩成分与非对角元素上的零点质量结合起来，并推导出后验收缩率和用于模型选择的拉普拉斯近似。 [17], 探索了采用双指数先验的尖峰与厚尾变体，[2] 则采用拟似然方法，获得了良好的后验收缩表现。尖峰与厚尾先验的一个局限是包含二元指示变量，这可能会妨碍后验探索。

图形马蹄先验 [21]通过对精度矩阵进行全局—局部收缩克服了这一挑战，既能实现强大的自适应正则化，其实证表现也优于频率学派和贝叶斯方法。近期研究进一步探索了马蹄先验的变体，包括伪似然方法 [42]和类马蹄先验 [38]。

在实践中，数据常常包含删失值或缺失值，尤其是在生物医学、环境或流行病学研究中，测量值可能低于检测限，参见例如 [13, 31,36, 34]。忽略删失或缺失可能导致有偏推断，然而，在高维贝叶斯高斯图模型中处理此类数据的方法仍然很少。 [22, 33], 提出了针对缺失或删失数据的早期协方差估计方法，但这些方法并非贝叶斯方法。 [3, 39], 提出了用于删失图形套索的频率学派方法， [4],提供了相应的软件实现，但贝叶斯替代方法仍未得到充分发展。

本文将图形马蹄先验扩展至可处理删失和任意缺失的高斯数据。我们的删失图形马蹄先验（CGHS）模型引入潜变量表述，以处理删失观测和缺失，同时利用马蹄先验的自适应收缩特性来恢复稀疏精度矩阵。这种方法能够在标准图形马蹄先验方法可能失效的复杂真实世界数据集中，稳健地推断条件依赖关系。

为对模型进行后验推断，我们开发了一种吉布斯采样器，每一步都会更新一组参数和潜变量。该采样器依次执行几个关键步骤：首先，对删失潜变量值进行插补，或对缺失数据进行插补；其次，更新回归参数和残差方差；第三，调整局部和全局收缩尺度；最后，重构精度矩阵。此过程从模型参数和潜变量的联合后验分布中生成样本，进而可用于对潜在图结构进行可靠估计。

我们还提出了一套全面的理论结果，用以刻画所考虑的每种情形下后验分布的性质。这一方面确属我们的创新之处；据我们所知，在涉及删失或缺失的复杂场景中，频率学派方法尚无相关理论结果。具体而言，我们推导了删失数据情形和缺失数据情形下精度矩阵的后验集中速率。这些结果为贝叶斯表述的稳定性和可靠性提供了严格的理论依据，并表明即使在具有挑战性的数据生成机制下，该方法也能适应潜在信号。

除理论贡献外，我们还开展了广泛的模拟研究，以评估所提出方法的有限样本表现。在涵盖多种稀疏度水平、噪声结构以及删失或缺失数据观测模式的广泛场景中，我们的贝叶斯估计量通常优于其基于套索的方法。最后，我们通过将该方法应用于单细胞数据，展示了其实际效用。该模型在适应这些数据集内在的稀疏性与不确定性的同时，成功捕捉到了关键的生物信号。

本文其余部分的安排如下。第2 节介绍删失数据情形下的模型，以及采用吉布斯采样器的方法。第3 节介绍缺失数据情形下的模型，以及采用吉布斯采样器的方法。第4 节和第5节给出数值模拟研究和真实数据应用。第6 节介绍理论结果，证明见附录B。第7节进行讨论并总结。附录A中给出了其他模拟结果和图表。

2 删失图形马蹄先验

2.1 模型设定

令 $Y = (Y_1, \ldots\ldots, Y_n) \in \mathbb{R}^{n\times p}$ 表示数据矩阵，其中每个 $Y_i \sim \mathcal{N}_p(0, \Omega^{-1})$。这里，$\Omega$ 表示精度矩阵（协方差矩阵的逆矩阵）。然而，在实际中，变量并非总能被完整观测；相反，测量值会受到各变量特定阈值处的左删失。对于每个坐标j，如果真实值 Y_{ij} 低于删失阈值c_j，我们只记录该阈值本身。因此，

观测值的形式为

$$\widetilde{Y}_{ij} = \begin{cases} Y_{ij}, & \text{if } Y_{ij} > c_j, \\ c_j, & \text{if } Y_{ij} \le c_j, \end{cases} \tag{1}$$

对于 $i \in \{1, \ldots, n\}$ 和 $j \in \{1, \ldots, p\}$。

为处理删失问题，我们引入潜在未删失变量 $Z = (Z_1, \ldots, Z_n)^\top \in \mathbb{R}^{n \times p}$，其分布与原始高斯模型相同，即 $Z_i \sim \mathcal{N}_p(0, \Omega^{-1})$。观测数据随后通过 $Y_{ij} = \max(Z_{ij}, c_j)$ 与潜在过程相关联。因此，推断的重点是基于删失观测 Y 估计稀疏精度矩阵 Ω。

对于每个观测值 i，我们定义集合

$$\mathcal{O}_i = \{j : \widetilde{Y}_{ij} > c_j\}, \quad \text{and} \quad \mathcal{C}_i = \{j : \widetilde{Y}_{ij} = c_j\}$$

分别表示未删失条目和删失条目的索引。相应地对 Z_i 进行划分后，观测数据似然可表示为未删失分量的高斯密度项与删失分量的多元高斯分布函数项的乘积：

$$L(\Omega; \widetilde{Y}) = \prod_{i=1}^{n} \phi_{|\mathcal{O}_i|}\left(\widetilde{Y}_{i,\mathcal{O}_i}; 0, \Sigma_{\mathcal{O}_i,\mathcal{O}_i}\right) \cdot \Phi_{|\mathcal{C}_i|}\left(c_{\mathcal{C}_i} \mid \mu_{i,\mathcal{C}_i|\mathcal{O}_i}, \Sigma_{\mathcal{C}_i|\mathcal{O}_i}\right), \tag{2}$$

其中 $\Sigma = \Omega^{-1}$，给定已观测分量时，删失分量的条件均值和协方差为

$$\begin{aligned} \mu_{i,\mathcal{C}_i|\mathcal{O}_i} &= \Sigma_{\mathcal{C}_i,\mathcal{O}_i} \Sigma_{\mathcal{O}_i,\mathcal{O}_i}^{-1} \widetilde{Y}_{i,\mathcal{O}_i}, \\ \Sigma_{\mathcal{C}_i|\mathcal{O}_i} &= \Sigma_{\mathcal{C}_i,\mathcal{C}_i} - \Sigma_{\mathcal{C}_i,\mathcal{O}_i} \Sigma_{\mathcal{O}_i,\mathcal{O}_i}^{-1} \Sigma_{\mathcal{O}_i,\mathcal{C}_i}. \end{aligned}$$

2.2 先验设定

为了使精度矩阵具有稀疏性，如 [21], 我们对非对角元素赋予图形马蹄先验。对于索引 $j < k$，先验设定采用以下层次形式

$$\begin{aligned} \omega_{jk} \mid \lambda_{jk}, \tau &\sim \mathcal{N}(0, \lambda_{jk}^2 \tau^2), \\ \lambda_{jk} &\sim \mathcal{C}^+(0, 1), \\ \tau &\sim \mathcal{C}^+(0, 1), \end{aligned} \tag{3}$$

其中，$\mathcal{C}^+$ 表示半柯西分布。对角元素 ω_{jj} 受到约束以确保正定性，通常赋予非信息先验，或在此类约束下进行估计。

为每个系数指定一个自身的尺度因子，记为 λ_{jk}，它充当局部收缩参数；同时，所有系数共用一个尺度因子 τ，作为全局收缩参数。这种层次先验结构带来两种理想效果：先验概率质量在零附近高度集中，从而强力抑制噪声变量；同时具有重尾，使较大且有信息量的信号基本不受影响 [12]。

2.3 用于删失的数据增强

删失结构天然适用于数据增广方法，其中潜在高斯变量 Z 被视为缺失数据。给定观测值数据 Y 和删失阈值（$c_1, \ldots, c_p$）后，从其条件分布中更新增广变量 Z_{ij}。如果某个观测值未被删失，那么 Z_{ij} 与 Y_{ij} 相同，并服从其条件高斯分布。如果某个观测值被删失，那么 Z_{ij} 从相同的高斯分布中抽取，但限定在区域（$-\infty, c_j$]内。形式上，条件分布的形式为

$$Z_{ij} \mid \widetilde{Y}_{ij}, \Omega, Z_{i,-j} \sim \begin{cases} \mathcal{N}(\mu_{ij}, \sigma_j^2), & \text{if } \widetilde{Y}_{ij} = Z_{ij}, \\ \mathcal{N}(\mu_{ij}, \sigma_j^2)\, \mathbb{I}(Z_{ij} \le c_j), & \text{if } \widetilde{Y}_{ij} = c_j, \end{cases}$$

其中，μ_{ij} 和σ_j^2 由多元正态模型下给定其余坐标时 Z_{ij} 的条件分布导出：

$$Z_{ij} \mid Z_{i,-j} \sim \mathcal{N}\left(-\frac{1}{\omega_{jj}} \sum_{k \neq j} \omega_{jk} Z_{ik}\,,\, \omega_{jj}^{-1}\right).$$

2.4 删失图形马蹄先验的 Gibbs 采样器

2.4.1 通过逐节点回归进行参数化

为了对 Ω 进行参数化，我们采用逐节点回归框架 [32, 40]。令 $Z_{\cdot j}$ 为 Z的第 j列，$Z_{\cdot,\,-j}$ Z 中去除第 j 列后的 $n \times(p-1)$ 子矩阵，则给定其余变量时，$Z_{\cdot j}$的条件分布建模为

$$Z_{\cdot j} \,\big|\, Z_{\cdot,-j},\ \theta_j,\ \sigma_j^2 \ \sim\ \mathcal{N}\big(Z_{\cdot,-j}\,\theta_j,\ \sigma_j^2 I_n\big),\ j = 1, \ldots, p,$$

