第50卷第6期 中国科学技术大学学报 Vol. 50, No. 6
2020年6月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Jun. 2020
文章编号:0253-2778(2020)06-0752-06

可延展的高维精度矩阵的置信区间

周慧婷,周 佳,郑泽敏

(中国科学技术大学管理学院国际金融研究院,安徽合肥 230601)

摘要:针对高维精度矩阵置信区间存在的计算效率低下的问题,提出了De-SCIO统计量.相比较其他方法,基于De-SCIO统计量构造的置信区间计算效率得到了很大的提升,并且它的平均覆盖率更接近于真实覆盖率.De-SCIO统计量构造简单,避免了复杂的理论推导.在合理的条件假设下,证明了De-SCIO统计量的渐近正态性.模拟实验以及实例分析展示了该方法在平均覆盖率和计算效率上的优势.

关键词:精度矩阵;置信区间;稀疏性;去偏统计量
中图分类号:O212.2 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2020.06.006
2010 Mathematics Subject Classification:Primary 62H30;Secondary 6207

引用格式:周慧婷,周佳,郑泽敏.可延展的高维精度矩阵的置信区间[J].中国科学技术大学学报,2020,50(6):752-757.
ZHOU Huiting, ZHOU Jia, ZHENG Zemin. Scalable confidence intervals of precision matrices in high dimensions[J]. Journal of University of Science and Technology of China, 2020,50(6):752-757.

Scalable confidence intervals of precision matrices in high dimensions

ZHOU Huiting, ZHOU Jia, ZHENG Zemin

(*International Instiure of Finance, School of Management, University of Science and Technology of China, Hefei* 230601, *China*)

Abstract: In order to solve the problem of the computational inefficiency in confidence intervals of high-dimensional precision matrices, the De-SCIO was proposed. Compared with other methods, the computational efficiency of the confidence intervals based on De-SCIO statistic are greatly improved, and their average coverage is closer to the true level. The construction of the De-SCIO statistic is simple and avoids complicated theoretical derivation. Under reasonable assumptions, the asymptotic normality of the De-SCIO statistic was proved. The advantages of this method in average coverage and computational efficiency were demonstrated by the numerical studies and real data example.

Key words: precision matrix; confidence intervals; sparsity; de-sparsified statistic

0 引言

随着信息化时代的到来,个体与个体之间的联系越来越广泛,这些广泛的联系形成了错综复杂的网络数据.网络数据发生在各个领域,例如现代医疗、线上营销、气候数据、社交网络等[1-3].研究这些由大量的个体组成、彼此间联系复杂的网络数据具有很现实的意义.图模型提供了一种灵活的方法来指定一组变量之间的条件关联性[4].在图模型中,其边的存在与变量间条件关联相互等价,并且图模型边的确定可以由它的精度矩阵(协方差的逆矩阵)来确定[5].特别在高斯图模型中,图模型的边完全由精度矩阵的非零元素来表示[6].例如,在基因网络数据中,精度矩阵中的非零元素对应基因或蛋白质之间的相互作用.精度矩阵还有很多其他的应用,例如分类和项目组合管理等.由此可见,精度矩阵对恢复图模型具有至关重要的作用.

现有的研究大多集中在精度矩阵的点估计上[7-12].然而,通过置信区间来量化统计的不确定性是很重要的,它可以用来对精度矩阵进行假设检验,去除点估计的噪音影响[13],也可以用来测试不同人群对应的网络是否相同.本文主要研究精度矩阵置信区间的估计.

近年来,高维统计推断主要发生在线性和广义线性模型中[14-17],对于精度矩阵的统计推断是最近才开始研究的.一般来说,目前对精度矩阵的区间估计大致可以分为两类:一类是基于似然函数的KKT(Karush-Kuhn-Tucker)转化,主要有文献[18-21].这些方法有个共同的特征,它们都是通过KKT转化方法得到一个去偏统计量,随后证明这

收稿日期:2019-12-26;**修回日期**:2020-06-03
基金项目:国家自然科学基金(72071187,11671374,71731010,71921001),中央高校基础研究基金(WK3470000017,WK2040000027)资助.
作者简介:周佳,女,1994年生,博士.研究方向:高维统计推断,变量选择.E-mail:tszhjia@mail.ustc.edu.cn
通讯作者:郑泽敏,博士/教授.E-mail:zhengzm@ustc.edu.cn

个去偏统计量对于精度矩阵的各个元素渐近正态，最后基于这个去偏统计量构造出精度矩阵各元素的置信区间. KKT 转化方法最早是由文献[18]提出，它是遵循文献[15]提出的线性模型统计推断的思想，从而构造出图模型的 De-Glasso（de-sparsified graphical Lasso）统计量. 并且其证明，在一定的条件下，De-Glasso 统计量中每个元素是精度矩阵元素的渐近正态估计. 还有一类是使用一个回归变换的方法获得一个估计量，从而进行统计推断. 最早是由文献[11]提出的基于成对节点回归的 ANT 方法，它是每两个变量对剩余的变量进行回归，从而得出一个渐近正态统计量.

