Gauss 关联结构图模型的充分降维方法

何勇¹, 季加东^{1*}, 张新生²

1. 山东财经大学统计学院, 济南 250014;

2. 复旦大学管理学院, 上海 200433

E-mail: heyong@sdufe.edu.cn, jiadong@sdufe.edu.cn, xszhang@fudan.edu.cn

收稿日期: 2016-12-29; 接受日期: 2018-01-20; 网络出版日期: 2018-07-26; * 通信作者

国家自然科学基金 (批准号: 11571080) 和山东省自然科学基金 (批准号: ZR2018BH033) 资助项目

摘要 本文提出了恢复 Gauss 关联结构 (copula) 图模型的充分降维方法, 该方法在超高维情形下具有很高的计算效率. 本质上, 充分降维是通过对利用非参数方法估计的相关系数矩阵进行截断来实现的. 本文给出了所提方法的理论性质, 保证其所估计的边集合以概率趋于 1 覆盖所有真实存在边的集合. 数值模拟研究发现, 本文所提方法与现存方法相比有相近的估计表现, 而计算效率却更高. 最后分析了一组基因数据来展示本文所提方法的实际应用表现.

关键词 图模型 Gauss 关联结构 充分降维 非参数方法

MSC (2010) 主题分类 62H99, 62G35

1 引言

图模型一直以来是研究随机变量间相关关系的重要工具, 并且在信号处理和网络建模等领域引起了广泛的关注. Gauss 图模型是最经典也是研究最广泛的一类图模型. 具体地, 假设 $\mathbf{X}$ 为 p 维正态分布随机向量, 即 $\mathbf{X} = (X_1, \dots, X_p)^\top \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, 则该分布的条件独立结构可以用图模型 $\mathcal{G} = (V, E)$ 来刻画, 其中 $V = \{1, \dots, p\}$ 是顶点集合, E 为 $V \times V$ 中边的集合. 顶点对 (i, j) 间存在一条边当且仅当 X_i 在给定所有剩余变量 $X_{V \setminus \{i, j\}}$ 的条件下与 X_j 相关. 顶点对 (i, j) 间不存在边当且仅当 X_i 在给定所有剩余变量 $X_{V \setminus \{i, j\}}$ 的条件下与 X_j 独立. 事实上, 对于 Gauss 图模型, 估计其结构等价于恢复精确矩阵的支撑集, 其中精确矩阵是协方差矩阵的逆矩阵, 即 $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1} = (\Omega_{i,j})$, $X_i \perp X_j \mid \mathbf{X}_{V \setminus \{i, j\}}$ 当且仅当 $\Omega_{i,j} = 0$. 在维数 p 比样本量 n 大很多的高维问题背景下, 关于 Gauss 图模型 (Gaussian graphical model, GGM) 的统计推断问题变得更加有挑战性, 过去几十年也出现了大量研究高维 Gauss 图模型估计的工作. 文献 [1] 提出了相邻选择 (neighborhood selection) 的方法来估计 Gauss 图模型, 该方法的基本思想是用 LASSO (least absolute shrinkage and selection operator) 来求解一系列的稀疏回归方程. 文献 [2] 提出了惩罚似然的方法并基于内点优化算法直接估计精确矩阵. 文献 [3, 4] 改进了内点优化算法

英文引用格式: He Y, Ji J D, Zhang X S. Sufficient dimension reduction for Gaussian copula graphical models (in Chinese). Sci Sin Math, 2018, 48: 1819–1830, doi: 10.1360/SCM-2016-0805

来求解文献 [2] 提出的惩罚优化问题. 文献 [5] 提出了图 LASSO (graphical LASSO, GLASSO) 的方法来估计稀疏的精确矩阵, 本质上是基于成块坐标下降算法求解带有 ℓ_1 范数惩罚的似然方法. 文献 [6] 提出了图 Dantzig selector 的方法; 文献 [7] 提出了 CLIME (constrained L_1 -minimization for inverse matrix estimation) 方法, 这两种惩罚优化方法都可以通过线性规划来求解.

在实际应用中, 联合多元正态的假设有一定的局限性. 为了放松这一假设, 文献 [8] 提出了 Gauss 关联结构 (copula) 图模型. 该文献中假定存在一组单变量单调函数使得每一维变量作变换后联合服从多元正态分布. 具体地, 对于随机向量 $\mathbf{X} = (X_1, \dots, X_p)^\top$, 如果 $f(\mathbf{X}) = (f_1(X_1), \dots, f_p(X_p))^\top$ 服从多元正态分布, 则称 $\mathbf{X}$ 服从 Gauss 关联结构分布, 变量 $\mathbf{X}$ 各个分量间的条件相关结构图称为 Gauss 关联结构图模型. Gauss 关联结构图模型包含了更大范围的一类图, 可以认为是 Gauss 图模型的更有利的代替. 文献 [9] 提出了一种半参数的方法来估计 Gauss 关联结构图模型. 在现有文献的基础上 (如文献 [6, 7, 10]), 文献 [9] 证明了所提方法在精确矩阵估计和图模型恢复方面, 能够达到与参数方法相同的最优收敛速度.

尽管现有文献已经提出了非常多的估计高维无向图模型的方法, 但是当变量的维数成千上万甚至更大时, 现有方法的计算负担是非常重的. 在这种超高维的情形下, 现有的算法甚至将不再适用. 为此, 本文提出超高维图模型的充分降维方法, 减轻计算负担, 同时保证方法有很好的估计表现.

文献 [11] 在线性回归模型下通过筛选协变量与响应变量的边际相关系数实现了充分降维目的, 并且能够保证以概率趋于 1 将所有重要变量留在模型中. 这种思想在之后被广泛应用, 有兴趣的读者可以参见文献 [12–14].

在变量维数可以是样本量指数阶的超高维情形下, 本文提出 Gauss 关联结构图模型的估计方法, 该方法便于实施, 计算效率更高. 本文证明了在一定的假设之下, 该方法所估计的边集合以概率趋于 1 覆盖所有真实存在边的集合. 数值模拟研究发现本文所提方法与现存方法相比, 有相近的估计表现, 而计算速度提高可达到其他方法的一倍.

本文的结构如下: 第 2 节给出基本的记号以及背景知识介绍. Gauss 关联结构图模型的充分降维方法在第 3 节中给出. 第 4 节给出本文所提方法的理论性质. 同时, 本文在第 5 节中通过数值模拟研究了提出方法在估计精度和计算速度方面的表现. 第 6 节用一组基因数据集验证了提出方法的实际应用价值. 第 7 节给出一个简单的小结以及今后可以拓展的方向. 第 4 节中定理的证明见附录.