其中，$\theta_j \in \mathbb{R}^{p-1}$ 表示回归系数，$\sigma_j^2 > 0$ 为条件方差。它与精度矩阵之间的关系明确如下：

$$\Omega_{jj} = \sigma_j^{-2}, \qquad \Omega_{-j,j} = -\sigma_j^{-2}\,\theta_j.$$

由于每个 p 回归都会产生一个独立的估计值，因此需要执行对称化步骤（如下所述），将它们整合为一个对称精度矩阵。在这种形式下，对于满足 $k \neq j$的每个系数 θ_{jk}，有：

$$\theta_{jk} \,\big|\, \lambda_{jk}^2, \tau_j^2, \sigma_j^2 \ \sim\ \mathcal{N}\big(0,\ \sigma_j^2 \tau_j^2 \lambda_{jk}^2\big),$$

局部收缩为 λ $_{jk}^2$和全局收缩 τ_j^2。遵循 Carvalho 等人 [12],，先验层级结构 is

$$\begin{aligned}
\lambda_{jk}^2 \mid \nu_{jk} &\sim \text{Inv-Gamma}(1,\, \nu_{jk}^{-1}), & \nu_{jk} &\sim \text{Inv-Gamma}(1,\, 1),\\
\tau_j^2 \mid \xi_j &\sim \text{Inv-Gamma}\Big(\tfrac{m+1}{2},\ \xi_j^{-1}\Big), & \xi_j &\sim \text{Inv-Gamma}(1,\, 1),
\end{aligned}$$

其中 $m = p - 1$。残差方差设为逆伽马先验，$\sigma_j^2 \sim$ Inv-Gamma(a_0，b_0)。注意，此处我们采用形状–率参数化，Inv-Gamma(a，b) 的密度与 $x^{-a-1}\exp(-b/x)$ 成正比。

给定 (θ_j，σ_j^2)，潜变量 Z_{ij}的条件均值为 $\mu_{ij} = z_{i\,\cdot,\,-j}^{\top}\theta_j$，方差为 σ_j^2。观测机制定义如下：

(a) 若 $Y_{ij} > c_j$ （未删失），则 $Z_{ij} = Y_{ij}$ 被完全观测。

(b) 若 $Y_{ij} = c_j$ （删失），则 Z_{ij} 服从截断正态分布，

$$Z_{ij} \,\big|\, \mathcal{C}_{ij} = 1 \ \sim\ \mathcal{N}(\mu_{ij}, \sigma_j^2) \ \text{ with support } (-\infty, c_j].$$

2.4.2 后验计算

后验推断通过吉布斯采样器进行，其中参数和潜变量的每个区块依次更新。采样器交替执行删失潜变量值的插补、回归参数的更新、局部和全局收缩尺度的调整，以及精度矩阵的重构。下面详细介绍该算法的各个组成部分。

采样器的第一步更新潜在高斯数据 Z。当某个观测值未被删失，即 $Y_{ij} > c_j$时，相应的潜在值将被确定性地设为 $Z_{ij} = Y_{ij}$。对于删失

观测值 $Y_{ij} = c_j$，潜在条目将从均值为$\mu_{ij} = z_{i,\,-j}^{\top}\theta_j$、方差为 σ_j^2，并在 c_j 处进行上截断的截断正态分布中抽取：

$$Z_{ij} \;\sim\; \mathcal{N}(\mu_{ij}, \sigma_j^2) \;\text{ restricted to } (-\infty, c_j]. \tag{4}$$

给定增广数据，回归系数 θ_j 从其高斯全条件分布中更新。对于每个节点 $j \in \{1, \ldots, p\}$，定义响应向量 $y_j = Z_{\cdot,\,j}$和预测变量矩阵 $X^{(j)} = Z_{\cdot,\,-j}$。条件高斯似然给出回归模型

$$y_j = X^{(j)}\theta_j + \varepsilon_j, \qquad \varepsilon_j \sim \mathcal{N}(0, \sigma_j^2 I_n).$$

回归系数采用马蹄先验：

$$\theta_{jk} \mid \lambda_{jk}, \tau_j, \sigma_j^2 \sim \mathcal{N}(0, \sigma_j^2\tau_j^2\lambda_{jk}^2), \quad \lambda_{jk} \sim \mathcal{C}^{+}(0,1), \quad \tau_j \sim \mathcal{C}^{+}(0,1).$$

后验更新步骤如下。回归系数的条件协方差和均值为

$$V_j = \left(X^{(j)\top} X^{(j)} + \operatorname{diag}\left(\frac{1}{\tau_j^2 \lambda_{j\cdot}^2} \right) \right)^{-1}, \qquad \mu_{\theta_j} = V_j X^{(j)\top} y_j,$$

并通过 $\theta_j \sim \mathcal{N}(\mu_{\theta_j},\ \sigma_j^2 V_j)$。更新系数。

残差方差由其逆伽马条件后验分布更新，

$$\sigma_j^2 \sim \text{Inv-Gamma}\left(a_0 + \frac{n}{2},\; b_0 + \frac{1}{2}\|y_j - X^{(j)}\theta_j\|^2 \right).$$

局部尺度采用 Makalic–Schmidt 参数化进行更新，

$$\lambda_{jk}^2 \sim \text{Inv-Gamma}\left(1,\; \frac{1}{\nu_{jk}} + \frac{\theta_{jk}^2}{2\sigma_j^2\tau_j^2} \right), \qquad \nu_{jk} \sim \text{Inv-Gamma}(1, 1 + 1/\lambda_{jk}^2),$$

并且全局尺度参数更新为

$$\tau_j^2 \sim \text{Inv-Gamma}\left(\frac{m+1}{2},\; \frac{1}{\xi_j} + \sum_k \frac{\theta_{jk}^2}{2\sigma_j^2\lambda_{jk}^2} \right), \qquad \xi_j \sim \text{Inv-Gamma}(1, 1 + 1/\tau_j^2).$$

更新所有节点回归后，重构精度矩阵 Ω。对角元素由残差方差的倒数确定：$\Omega_{jj} = 1/\sigma_j^2$，非对角元素则由回归系数得到：$\Omega_{-j,\,j} = -\theta_j/\sigma_j^2$. 由于每个回归是分别更新的，重构后的矩阵可能并非完全对称。因此需要执行对称化步骤，例如对上、下三角元素取平均。如有必要，还可以将所得矩阵投影到对称正定矩阵锥中，以确保其有效性。

舍弃初始预热期后，计算后验汇总统计量，例如精度矩阵均值与中位数，具体为

$$\hat{\Omega}_{\text{mean}} = \frac{1}{n_{\text{keep}}} \sum_{t \in \text{kept}} \Omega^{(t)}, \qquad \hat{\Omega}_{\text{median}} = \text{median}_{t \in \text{kept}}\, \Omega^{(t)}.$$

2.4.3 初始化与实际注意事项

潜变量 Z 通过以下方式初始化：未删失时设为 $Z_{ij} = Y_{ij}$，删失条目则设为 $Z_{ij} = c_j - \epsilon_{ij}$，其中 $\epsilon_{ij} \sim |\mathcal{N}(0,1)|$。回归参数初始化为中性值（$\theta_j = 0$、$\sigma_j^2 = 1$、$\lambda_{jk}^2 = \nu_{jk} = \tau_j^2 = \xi_j = 1$）。默认超参数 $(a_0, b_0) = (10^{-2}, 10^{-2})$ 为方差项提供弱信息先验，而层次马蹄先验则能自适应地收缩小系数，同时避免过度收缩强信号。高效的截断正态抽样依赖于稳定的逆累积分布函数法实现，并针对极端截断采用指数倾斜。线性代数计算采用 A 的乔列斯基分解，并通过 $\theta_j = \mu_\theta + \sigma_j\, L^{-\top}U$ 进行高斯抽样，其中 $LL^{\top} = A$ 且 $U \sim \mathcal{N}(0, I)$。

2.5 扩展至右删失数据

扩展到右删失数据的处理方式类似，即重点关注 $(-\infty, c]$，我们使用 $[c, \infty)$。形式上，条件分布的形式为

$$Z_{ij} \mid \widetilde{Y}_{ij}, \Omega, Z_{i,-j} \sim \begin{cases} \mathcal{N}(\mu_{ij}, \sigma_j^2), & \text{if } \widetilde{Y}_{ij} = Z_{ij}, \\ \mathcal{N}(\mu_{ij}, \sigma_j^2)\, \mathbb{I}(Z_{ij} \geq c_j), & \text{if } \widetilde{Y}_{ij} = c_j, \end{cases}$$

因此，（4）被替换为

$$Z_{ij} \sim \mathcal{N}(\mu_{ij}, \sigma_j^2) \text{ restricted to } [c_j, \infty). \tag{5}$$

算法的另一步保持不变。

3 含缺失数据的图形马蹄先验

3.1 模型设定与先验设定

考虑来自一个均值为零、精度矩阵为 $\Omega \succ 0$ 的 p 维高斯分布的独立同分布样本 $Y = (Y_1, \ldots, Y_n) \in \mathbb{R}^{n \times p}$。在实践中，这类数据集通常包含缺失值。令 $\mathcal{O} \subseteq \{1, \ldots, n\} \times \{1, \ldots, p\}$ 表示观测条目的集合，$\mathcal{M} = \mathcal{O}^c$ 表示缺失条目的集合。对于 $(i, j) \in \mathcal{O}$，我们观测到 Y_{ij}，而 $\mathcal{M}$ 中的值仍未观测到。推断目标是在此类不完整数据下恢复稀疏精度矩阵 Ω。

我们始终假设数据是随机缺失的，即缺失概率仅取决于观测到的变量，而不取决于未观测值本身。在此假设下，基于似然的推断可以忽略缺失机制，从而使我们能够通过数据增广来构建贝叶斯估计程序。

观测数据似然通过对每个 Y_i 的缺失分量进行边际化得到。令 $\mathcal{O}_i$ 表示第 i 行中观测到的坐标集合，则 Y_i 服从均值为零、协方差矩阵为 $\Sigma = \Omega^{-1}$ 对应主子矩阵的多元高斯分布。因此，

$$L(\Omega; \{\widetilde{Y}_i\}_{i=1}^n) = \prod_{i=1}^n p(\widetilde{Y}_i \mid \Omega),$$

其中 $\widetilde{Y}_i \sim \mathcal{N}_{|\mathcal{O}_i|}\big(0,\ \Sigma_{\mathcal{O}_i,\ \mathcal{O}_i}\big)$. 等价地，似然可写为

$$L(\Omega; \widetilde{Y}) = \prod_{i=1}^n (2\pi)^{-|\mathcal{O}_i|/2} \, |\Sigma_{\mathcal{O}_i, \mathcal{O}_i}|^{-1/2} \exp\left(-\frac{1}{2} \widetilde{Y}_i^\top \Sigma_{\mathcal{O}_i, \mathcal{O}_i}^{-1} \widetilde{Y}_i \right), \tag{6}$$

其中 $\Sigma = \Omega^{-1}$。这种表示强调，只有观测条目会对似然作出贡献，而未观测条目则被积分消去。

这里，我们采用与上一节相同的图形马蹄先验，见（3）.