当精度矩阵的维数适中时，上述方法可以很有效地得出精度矩阵的置信区间，然而当维数变高时，用这些方法计算精度矩阵置信区间的速度开始变慢. 为了解决效率低下的问题，本文基于 SCIO (sparse column-wise inverse operator)点估计提出了 De-SCIO 统计量，并且证明 De-SCIO 统计量对于精度矩阵的每个元素渐近正态，据此得出精度矩阵元素基于 De-SCIO 统计量构造的置信区间.

符号约定：对于矩阵 $A=(a_{ij})\in\mathbb{R}^{p\times q}$，定义 $|A|_\infty=\max_{ij}|A_{ij}|$，$\|A\|_\infty=\max_{1\leqslant i\leqslant p}\sum_{j=1}^{q}|a_{ij}|$ 为矩阵的∞-范数，$\|A\|_1=\|A^{\mathrm{T}}\|_\infty$ 为矩阵的 1-范数. 对于序列 f_n，g_n，如果存在独立于 n 的常数 $C>0$，$\forall$ 的 $n>C$ 有 $|f_n|\leqslant C|g_n|$，则定义 $f_n=O(g_n)$. 如果存在独立于 n 的常数 C_1,C_2，并且 $C_1<C_2$，$\forall$ 的 $n>C_2$ 有 $C_1|g_n|\leqslant|f_n|\leqslant C_2|g_n|$，则定义 $f_n\asymp g_n$. 若如果有 $\lim_{n\to\infty}f_n/g_n=0$，则定义 $f_n=o(g_n)$. 对于随机变量序列 x_n，如果 x_n 依概率有界，则定义为 $x_n=O_{\mathbf{P}}(1)$. 如果 $x_n/r_n=O_{\mathbf{P}}(1)$，则定义 $x_n=O_{\mathbf{P}}(r_n)$. 如果 x_n 依概率收敛到 0，则定义 $x_n=o_{\mathbf{P}}(1)$.

1　精度矩阵每个元素的置信区间

1.1　模型建立

设 X 为 p 维服从高斯多元分布的随机变量，即：

$$X=(x_1,\cdots,x_p)^{\mathrm{T}}\sim N(\mu,\Sigma^*)\tag{1}$$

式中，μ 为 p 维均值向量，$\Sigma^*=(\delta_{ij})_{p\times p}$ 为协方差矩阵. 设 $G=(V,E)$ 为高斯无向图，$V=\{x_1,\cdots,x_p\}$ 为 G 的顶点集，$E=\{(i,j)\}$ 为各顶点之间边的集合. 它们满足以下的性质：

$$x_i\perp x_j\mid X_{-\{i,j\}}\Leftrightarrow(i,j)\notin E,$$

这表明 x_i 和 x_j 对其余 $p-2$ 个顶点条件独立以及 x_i 和 x_j 之间不存在边相互等价.

在高斯图模型(1)中，$\Theta^*=(\Sigma^*)^{-1}$ 为其精度矩阵. 文献[6]提出 x_i 和 x_j 之间的边可以由精度矩阵的元素 Θ_{ij}^* 来表示. 当 $\Theta_{ij}^*=0$ 时，x_i 和 x_j 之间无边. 当 $\Theta_{ij}^*\neq0$ 时，x_i 和 x_j 之间有边. 设 $v=\{1,\cdots,p\}$，定义 $S=\{(i,j)\in v\times v\mid\Theta_{ij}^*\neq0\}$ 为精度矩阵的支撑集. 因此，恢复图模型 G 的问题等价于恢复精度矩阵的支撑集.

设 $X_1,\cdots,X_n\in\mathbb{R}^p$ 独立同分布于高斯图模型(1). 不失一般性，在全文中，本文假设均值向量 $\mu=0$. 在下文中将介绍精度矩阵置信区间的求解方法.