2 背景介绍

2.1 记号

对于向量 $\boldsymbol{\mu} = (\mu_1, \dots, \mu_d) \in \mathbb{R}^d$, 记 $\|\boldsymbol{\mu}\|_0 = \sum_{i=1}^d I(\mu_i \neq 0)$, $\|\boldsymbol{\mu}\|_1 = \sum_{i=1}^d |\mu_i|$, $\|\boldsymbol{\mu}\|_2 = \sqrt{\sum_{i=1}^d \mu_i^2}$, $\|\boldsymbol{\mu}\|_\infty = \max_i |\mu_i|$, 记 $\boldsymbol{\mu}_{\setminus i}$ 为将 $\boldsymbol{\mu}$ 的第 i 个元素去掉所得的 $d-1$ 维向量. 对于矩阵 $\mathbf{A} = [a_{ij}] \in \mathbb{R}^{d \times d}$, 记 $\|\mathbf{A}\|_{\max} = \max_{i,j} |a_{ij}|$ 和 $\|\mathbf{A}\|_1 = \sum_{i=1}^d \sum_{j=1}^d |a_{ij}|$. 记 $\lambda_{\min}(\mathbf{A})$ 和 $\lambda_{\max}(\mathbf{A})$ 分别为矩阵 $\mathbf{A}$ 的最小特征值和最大特征值, $\text{Tr}(\mathbf{A})$ 为矩阵 $\mathbf{A}$ 的迹. 对于集合 $\mathcal{H}$, 记 $|\mathcal{H}|$ 为 $\mathcal{H}$ 中元素的个数. 对于矩阵 $\mathbf{B} = [b_{ij}] \in \mathbb{R}^{q \times d}$, 记 $\mathbf{B}_{i,\setminus j}$ 为 $\mathbf{B}$ 的第 i 行去掉第 j 个元素所得的向量, 记 $\mathbf{B}_{\setminus i,j}$ 的第 j 列去掉第 i 个元素所得的向量, 记 $\mathbf{B}_{\setminus i,\setminus j}$ 为 $\mathbf{B}$ 去掉第 i 行和 j 列得到的 $(q-1) \times (d-1)$ 维矩阵. 对于两个实数序列 $\{a_n\}$ 和 $\{b_n\}$, 如果存在常数 C , 使得 $|a_n| \leq C|b_n|$ 对任意的 n 成立, 则记为 $a_n = O(b_n)$; 如果 $\lim_{n \rightarrow \infty} a_n/b_n = 0$, 则记为 $a_n = o(b_n)$; 如果存在常数 c 和 C 使得 $c \leq a_n/b_n \leq C$ 对任意的 n 成立, 则记为 $a_n \asymp b_n$. 对于随机变量 X , $E(X)$ 为其期望, $\text{Var}(X)$ 为其方差.

2.2 Gauss 关联结构图模型及 SKEPTIC 估计方法

Gauss 关联结构分布在有些文献中也称之为非参数正态 (Nonparanormal, 简称为 NPN) 分布, 实际上是 Gauss 分布的半参数推广. 令 $f = \{f_1, \dots, f_p\}$ 为单调单变量函数集合, Σ 为正定相关系数矩阵. 一个 p 维随机向量 $\mathbf{X}$ 服从 Gauss 关联结构分布当且仅当 $f(\mathbf{X}) = (f_1(X_1), \dots, f_p(X_p))^T \sim N(\boldsymbol{\mu}, \Sigma)$, 记作 $\mathbf{X} \sim \text{NPN}(\boldsymbol{\mu}, \Sigma, f)$. 令 $\Omega = (\Omega_{jk})$ 为 Σ 的逆矩阵, 文献 [8] 证明了估计 Gauss 关联结构变量的条件相依图结构等价于恢复 Ω 的支撑集, 即 $\Omega_{jk} = 0$ 当且仅当 $X_j \perp X_k \mid X_{\setminus\{j,k\}}$.

文献 [9] 提出了一种估计 Gauss 关联结构图模型的方法, 其方法的基本思想是利用 Kendall's tau 及 Spearman's rho 来估计相关系数矩阵 Σ , 绕开了估计边界的单调变换函数, 文中称该方法为 SKEPTIC (Spearman/Kendall estimates preempt transformations to infer correlation) 方法. 之后再利用估计的相关系数矩阵代入到现有的方法 (如 GLASSO 方法) 中来估计相关系数矩阵的逆矩阵, 从而恢复图结构. 具体地, 若 $\mathbf{X} \sim \text{NPN}(\boldsymbol{\mu}, \Sigma, f)$, 则有如下函数关系式: $\Sigma_{ij} = 2 \sin\{(\pi/6)\rho_{jk}\} = \sin\{(\pi/2)\tau_{jk}\}$, 其中 ρ_{jk} 和 τ_{jk} 分别为 x_j 和 x_k 的 Spearman's rho 以及 Kendall's tau 相关系数. 因此可以通过下式估计未知的相关系数矩阵:

$$\widehat{S}_{jk}^{\rho} = \begin{cases} 2 \sin\left(\frac{\pi}{6}\widehat{\rho}_{jk}\right), & j \neq k, \\ 1, & j = k, \end{cases} \quad \widehat{S}_{jk}^{\tau} = \begin{cases} \sin\left(\frac{\pi}{2}\widehat{\tau}_{jk}\right), & j \neq k, \\ 1, & j = k. \end{cases}$$

文献 [9] 证明了基于 Spearman's rho 和 Kendall's tau 统计量最终得到的图模型的估计具有相似的理论性质. 下文介绍时省略上标 ρ 和 τ , 将基于秩的相关系数矩阵的估计简记为 $\widehat{S}$.

2.2.1 基于 GLASSO 的 SKEPTIC 方法

对于经典的 Gauss 图模型, 即 $\mathbf{X} = (X_1, \dots, X_p) \sim N(\boldsymbol{\mu}, \Sigma)$, GLASSO 方法实际上是求解以下优化问题:

$$\widehat{\Omega} = \arg \min_{\Omega \succ 0} \{\text{Tr}(\widehat{\Sigma}_n \Omega) - \log |\Omega| + \lambda \|\Omega\|_1\},$$

其中 $\widehat{\Sigma}_n$ 为样本协方差. 在 Gauss 关联结构图模型下, 基于 GLASSO 的 SKEPTIC 方法实际上是用基于秩的相关系数矩阵的估计 $\widehat{S}$ 代替上式中的 $\widehat{\Sigma}_n$:

$$\widehat{\Omega}^{\text{GLASSO}} = \arg \min_{\Omega \succ 0} \{\text{Tr}(\widehat{S} \Omega) - \log |\Omega| + \lambda \|\Omega\|_1\}. \quad (2.1)$$