3.2 通过吉布斯抽样进行后验计算

在存在缺失数据的情况下，采用吉布斯抽样算法，在图形马蹄先验下进行后验推断。该采样器交替执行两步：填补潜在高斯矩阵中的缺失条目，以及在图形马蹄先验下更新精度矩阵。

3.2.1 初始化

吉布斯采样器首先为潜在数据矩阵 Z 和精度矩阵 Ω 设置初始值。对于缺失条目，可以采用简单的填补方法，例如使用列均值，或从根据每个变量的观测分量拟合的高斯分布中抽样，以获得合理的初始值。因此，初始矩阵为

$$Z_{ij}^{(0)} = \begin{cases} \widetilde{Y}_{ij}, & (i,j) \in \mathcal{O}, \\ \text{imputed value}, & (i,j) \in \mathcal{M}. \end{cases}$$

精度矩阵 Ω 必须初始化为对称正定矩阵。单位矩阵是最简单的选择，而岭正则化逆协方差估计可以加快收敛。在所有情况下，初始化都必须满足 $\Omega^{(0)} \succ 0$。

3.2.2 缺失值插补

在每次迭代中，根据观测分量和当前协方差估计，从条件高斯分布中重新抽取 Z 中的缺失条目。令 $\Sigma^{(t-1)} = (\Omega^{(t-1)})^{-1}$，并将其划分为对应于给定行 i的观测索引 o 和缺失索引 m 的分块。缺失值的条件均值和协方差为

$$\mu_m = \Sigma_{mo}\Sigma_{oo}^{-1} Z_{i,o}, \qquad \Sigma_m = \Sigma_{mm} - \Sigma_{mo}\Sigma_{oo}^{-1}\Sigma_{om}.$$

然后将缺失条目插补为

$$Z_{i,m} \sim \mathcal{N}(\mu_m, \Sigma_m).$$

3.2.3 逐节点回归更新

获得完整的潜在数据矩阵 Z 后，我们按照第2.4节所述，在图形马蹄先验下通过节点回归更新精度矩阵。为简洁起见，此处不再重复介绍。

完整的潜在数据矩阵 Z 可同时获得，从而为高维高斯情形下的插补与稀疏图模型构建提供统一且连贯的贝叶斯框架。

4 数值研究

针对删失数据和缺失数据两种情形的方法均已通过R 软件包GHScenmis实现，可从 GitHub 获取：
https://github.com/tienmt/ghscenmis。

4.1 模拟设置与比较方法

我们将所提出的方法与针对删失数据和缺失数据调整的套索方法进行比较，如 [3, 39], 所述；其软件实现可通过R 软件包cglasso [4]获取。我们的吉布斯采样器运行 5000 次采样，并将前 1000 次作为预热期。cglasso 方法采用默认选项运行，其调优参数通过 AIC 信息准则选取。

在我们的模拟研究中，我们考虑两种生成真实精度矩阵（Ω_{true}）及其对应协方差矩阵（Σ_{true}）的设置。

设置 1： 三对角精度矩阵。
在第一种情形中，我们构造一个三对角精度矩阵，其非对角元素表示

相邻变量之间的弱条件依赖关系。形式上，令 p 表示变量的数量，并定义一个 $p \times p$ 矩阵 A，其中

$$A_{ij} = \begin{cases} 0.3, & \text{if } |i-j| = 1 \\ 0, & \text{otherwise} \end{cases}$$

因此，真实精度矩阵为 $\Omega_{\text{true}} = \text{diag}(1, \ldots, 1) + A$。相应的协方差矩阵通过矩阵求逆得到：$\Sigma_{\text{true}} = \Omega_{\text{true}}^{-1}$. 该结构模拟了一个简单的链图，其中每个变量都依赖于其紧邻变量。

设置 2： 小型全连接子集。
在第二种情景中，我们引入一个小型、密集连接的变量子集，以研究少数变量之间的中等相关性所产生的影响。我们将协方差矩阵 Σ_{true} 定义为单位矩阵，并在前三个变量之间添加值为 0.5 的非对角元素：

$$\Sigma_{\text{true}} = \begin{pmatrix} 1 & 0.5 & 0.5 & 0 & \cdots & 0 \\ 0.5 & 1 & 0.5 & 0 & \cdots & 0 \\ 0.5 & 0.5 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & & 0 \\ \vdots & \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{pmatrix}.$$

随后计算得到相应的精度矩阵 $\Omega_{\text{true}} = \Sigma_{\text{true}}^{-1}$, 它仍然是一个非常稀疏的矩阵。这种设置使我们能够考察存在小型但强相关性块时的估计性能，并将其与设置 1 中的稀疏三对角结构进行对比。

我们评估每种方法模拟设置的 $\|\Omega - \Omega_{\text{true}}\|_2$，重复 100 次。$\widehat{\Omega}$，误差定义为 $\|\widehat{\Omega}$

4.2 删失数据的模拟结果

我们首先展示在多种情景下对数据进行删失的模拟研究。

维度递增时的结果。 我们将样本量设为 $n = 200$，并改变维度$p \in \{10, 20, 30\}$。数据由多元高斯分布生成，其真实精度矩阵对应于设置 I 或设置 II（如上所述）。随后，在固定阈值 $c = (-0.5,\ 0.5,\ -0.5,\ 0.5,\ \ldots\ldots,\ -0.5,\ 0.5)$ 处对生成的数据进行左删失。此过程重复 100 次，并在表 1 中汇总所得估计误差。

结果显示出明显趋势：随着维度增加，HScen 和删失图形套索（cglasso）的误差均增大。尽管如此，我们的贝叶斯方法始终能取得更小的误差，凸显了其在高维删失情形下的优势。

数据删失比例逐渐增加时的结果。接下来，我们考虑样本量固定为 $n = 200$ 且维度为 $p = 10$的情形。我们不再对固定向量进行删失，而是考察不同删失比例下的影响：数据的 10%、20% 和 30%。具体而言，数据由多元高斯分布生成，其真实精度矩阵采用设置 I 或设置 II（如上所述），并在与目标比例相对应的经验分位数处进行删失。该过程重复 100 次，所得估计误差汇总于表 2。

结果表明，HScen 和 cglasso 都会受到删失程度的影响。然而，我们的贝叶斯方法 HScen 在所有情景下始终优于 cglasso；当 50% 的数据被删失时，二者之间的性能差距尤为明显。

表1：删失数据的模拟结果，变量维度逐渐增加（变量数量）$n = 200$，删失向量固定为 $c = (-0.5,\ 0.5,\ \ldots\ldots,\ -0.5,\ 0.5)$。TPR：真正率；FDR：假阳性率。

方法	$\|\widehat{\beta} - \beta_0\|_2^2$	TPR	FDR	$\|\widehat{\beta} - \beta_0\|_2^2$	TPR	FDR
	设置 I，$p = 10$			设置 II，$p = 10$		
cgLasso	2.13 (0.21)	0.00 (0.01)	0.00 (0.00)	1.91 (2.02)	1.00 (0.00)	0.59 (0.40)
cenGHS	0.65 (0.23)	0.82 (0.11)	0.01 (0.02)	0.65 (0.38)	0.92 (0.15)	0.00 (0.01)
	设置 I，$p = 20$			设置 II，$p = 20$		
cgLasso	4.44（0.19）	0.00（0.01）	0.00（0.00）	2.44 (2.47)	1.00 (0.03)	0.27 (0.35)
cenGHS	1.60 (0.42)	0.71 (0.10)	0.01 (0.01)	0.96 (0.37)	0.87 (0.16)	0.00 (0.00)
	设置 I，$p = 30$			设置 II，$p = 30$		
cgLasso	6.69 (0.26)	0.00 (0.01)	0.00 (0.00)	3.04 (4.63)	1.00 (0.03)	0.11 (0.26)
cenGHS	2.53 (0.44)	0.61 (0.10)	0.00 (0.00)	1.42 (0.75)	0.79 (0.16)	0.00 (0.00)

表2：$n = 200$，$p = 10$ 删失数据的模拟结果，其中数据删失比例依次增加至 10%、20% 和 30%。真正率（TPR）：真正率；假发现率（FDR）：假阳性率。