1.2　De-SCIO 统计量

本节将基于 SCIO 点估计构造 De-SCIO 统计量. SCIO 估计量是由文献[10]提出，它是精度矩阵点估计的一种估计方法，由于其逐列计算以及不需外部迭代等优点而以计算高效著称. 它的求解如下：

设 $\beta=(\beta_1,\cdots,\beta_p)^{\mathrm{T}}$，为估计 β_i，求解下列最优化问题：

$$\hat{\beta}_i=\arg\min_{\beta\in\mathbb{R}^p}\left\{\frac{1}{2}\beta^{\mathrm{T}}\hat{\Sigma}\beta-e_i^{\mathrm{T}}\beta+\lambda_{ni}\|\beta\|_1\right\}.$$

则由上式可得 β_i 的估计值为 $\hat{\beta}_i=(\hat{\beta}_{i1},\cdots,\hat{\beta}_{ip})^{\mathrm{T}}$. 由此得出精度矩阵 Θ^* 的 SCIO 估计值 $\hat{\Theta}=(\hat{w}_{ij})_{p\times p}$，其中 $\hat{w}_{ij}$ 通过下式得到：

$$\hat{w}_{ij}=\hat{w}_{ji}=\hat{\beta}_{ij}I\{|\hat{\beta}_{ij}|<|\hat{\beta}_{ji}|\}+\hat{\beta}_{ji}I\{|\hat{\beta}_{ij}|\geqslant|\hat{\beta}_{ji}|\}\tag{2}$$

遵循文献[15,18]去偏的思想，为了消除正则化惩罚带来的偏差，本文基于以下固定分解式来构造基于 SCIO 的渐近正态估计量：

$$2\hat{\Theta}-\hat{\Theta}\hat{\Sigma}\hat{\Theta}-\Theta^*=-\Theta^*W\Theta^*+\mathrm{rem}\tag{3}$$

式中，$\mathrm{rem}=-(\hat{\Theta}-\Theta^*)W\Theta^*-(\hat{\Theta}\hat{\Sigma}-I)(\hat{\Theta}-\Theta^*)$ 为余项，$W=\hat{\Sigma}-\Sigma^*$，可以得到 De-SCIO 统计量为

$$\hat{T}=2\hat{\Theta}-\hat{\Theta}\hat{\Sigma}\hat{\Theta}\tag{4}$$

在节 2 定理 2.1 中，可以证明，$\hat{T}_{ij}$ 是 Θ_{ij}^* 的渐近正态统计量. 在引理 2.1 中，可以得到其理论方差的相合估计. 由此得出 Θ_{ij}^* 的 $1-\alpha$ 的置信区间为

$$I_{ij,\alpha}\equiv I_{ij}(\hat{\Theta}_{ij},\alpha,n)=[\hat{T}_{ij}-\Phi^{-1}(1-\alpha/2)\hat{\sigma}_{ij}/\sqrt{n},\ \hat{T}_{ij}+\Phi^{-1}(1-\alpha/2)\hat{\sigma}_{ij}/\sqrt{n}]\tag{5}$$

由于 $\hat{T}$ 是 $\hat{\Theta}$ 的简单四则运算，所以 $\hat{T}$ 的计算复杂度主要来自于估计 $\hat{\Theta}$，又因为 SCIO 方法对 $\hat{\Theta}$ 的估计具有高效的计算效率，这使得式(5)可以快速地得到 Θ_{ij}^* 的置信区间. 在节 3 的模拟实验中可以看出，相比于 De-Glasso 统计量，De-SCIO 统计量得出的置信区间不仅覆盖率准确，计算速度也更为快速.

2　理论性质

本节将为 De-SCIO 统计量的渐近正态性提供理论保证. 在给出理论性质之前，首先给出 SCIO 估计量需要满足的条件.

条件 2.1　存在正的常数 $M_p=O(1)$，$L\asymp1$，使得 $\|\Theta^*\|_1\leqslant M_p$，$1/L\leqslant\Lambda_{\min}(\Theta^*)\leqslant\Lambda_{\max}(\Theta^*)\leqslant L$ 成立. Θ^* 的最大行基数满足 $d=o(\sqrt{n\log p})$.

条件 2.2　存在 $0<\alpha<1$，使得

$$\max_{1\leqslant i\leqslant p}\|\Sigma^*_{S_i^c\times S_i}\Sigma^{*-1}_{S_i\times S_i}\|_\infty\leqslant1-\alpha.$$

在条件 2.1 中，$\Lambda_{\min}(\cdot)$ 和 $\Lambda_{\max}(\cdot)$ 分别为给定的对称矩阵的最小和最大特征值. $d=\max_{i\in v}|\{j\in v:\Theta_{ij}^*\neq0\}|$ 为 Θ^* 中行非零元素最多的个数，它

满足一定的稀疏性. 条件 2.1 是大部分文献需要满足的条件，例如文献[11-13]. 条件 2.2 是类似于文献[18]中的不可表示性条件，它是很多文献通过惩罚方法支持恢复的必要条件[9,13]. 在这里，条件 2.1 和条件 2.2 是 SCIO 估计量满足其理论性质的条件保证[10].

