计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(5):28−36 [doi: 10.15888/j.cnki.csa.009504]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

具有错误发现率控制的网络连接数据变量选择①

卢 滢, 李 阳

(中国科学技术大学 管理学院 统计与金融系, 合肥 230026)
通信作者: 李 阳, E-mail: tjly@ustc.edu.cn

摘 要: 网络连接数据的统计推断问题已成为近年来统计学研究的热点问题. 传统模型中样本数据间的独立性假设通常不能满足现代网络连接数据的分析需求. 本文研究了网络连接数据中每个节点的独立效应, 并借助融合惩罚的思想, 使得相互连接节点的独立效应趋同. 同时借助仿变量方法 (Knockoff) 仿冒原始变量的数据依赖结构、构造与目标变量无关的属性特征, 提出了针对网络连接数据进行变量选择的仿变量方法 (NLKF). 从理论上证明了 NLKF 方法将变量选择的错误发现率 (*FDR*) 控制在目标水平. 对于原始数据协方差未知的情形, 使用估计的协方差矩阵仍具有上述良好的统计性质. 通过与传统变量选择方法 Lasso 对比, 说明了本文方法的可靠性. 最后结合因子投资领域 2022 年 1–12 月中国 A 股市场 4 000 只股票的 200 个因子数据及每只股票所属申万一级行业构造的网络关系, 给出模型的应用实例.

关键词: 网络连接数据; 变量选择; Knockoff 方法; 错误发现率

引用格式: 卢滢,李阳.具有错误发现率控制的网络连接数据变量选择.计算机系统应用,2024,33(5):28–36. http://www.c-s-a.org.cn/1003-3254/9504.html

Variable Selection in Network-linked Data with FDR Control

LU Ying, LI Yang

(Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei 230026, China)

Abstract: The statistical inference of network data has become a hot topic in statistical research in recent years. The independence assumption among sample data in traditional models often fails to meet the analytical demands of modern network-linked data. This work studies the independent effect of each network node in the network-linked data, and based on the idea of fusion penalty, the independent effect of the associated nodes is converged. Knockoff variables construct covariates independent of the target variable by imitating the structure of the original variable. With the help of Knockoff variables, this study proposes a general method framework for variable selection for network-linked data (NLKF). The study proves that NLKF can control the false discovery rate (*FDR*) at the target level and has higher statistical power than the Lasso variable selection method. When the covariance of the original data is unknown, the covariance matrix using the estimation still has good statistical properties. Finally, combining the 200 factor samples of more than 4 000 stocks in the A-share market and their network relations constructed by Shenyin Wanguo's first-level industry classification, an example of the application in the field of financial engineering is given.

Key words: network-linked data; variable selection; Knockoff method; false discovery rate (*FDR*)

网络连接数据即为包含数据点之间关联信息的网状结构数据. 1991 年, CERFNet、PSINet 及 AlterNet 商业化, 世界步入互联网时代. 随着计算机技术的发展和物联网的广泛应用, 数据的增长速度和应用范围已达

① 基金项目: 国家自然科学基金 (12101584)
收稿时间: 2023-11-29; 修改时间: 2023-12-29; 采用时间: 2024-01-12; csa 在线出版时间: 2024-04-01
CNKI 网络首发时间: 2024-04-03

到前所未有的高度. 根据思科 (Cisco) 的 2023 年全球网络趋势报告, 全球数据流量在过去 5 年中增长了近 4 倍. IDC 预测到 2025 年将达到每年 163 ZB[1]. 伴随数据量的爆炸式扩张, 数据的种类和结构也日益多元化. 不止单个人或物的属性信息被详细记录, 人与人、人与物的关联信息也被以网络结构记录下来. 常见的数据网络包括社交网络、电力网络、生物网络、病毒传播网络、道路交通网络等. 网络数据的记录为其统计分析提供有效的增量信息[2]. 例如, 在社交网络中, 对用户之间互动模式的分析, 可以深入理解社会趋势和信息传播机制[3]. 在生物信息学领域, 对网络数据的分析有助于揭示生物分子间的相互作用, 从而推动对疾病机理的理解和新药的开发[4]. 在金融领域, 对资产之间网络关系的分析, 可以更好地理解市场风险和指导投资决策[5]. 在物联网技术中, 网络分析对于提高系统效率和设备安全也至关重要.

面对如此快速增长和多样化的数据环境, 各领域的模型和方法也在迅速演变. 传统的统计模型, 如基于独立同分布假设的模型, 面临着新的挑战. 这些模型往往假设数据点之间是独立的, 但在网络连接数据中数据点之间的相互关联和依赖是核心特征. 因此需要新的分析方法来解读这些复杂的关系, 尤其是在面对网络数据的多维特征和动态关联时. 这种转变要求各领域不断创新, 开发出能够有效处理网络连接数据的新方法和技术. 网络连接数据的建模和有效利用已成为当前计算机科学、社会科学和统计学等多个领域的重点研究方向之一.