方法	$\|\widehat{\beta} - \beta_0\|_2^2$	TPR	FDR	$\|\widehat{\beta} - \beta_0\|_2^2$	TPR	FDR
	设置 I，10% 删失			设置 II，10% 删失		
cgLasso	0.45 (0.34)	1.00 (0.00)	0.51 (0.24)	0.41 (0.19)	1.00 (0.00)	0.28 (0.15)
cenGHS	0.38 (0.13)	0.99 (0.04)	0.03 (0.03)	0.39 (0.18)	0.99 (0.05)	0.01 (0.02)
	设置 I，20% 删失			设置 II，20% 删失		
cgLasso	0.54 (0.26)	1.00 (0.00)	0.62 (0.29)	0.55 (0.41)	1.00 (0.00)	0.43 (0.29)
cenGHS	0.41 (0.14)	0.98 (0.04)	0.04 (0.03)	0.42 (0.18)	1.00 (0.03)	0.01 (0.02)
	设置 I，30% 删失			设置 II，30% 删失		
cgLasso	0.66 (0.22)	0.99 (0.03)	0.54 (0.35)	0.87 (0.60)	1.00 (0.00)	0.62 (0.35)
cenGHS	0.46 (0.14)	0.96 (0.06)	0.03 (0.03)	0.46 (0.23)	0.97 (0.09)	0.01 (0.02)

样本量增加时的结果。接下来，我们研究增加样本量对估计精度的影响。维度固定为 $p = 10$，并在固定向量（-0.5、0。5、-0.5、0。5、……、-0.5、0。5）处施加删失。样本量在 $n = 200$、500、1000 之间变化。数据由多元高斯分布生成，其真实精度矩阵分别对应设置 I 或设置 II（如上文所述）。100 次重复试验的结果汇总于表3。

正如预期，随着样本量增加，HScen 和 cglasso 的估计误差均有所降低。然而，贝叶斯方法 HScen 展现出明显优势，误差显著更小，尤其是在样本量较大时。此外，在设置 I 下，底层图结构为链时，cglasso 的表现较差，这进一步凸显了我们方法的稳健性。

4.3 缺失数据模拟

维度递增时的结果。我们将样本量固定为 $n = 200$，并改变维度$p \in \{10, 20, 30\}$。数据由多元高斯分布生成，其真实精度矩阵由设置 I 或设置 II 指定（如上所述）。为模拟缺失情况，我们从模拟得到的观测值中均匀随机地删除 10%。重复此过程 100 次，并将所得估计误差列于表4中。

结果表明，两种方法的估计误差都会随着维度增加而增大。尽管如此，贝叶斯方法 HScen 始终比 cgLasso 的误差更小；当真实精度矩阵对应设置 II 时，两者的性能差距最为明显。

表 3：对于 $p = 10$ 的删失数据，在删失向量固定为 $c = (-0.5,\ 0。5,\ \ldots\ldots,\ -0.5,\ 0。5)$ 时，随着样本量增加的模拟结果。TPR：真正率；FDR：假发现率。

方法	$\|\widehat{\beta} - \beta_0\|_2^2$	TPR	FDR	$\|\widehat{\beta} - \beta_0\|_2^2$	TPR	FDR
	设置 I，$n = 200$			设置 II，$n = 200$		
cgLasso	2.13 (0.21)	0.00 (0.01)	0.00 (0.00)	1.91 (2.02)	1.00 (0.00)	0.59 (0.40)
cenGHS	0.65 (0.23)	0.82 (0.11)	0.01 (0.02)	0.65 (0.38)	0.92 (0.15)	0.00 (0.01)
	设置 I，$n = 500$			设置 II，$n = 500$		
cgLasso	2.01（0.09）	0.00（0.00）	0.00（0.00）	1.00 (0.63)	1.00 (0.00)	0.72 (0.41)
cenGHS	0.24 (0.08)	1.00 (0.00)	0.02 (0.02)	0.21 (0.10)	1.00 (0.00)	0.01 (0.01)
	设置 I，$n = 1000$			设置 II，$n = 1000$		
cgLasso	1.97 (0.08)	0.04 (0.15)	0.00 (0.00)	0.82 (0.51)	1.00 (0.00)	0.76 (0.41)
cenGHS	0.11 (0.03)	1.00 (0.00)	0.02 (0.02)	0.10 (0.05)	1.00 (0.00)	0.01 (0.01)

表 4：$n = 200$ 和维度增加条件下 10% 缺失数据的模拟结果。TPR：真正率，FDR：假发现率。

方法	$\|\widehat{\beta} - \beta_0\|_2^2$	TPR	FDR	$\|\widehat{\beta} - \beta_0\|_2^2$	TPR	FDR
	设置 I，$p = 10$			设置 II，$p = 10$		
cgLasso	0.62 (0.37)	1.00 (0.00)	0.66 (0.30)	0.45 (0.30)	1.00 (0.00)	0.34 (0.20)
cenGHS	0.38 (0.12)	0.96 (0.06)	0.02 (0.03)	0.37 (0.17)	0.97 (0.09)	0.01 (0.01)
	设置 I，$p = 20$			设置 II，$p = 20$		
cgLasso	0.93 (0.14)	1.00 (0.00)	0.42 (0.07)	0.71 (0.20)	1.00 (0.00)	0.09 (0.09)
cenGHS	0.91 (0.18)	0.90 (0.07)	0.01 (0.01)	0.58 (0.23)	0.95 (0.12)	0.00 (0.00)
	设置 I，$p = 30$			设置 II，$p = 30$		
cgLasso	1.66 (0.20)	1.00 (0.01)	0.33 (0.12)	0.98 (0.21)	1.00 (0.00)	0.03 (0.04)
cenGHS	1.42 (0.25)	0.85 (0.07)	0.00 (0.00)	0.78 (0.31)	0.92 (0.15)	0.00 (0.00)

样本量递增时的结果。接下来，我们考察样本量增加对估计精度的影响。维度固定为 $p = 10$，样本量设为 $n = 200$、500和1000。数据从多元高斯分布生成，真实精度矩阵对应设置 I 或设置 II（如上所述），随后以均匀随机方式移除 10% 的数据。100 次重复试验的结果见表5。

正如预期，随着样本量增加，两种方法的误差均会降低。然而，在所有情景下，贝叶斯方法 HScen 始终优于 cglasso，表现出对缺失数据更强的稳健性。

缺失数据比例增加时的结果。现在将样本量固定为 $n = 200$，维度固定为 $p = 10$，并研究缺失数据比例变化的影响。具体而言，我们考虑 10%、20% 和 30% 的缺失水平。数据生成自多元高斯分布，其真实精度矩阵分别对应设置 I 或设置 II（如上所述）。随后，按照指定比例的条目以完全随机的方式移除。该过程重复 100 次，所得估计误差如表6所示。

正如预期，随着缺失数据比例上升，两种方法的估计误差均略有增加。然而，贝叶斯方法 cen GHS 始终优于 cglasso，在各个缺失程度下都能获得更小且更稳定的误差。

附录中给出了一些轨迹图和自相关函数图A ，并列出了 MCMC 样本的有效样本量，以展示该算法的行为。

表5：样本量增加 $p = 10$ 时10%缺失数据的模拟结果。TPR：真正率；FDR：错误发现率。

方法	$\|\widehat{\beta} - \beta_0\|_2^2$	TPR	FDR	$\|\widehat{\beta} - \beta_0\|_2^2$	TPR	FDR
	设置 I，$n = 200$			设置 II，$n = 200$		
cgLasso	0.62（0.37）	1.00 (0.00)	0.66 (0.30)	0.45 (0.30)	1.00 (0.00)	0.34 (0.20)
cenGHS	0.38 (0.12)	0.96 (0.06)	0.02 (0.03)	0.37 (0.17)	0.97 (0.09)	0.01 (0.01)
	设置 I，$n = 500$			设置 II，$n = 500$		
cgLasso	0.27 (0.08)	1.00 (0.00)	0.86 (0.29)	0.19 (0.12)	1.00 (0.00)	0.28 (0.24)
cenGHS	0.14 (0.05)	1.00 (0.00)	0.03 (0.03)	0.14 (0.06)	1.00 (0.00)	0.01 (0.02)
	设置 I，$n = 1000$			设置 II，$n = 1000$		
cgLasso	0.13 (0.03)	1.00 (0.00)	1.00 (0.00)	0.17 (0.05)	1.00 (0.00)	0.58 (0.46)
cenGHS	0.07 (0.02)	1.00 (0.00)	0.03 (0.03)	0.07 (0.03)	1.00 (0.00)	0.01 (0.02)

表6：使用 $n = 200$，$p = 20$ 对删失数据进行模拟，删失数据量依次增加至 10%、20% 和 30%。TPR：真正率；FDR：假发现率。

方法	$\|\widehat{\beta} - \beta_0\|_2^2$	TPR	FDR	$\|\widehat{\beta} - \beta_0\|_2^2$	TPR	FDR
	设置 I，缺失 10%			设置 II，缺失 10%		
cgLasso	0.93 (0.14)	1.00 (0.00)	0.42 (0.07)	0.71 (0.20)	1.00 (0.00)	0.09 (0.09)
cenGHS	0.91 (0.18)	0.90 (0.07)	0.01 (0.01)	0.58 (0.23)	0.95 (0.12)	0.00 (0.00)
	设置 I，缺失 20%			设置 II，缺失 20%		
cgLasso	1.65 (1.24)	1.00 (0.01)	0.53 (0.22)	0.89 (0.22)	1.00 (0.00)	0.08 (0.11)
cenGHS	1.15 (0.21)	0.75 (0.09)	0.00 (0.00)	0.67 (0.31)	0.90 (0.15)	0.00 (0.00)
	设置 I，缺失 30%			设置 II，缺失 30%		
cgLasso	1.94 (0.78)	0.99 (0.02)	0.61 (0.26)	1.17 (0.26)	0.99 (0.06)	0.05 (0.06)
cenGHS	1.71 (0.26)	0.49 (0.10)	0.00 (0.00)	0.81 (0.30)	0.72 (0.16)	0.00 (0.00)

5 单细胞数据的应用：巨核细胞-红系祖细胞

在一项关于血细胞生成的研究中，Psaila 等人 [36] 通过分化，发现了三种由造血干细胞衍生的不同细胞亚群。其中一个亚群被称为 MK-MEP，是一种此前未被表征的罕见双能细胞群，主要产生巨核细胞后代。数据可从`cglasso` 软件包获取