定理 2.1 设 $X_1,\cdots,X_n \in \mathbb{R}^p$ 独立同分布于高斯图模型(1). 假设 $\Theta^* = (\Sigma^*)^{-1}$ 存在并且满足条件 2.1 和条件 2.2，$1/\sigma_{ij} = O(1)$. 设 $\hat{\Theta}$ 为式(2)定义的 SCIO 估计量，调节参数 $\lambda_{ni} \asymp \sqrt{\log p/n}$. 令 $\hat{T}$ 为式(4)定义的 De-SCIO 统计量. 则在稀疏性假设 $d = o(\sqrt{n}\log p)$ 下，对于所有的 $(i,j) \in v \times v$，有下式成立：

$$\frac{\sqrt{n}(\hat{T}ij - \Theta^*_{ij})}{\sigma_{ij}} \xrightarrow{\mathrm{d}} N(0,1) \tag{6}$$

此处定义 $\sigma^2_{ij} = \mathrm{Var}(\Theta^{*\mathrm{T}} X_1 X_1^{\mathrm{T}} \Theta^*_j)$ 为理论方差，$\xrightarrow{\mathrm{d}}$ 为依分布收敛. 稀疏性假设与文献[18]类似，定理 2.1 展示在稀疏性假设下，$\hat{T}_{ij}$ 是 Θ^*_{ij} 的渐近正态分布. 这个结论与文献[18]中的定理 1 相似，不同的是文献[18]中的 De-Glasso 统计量是通过对 Glasso 估计量 KKT 转化推导得到，而本文提出的 De-SCIO 统计量是基于固定分解，一定程度上简化了求导推导过程.

由于在定理 2.1 中 σ^2_{ij} 是未知的，为了得到 Θ^*_{ij} 的置信区间，需要得出 σ^2_{ij} 的相合估计 $\hat{\sigma}^2_{ij} > 0$. 在高斯设定下，引理 2.1 可以很容易地推导出理论方差，并且得到理论方差的相合估计.

引理 2.1 设 $X_1,\cdots,X_n \in \mathbb{R}^p$ 独立同分布于高斯图模型(1). 假设 $\Theta^* = (\Sigma^*)^{-1}$ 存在并且满足条件 2.1 和条件 2.2. 设 $\hat{\Theta}$ 为式(2)定义的 SCIO 估计量，调节参数 $\lambda_{ni} \asymp \sqrt{\log p/n}$. 则 $1/\sigma_{ij} = O(1)$，$\sigma^2_{ij} = \mathrm{Var}(\Theta^{*\mathrm{T}} X_1 X_1^{\mathrm{T}} \Theta^*_j) = \Theta^*_{ii}\Theta^*_{jj} + \Theta^{*2}_{ij}$ 和 $\hat{\sigma}^2_{ij} = \hat{\Theta}_{ii}\hat{\Theta}_{jj} + \hat{\Theta}^2_{ij}$，且有下式成立：

$$|\hat{\sigma}^2_{ij} - \sigma^2_{ij}| = O_{\mathbf{P}}(\sqrt{\log p/n}).$$

引理 2.1 可以得出 $\hat{\sigma}^2_{ij}/\sigma^2_{ij} = o_{\mathbf{P}}(1)$，即 $\hat{\sigma}^2_{ij}$ 是 σ^2_{ij} 的相合估计. 因此，可以用 $\hat{\sigma}^2_{ij}$ 来替换定理 2.1 中的理论方差 σ^2_{ij}，从而得到式(5)中 Θ^*_{ij} 的 $1-\alpha$ 的置信区间.

限于篇幅，本文中的定理和引理的证明不再给出. 如对证明有疑惑，请发邮件联系作者.

3 模拟实验

为了验证 De-SCIO 统计量构成的置信区间的覆盖准确性与计算效率，本节设计模拟实验，通过与 De-Glasso 统计量[18]对比来展开讨论.