令 $\mathbf{U}$ 为 Σ 的初始估计. 文献 [5] 提出通过成块坐标下降方式优化 $\mathbf{U}$ 的行和对应的列来求解 (2.1) 中的优化问题. 具体地, 将矩阵 $\mathbf{U}$ 和 $\widehat{S}$ 分解为

$$\mathbf{U} = \begin{pmatrix} \mathbf{U}_{\setminus i, \setminus i} & \mathbf{U}_{\setminus i, i} \\ \mathbf{U}_{i, \setminus i}^T & u_{ii} \end{pmatrix}, \quad \widehat{S} = \begin{pmatrix} \widehat{S}_{\setminus i, \setminus i} & \widehat{S}_{\setminus i, i} \\ \widehat{S}_{i, \setminus i}^T & s_{ii} \end{pmatrix}. \quad (2.2)$$

每次通过下面的优化问题来更新 $\mathbf{U}_{i, \setminus i}$:

$$\mathbf{U}_{i, \setminus i} = \arg \min_{\mathbf{y}} \{\mathbf{y}^T \mathbf{U}_{\setminus i, \setminus i}^{-1} \mathbf{y} : \|\mathbf{y} - \widehat{S}_{i, \setminus i}\|_{\infty} \leq \lambda\}. \quad (2.3)$$

置换行和列从而保证下一步要优化的列 $\mathbf{U}_{i, \setminus i}$ 总是在最后一列, 从而逐步更新矩阵 $\mathbf{U}$ 的元素直至收敛. 在这个过程中, 只要初始估计 $\mathbf{U}$ 是正定矩阵, 则每一步更新得到的矩阵也始终是正定矩阵. 文献 [3]

指出了优化问题 (2.3) 与下列对偶问题等价:

$$\min_{\beta} \left\{ \frac{1}{2} \|U_{\setminus i, \setminus i}^{1/2} \beta - c\|_2^2 + \lambda \|\beta\|_1 \right\}, \quad (2.4)$$

其中 $c = U_{\setminus i, \setminus i}^{-1/2} \widehat{S}_{i, \setminus i}$. 如果 β 是优化问题 (2.4) 的解, 则 $U_{i, \setminus i} = U_{\setminus i, \setminus i} \beta$ 为优化问题 (2.3) 的解.

2.2.2 基于相邻选择的 SKEPTIC 方法

文献 [1] 提出了相邻选择的方法来估计高维 Gauss 图模型, 为了介绍充分降维方法, 本节将简单回顾下相邻选择方法. 考虑 Gauss 图模型, $\mathbf{X} = (X_1, \dots, X_p) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, 对于每个节点 $i \in V$, 考虑 X_i 在给定其他变量时的最优预测. 令 $\beta^i \in \mathbb{R}^p$ 为最优预测的回归系数向量:

$$\beta^i = \arg \min_{\beta: \beta_i = 0} E \left(X_i - \sum_{j \in V} \beta_j X_j \right)^2. \quad (2.5)$$

对于节点 $j \in V$ 和精确矩阵 $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1} = (\Omega_{ij})$, 有 $\beta_j^i = -\Omega_{ij}/\Omega_{ii}$. 因此, β^i 的非零元素的指标集合与精确矩阵对应行 (第 i 行) 的非零元素指标集合相同, 该指标集合也正是与第 i 个节点相连接的节点集合 (第 i 个节点的邻居). 由此可知, 对变量 X_i 的最优预测实际上是第 i 个节点邻居对应的变量的线性组合, 这是相邻选择方法的基础. 记第 i 个节点的邻居的集合为 $\text{ne}_i = \{j \in V : \beta_j^i \neq 0\}$. 记 $\mathbf{X}_{n \times p}$ 为观测矩阵, 第 i 行对应 $\mathbf{X}$ 第 i 次观测, 第 j 列对应 X_j 的 n 次独立观测. 回归系数向量 β^i 的 LASSO 估计由下式给出:

$$\widehat{\beta}^i = \arg \min_{\beta: \beta_i = 0} \left(\frac{1}{n} \|\mathbf{X}_i - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_1 \right),$$

其中 $\mathbf{X}_i$ 为矩阵 $\mathbf{X}$ 的第 i 列. 因此, 第 i 个节点的邻居的集合的估计为 $\widehat{\text{ne}}_i = \{j \in V : \widehat{\beta}_j^i \neq 0\}$. 文献 [5] 指出样本协方差矩阵 $\widehat{\boldsymbol{\Sigma}}_n$ 为得到 LASSO 估计的充分统计量, 从而也是估计 Gauss 图模型的充分统计量 ($\beta_j^i = -\Omega_{ij}/\Omega_{ii}$). 对于 Gauss 关联结构图模型, 类似可知, 基于秩的相关系数矩阵的估计 $\widehat{\mathbf{S}}$ 也是估计其图结构的充分统计量. 具体地, 在计算时, 我们只需要在 (2.4) 中用 $\widehat{\mathbf{S}}_{\setminus i, \setminus i}$ 代替 $U_{\setminus i, \setminus i}$.

3 充分降维方法

本节给出超高维 Gauss 关联结构图模型的充分降维方法. 我们观察到 (2.4) 本质上是求解 LASSO 优化问题, 这启发我们可以采取文献 [11] 中的充分降维思想来提高优化问题的求解速度.

3.1 基于 GLASSO 的充分降维方法

令 $\mathcal{M}_*^i = \{j \neq i : \beta_j^i \neq 0\}$ 为 β^i 的稀疏非零元素的指标集合, 集合中元素的个数记为 $s = |\mathcal{M}_*^i|$, 即 $\{X_j : j \in \mathcal{M}_*^i\}$ 中的 s 个变量为节点 i 的邻居. $\widehat{\mathbf{S}}$ 为第 2.2 小节定义的基于秩的相关系数矩阵的估计. 对任意给定的 $\gamma \in (0, 1)$, 我们对向量 $\widehat{\mathbf{S}}_{i, \setminus i} = (\widehat{S}_{i1}, \dots, \widehat{S}_{i, i-1}, \widehat{S}_{i, i+1}, \dots, \widehat{S}_{ip})$ 按照绝对值的大小排序, 进而定义子模型

$$\mathcal{M}_\gamma^i = \{1 \leq j \neq i \leq p : |\widehat{S}_{i,j}| \text{ 为排在前 } [\gamma n] \text{ 个最大的元素}\}, \quad (3.1)$$

其中 $[\gamma n]$ 代表 γn 的整数部分. 这就将维数从 $p-1$ 降到了一个合适的大小 $[\gamma n]$, 这比样本量 n 还小 ($\gamma < 1$). 定义 $\mathcal{A} = \{\beta \in \mathbb{R}^{p-1} : \beta_j = 0, j \notin \mathcal{M}_\gamma^i\}$, 则现在只要求解以下优化问题:

$$\hat{\beta}^i = \arg \min_{\beta \in \mathcal{A}} \left\{ \frac{1}{2} \|U_{\setminus i, \setminus i}^{1/2} \beta - c\|_2^2 + \lambda \|\beta\|_1 \right\}, \quad (3.2)$$

其中 $c = U_{\setminus i, \setminus i}^{-1/2} \hat{S}_{i, \setminus i}$. 称该充分降维方法为 SSG (SKEPTIC-SIS-GLASSO) 方法, 具体的步骤在算法 1 中给出.