使用多重RT-qPCR分析了48个人源单个MK-MEP细胞中63个基因的表达情况。RT-qPCR测量值通常为右删失，制造商规定的检测限为40（即删失水平）。按照Pipelers等人提出的方法，对原始表达值进行均值归一化 [34]。更多详情请参阅Augugliaro等人 [3]。

我们运行 GHScen方法 10,000 次迭代，并丢弃前 1,000 次迭代作为预热期。图 1展示了不同方法得到的估计图。对于 GHScen，使用不包含零的 95%可信区间来识别精度矩阵中的非零元素。与 cgLasso 相比，GHScen 估计的图明显更加稀疏，并揭示了五个不同的连通子群，而 cgLasso 则产生了一个较大的连通分量。

6 理论结果

本节研究所提出的删失图形马蹄先验在存在缺失或删失

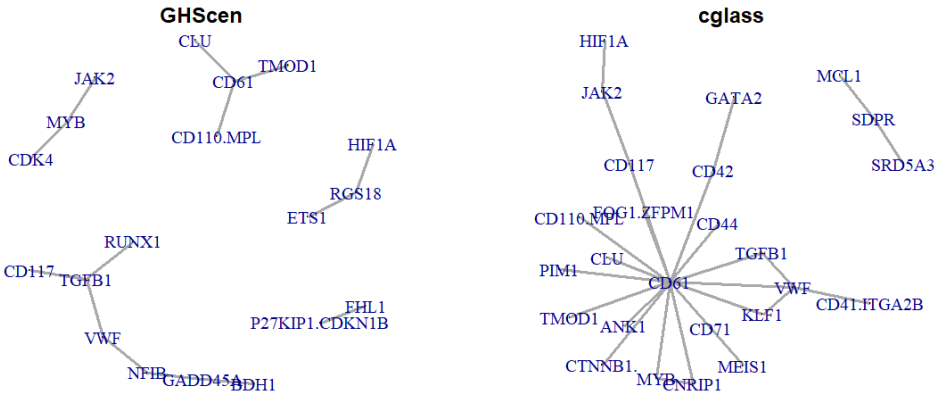

图1：真实数据集 MKMEP 上不同方法得到的图。与 cglasso 相比，我们看到 GHSsen 返回的网络更加稀疏。

观测值的高维高斯图模型中，温度化（或分数）后验的后验收缩。主要技术难点来自观测数据似然：缺失会产生边际高斯似然，而删失则会引入非线性对数累积分布函数项，从而扭曲曲率并使全局对数凹性不再成立。我们证明，尽管存在这些复杂因素，图形马蹄先验的全局—局部收缩结构仍能在标准稀疏性和谱正则条件下带来精确的后验集中。

我们的分析基于 R´enyidivergence 进行，该散度为温度化后验提供了一种自然的信息论损失，并且非常适用于具有非二次结构的似然。结果表明，后验以与完全观测的高斯图模型相同的阶集中于真实精度矩阵附近，其差异仅在于依赖删失机制的常数。

令 p_Ω 和 P_Ω 分别表示高斯 $\mathcal{N}(0,\ \Omega^{-1})$随机变量的密度和分布。用 $\pi_{HS}(\Omega)$表示 Ω 上的先验，温度后验 [8, 1, 25, 29] 定义为

$$\pi_{n,\alpha}(\Omega) \propto L^{\alpha}(\Omega;\tilde{Y})\pi_{HS}(\Omega).$$

我们考虑关于真实参数的以下假设。

假设1(有界谱**)**。精度矩阵Ω、Ω_0 属于以下谱带

$$\mathcal{S} = \{\Omega \succ 0 :\ 0 < m \le \lambda_{\min}(\Omega) \le \lambda_{\max}(\Omega) \le M < \infty\}.$$

假设2。真实精度矩阵Ω_0 是稀疏的，具体而言

$$\Omega_0 \in \mathcal{S}(s) := \left\{ \Omega \in \mathcal{S} : \sum_{1\le i<j\le p} \mathbf{1}_{(\omega_{ij}\neq 0)} \le s,\ s < n \right\}.$$

假设3 （非退化删失概率。）。 存在有限常数 $K_1 = K_1\,(m, M, c)$ ，使得对于每个 $\Omega_t \in \mathcal{S}$ 以及每个矩阵 $A \in \mathbb{R}^{p\times p}$，涉及期望海森矩阵的以下二次型有界：

$$|\mathrm{trace}(A^{\top}\mathbb{E}_{\Omega_0}[\nabla^2 \ell(\Omega_t;\widetilde{Y})]\ A)| \le K_1 \|A\|_F^2.$$

假设1是频率学派和贝叶斯文献中关于稀疏精度矩阵估计的标准假设， [10, 19, 6, 38, 42]。假设2 中的稀疏性条件在高维情形下很常见 [7, 18, 5]。假设3 对于删失数据而言是一项非平凡的假设，因为似然中包含对数累积分布函数项。它的作用是控制观测数据对数似然的局部曲率。由于存在删失，观测数据对数似然包含的是高斯分布函数项，而非纯粹的二次型。这些项会破坏全局对数凹性，并且原则上可能导致似然在某些区域近乎平坦。该假设

要求删失对数似然的期望海森矩阵在真值的某个谱邻域内，其弗罗贝尼乌斯范数保持一致有界。该条件确保删失造成的信息损失受到控制，并且似然保留足够的局部可识别性，以支持后验集中。具体而言，它排除了退化的删失情形；在这类情形中，数据的大部分内容对精度矩阵几乎不提供信息。

从统计学角度看，假设3 可解释为一种信息非退化条件：它保证删失不会消除与非零精度矩阵元素相对应方向上的曲率。频率学派对删失图模型进行分析时，也隐含地要求类似的曲率条件或受限特征值条件，尽管这些条件很少被明确提出。在此，假设使贝叶斯后验能够以与完全观测情形相同的速率集中，差异仅在于常数取决于删失阈值。

我们首先给出 $p > n$ 设定下温度后验的一致性结果。

给定两个概率测度 P 和 R，它们都由一个共同测度 ν支配；对于任意 $\alpha \in (0,\ 1)$，P 与 R 之间的 α-Rényi散度由以下公式给出：

$$\mathcal{D}_\alpha(P,R) = \frac{1}{\alpha-1}\log\int\left(\frac{\mathrm{d}P}{\mathrm{d}\nu}\right)^{\alpha}\left(\frac{\mathrm{d}R}{\mathrm{d}\nu}\right)^{1-\alpha}\mathrm{d}\nu.$$

此外，Kullback-Leibler散度定义为 $\mathcal{K}(P,\ R) = \int\log\left(\frac{\mathrm{d}P}{\mathrm{d}R}\right)\mathrm{d}P$。该定义成立的前提是 P 相对于 R 绝对连续（记作 $P \ll R$）。如果不满足这一条件，则将散度取为 $+\infty$。下面我们给出删失数据情形下的后验收缩结果，随后给出缺失数据情形下的相应结果。

定理 1 (删失数据的后验收缩)。 考虑第 2 节中删失数据的模型设定。对于任意 $\alpha \in (0,1)$，在假设 1、2、3成立时，我们有

$$\mathbb{E}\left[\int \mathcal{D}_\alpha(P_\Omega, P_{\Omega_0})\pi_{n,\alpha}(\mathrm{d}\Omega)\right] \le \frac{1+\alpha}{1-\alpha}\varepsilon_n.$$

其中 $\varepsilon_n = Kn^{-1}s\log p,\ \ K > 0$ 是仅取决于 m、M 和删失值 c 的某个通用常数。

定理 2 (缺失数据的后验收缩)。 考虑第 3 节中缺失数据的模型设定。对于任意 $\alpha \in (0,1)$，在假设 1、2成立时，我们有

$$\mathbb{E}\left[\int \mathcal{D}_\alpha(P_\Omega, P_{\Omega_0})\pi_{n,\alpha}(\mathrm{d}\Omega)\right] \le \frac{1+\alpha}{1-\alpha}\varepsilon_n.$$

其中 $\varepsilon_n = Kn^{-1}s\log p,\ \ K > 0$ 是某个不依赖于 n、p、s 的通用常数。

结果讨论。 定理 1 和 2 表明，对于缺失数据和删失数据两种情况，温度后验均以 $\varepsilon_n \asymp n^{-1}s\log p$ 的速率收缩，该速率与完整观测下稀疏精度矩阵估计的最优速率 [38]一致。因此，不完整观测并未从根本上改变这一问题的统计难度。

这两种情形的区别在于似然几何。对于缺失数据，边缘化后可通过标准的 Kullback–Leibler 控制得到收缩。对于删失数据，则需要额外的正则性条件，以控制高斯累积分布函数项所诱导的对数似然曲率。

重要的是，这些结果具有自适应性：收缩速率取决于未知的稀疏度 s，且无需对马蹄先验进行调优。这证实了，在删失和缺失情况下，全局-局部收缩在理论上仍然稳定。据我们所知，这是首个针对存在删失时稀疏精度矩阵估计的后验收缩结果；对于删失图形套索等惩罚似然方法，目前尚无可比的保证。

7 讨论与结论

本工作通过引入删失图形马蹄先验（CGHS）模型，将贝叶斯精度矩阵估计扩展到涉及删失和缺失的（高维）高斯数据场景。在图形马蹄先验的自适应收缩特性基础上，我们的模型纳入了潜变量表示，能够自然地处理删失观测以及任意缺失条目。这一推广显著拓宽了马蹄框架的适用范围，使其能够在现有图形建模工具无法直接处理不完整或删失测量的问题中进行有理论依据的推断。

本文的一项重要贡献是开发了一种用于后验计算的高效吉布斯抽样算法。通过交替进行潜在数据插补、更新回归系数和误差方差、调整局部与全局收缩尺度，以及重构精度矩阵，该抽样器为模型提供了一条可行的完整贝叶斯推断途径；否则，这些模型在计算上将难以处理。