首先，生成模拟数据集. 在数据设计中，设置真实精度矩阵 $\Theta^* = \mathrm{tridiag}(\rho,1,\rho)$ 为参数 $\rho > 0$ 的三对角矩阵，从 $N_p(0,\Sigma^*)$ 随机选取 n 个样本得到 $n \times p$ 的样本矩阵 X. 随后设计样本量 $n = 200$，维度 $p = 100,200,300,400,500,600,1000$，分别使用 De-SCIO 统计量、De-Glasso 统计量来估计 Θ^*_{ij} 的置信区间. 最后为了方便比较，仅仅考虑置信水平 90%，95%，99%的置信区间.

再次，在整个模拟计算中，在计算 De-Glasso 统计量时，使用 Glasso R 包(版本 1.10). 文献[18]提出 Glasso 估计量的理论调节参数 $\lambda \asymp \sqrt{\log p/n}$. 在计算 Glasso 时，本文采用文献[18]提出的调节参数 $\lambda = \sqrt{\log p/n}$. 同理，文献[10]提出 SCIO 估计量的理论调节参数 $\lambda_{ni} \asymp \sqrt{\log p/n}$. 在计算 De-SCIO 时，使用 SCIO R 包(版本 0.6.1)，使用调节参数 $\lambda_{ni} = \sqrt{\log p/n}$.

最后，本节通过平均覆盖率、平均计算时间两部分分别对比研究。

3.1 平均覆盖率

在平均覆盖率部分，使用四个指标来比较，分别为：Avgcov_S，Avgcov_{S^c}，$\mathrm{Avglength}_S$ 和 $\mathrm{Avglength}_{S^c}$. 其中 Avgcov_S 为在支撑集 S 上的平均覆盖率，它的计算公式为

$$\mathrm{Avgcov}_S = \frac{1}{|S|}\sum_{(i,j)\in S}\hat{\alpha}_{ij},$$

这里 $\hat{\alpha}_{ij} = \frac{1}{N}\sum_k^N \mathbf{1}_k\{\Theta^*_{ij} \in I_{ij,\alpha}\}$ 为重复迭代 N 次，Θ^*_{ij} 被式(5)定义的置信区间 $I_{ij,\alpha}$ 覆盖的概率. 同理，Avgcov_{S^c} 为在非支撑集 S^c 上的平均覆盖率. $\mathrm{Avglength}_S$ 为在 S 上的平均置信区间长度，它的计算公式为

$$\mathrm{Avglength}_S = \frac{1}{|S|}\sum_{(i,j)\in S}\hat{C}_{ij},$$

这里 $\hat{C}_{ij} = \frac{1}{N}\sum_k^N\{2\Phi^{-1}(1-\alpha/2)\hat{\sigma}_{ij}/\sqrt{n}\}_k$ 为重复迭代 N 次置信区间 $I_{ij,\alpha}$ 的平均长度. 同理，$\mathrm{Avglength}_{S^c}$ 为在 S^c 上的平均置信区间长度.

为了证明 De-SCIO 统计量具有广泛的适用性，在精度矩阵中，本文设置两个不同的 ρ 值，分别为 $\rho = 0.3$ 和 $\rho = 0.2$，重复迭代次数都为 $N = 100$，计算结果分别为表 1 和表 2.

在表 1 中，可以看出，De-SCIO 统计量相比于 De-Glasso 统计量，在 S 和 S^c 上的平均覆盖率均更接近于真实值. De-Glasso 统计量在 S 上平均覆盖率很低，而在 S^c 上平均覆盖率很高，它更容易将精度矩阵的非零元素识别为零元素，也就是更容易遗漏变量间的联系，这对恢复变量间的联系很不利. 事实上，在很多实际应用中，研究者们更倾向于尽可能多的找出变量之间的联系. 例如在基因检测中，更感兴趣的是基因间的联系，从而能发现对疾病的影响. 而 De-SCIO 统计量更好地实现了这一点，它对于变量的联系恢复很准确，基本接近我们设置的真实的置信水平.

并且本文还发现，De-SCIO 统计量在维度 p 越来越大时，它的估计始终维持很好的覆盖率，不随 p 的增加而减小. 而 De-Glasso 统计量对 p 较为敏感，可以看出，当 p 越来越大时，它在 S 上的平均覆盖率越来越小. 相比于 De-Glasso 统计量，De-SCIO 统计量适合处理更高维的精度矩阵.

与表 1 相比，表 2 设置了一个更小的参数值 $\rho = 0.2$，可以发现，其结论与表 1 的保持一致，并且 De-SCIO 统计量在 S 和 S^c 上依旧保持精确的平

均覆盖率，可见参数值的变小并不影响 De-SCIO　　统计量的估计精度.