现有的网络连接模型的研究, 根据研究方向的不同, 可分为两类. 一类聚焦于识别关系数据中隐含的网络结构, 例如机器学习领域的社区检测[6], 研究了如何在复杂网络中识别出具有紧密联系的节点群组. 另一类旨在发展一般的数据分析框架以利用网络数据进行回归和预测. 网络连接数据的回归问题被关注初期, 一些学者针对某些特定情况下的网络结构加以利用, 例如自回归模型[7]及它的变体组间交互效应和组间固定效应模型[8]. 这些模型先假定网络对数据分析的影响的具体形式, 比如内生效应、外生效应或是相关效应等, 之后将研究的重点放在效应的识别上. 融合惩罚的提出为网络连接数据的统计处理提供了新的视角. 这一思想最初由 Stephanie 等[9]和 Tibshirani 等[10]引入, 旨在缩减网络分析中相互关联节点之间回归系数的差异, 被视为点估计的回归版本, 或是在贝叶斯框架中被视作马尔可夫随机场的先验回归. 基于这一理论基础, Li 等[11]提出网络一致性模型, 为网络连接数据回归分析提供了一个创新而实用的框架.

针对网络连接数据的复杂性, 一个关键的任务是如何从庞大的网络数据中提取出最有意义的特征. 这不仅涉及识别数据点间的关联关系, 还要理解这些关系如何影响整体网络的行为和性质. 在这个过程中, 变量选择[12]成为一个重要的环节, 它有助于简化模型, 增强其解释性和预测能力. 例如, 在计算机科学领域许多实际应用如自然语言处理[13]、图像识别[14,15]和推荐系统[16]等都涉及大量特征, 并非所有特征都对最终任务有实质性贡献. 有效的变量选择方法能显著提高模型效率和性能, 有助于更深入地理解和利用网络结构数据. 近些年涌现了许多创新的变量选择方法, 包括了经典的前向选择、后向选择、逐步回归[17]以及基于正则化的方法, 如 Lasso、岭回归和弹性网[18–20]等. 机器学习技术[21,22]也为变量选择提供了广泛思路.

在现有的网络连接数据变量选择研究中, 通常采用基于 p 值的方法, 但面临着明显的局限性. 这主要是因为网络连接数据通常具有高度的结构依赖性和复杂性, 传统的基于 p 值的变量选择方法大多建立在数据点相互独立的假设之上. 在这种情况下, p 值方法无法准确反映变量之间的真实关系, 尤其是在面对高维网络数据时, 这种方法很容易导致错误的发现, 从而影响模型的准确性和可靠性.

Benjamini 等[23]最早引入错误发现率 (false discovery rate, *FDR*) 的概念, 用于描述在所有选出特征中错选比例的期望. 2015 年 Barber 等[24]提出 Knockoff 仿变量方法以有效地控制多重假设检验中的错误发现率 (*FDR*), 对于避免过度发现与伪发现相关问题具有重要意义. Knockoff 仿变量方法通过构建用作控制变量的独立于原始变量的仿冒变量, 模仿原始协变量的依赖结构. 通过比较原始变量与仿冒变量的表现情况, 挑选出真正与响应变量相关的特征. 为了适应因变量对协变量的任意依赖结构, Candes 等[25]在一般高维非线性模型中引入 Knockoff 框架, 提出 Model-X Knockoff 方法. Fan 等[26]在 Model-X Knockoff 模型的基础上做了改进, 将 Knockoff 方法拓展到协变量的联合分布未知的情况, 提出 Knockoff 变量选择框架下的 RANK 方法.

为了克服相互依赖数据无法精确计算 p 值这一局限性, 本文提出一种基于错误发现率控制的网络连接

数据变量选择方法 (NLKF). NLKF 方法通过模拟原始数据的结构创建原始特征的“Knockoff 特征”, 并利用网络内聚性惩罚机制, 使相互连接的节点在独立效应上趋于一致, 进而挑选出与目标变量相关的重要特征. 文章从理论和实验两方面说明本文提出的 NLKF 方法能够将网络连接数据变量选择的错误发现率控制在预设水平. 该方法不依赖于网络结构的先验信息, 广泛适用于网络数据分析的实际应用场景. NLKF 方法在处理复杂和高维数据方面表现出色, 能够更有效地满足现代数据分析的需求.

1 具有 *FDR* 控制的网络连接数据变量选择方法

在实际回归及预测问题中, 通常只有少数几个样本属性与目标变量相关. 在线性模型问题中, 意味着回归系数向量β中只有少数几个分量是非零的. 本文研究的核心问题在于设计有可靠性保证的网络连接数据的变量选择方法——网络连接数据仿变量方法 (NLKF).