这项工作的理论分析或许是其最具特色之处。我们推导了删失数据和缺失数据情形下的后验集中速率，并证明在标准稀疏性条件下，CGHS 能以近乎最优的速率恢复潜在的精度结构。据我们所知，在这些情形下，频率学派或贝叶斯版本的图形套索均无类似结果。因此，我们的理论填补了文献中的一项重要空白，表明即使删失或不完整观测扭曲了似然，全局—局部收缩先验仍然稳定且能够自适应地提供信息。

未来有几个值得探索的方向。一是将 CGHS 扩展到高斯设定之外，例如扩展至基于 copula 的图模型或半参数图形模型，从而拓宽其在非正态数据结构中的适用范围。二是开发可扩展的后验近似方法，例如变分推断或随机梯度马尔可夫链蒙特卡罗，以提高基因组学和神经影像学等超高维领域的计算效率。将非随机缺失（MNAR）机制纳入潜变量框架也是一项挑战，且具有重要的实际意义；尤其是在生物医学研究中，缺失可能取决于未观测值。最后，建立正式的模型选择策略或可信图程序，将有助于对推断出的网络结构进行有原则的不确定性量化。

致谢

本文所呈现的观点、研究结果和意见仅代表作者本人，并不反映挪威公共卫生研究所或 IIM Indore 的官方立场。

利益冲突/竞争性利益

作者声明不存在任何潜在利益冲突。

参考文献

[1] Alquier, P. 和 Ridgway, J.（2020）。温度化后验及其变分近似的集中性。《统计年刊》，48(3):1475–1497。[2] Atchad´e, Y. F.（2019）。大型高斯图模型的准贝叶斯估计。《多元分析杂志》，173:656–671。[3] Augugliaro, L., Abbruzzo, A. 和 Vinciotti, V.（2020）。 ℓ_1-惩罚的删失高斯图模型。《生物统计学》，21(2):e1–e16。

[4] Augugliaro, L., Sottile, G., Wit, E. C. 和 Vinciotti, V.（2023）。cglasso：用于删失值和缺失值条件图形套索推断的 R 软件包。《统计软件杂志》，105:1–58。[5] Banerjee, S., Castillo, I. 和 Ghosal, S.（2021）。高维模型中的贝叶斯推断。arXiv 预印本 arXiv:2101.04491。[6] Banerjee, S. 和 Ghosal, S.（2015）。图模型中的贝叶斯结构学习。《多元分析杂志》，136:147–162。[7] Bellec, P. C., Lecu´e, G. 和 Tsybakov, A. B.（2018）。斜率与套索相遇：改进的预言界及最优性。《统计年刊》，46(6B):3603–3642。[8] Bhattacharya, A., Pati, D. 和 Yang, Y.（2019）。贝叶斯分数后验。《统计年鉴》，47(1):39–66。[9] Cai, T., Liu, W. 和 Luo, X.（2011）。稀疏精度矩阵估计的一种约束 l-1 最小化方法。《美国统计学会杂志》，106(494):594–607。[10] Cai, T. T., Ren, Z. 和 Zhou, H. H.（2016）。结构化高维协方差矩阵与精度矩阵估计：最优速率与自适应估计。电子统计学期刊，10:1–59。[11] Callot, L., Caner, M., Onder, A. O. 和 Ula¸san, E.（2021）。估计大型投资组合的逐节点回归方法。《商业与经济统计杂志》，39(2):520–531。[12] Carvalho, C. M., Polson, N. G. 和 Scott, J. G.（2010）。稀疏信号的马蹄估计量。《生物计量学》，第465–480页。[13] Derveaux, S., Vandesompele, J. 和 Hellemans, J.（2010）。如何利用实时 PCR 成功开展基因表达分析。《方法》，50(4):227–230。[14] Fan, J., Feng, Y. 和 Wu, Y.（2009）。通过自适应套索和 SCAD 惩罚进行网络探索。《应用统计年鉴》，3(2):521–541。[15] Fan, J., Liao, Y. 和 Liu, H.（2016）。大型协方差矩阵与精度矩阵估计概述。《计量经济学杂志》，19(1):C1–C32。[16] Friedman, J., Hastie, T. 和 Tibshirani, R.（2008）。利用图形套索进行稀疏逆协方差估计。《生物统计学》，9(3):432–441。[17] Gan, L., Narisetty, N. N. 和 Liang, F.（2019）。具有不等收缩的图模型贝叶斯正则化。《美国统计学会杂志》，114(527):1218–1231。[18] Gao, C., van der Vaart, A. W. 和 Zhou, H. H.（2020）。贝叶斯结构化线性模型的一般框架。《统计年鉴》，48(5):2848–2878。[19] Jankov´a, J. 和 van de Geer, S.（2017）。高维精度矩阵估计中诚实的置信域与最优性。《检验》，26:143–162。[20] Lauritzen, S. L.（1996）。《图模型》。牛津大学出版社。[21] Li, Y., Craig, B. A. 和 Bhadra, A.（2019）。逆协方差矩阵的图形马蹄估计量。《计算与图形统计学杂志》，28(3):747–757。[22]Lounici, K.（2014）。含缺失观测的高维协方差矩阵估计。《伯努利》，20(3):1029–1058。

[23] Mai, T. T. (2024). 使用马蹄先验估计高维精度矩阵时稀疏贝叶斯模型的集中。*Stat*, 13(4):e70008.
[24] Mai, T. T. (2025a). 随机设计和未知误差方差下稀疏高维线性回归的自适应后验集中速率。韩国统计学会期刊, pages 1–35.[25] Mai, T. T. (2025b). 1位矩阵补全中分数后验的集中性质。机器学习, 114(1):7.[26] Mai, T. T. (2025c). 高维场景下有界响应的处理：一种基于马蹄先验贝叶斯Beta回归的方法。*arXiv*预印本 *arXiv:2505.22211*.[27] Mai, T. T. (2025d). 基于马蹄先验的删失响应高维贝叶斯托比特回归。*arXiv*预印本 *arXiv:2505.08288*.[28] Mai, T. T. (2025e). 稀疏广义贝叶斯逻辑回归的高维分类性质。统计学论文.[29] Mai, T. T. (2025f). 广义降秩回归中分数后验的性质。《多元分析杂志》, page 105481.[30] Mai, T. T. (2025g). 一种用于高维分位数预测的稀疏PAC-贝叶斯方法。统计与计算, 35(4):93.[31] McCall, M. N., McMurray, H. R., Land, H., and Almudevar, A. (2014). qPCR数据中的未检出值。生物信息学, 30(16):2310–2316.[32] Meinshausen, N. and B¨uhlmann, P. (2006). 使用套索进行变量选择和高维图分析。统计年鉴, 34:1436–1462.[33] Pesonen, M., Pesonen, H., and Nevalainen, J. (2015). 左删失数据的协方差矩阵估计。计算统计与数据分析, 92:13–25.[34] Pipelers, P., Clement, L., Vynck, M., Hellemans, J., Vandesompele, J., and Thas, O. (2017). 用于qPCR差异基因表达分析的统一删失正态回归模型。*PLOS ONE*, 12(8):e0182832.[35] Pourahmadi, M. (2011). 协方差估计：广义线性模型与正则化视角。统计科学, 26(3):369–387.[36] Psaila, B., Barkas, N., Iskander, D., Roy, A., Anderson, S., Ashley, N., Caputo, V. S., Lichtenberg, J., Loaiza, S., Bodine, D. M., et al. (2016). 人巨核细胞-红系祖细胞的单细胞分析揭示了不同的巨核细胞和红系分化通路。基因组生物学, 17(1):83.[37] Ryali, S., Chen, T., Supekar, K., and Menon, V. (2012). 使用基于稳定性选择的稀疏偏相关和弹性网络惩罚项估计功能性磁共振成像数据中的功能连接。神经影像, 59(4):3852–3861.[38] Sagar, K., Banerjee, S., Datta, J., and Bhadra, A. (2024). 类马蹄先验—惩罚对偶下的精度矩阵估计。电子统计学期刊, 18(1):1–46.[39] St¨adler, N. and B¨uhlmann, P. (2012). 缺失值：稀疏逆协方差估计及其在稀疏回归中的扩展。统计与计算, 22(1):219–235.

[40] Van de Geer, S., B¨uhlmann, P., Ritov, Y., Dezeure, R., et al. (2014). 高维模型中渐近最优置信域和检验。《统计年刊》, 42(3):1166–1202.[41] Wang, H. (2012). 贝叶斯图形套索模型及高效后验计算。《贝叶斯分析》, 7(4):867–886.[42] Zhang, R., Yao, Y., and Ghosh, M. (2022). 精度矩阵估计中带收缩先验的准贝叶斯模型的收缩性质。《统计规划与推断杂志》, 221:154–171.