表 1　$p=100,500,1000, n=200, \rho=0.3, N=100$ 时，De-SCIO 与 De-Glasso 的比较

Tab. 1　The comparision of De-SCIO and De-Glasso when $p=100,500,1000$ and $n=200, \rho=0.3, N=100$

		0.90		0.95		0.99	
		De-SCIO	De-Glasso	De-SCIO	De-Glasso	De-SCIO	De-Glasso
$p=100$	Avgcov_S	0.881	0.770	0.936	0.851	0.984	0.947
	Avglength_S	0.253	0.210	0.302	0.250	0.396	0.328
	Avgcov_{S^c}	0.915	0.961	0.960	0.983	0.993	0.997
	Avglength_{S^c}	0.220	0.182	0.262	0.217	0.344	0.285
$p=500$	Avgcov_S	0.880	0.720	0.935	0.806	0.983	0.920
	Avglength_S	0.254	0.204	0.303	0.243	0.398	0.320
	Avgcov_{S^c}	0.913	0.979	0.959	0.992	0.993	0.999
	Avglength_{S^c}	0.221	0.178	0.264	0.212	0.347	0.279
$p=1000$	Avgcov_S	0.882	0.662	0.935	0.758	0.982	0.892
	Avglength_S	0.257	0.203	0.306	0.242	0.403	0.318
	Avgcov_{S^c}	0.910	0.986	0.957	0.995	0.992	0.999
	Avglength_{S^c}	0.224	0.177	0.267	0.211	0.351	0.278

表 2　$p=100,500,1000, n=200, \rho=0.2, N=100$ 时，De-SCIO 与 De-Glasso 的比较

Tab. 2　The comparision of De-SCIO and De-Glasso when $p=100,500,1000$ and $n=200, \rho=0.2, N=100$

		0.90		0.95		0.99	
		De-SCIO	De-Glasso	De-SCIO	De-Glasso	De-SCIO	De-Glasso
$p=100$	Avgcov_S	0.891	0.812	0.944	0.893	0.987	0.974
	Avglength_S	0.259	0.219	0.309	0.261	0.406	0.343
	Avgcov_{S^c}	0.906	0.954	0.954	0.980	0.991	0.996
	Avglength_{S^c}	0.227	0.192	0.270	0.229	0.355	0.301
$p=500$	Avgcov_S	0.893	0.785	0.944	0.873	0.987	0.969
	Avglength_S	0.261	0.214	0.312	0.255	0.409	0.336
	Avgcov_{S^c}	0.904	0.967	0.953	0.987	0.991	0.998
	Avglength_{S^c}	0.229	0.188	0.273	0.224	0.359	0.294
$p=1000$	Avgcov_S	0.896	0.766	0.946	0.858	0.987	0.966
	Avglength_S	0.265	0.213	0.315	0.254	0.414	0.333
	Avgcov_{S^c}	0.902	0.973	0.952	0.990	0.991	0.999
	Avglength_{S^c}	0.232	0.187	0.276	0.222	0.363	0.292

3.2　平均 CPU 时间

在平均计算时间部分，设置参数为 $n=200$，$\rho=0.3$，$\alpha=0.05$，$N=100$，随后分别记录置信水平为 $1-\alpha$ 的 De-SCIO 置信区间和 De-Glasso 置信区间的平均 CPU 时间. 计算结果见图 1，其中 x 轴表示维度 p，y 轴表示平均 CPU 时间.

在图 1 中可见，随着维度 p 的变化，通过 De-SCIO 方法得到置信区间的平均 CPU 时间呈线性增长，而通过 De-Glasso 方法得到置信区间的平均 CPU 时间呈指数增长. 在 $p=600$ 时，De-Glasso 方法的平均 CPU 时间是 De-SCIO 的 19 倍，并且随着 p 的增大相差越来越多. 在计算更高维精度矩阵时，De-SCIO 统计量比 De-Glasso 统计量计算更有效.