算法 1 SSG 算法

Input: $X_1, \dots, X_n$

```

1: procedure
2:   令  $U = \hat{S} + \lambda I$ . 矩阵  $U$  的对角线元素在接下来的步骤中保持不变.
3:   for  $i \in 1, \dots, p$ , do
4:     根据 (3.1) 挑选指标集合  $\mathcal{M}_\gamma^i$ .
5:     求解 (3.2) 中的 LASSO 估计.
6:   end for
7:   根据  $U_{i, \setminus i} = U_{\setminus i, \setminus i} \hat{\beta}^i$  按顺序更新矩阵  $U$  的行和列.
8:   如果矩阵  $U$  没有收敛, 则返回步骤 3, 否则执行下一步.
9:   算法结束.
10: end procedure

```

至此, 我们得到了 Σ 的估计 U . 为了恢复图模型结构, 我们需要得到 Σ 的逆矩阵的估计, 为此可以提出一种简单得到 $\hat{\Omega} = U^{-1}$ 的方法. 事实上, 我们有

$$\Omega_{i, \setminus i} = -U_{\setminus i, \setminus i}^{-1} U_{i, \setminus i} \Omega_{i, i}, \quad (3.3)$$

$$\Omega_{i, i} = (u_{i, i} - U_{i, \setminus i}^\top U_{\setminus i, \setminus i}^{-1} U_{i, \setminus i})^{-1}. \quad (3.4)$$

由于 $\hat{\beta}^i = U_{\setminus i, \setminus i}^{-1} U_{i, \setminus i}$, 则有 $\hat{\Omega}_{i, i} = (u_{i, i} - U_{i, \setminus i}^\top \hat{\beta}^i)^{-1}$ 和 $\hat{\Omega}_{i, \setminus i} = -\hat{\beta}^i \hat{\Omega}_{i, i}$. 因此可以将所有的回归系数向量 $\hat{\beta}^i$ 存储在 $p \times p$ 维矩阵 $\hat{B}$ 中, 待算法收敛后代入上式计算 $\hat{\Omega}$. 对节点 $i \in V$, 定义 $\widehat{\text{ne}}_i = \{j \neq i : \hat{\Omega}_{i, j} \neq 0\}$. 图模型边的集合自然可以由下式估计:

$$\hat{E}_\lambda^\wedge = \{(i, j) : i \in \widehat{\text{ne}}_j \wedge j \in \widehat{\text{ne}}_i\} \quad \text{或者} \quad \hat{E}_\lambda^\vee = \{(i, j) : i \in \widehat{\text{ne}}_j \vee j \in \widehat{\text{ne}}_i\}. \quad (3.5)$$

3.2 基于相邻选择的充分降维方法

事实上, 如果 $U_{\setminus i, \setminus i} = \hat{S}_{\setminus i, \setminus i}$, 则 (2.4) 的解就是相邻选择方法中回归系数向量的 LASSO 估计. 令 $\mathcal{M}_*^i = \{j \neq i : \beta_j^i \neq 0\}$ 为 β^i 的稀疏非零元素的指标集合, 集合中元素的个数记为 $s = |\mathcal{M}_*^i|$, 即 $\{X_j : j \in \mathcal{M}_*^i\}$ 中的 s 个变量为节点 i 的邻居. $\hat{S}$ 为第 2.2 小节定义的基于秩的相关系数矩阵的估计. 对任意给定的 $\gamma \in (0, 1)$, 对向量 $\hat{S}_{i, \setminus i} = (\hat{S}_{i1}, \dots, \hat{S}_{i, i-1}, \hat{S}_{i, i+1}, \dots, \hat{S}_{ip})$ 按照绝对值的大小排序, 进而定义子模型

$$\mathcal{M}_\gamma^i = \{1 \leq j \neq i \leq p : |\hat{S}_{i, j}| \text{ 为排在前 } [\gamma n] \text{ 个最大的元素}\}, \quad (3.6)$$

其中 $[\gamma n]$ 代表 γn 的整数部分. 这就将维数从 $p-1$ 降到了一个合适的大小 $[\gamma n]$, 这比样本量 n 还小 ($\gamma < 1$). 定义

$$\mathcal{A} = \{\beta \in \mathbb{R}^{p-1} : \beta_j = 0, j \notin \mathcal{M}_\gamma^i\},$$

则现在只需要求解以下优化问题:

$$\hat{\beta}^i = \arg \min_{\beta \in \mathcal{A}} \left\{ \frac{1}{2} \|\hat{\mathbf{S}}_{i, \setminus i}^{1/2} \beta - \mathbf{c}\|_2^2 + \lambda \|\beta\|_1 \right\}, \quad (3.7)$$

其中 $\mathbf{c} = \hat{\mathbf{S}}_{i, \setminus i}^{-1/2} \hat{\mathbf{S}}_{i, \setminus i}$. 称该充分降维方法为 SSM (SKEPTIC-SIS-MB) 方法, 具体的步骤在算法 2 中给出. 图模型边的集合类似地可由 (3.5) 给出.

算法 2 SSM 算法

Input: $\mathbf{X}_1, \dots, \mathbf{X}_n$

- 1: procedure
  - 2: 令 $\mathbf{U} = \hat{\mathbf{S}}$. 矩阵 $\mathbf{U}$ 的对角线元素在接下来的步骤中保持不变.
  - 3: for $i \in 1, \dots, p$, do
  - 4: 根据 (3.6) 挑选指标集合 $\mathcal{M}_\gamma^i$.
  - 5: 求解 (3.7) 中的 LASSO 估计.
  - 6: end for
  - 7: 根据 $\mathbf{U}_{i, \setminus i} = \mathbf{U}_{i, \setminus i} \hat{\beta}^i$ 按顺序更新矩阵 $\mathbf{U}$ 的行和列.
  - 8: end procedure
-

注 3.1 在第 2.2 小节基于 Spearman's rho 和 Kendall's tau 得到的相关系数矩阵的估计 $\hat{\mathbf{S}}$ 并不一定为非负定矩阵, 而算法中要求初始矩阵为非负定矩阵. 为此, 本文采用文献 [9] 中提出的方法, 即求解下面的优化问题得到 $\tilde{\mathbf{S}}$ 的代替 $\tilde{\mathbf{S}}$: $\tilde{\mathbf{S}} = \arg \min_{\mathbf{S} \succeq 0} \|\tilde{\mathbf{S}} - \mathbf{S}\|_{\max}$. 可以证明, 用 $\tilde{\mathbf{S}}$ 代替 $\hat{\mathbf{S}}$ 仍能保证相应的理论结果成立. 在实际模拟中, 为了简便, 我们直接将矩阵 $\hat{\mathbf{S}}$ 进行奇异值分解, 将所有负的奇异值截断成 0.