补充材料

“删失图形马蹄先验：基于删失和缺失数据的贝叶斯稀疏精度矩阵估计”The Tien Mai、Sayantan Banerjee

A 附加模拟结果

现在，我们提供一些轨迹图和自相关函数图，以及吉布斯采样器的有效样本量，以展示我们算法的行为。

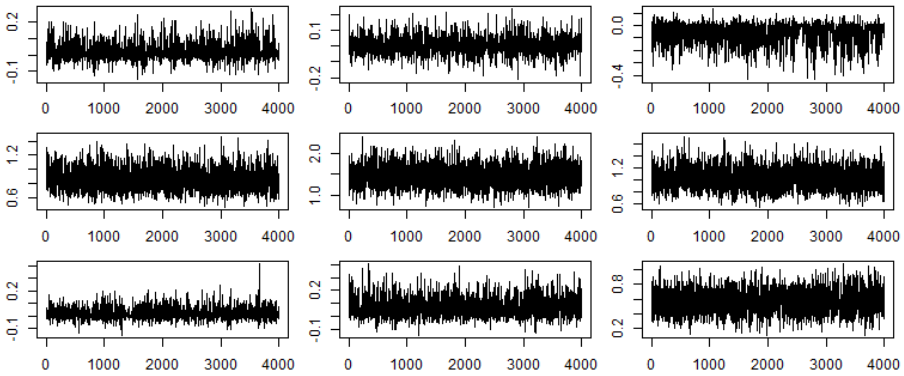

图 2：吉布斯采样器中选定参数项的轨迹图。顶行：随机选取的三个真实值为 0 的参数项。中行：随机选取的三个真实值为 1 的参数项。底行：随机选取的三个真实值为 0.3 的参数项。真实精度矩阵采用设置 I，其中$p = 10,\ \ n = 100$。10% 的数据缺失。

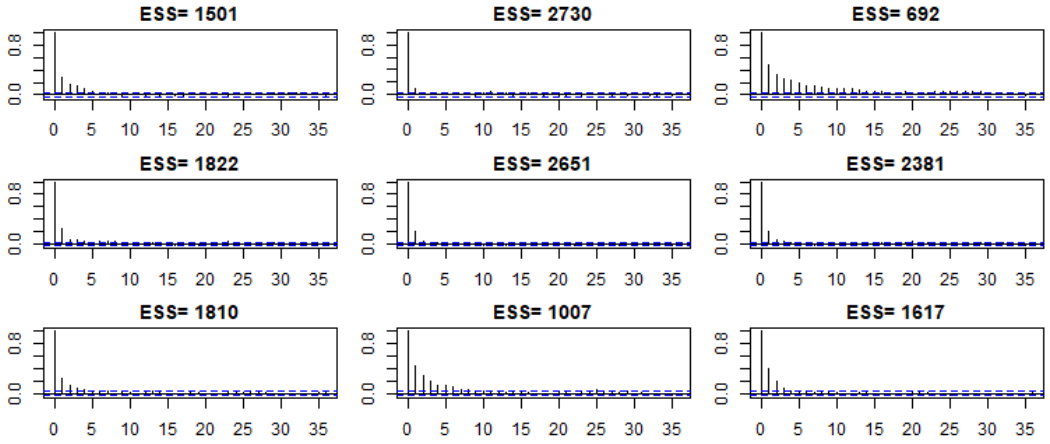

图 3：与图 4中相同的一些随机参数项的吉布斯采样器自相关函数图。顶行（三幅图）：随机选取的 3 个真实值为 0 的参数项。中行（三幅图）：随机选取的 3 个真实值为 1 的参数项。底行（三幅图）：随机选取的 3 个真实值为 0.3 的参数项。图中还给出了 ESS（有效样本量）。10% 的数据缺失。

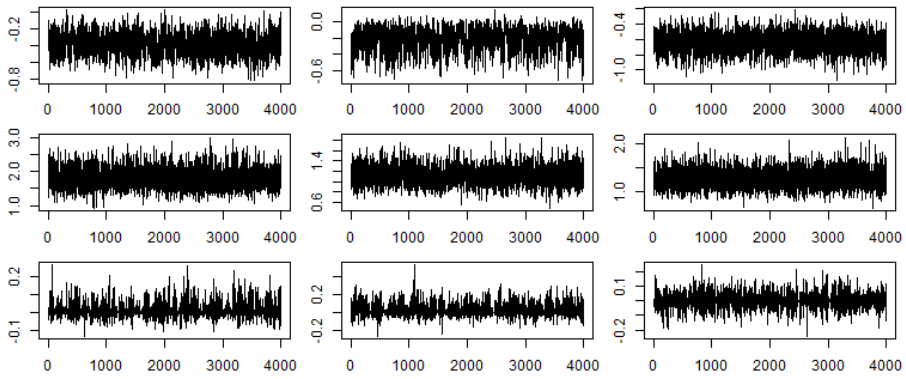

图4：吉布斯采样器对选定参数项的轨迹图。第一行：随机选取的三个真实值为 -0.5的参数项。第二行：随机选取的三个真实值为 1 的参数项。第三行：随机选取的三个真实值为 0 的参数项。真实精度矩阵对应设置 II，其中$p = 10,\ \ n = 100$。10% 的数据缺失。

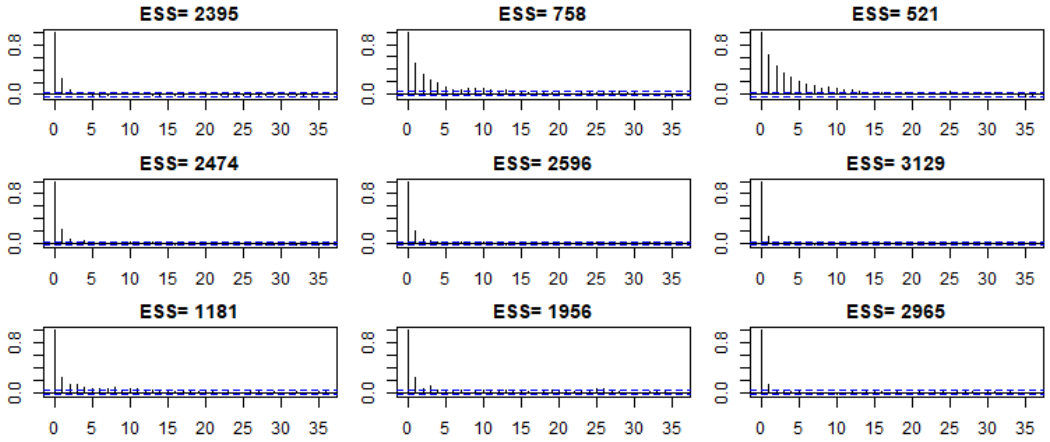

图5：与图4中相同，吉布斯采样器对一些随机选取参数项的自相关函数图。第一行：随机选取的三个真实值为 -0.5的参数项。第二行：随机选取的三个真实值为 1 的参数项。第三行：随机选取的三个真实值为 0 的参数项。真实精度矩阵对应设置 II，其中 $p = 10,\ \ n = 100$。同时给出了 ESS（有效样本量）。10% 的数据缺失。

B 证明

B.1 主要证明

在证明中， K 表示一个正的通用常数。其取值可能逐行变化，但为简化表述，我们不必每次都指定一个新的常数。

定理的证明 *1*。 由引理 2，可得 $\mathcal{K}(P_{\Omega_0}, P_\Omega) \leq K\|\Omega - \Omega_0\|_F^2$. 定义 $\rho_n \propto \mathbf{1}_{\|\Omega-\Omega_0\|_2<\delta}\Pi_{HS}$，其中 $\delta = [s\log(p/s)/n]^{1/2}$。于是，

$$\int \mathcal{K}(P_{\omega_0}, P_\omega)\rho_n(\mathrm{d}\omega) \leq K\delta^2 = K\frac{s\log(p/s)}{n}.$$

由引理1，可得，

$$\frac{1}{n}\mathcal{K}(\rho_n, \pi) \leq K\frac{s\log(p(p-1)/s)}{n} \lesssim K\frac{s\log(p)}{n}.$$

因此，我们可以应用定理 3 得到，对于任意 $\alpha \in (0,1)$,

$$\mathbb{E}\left[\int \mathcal{D}_\alpha(P_\Omega, P_{\Omega_0})\pi_{n,\alpha}(\mathrm{d}\Omega)\right] \leq \frac{1+\alpha}{1-\alpha}\varepsilon_n,$$

其中 $\varepsilon_n = Kn^{-1}s\log p$， $K > 0$ 为某个不依赖于 n， p， s 的通用常数。

□

定理 *2* 的证明。由引理 3，可得 $\mathcal{K}(P_{\Omega_0}, P_\Omega) \leq K\|\Omega - \Omega_0\|_F^2$. 其余证明与上文完全相同。

□

B.2 辅助结果

引理 1 (引理 3，见 [23])。 假设 $\omega_0 \in \mathbb{R}^d$ 满足 $\|\omega_0\|_0 = s^*$ 且 $s^* < n < d$ 且$\|\omega_0\|_\infty \leq C_1$。假设 $\omega \sim \pi_{HS}$。定义 $\delta_n = \{s^*\log(d/s^*)/n\}^{1/2}$。那么，对于某个常数$K > 0$，有

$$\pi_{HS}(\|\omega - \omega_0\|_2 < \delta_n) \geq e^{-Ks^*\log(d/s^*)}.$$

且 $\mathcal{K}(\rho_n, \pi_{HS}) \leq \log \frac{1}{\pi_{HS}(\|\omega-\omega_0\|_2<\delta_n)} \leq Ks^*\log(d/s^*)$。

我们将使用以下一般结果，另见 [24, 28, 26, 27, 30]。

定理 3 （ [1]中的定理 2.6）。假设存在一个序列 $\varepsilon_n > 0$ 和一个分布 ρ_n， 使得

$$\int \mathcal{K}(P_{\omega_0}, P_\omega)\rho_n(\mathrm{d}\omega) \leq \varepsilon_n; \text{ and } \mathcal{K}(\rho_n, \pi) \leq n\varepsilon_n.$$

那么，对于任意 $\alpha \in$ （0、1），

$$\mathbb{E}\left[\int \mathcal{D}_\alpha(P_\omega, P_{\omega_0})\pi_{n,\alpha}(\mathrm{d}\omega)\right] \leq \frac{1+\alpha}{1-\alpha}\varepsilon_n.$$

B.2.1 删失数据的引理

由（2）可知，在删失数据的情况下， $\widetilde{y}$ 的观测数据对数似然为

$$\ell(\Omega; \widetilde{y}) = \frac{1}{2}\log\det(\Omega_{\mathcal{O}\mathcal{O}}) - \frac{1}{2}\widetilde{y}_{\mathcal{O}}^\top\Omega_{\mathcal{O}\mathcal{O}}\widetilde{y}_{\mathcal{O}} + \log\Phi_{|\mathcal{C}|}\big(c_{\mathcal{C}};\, \mu_{\mathcal{C}|\mathcal{O}}(\widetilde{y}_{\mathcal{O}}), \Sigma_{\mathcal{C}|\mathcal{O}}\big),$$