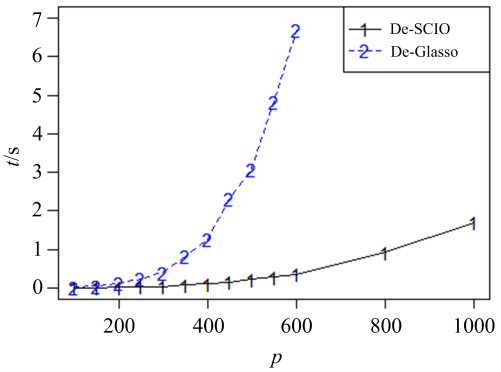

图 1　De-SCIO 和 De-Glasso 平均 CPU 时间

Fig. 1　The average CPU times of De-SCIO and De-Glasso

4　实例分析

为了展示 De-SCIO 方法在实际数据中的作用，本文选用人类基因数据集，分别运用 De-SCIO 及 De-Glasso 两种方法来恢复基因数据图模型的边，进而来研究基因间的相互作用. 基因数据集为公开数据集，我们是从 BDgraph R 包中获得，它包含 $n=60$ 个来自于美国不同州居民的 B 淋巴细胞中的

基因表达数据[23]. 为了更方便地分析，本文依据文献[24]，从 47293 个基因中选用 $p=100$ 个基因表达数据，并且对它们进行标准化处理.

本文分别使用 De-SCIO、De-Glasso 方法估计基因表达数据精度矩阵 Θ_{ij}^{*} 的置信区间，进而来恢复基因间的联系，显著性水平设置为 $\alpha=0.05$. 在计算 De-SCIO、De-Glasso 统计量时，都选用五折交叉验证从 $[a\sqrt{\log p/n}, b\sqrt{\log p/n}]$ 的路径中选择使 Bregman 损失函数最小的调节参数值. Bregman 损失函数[13]定义为 $L(\hat{\Sigma},\hat{\Theta})=\langle\hat{\Theta},\hat{\Sigma}\rangle-\log\det(\hat{\Theta})$，其中 $\hat{\Sigma}$ 为测试集的样本协方差矩阵，$\hat{\Theta}$ 为训练集上计算的精度矩阵的点估计值.

计算结果显示，De-SCIO 方法得出 77 条边为显著的，De-Glasso 方法得出 56 条边为显著的. 与前人研究的基因联系[24]对比，De-SCIO 方法得出的边集大部分与前人的结果保持一致. 例如 De-SCIO 方法得出的{GI_33356162-S 与 GI_17981706-S，Hs.512137-S 与 GI_41190507-S，GI_13514808-S 与 GI_33356162-S，GI_18641317-S 与 GI_24797066-S}等基因对有显著性关系，均与文献[24]提出的边相同. 然而这其中的 GI_18641317-S 与 GI_24797066-S 等基因对的联系，却不包括在 De-Glasso 方法得出的边集中. 这样的结论与模拟实验类似，相比较于 De-SCIO 方法，De-Glasso 方法容易将精度矩阵的非零元素识别为零元素，容易遗漏变量间的联系. 而 De-SCIO 方法估计的边集更接近于真实边集.

另一方面，由 De-SCIO 方法得出的基因 GI_24797066-S 与 GI_37546026-S 之间有显著性关系，然而在先前的研究中并没有被提出. 这可能是一个潜在的基因间的相互作用，可以作为分析人类基因间相互作用的补充.

5 结论

为了解决现有精度矩阵推断方法在面对高维精度矩阵时存在的计算效率低下的问题，本文提出了基于 SCIO 的去稀疏估计量，称为 De-SCIO. 我们严格证明了该统计量的渐进正态性，从而为置信区间的构造提供了依据. 相比较 De-Glasso 方法，De-SCIO 方法继承了 SCIO 估计的逐列求解以及不需要外部迭代等优点而具有显著的计算效率上的优势，从而可以有效地处理高维精度矩阵. 在模拟计算中可以看出，基于 De-SCIO 统计量构造的置信区间不仅对精度矩阵的覆盖准确率高，并且计算效率优于 De-Glasso 方法，并且在维度很高的情形下尤为显著. 在当今大数据的背景下，研究的变量数目日益增多，变量与变量之间的联系越来越复杂，本文基于当前大数据的背景，为揭示变量之间的复杂联系提供了一种可供选择的计算高效的方法.

然而，随着研究变量的增多，精度矩阵统计推断的计算依旧是一个主要问题. 随着研究的深入，计算快速的点估计相继出现，例如 ADMM (alternating direction method of multipliers)算法[26]. 研究这些计算快速的点估计的渐近性质进而来构造精度矩阵的置信区间将会是一件很有意义的事，这也会是本文进一步研究的问题. 此外，本文提出的去偏统计量是在高斯分布下得到的，还可以用 Lindeberg 中心极限定理[25]证明在次高斯分布的情况下也同样具有此优良的性质.

参考文献(References)

[1] SCHÄFER J, STRIMMER K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics [J]. Statistical Applications in Genetics and Molecular Biology, 2005, 4(1): Article 32.