4 理论性质

本节给出 Gauss 关联结构图模型充分降维方法的理论性质. 为此, 首先给出两个条件. 条件 (C1) 给出真实存在边的两个节点间的相关系数的最小值下界. 条件 (C2) 要求协方差矩阵 Σ 的最大值不能随样本量的增加而发散太快, 该条件在许多实际应用中都很常见, 例如, 平稳时间序列的协方差矩阵满足该条件^[11].

(C1) 存在常数 $C_1 > 0$ 和 $0 < \kappa < 1/2$, 使得 $\min_{(i,j) \in E} |\Sigma_{i,j}| \geq C_1 n^{-\kappa}$.

(C2) 存在常数 $\tau \geq 0$ 和 $C_2 > 0$ 使得 $\lambda_{\max}(\Sigma) \leq C_2 n^\tau$.

定理 4.1 假定条件 (C1) 成立, 且存在常数 $C_3 > 0$ 和 $\xi \in (0, 1 - 2\kappa)$, 满足 $\log(p) = C_3 n^\xi$. 对某常数 $C_4 > 0$, 令 $\gamma = C_4 n^{-(\kappa+1)}$, 则存在常数 C_5 和 C_6 使得

$$\mathbb{P}(\mathcal{M}_*^i \subseteq \mathcal{M}_\gamma^i) \geq 1 - C_5 \exp(-C_6 n^{1-2\kappa}).$$

定理 4.1 保证了真实存在的边所构成的集合, 以很大的概率包含于 $\mathcal{M}_\gamma^i$, 即假阴性率很小.

定理 4.2 对某常数 $C_4 > 0$, 令 $\gamma = C_4 n^{-(\kappa+1)}$. 假定条件 (C1) 和 (C2) 成立, 且存在常数 $C_3 > 0$ 和 $\xi \in (0, 1 - 2\kappa)$, 满足 $\log(p) = C_3 n^\xi$, 则

$$\mathbb{P}\{|\mathcal{M}_\gamma^i| \leq O(n^{2\kappa+\tau})\} \geq 1 - C_5 \exp(-C_6 n^{1-2\kappa}).$$

当条件 (C2) 成立时, 集合 $\mathcal{M}_\gamma^i$ 中元素的个数是 $n^{2\kappa+\tau}$ 阶的, 若 $2\kappa + \tau < 1$, 则定理 4.2 保证充分降维方法可以将维数 p 降到比样本量 n 还小. 结合定理 4.1 和 4.2, 说明所提的充分降维方法可以实

现以很大的概率将真实存在边的集合全部筛选出来, 并且在做每一步回归时可以将维数降到比样本量还小, 极大减轻了计算的负担.

5 模拟研究

本节将进行模拟实验来研究所提充分降维方法的有限样本表现. 考虑以下三种方式来生成边的集合 E :

- (1) Band 结构: 对于 $|i - j| \leq 2$, 令 $(i, j) \in E$. 否则, 令 $(i, j) \notin E$.
- (2) Hub 结构: 将 p 个变量分成 20 个互不相交且等大小的子集: $C_1 \cup C_2 \cup \cdots \cup C_{20} = \{1, \dots, p\}$, $|C_k| = p/20$, $C_i \cap C_j = \emptyset$. 对于 $i \in C_k$ 中最小的指标, 对所有的 $\{j \neq i : j \in C_k\}$, 令 $(i, j) \in E$.
- (3) Cluster 结构: 将 p 个变量分成 20 个互不相交且等大小的子集: $C_1 \cup C_2 \cup \cdots \cup C_{20} = \{1, \dots, p\}$, $|C_k| = p/20$, $C_i \cap C_j = \emptyset$. 对 $i \in C_k$, $j \in C_k$, $i < j$, 以一定的概率令 $(i, j) \in E$, 而对于 $i \in C_k$, 有 $j \in C_l$, $k \neq l$, $(i, j) \notin E$.

当生成完边的集合 E 时, 按下述步骤产生精确矩阵 Ω :

步骤 1 生成 $p \times p$ 维矩阵 $\Theta = (\theta_{ij})$, 其中对 $(i, j) \in E$, $\theta_{ij} = \theta_{ji} = 0.5$, 否则, 令 $\theta_{ij} = \theta_{ji} = 0$.

步骤 2 为了保证所生成的矩阵 Ω 正定, 令 $\Omega = \Theta + (0.2 + |\lambda_{\min}(\Theta)|)I$.

生成精确矩阵之后, 协方差矩阵 Σ 通过 Ω^{-1} 产生, 之后我们标准化 Σ 使其对角线元素为 1. 有了 Σ , 我们从多元正态分布 $N(\mathbf{0}, \Sigma)$ 中产生 n 个独立观测样本. 为了获得来自 Gauss 关联结构分布的样本观测, 我们需要有变换函数 f 的反函数 $g = f^{-1}$. 为了简便, 我们每一维度都采用相同的单变量单调函数 $f_1 = f_2 = \cdots = f_p = f$. 本文采用以下两种 g 的函数形式: Gauss 累积分布函数变换和幂变换, 具体形式参见文献 [8]. 为了与文献 [8] 中的模拟保持一致, 这里采用相同的参数 $u_g = 0.05$, $\sigma_g = 0.4$, $\alpha = 3$. 在该模拟研究中, 令变量的维数 $p = 1,000$, 样本量设置为 $n = 100$ 和 $n = 200$.