其中 $\mu_{\mathcal{C}|\mathcal{O}}(\widetilde{y}_{\mathcal{O}}) = \Sigma_{\mathcal{C}\mathcal{O}}\Sigma_{\mathcal{O}\mathcal{O}}^{-1}\widetilde{y}_{\mathcal{O}}$ 和 $\Sigma_{\mathcal{C}|\mathcal{O}} = \Sigma_{\mathcal{C}\mathcal{C}} - \Sigma_{\mathcal{C}\mathcal{O}}\Sigma_{\mathcal{O}\mathcal{O}}^{-1}\Sigma_{\mathcal{O}\mathcal{C}}$。

~**引理 2.** 在假设 *1* 和 *3* 下，对于删失高斯模型及其对数似然 $\ell(\Omega;Y)$，存在一个常数 $K>0$，它仅取决于 m，M、以及 c，使得对于所有 $\Omega \in \mathcal{S}$，

$$\mathbb{E}_{\Omega_0}\left[\ell(\Omega_0;\widetilde{Y})-\ell(\Omega;\widetilde{Y})\right] \le K\left\|\Omega-\Omega_0\right\|_F^2.$$

~引理的证明 *2.* 步骤 *1*：期望对数似然的泰勒展开。令 $L(\Omega):=\mathbb{E}_{\Omega_0}[\ell(\Omega;Y)]$表示真实精度矩阵 Ω 下的期望对数似然$_{0}$。由于对数似然函数~$\ell(\Omega;Y)$ 关于 Ω 在谱带 $\mathcal{S}$ 上二阶连续可微，因此我们对 $L(\Omega)$ 在 Ω_0 处进行二阶泰勒展开。

利用拉格朗日余项形式，存在某个 $\Omega^*=\Omega_0+t(\Omega-\Omega_0)$，其中 $t\in\in(0,1)$，使得：

$$L(\Omega)-L(\Omega_0)=\langle\Omega-\Omega_0,\nabla L(\Omega_0)\rangle+\frac{1}{2}\langle\Omega-\Omega_0,\nabla^2L(\Omega^*)[\Omega-\Omega_0]\rangle,$$

其中 $\langle A,B\rangle=\mathrm{tr}(A^\top B)$ 表示弗罗贝尼乌斯内积。

步骤 *2*：一阶项为零。期望对数似然的梯度 $\nabla L(\Omega)$ 由~以下公式给出：$\mathbb{E}_{\Omega_0}[\nabla\ell(\Omega;Y)]$。最大似然估计的一个基本性质是，在真实参数处的期望得分函数为零。因此：

$$\nabla L(\Omega_0)=\mathbb{E}_{\Omega_0}[\nabla\ell(\Omega_0;\widetilde{Y})]=0.$$

将其代入展开式可得：

$$L(\Omega)-L(\Omega_0)=\frac{1}{2}\langle\Omega-\Omega_0,\nabla^2L(\Omega^*)[\Omega-\Omega_0]\rangle.$$

重新整理，使其与关注的量相匹配：

$$\mathbb{E}_{\Omega_0}\left[\ell(\Omega_0;\widetilde{Y})-\ell(\Omega;\widetilde{Y})\right]=L(\Omega_0)-L(\Omega)=-\frac{1}{2}\langle\Omega-\Omega_0,\nabla^2L(\Omega^*)[\Omega-\Omega_0]\rangle.$$

步骤 *3*：海森矩阵的一致有界性。由于 $\mathcal{S}$ 是凸的，中间矩阵 Ω^* 位于 $\mathcal{S}$内。根据假设 3，期望对数似然的海森矩阵在 $\mathcal{S}$上是一致有界的。具体而言，对于任意矩阵 Δ，其二次型满足：

$$\left|\langle\Delta,\nabla^2L(\Omega^*)[\Delta]\rangle\right|\le K_1\|\Delta\|_F^2,$$

其中常数 $K_1>0$ 取决于 m，M、以及 c。

由于期望对数似然是凹的，项 $\langle\Omega-\Omega_0,\ \nabla^2L(\Omega^*)[\Omega-\Omega_0]\rangle$ 非正。因此：

$$-\frac{1}{2}\langle\Omega-\Omega_0,\nabla^2L(\Omega^*)[\Omega-\Omega_0]\rangle\le\frac{1}{2}\left|\langle\Omega-\Omega_0,\nabla^2L(\Omega^*)[\Omega-\Omega_0]\rangle\right|\le\frac{1}{2}K_1\|\Omega-\Omega_0\|_F^2.$$

令 $K=\frac{1}{2}K_1$,~即可得到所需结果：

$$\mathbb{E}_{\Omega_0}\left[\ell(\Omega_0;\widetilde{Y})-\ell(\Omega;\widetilde{Y})\right]\le K\|\Omega-\Omega_0\|_F^2.$$

常数 K >0 取决于 m，M，以及删失阈值 c。证明完成 .

□

B.2.2 缺失数据的引理

我们假设（可能是随机的）观测模式 $(\mathcal{O}_i)_{i=1}^n$ 是任意的，但与未观测值相互独立（随机缺失）。由（6）可知，对于缺失数据情形，对数似然为

$$\ell_i(\Omega; \widetilde{Y}_i) \;=\; -\frac{1}{2}\log\det(\Sigma_{\mathcal{O}_i}) - \frac{1}{2}\widetilde{Y}_i^\top \Sigma_{\mathcal{O}_i}^{-1}\widetilde{Y}_i.$$

引理 3. 在假设 *1*下，存在一个有限常数 $K > 0$，它仅取决于 m 和 M，使得：

$$E_{\Omega_0}\big[\,\ell(\Omega_0; \widetilde{Y}) - \ell(\Omega; \widetilde{Y})\,\big] \le K\,\|\Omega - \Omega_0\|_F^2.$$

引理 *3* 的证明。我们逐步证明该结论，并跟踪各常数 .

步骤 *1*。固定一行 i 及其观测到的索引集 $O := \mathcal{O}_i$。令 $A := \Sigma_{0,\,O}$ 和 $B := \Sigma_O$ 表示 Σ_0 和 Σ 的~对应主子矩阵。给定观测模式，在真实模型 (d:= $|O|$) 下，$Y_i \sim \mathcal{N}_d(0,\ A)$。第 i 行的期望贡献等于这两个中心高斯分布之间的 KL散度：

$$\delta_i \;:=\; \mathbb{E}_{\Omega_0}\big[\ell_i(\Omega_0; \widetilde{Y}_i) - \ell_i(\Omega; \widetilde{Y}_i)\big] = \mathcal{K}\big(\mathcal{N}(0, A)\,\big\|\,\mathcal{N}(0, B)\big).$$

标准闭式表达式为

$$\delta_i \;=\; \frac{1}{2}\Big\{\,\mathrm{tr}(B^{-1}A) - d + \log\frac{\det B}{\det A}\Big\}.$$

步骤 *2*。令 $E := A^{-1/2}BA^{-1/2}$（因此 $E \succ 0$）。那么

$$\delta_i = \frac{1}{2}\Big\{\,\mathrm{tr}(E^{-1}) - d + \log\det E\Big\} =: \frac{1}{2}\phi(E).$$

注 $\phi(I) = 0$ 且 $\nabla\phi(I) = 0$（一阶项相消）。根据假设 3中的谱界，我们有

$$\lambda_{\min}(A) \ge \lambda_{\min}(\Sigma_0) \ge 1/M, \qquad \lambda_{\max}(A) \le \lambda_{\max}(\Sigma_0) \le 1/m,$$

对于 B，同理可得。因此，E 的特征值位于紧区间 $[m/M,\, M/m]$内。令 $\mathcal{S}$ 表示这类 E 的集合（维度至多为 p）。函数 ϕ在 $\mathcal{S}$上二阶连续可微，因此其海森矩阵在该处有界；记为

$$H_{\max} := \sup_{E\in\mathcal{S}} \|\mathrm{Hess}\,\phi(E)\|_{\mathrm{op}} < \infty.$$

对 I应用带积分余项的二阶泰勒公式（或余项的中值形式），可得

$$\phi(E) \;=\; \tfrac{1}{2}\langle E - I,\ \mathrm{Hess}\,\phi(\xi)[E - I]\rangle \le \tfrac{1}{2}H_{\max}\|E - I\|_F^2$$

对于介于 I 和 E之间的线段上的某个 ξ。因此

$$\delta_i \;=\; \tfrac{1}{2}\phi(E) \;\le\; \tfrac{H_{\max}}{4}\,\|E - I\|_F^2.$$

步骤 *3*。注意到 $E - I = A^{-1/2}(B - A)A^{-1/2}$。因此

$$\|E - I\|_F \le \|A^{-1/2}\|_2^2\,\|B - A\|_F = \frac{1}{\lambda_{\min}(A)}\|B - A\|_F.$$

因为 $\lambda_{\min}(A) \ge 1/M$ 我们得到 $\|E - I\|_F \le M\,\|B - A\|_F$。因此

$$\delta_i \le \frac{H_{\max}}{4}\,M^2\,\|B - A\|_F^2.$$

步骤 *4.* 我们使用精确恒等式

$$\Sigma - \Sigma_0 \;=\; \Omega^{-1} - \Omega_0^{-1} \;=\; \Omega^{-1}(\Omega_0 - \Omega)\Omega_0^{-1}.$$

因此，根据算子范数界 $\|\Omega^{-1}\|_2,\ \|\Omega_0^{-1}\|_2 \le 1/m$,

$$\|\Delta\|_F \le \|\Omega^{-1}\|_2 \, \|\Omega_0^{-1}\|_2 \, \|\Omega - \Omega_0\|_F \le \frac{1}{m^2} \, \|\Omega - \Omega_0\|_F.$$

结合前一步可得

$$\Delta(\Omega) \le \frac{H_{\max}}{4} \, \frac{M^2}{m^4} \, \|\Omega - \Omega_0\|_F^2.$$

这就证明了所述的上界。 □