1.1 模型设定

考虑下面一个网络内聚性回归模型:

$$\boldsymbol{Y}=\boldsymbol{\alpha}+\boldsymbol{X\beta}+\boldsymbol{\varepsilon} \tag{1}$$

其中, 截距项$\boldsymbol{\alpha}=(\alpha_1,\cdots,\alpha_n)^{\mathrm{T}}\in R^n$是未知的$n$维独立节点效应向量, α_i, $i\in[1,n]$为节点间各不相同的点效应. $\boldsymbol{Y}=(y_1,\cdots,y_n)^{\mathrm{T}}\in R^n$是 n 维响应变量, $\boldsymbol{X}=(x_1,\cdots,x_n)^{\mathrm{T}}\in R^{n\times p}$是$n\times p$维设计矩阵, x_i是第i个p维协变量, $\boldsymbol{\beta}=(\beta_1,\cdots,\beta_p)^{\mathrm{T}}\in R^p$是未知的$p$维系数向量, $\boldsymbol{\varepsilon}$是n维误差向量, 且$E(\boldsymbol{\varepsilon})=0$, $Var(\boldsymbol{\varepsilon})=\sigma^2\boldsymbol{I}$.

无向连接图$\mathcal{G}$记录数据间的网络连接信息. $\mathcal{G}=(V,E)$, $V=1,2,\cdots,n$为$\mathcal{G}$的顶点集, $E\subset V\times V$为$\mathcal{G}$的边集. 连接矩阵$\boldsymbol{A}\in R^{n\times n}$记录连接图$\mathcal{G}$的节点关系信息. 当节点$u$、$v$相连, 即$(u,v)\in E$时, $A_{uv}=1$; 当节点u、v不相连, 即$(u,v)\notin E$时, $A_{uv}=0$. 对于无向连接图$\mathcal{G}$, 连接矩阵满足$A_{uv}=A_{vu}$. 同时假定在连接图中没有环, 即对于任意$v\in V$有 $A_{vv}=0$. $\boldsymbol{D}$为连接图$\mathcal{G}$的度矩阵, $\boldsymbol{D}=\mathrm{diag}(d_1,d_2,\cdots,d_n)$. 其中$d_u$为顶点$u$的度, $d_u=\sum_{v\in V}A_{uv}$. $\boldsymbol{L}$为连接图$\mathcal{G}$的拉普拉斯矩阵, $\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{A}$.

上述模型中, 网络数据节点x_i, $i\in[1,n]$的节点效应α_i, $i\in[1,n]$为模型增加了n个未知参数. 在不对$\boldsymbol{\alpha}$的结构做任何假设的情况下, 引入网络内聚性惩罚, 认为相互连接的网络节点应具有相似的节点效应. 即引入包含网络内聚性惩罚$\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{L\alpha}$的损失函数:

$$Loss(\boldsymbol{\alpha},\boldsymbol{\beta})=\|\boldsymbol{Y}-[\boldsymbol{X}\tilde{\boldsymbol{X}}]\boldsymbol{b}-\boldsymbol{\alpha}\|^2+\mu\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{L\alpha}+\lambda\|\boldsymbol{b}\|_1 \tag{2}$$

其中, λ和μ为惩罚系数, $\|\cdot\|_1$为L_1范数, $\boldsymbol{L}$为数据网络结构的拉普拉斯矩阵, $\boldsymbol{\alpha}=\{\alpha_1,\cdots,\alpha_n\}^{\mathrm{T}}$为节点效应向量, $\boldsymbol{\beta}=\{\beta_1,\cdots,\beta_p\}^{\mathrm{T}}$为模型回归系数.

网络内聚性惩罚项$\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{L\alpha}$可以写成更直观的形式:

$$\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{L\alpha}=\sum_{(u,v)\in E}(\alpha_u-\alpha_v)^2$$

由上式可知, 网络内聚性惩罚思想, 是通过惩罚相互连接的节点之间的差异, 使得相互连接节点的独立效应$\boldsymbol{\alpha}$趋于相同.

1.2 网络连接数据的变量选择方法 (NLKF)

为协变量$\boldsymbol{X}$构建模仿其内部依赖结构且与目标变量$\boldsymbol{Y}$无关的 Knockoff 特征$\tilde{\boldsymbol{X}}\in R^{n\times p}$如下.