记 $|E|$ 为图 $\mathcal{G}$ 中边的数目. 为了评价图模型估计的表现, 本文采用两个指标: 假阳性率和假阴性率. 令 $\hat{\mathcal{G}}_\lambda = (V, \hat{E}_\lambda)$ 为调节参数为 λ 时估计的图. 假阳性数目 (false positive number) FP_λ 的定义为在 $\hat{E}_\lambda$ 集合却不在集合 E 中边的条数. 假阴性数目 (false negative number) FN_λ 的定义为在集合 E 却不在集合 $\hat{E}_\lambda$ 中边的条数. 进而定义假阴性率 (false negative rate, FNR) 和假阳性率 (false positive rate, FPR) 分别为

$$\text{FNR}_\lambda = \frac{\text{FN}_\lambda}{|E|} \quad \text{和} \quad \text{FPR}_\lambda = \text{FP}_\lambda / \left(\binom{p}{2} - |E| \right).$$

记 Λ 为所有调节参数的集合, 定义最优的调节参数 λ^* 为

$$\lambda^* = \arg \min_{\lambda \in \Lambda} \{ \text{FNR}_\lambda + \text{FPR}_\lambda \}.$$

R 软件包 “huge” 提供了估计无向图模型的各类函数, 本文的模拟结果基于以上软件包. 为了比较原有方法和充分降维方法估计图模型所需要的时间, 我们考虑了 4 种维数情形: p 分别为 1,000、1,500、2,000 和 2,500, 样本量均为 100、产生的图结构为 “huge” 包中的 E-R 随机图, 调节参数的个数均设置为 10 个. 计算设备为 2.6 GHz CPU 和 4 GB 内存的笔记本电脑. 所需时间 (秒) 结果在表 1 中给出, SSG 和 SSM 方法为本文提出的基于 GLASSO 以及基于相邻选择的充分降维方法, SG (SKEPTIC-GLASSO) 和 SM (SKEPTIC-MB) 分别代表第 2 节介绍的基于 GLASSO 的 SKEPTIC 方法和基于相邻选择的 SKEPTIC 方法. 从表 1 可以看出, 充分降维方法可以使计算速度提高一倍.

基于 200 次重复的假阳性率和假阴性率结果在表 2 中给出. 模拟结果是根据 Kendall's tau 得到相关系数矩阵的估计, 进而代入本文所提方法得到的, 基于 Spearman's rho 的模拟效果相当, 在这里不再给出. 根据表 2 中的结果, 我们可以得到如下结论: (1) 充分降维的估计方法与原始方法的表现是

表 1 不同方法计算时间比较 (括号中为标准差)

d	SSG	SG	SSM	SM
1,000	1121.48 (26.01)	2181.23 (48.09)	114.04 (1.00)	213.96 (0.63)
1,500	1622.92 (157.90)	2606.07 (154.64)	133.42 (2.23)	232.44 (0.19)
2,000	11213.62 (21.10)	21520.20 (30.82)	161.85 (0.28)	261.93 (1.11)
2,500	12144.97 (25.23)	22730.55 (49.15)	197.82 (1.52)	334.68 (1.81)

表 2 不同方法的 FPR 以及 FNR 模拟结果 (括号中为标准差)

			SSG		SSM		
			FPR	FNR	FPR	FNR	
n							
Gauss 分布函数变换	Band	100	0.1242 (0.01)	0.0991 (0.01)	0.0622 (0.01)	0.1327 (0.02)	
		200	0.0630 (0.02)	0.0514 (0.01)	0.0217 (0.00)	0.0194 (0.00)	
	Hub	100	0.0191 (0.00)	0.0161 (0.00)	0.0093 (0.00)	0.0428 (0.01)	
		200	0.0027 (0.00)	0.0014 (0.00)	0.0013 (0.00)	0.0007 (0.00)	
	Cluster	100	0.0878 (0.01)	0.1619 (0.01)	0.0615 (0.01)	0.3320 (0.01)	
		200	0.0528 (0.01)	0.0887 (0.01)	0.0505 (0.01)	0.0786 (0.01)	
	幂次函数变换	Band	100	0.1214 (0.01)	0.1035 (0.01)	0.0552 (0.00)	0.1432 (0.01)
			200	0.0630 (0.02)	0.0514 (0.01)	0.0217 (0.00)	0.0194 (0.00)
Hub		100	0.0191 (0.00)	0.0161 (0.00)	0.0038 (0.00)	0.0780 (0.00)	
		200	0.0027 (0.00)	0.0014 (0.00)	0.0004 (0.00)	0.0019 (0.00)	
Cluster		100	0.1028 (0.01)	0.2153 (0.02)	0.0640 (0.01)	0.3541 (0.01)	
		200	0.0475 (0.01)	0.0871 (0.01)	0.0427 (0.01)	0.0735 (0.01)	
			SG		SM		
n			FPR	FNR	FPR	FNR	
Gauss 分布函数变换	Band	100	0.1119 (0.00)	0.0986 (0.01)	0.0541 (0.00)	0.1432 (0.01)	
		200	0.0528 (0.00)	0.0373 (0.01)	0.0523 (0.00)	0.0377 (0.01)	
	Hub	100	0.0190 (0.00)	0.0160 (0.00)	0.0087 (0.00)	0.0439 (0.01)	
		200	0.0050 (0.00)	0.0002 (0.00)	0.0015 (0.00)	0.0004 (0.00)	
	Cluster	100	0.0953 (0.01)	0.2141 (0.01)	0.0638 (0.00)	0.3759 (0.01)	
		200	0.0550 (0.01)	0.0690 (0.01)	0.0347 (0.00)	0.0923 (0.01)	
	幂次函数变换	Band	100	0.1132 (0.00)	0.0994 (0.01)	0.0542 (0.00)	0.1444 (0.01)
			200	0.0535 (0.00)	0.0368 (0.01)	0.0211 (0.00)	0.0198 (0.00)
Hub		100	0.0191 (0.00)	0.0156 (0.00)	0.0088 (0.00)	0.0433 (0.01)	
		200	0.0049 (0.00)	0.0003 (0.00)	0.0015 (0.00)	0.0005 (0.00)	
Cluster		100	0.1003 (0.01)	0.2453 (0.01)	0.0617 (0.00)	0.3216 (0.01)	
		200	0.0571 (0.01)	0.0794 (0.01)	0.0593 (0.01)	0.0603 (0.01)	

相当的,甚至在有些情形下充分降维的方法表现还要好. (2) 对于 Band 结构的图, SSM 方法的表现要比 SSG 方法的表现好很多. (3) 对于 Hub 结构图, SSG 方法的表现要比 SSM 方法的表现好一些. (4) 对于 Cluster 结构图, 由于图的结构更加复杂, 所有方法的假阳性率和假阴性率比在其他两种图结构下的要大.

6 实例分析

本节将本文所提方法应用到基因数据集, 该数据集是公开数据, 可以在 Gene Expression Omnibus 网站下载, 登录号为 GDS2771. 文献 [15,16] 曾研究过该数据集. 这组基因数据集由 97 个肺癌患者、90 个对照和 5 个疑似肺癌患者的呼吸道上皮细胞取样得到的 22,282 个微阵列基因表达值构成. 在该实例分析中, 我们仅考虑 97 个肺癌患者、90 个对照的样本和样本方差排在前 2,000 名的基因. 每一个基因的表达值数据都经过组内标准化处理, 均值为 0, 方差为 1.