[2] CHEN X, LIU Y, LIU H, et al. Learning spatial-temporal varying graphs with applications to climate data analysis[C]// Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence. Palo Alto, CA: Association for the Advancement of Artificial Intelligence, 2010: 425-430.

[3] FAN J, LIAO Y, LIU H. An overview of the estimation of large covariance and precision matrices [J]. Econometrics Journal, 2016, 19(1): C1-C32.

[4] WAINWRIGHT M J, JORDAN M I. Graphical Models, Exponential Families, and Variational Inference [M]. Hanover, MA: Now, 2008.

[5] LIU H, LAFFERTY J, WASSERMAN L. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs [J]. Journal of Machine Learning Research, 2009, 10(3): 2295-2328.

[6] LAURITZEN S. Graphical Models [M]. New York: Oxford Univ Press, 1996.

[7] MEINSHAUSEN N, BÜHLMANN P. High-dimensional graphs and variable selection with the Lasso [J]. The Annals of Statistics, 2006, 34(3): 1436-1462.

[8] BICKEL P J, LEVINA E. Covariance regularization by thresholding[J]. The Annals of Statistics, 2008, 36(6): 2577-2604.

[9] RAVIKUMAR P, WAINWRIGHT M J, RASKUTTI G, et al. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence [J]. Electronic Journal of Statistics, 2011, 5: 935-980.

[10] LIU W, LUO X. Fast and adaptive sparse precision matrix estimation in high dimensions[J]. Journal of Multivariate Analysis, 2015, 135:153-162.

[11] REN Z, SUN T, ZHANG C, et al. Asymptotic normality and optimalities in estimation of large Gaussian graphical models [J]. The Annals of Statistics, 2015, 43(3): 991-1026.

[12] FAN Y, LV J. Innovated scalable efficient estimation in ultra-large Gaussian graphical models [J]. The Annals of Statistics, 2016, 44(5): 2098-2126.

[13] CAI T, LIU W, LUO X. A constrained ℓ_1 minimization approach to sparse precision matrix estimation [J]. Journal of the American Statistical Association, 2011, 106: 594-607.

[14] NICKL R, VAN DE GEER S. Confidence sets in sparse regression[J]. The Annals of Statistics, 2013, 41(6): 2852-2876.

[15] VAN DE GEER S, B UHLMANN P, RITOV Y, et al. On asymptotically optimal confidence regions and

tests for high dimensional models[J]. The Annals of Statistics, 2014, 42(3): 1166-1202.
[16] MEINSHAUSEN N. Assumption-free confidence intervals for groups of variables in sparse high-dimensional regression [DB/OL]. [2019-12-01]. https://arxiv.org/abs/1309.3489.
[17] ZHANG C H, ZHANG S S. Confidence intervals for low-dimensional parameters in high dimensional liner models[J]. Journal of the Royal Statistical Society: Series B, 2014, 76: 217-242.
[18] JANKOVA J, VAN DE GEER S. Confidence intervals for high-dimensional inverse covariance estimation [J]. Electronic Journal of Statistics, 2015, 9 (1): 1205-1229.
[19] JANKOVA J, VAN DE GEER S. Honest confidence regions and optimality in high-dimensional precision matrix estimation[J]. TEST: An Official Journal of the Spanish Society of Statistics and Operations Research, 2017, 26(1): 143-162.
[20] HUANG X, LI M. Confidence intervals for sparse precision matrix estimation via Lasso penalized D-trace loss[J]. Communications in Statistics: Theory and Methods, 2017, 46(24): 12299-12316.
[21] JANKOVA J, VAN DE GEER S. Inference in high dimensional graphical models [DB/OL]. [2019-12-01]. https://arxiv.org/abs/1801.08512.
[22] YUAN M, LIN Y. Model selection and estimation in the Gaussian graphical model[J]. Biometrika, 2007, 94(1): 19-35.
[23] STRANGER B E, NICA A C, FORREST M S, et al. Population genomics of human gene expression[J]. Nature Genetics, 2007, 39(10): 1217-1224.
[24] BHADRA A, MALLICK B K. Joint high-dimensional Bayesian variable and covariance selection with an application to eQTL analysis[J]. Biometrics, 2013, 69 (2): 447-457.
[25] DURRETT R. Probability: Theory and Examples [M]. Cambridge: Cambridge University Press, 2010.
[26] WANG C, JIANG B. An efficient ADMM algorithm for high dimensional precision matrix estimation via penalized quadratic loss[J]. Computational Statistics & Data Analysis, 2020, 142: Article 106812.