在处理基因表达实际数据时, 由于基因真实的条件相关性未知, 无法通过类似于假阳性率和假阴性率等准则来判定方法估计图模型的好坏. 该实证研究的目的在于, 通过比较癌症组和对照组估计的图结构, 找出与肺癌有关的关键基因. 本文采用基于相邻选择的充分降维方法来分别估计病例组和对照组的图结构. 在调节参数的选择上, 我们选择使得两组图结构差异最大的调节参数. 这样就得到了两组的估计图, 我们选出在两个估计图中“邻居”变化最大的 10 个基因节点, 这些节点最有可能与肺癌有关. 在这 10 个基因中, 有 3 个基因已经通过生物实验被证实确实与肺癌有关, 分别为“PLS1”^[17]、“DAD1”^[18]和“C1QBP”^[19]. 本文认为另外 7 个基因也很可能与癌症有关, 值得进行深入实验研究验证. 在图 1 中, 我们围绕“PLS1”基因给出了两组估计出的图结构. 白色的节点是两组中都与基因“PLS1”连接的节点, 而灰色的点是只在其中一组中与基因“PLS1”连接的节点. 图 1 也确实显示“PLS1”基因在两组中的“邻居”变化非常大, 因此有充分的理由相信该基因确实与肺癌有关.

图 1 癌症组 (a) 和对照组 (b) 估计的部分图结构

7 总结

图模型一直以来是研究随机变量间相关关系的重要工具, 并且在信号处理和网络建模等领域引起了广泛的关注. 在考虑图模型时, Gauss 假设在实际应用中有一定的局限性. 为了放松这一假设, 文献 [8] 提出了 Gauss 关联结构图模型. Gauss 关联结构图模型包含了更大范围的一类图, 可以认为是 Gauss 图模型的更好的代替. 然而, 在变量维数可以是样本量指数阶的超高维情形下, 现存的方法都会面临严重的计算问题. 为此, 本文提出了超高维 Gauss 关联结构图模型的充分降维方法, 该方法简单有效, 能够极大地提高计算效率. 本文证明了所提方法能够以概率趋于 1 选出所有真实存在边的集合. 数值模拟和实际数据验证了本文所提方法的优势. 这种充分降维的方法可以应用到超高维的半参数回归模型中去. 例如, 文献 [20] 提出了 Gauss 关联结构回归模型, 在超高维的情形下, 同样可以采用相似的充分降维方法来减轻计算负担, 这超出了本文的研究范围, 我们将在后续研究中讨论该问题.

致谢 作者非常感谢审稿人以及副主编具有建设性的意见和建议, 让文章的质量有了很大的提升.

参考文献

- 1 Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the Lasso. *Ann Statist*, 2006, 34: 1436–1462
- 2 Yuan M, Lin Y. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 2007, 94: 19–35
- 3 Banerjee O, El Ghaoui L, D’Aspremont A. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J Mach Learn Res*, 2008, 9: 485–516
- 4 Dahl J, Vandenberghe L, Roychowdhury V. Covariance selection for nonchordal graphs via chordal embedding. *Optim Methods Softw*, 2008, 23: 501–520
- 5 Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 2008, 9: 432–441
- 6 Yuan M. High dimensional inverse covariance matrix estimation via linear programming. *J Mach Learn Res*, 2010, 11: 2261–2286
- 7 Cai T, Liu W, Luo X. A constrained l_1 minimization approach to sparse precision matrix estimation. *J Amer Statist Assoc*, 2011, 106: 594–607
- 8 Liu H, Lafferty J, Wasserman L. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *J Mach Learn Res*, 2009, 10: 2295–2328
- 9 Liu H, Han F, Yuan M, et al. High-dimensional semiparametric Gaussian copula graphical models. *Ann Statist*, 2012, 40: 2293–2326
- 10 Lam C, Fan J. Sparsistency and rates of convergence in large covariance matrix estimation. *Ann Statist*, 2009, 37: 4254–4278
- 11 Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. *J R Stat Soc Ser B Stat Methodol*, 2008, 70: 849–911
- 12 Fan J, Song R. Sure independence screening in generalized linear models with NP-dimensionality. *Ann Statist*, 2010, 38: 3567–3604
- 13 Fan J, Feng Y, Song R. Nonparametric independence screening in sparse ultra-high dimensional additive models. *J Amer Statist Assoc*, 2011, 106: 544–557
- 14 Li G, Peng H, Zhang J, et al. Robust rank correlation based screening. *Ann Statist*, 2012, 40: 1846–1877
- 15 Danaher P, Wang P, Witten D M. The joint graphical lasso for inverse covariance estimation across multiple classes. *J R Stat Soc Ser B Stat Methodol*, 2014, 76: 373–397
- 16 Luo S, Song R, Witten D. Sure screening for Gaussian graphical models. *ArXiv:1407.7819*, 2014
- 17 Erdogan E, Klee E W, Thompson E A, et al. Meta-analysis of oncogenic protein kinase C α signaling in lung adenocarcinoma. *Clinical Cancer Res*, 2009, 15: 1527–1533
- 18 Kim Y H, Girard L, Giacomini C P, et al. Combined microarray analysis of small cell lung cancer reveals altered apoptotic balance and distinct expression signatures of MYC family gene amplification. *Oncogene*, 2006, 25: 130–138
- 19 McGee A M, Douglas D L, Liang Y, et al. The mitochondrial protein C1qbp promotes cell proliferation, migration and resistance to cell death. *Cell Cycle*, 2011, 10: 4119–4127
- 20 Cai T T, Zhang L. High-dimensional Gaussian copula regression: Adaptive estimation and statistical inference. *Statist*

Sinica, 2018, 28: 963–993

附录 A 主要理论结果的证明

在给出主要定理证明之前, 我们先给出三个引理, 其中引理 A.1 和 A.2 是文献 [9] 中的结论, 其证明可参见文献 [9].

引理 A.1 对任意 $n \geq \frac{21}{\log p} + 2$, 以概率大于等于 $1 - 1/p^2$, 有

$$\sup_{j,k} |\hat{S}_{jk}^\rho - \Sigma_{jk}| \leq 8\pi \sqrt{\frac{\log p}{n}}.$$

引理 A.2 对任意 $n \geq 2$, 以概率大于等于 $1 - 1/p$, 有

$$\sup_{j,k} |\hat{S}_{jk}^\tau - \Sigma_{jk}| \leq 2.45\pi \sqrt{\frac{\log p}{n}}.$$

引理 A.3 令 $X = (X_1, \dots, X_p)^\top$ 为一个 p 维随机向量, 均值为零, 协方差矩阵为 Σ , $\text{diag}(\Sigma) = I$. 假定 X_i 和 $\mathbf{X}_{\setminus i} = (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_p)^\top$ 满足线性模型 $X_i = \mathbf{X}_{\setminus i}^\top \beta^i + \epsilon$, 其中 ϵ 与 X_i 不相关. 对某常数 C , 定义

$$\mathcal{S}^i = \{j \neq i : |\Sigma_{ij}| > Cn^{-\kappa}\},$$

则 $|\mathcal{S}^i|$ 满足

$$|\mathcal{S}^i| \leq C^{-2} n^{2\kappa} \lambda_{\max}(\Sigma).$$

证明 $X_i = \mathbf{X}_{\setminus i}^\top \beta^i + \epsilon$, 上式两边同乘以 $\mathbf{X}_{\setminus i}$, 然后取期望, 可以得到 $\beta^i = \Sigma_{\setminus i, \setminus i}^{-1} \mathbb{E}(X_i \mathbf{X}_{\setminus i})$. 因此, $\mathbb{E}(X_i X_j) = (\Sigma_{\setminus i, \setminus i} \beta^i)_j$, 等式右边为向量 $\Sigma_{\setminus i, \setminus i} \beta^i$ 的第 j 个元素. 从而,

$$\mathcal{S}^i = \{j \neq i : |(\Sigma_{\setminus i, \setminus i} \beta^i)_j| > Cn^{-\kappa}\} = \{j \neq i : (\Sigma_{\setminus i, \setminus i} \beta^i)_j^2 > C^2 n^{-2\kappa}\}.$$

此外,

$$\|\Sigma_{\setminus i, \setminus i} \beta^i\|_2^2 = (\Sigma_{\setminus i, \setminus i} \beta^i)^\top \Sigma (\Sigma_{\setminus i, \setminus i} \beta^i) \leq \lambda_{\max}(\Sigma_{\setminus i, \setminus i}) \|\Sigma_{\setminus i, \setminus i}^{1/2} \beta^i\|_2^2 = \lambda_{\max}(\Sigma_{\setminus i, \setminus i}) \beta^{i\top} \Sigma \beta^i.$$

而且, 注意到 X_{-i} 与 ϵ 不相关, 我们有

$$\beta^{i\top} \Sigma \beta^i = \text{Var}(X_{\setminus i}^\top \beta^i) = \text{Var}(X_i) - \text{Var}(\epsilon) \leq 1.$$

因此有 $\|\Sigma_{\setminus i, \setminus i}^{1/2} \beta^i\|_2^2 \leq \lambda_{\max}(\Sigma_{\setminus i, \setminus i})$, 进而可以得到

$$|\mathcal{S}^i| \leq \frac{\lambda_{\max}(\Sigma_{\setminus i, \setminus i})}{C^2 n^{-2\kappa}} \leq C^{-2} n^{2\kappa} \lambda_{\max}(\Sigma).$$

定理得证. □

定理 4.1 的证明 我们将证明定理对 $\hat{S}^\tau$ 成立, 类似地可以证明定理对 $\hat{S}^\rho$ 成立. 引理 A.2 保证下式成立:

$$\mathbb{P}(|\hat{S}_{ij}^\tau - \Sigma_{ij}| \geq 2.45\pi C_3 n^{-\kappa}) \leq \exp(-C_3 n^{(1-2\kappa)}).$$

令 $C_4 = \frac{49}{30}\pi C_3$, $C_1 = 2.45\pi C_3$, 注意到

$$\begin{aligned} P(\mathcal{M}_*^i \subsetneq \mathcal{M}_\gamma^i) &= P\left(\bigcup_{j \in \text{ne}_i} \{|\widehat{S}_{i,j}^\tau| < C_4 n^{-\kappa}\}\right) \\ &\leq \sum_{j \in \text{ne}_i} P(|\widehat{S}_{i,j}^\tau| < C_4 n^{-\kappa}). \end{aligned}$$

而且 $|\text{ne}_i| < p$, 对 $j \in \text{ne}_i$, 条件 (C1) 说明

$$P(|\widehat{S}_{i,j}| < C_4 n^{-\kappa}) \leq P\left(|\widehat{S}_{i,j}^\tau - \Sigma_{ij}| \geq \frac{1}{3}C_1 n^{-\kappa}\right).$$

从而,

$$P(\mathcal{M}_*^i \subsetneq \mathcal{M}_\gamma^i) \leq p \exp(-C_3 n^{(1-2\kappa)}) = \exp(-2C_3 n^{(1-2\kappa)}).$$

定理得证. □

定理 4.2 的证明 令

$$\Delta^i = \bigcap_{j \neq i} \left\{ |\widehat{S}_{ij} - \Sigma_{ij}| \leq \frac{1}{3}C_1 n^{-\kappa} \right\},$$

$\mathcal{M}_\gamma^i = \{j \neq i : |\widehat{S}_{ij}| > \frac{2}{3}C_1 n^{-\kappa}\}$, 则在事件 Δ^i 下, 如果 j 属于 $\mathcal{M}_\gamma^i$, 则它一定属于集合 $\mathcal{S}^i$. 因此有 $P(\mathcal{M}_\gamma^i \subseteq \mathcal{S}^i) \geq P(\Delta^i)$. 此外, 可以证明

$$P(\Delta^i) \geq 1 - C_5 \exp(-C_6 n^{1-2\kappa}).$$

上式进一步说明

$$P(\mathcal{M}_\gamma^i \subseteq \mathcal{S}^i) \geq 1 - C_5 \exp(-C_6 n^{1-2\kappa}).$$

结合引理 A.3 的结论, 定理得证. □

Sufficient dimension reduction for Gaussian copula graphical models

Yong He, Jiadong Ji & Xinsheng Zhang

Abstract We propose a sufficient dimension reduction method for recovering Gaussian copula graphical model which is very simple and computationally-efficient in the ultrahigh-dimensional setting. The screening of the conditional dependence graph is obtained by thresholding the elements of the rank-based correlation matrix estimator. The proposed approach possesses the sure screening property: with probability tending to 1, the estimated edge set contains the true edge set. We illustrate the performance of the proposed method in a simulation study and on a gene expression data. The result shows that in practice it performs competitively with more complex and computationally-demanding techniques for graph estimation.

Keywords graphical model, Gaussian copula, sufficient dimension reduction, non-parametric method

MSC(2010) 62H99, 62G35

doi: 10.1360/SCM-2016-0805