定义 1. Network-Knockoff 特征. 对于网络连接数据$\boldsymbol{X}=(x_1,\cdots,x_n)^{\mathrm{T}}\in R^{n\times p}$, 定义其 Network-Knockoff 变量为:

$$\tilde{\boldsymbol{X}}=(\tilde{x}_1,\cdots,\tilde{x}_n)\in R^{n\times p} \tag{3}$$

其中, $\tilde{x}_i, i\in[1,n]$独立产生于条件分布:

$$\tilde{x}_i|x_i\sim N(\boldsymbol{C}x_i,\boldsymbol{B}^2)$$

$$\boldsymbol{C}=\boldsymbol{I}_p-\mathrm{diag}\{s\}\boldsymbol{\Omega}$$

$$\boldsymbol{B}=(2\mathrm{diag}\{s\}-\mathrm{diag}\{s\}\boldsymbol{\Omega}\mathrm{diag}\{s\})^{1/2}$$

其中, $\boldsymbol{\Omega}$为网络连接数据$\boldsymbol{X}$的精度矩阵.

由上述定义可知 $2p$ 维随机向量$(x_i^{\mathrm{T}},(\tilde{x}_i)^{\mathrm{T}})^{\mathrm{T}}$独立同分布于均值 0, 协方差如下的高斯分布:

$$\begin{cases} cov(x_i)=\boldsymbol{\Sigma}_0 \\ cov(x_i,\tilde{x}_i)=\boldsymbol{\Sigma}_0\mathbf{C} \\ cov(\tilde{x}_i)=\boldsymbol{B}^2+\boldsymbol{C\Sigma}_0\mathbf{C}^{\mathrm{T}} \end{cases}$$

则$\tilde{\boldsymbol{X}}=\boldsymbol{XC}^{\mathrm{T}}+\mathbf{Z}\boldsymbol{B}^2$, $\boldsymbol{Z}\overset{\text{i.i.d.}}{\sim}N(0,1)$. s 取$\boldsymbol{\Omega}^{-1}-2^{-1}\mathrm{diag}\{s\}$为正定的尽可能大的值.

实际应用中, 一种更符合实际的情况是网络连接数据的协方差未知, 需要提前估计数据的协方差$\boldsymbol{\Omega}^{-1}$. 本文方法对于估计的协方差矩阵同样适用.

求解式 (1) 中$\boldsymbol{\beta}$为:

$$\hat{\boldsymbol{\beta}}(\lambda)=\mathrm{argmin}\{\|(\boldsymbol{Y}-\boldsymbol{\alpha})-[\boldsymbol{X}\tilde{\boldsymbol{X}}]^{\mathrm{T}}\hat{\boldsymbol{b}}\|_2^2+\mu\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{L\alpha}+\lambda\|\hat{\boldsymbol{b}}\|_1\} \tag{4}$$

其中, μ和λ为惩罚系数, $\|\cdot\|_2$为L_2范数, $\|\cdot\|_1$为L_1 范数. $\hat{\boldsymbol{b}}=\left[\hat{\boldsymbol{\beta}}^{\mathrm{T}}\ \hat{\tilde{\boldsymbol{\beta}}}^{\mathrm{T}}\right]^{\mathrm{T}}\in R^{2p}$, $\hat{\boldsymbol{\beta}}^{\mathrm{T}}$和$\hat{\tilde{\boldsymbol{\beta}}}^{\mathrm{T}}$分别对应$\boldsymbol{X}$和$\tilde{\boldsymbol{X}}$的回归系数. 基于图拉普拉斯矩阵$\boldsymbol{L}$的基本性质, 此时, 式 (4) 可写为:

$$\hat{\boldsymbol{\beta}}(\lambda)=\operatorname{argmin}\{\|(\boldsymbol{Y}-\boldsymbol{\alpha})-[\boldsymbol{X}\ \tilde{\boldsymbol{X}}]^{\mathrm{T}}\hat{\boldsymbol{b}}\|_2^2+\mu\sum_{(u,v)\in E}(\alpha_u-\alpha_v)^2+\lambda\|\hat{\boldsymbol{b}}\|_1\} \tag{5}$$

变量选择的可靠性由错误发现率和统计功效 (*Power*) 二者度量, 计算方法如下:

$$FDR:=E\left[\frac{\#\left\{j:\beta_j=0 \text{ and } j\in\hat{S}\right\}}{\#\{j:j\in\hat{S}\}\vee 1}\right] \tag{6}$$

$$Power:=E\left[\frac{\#\left\{j:\beta_j\neq 0 \text{ and } j\in\hat{S}\right\}}{\#\left\{j:\beta_j\neq 0\right\}}\right] \tag{7}$$

其中, "#{*j*:条件}"表示满足条件的 j 的数量, $\hat{S}\subset\{1,\cdots,p\}$ 为被选出特征构成的集合, $a\vee b=\max\{a,b\}$. 由式 (6) 可知, *FDR* 定义为所有被选出特征中错选特征所占比例的期望. 由式 (7) 可知, *Power* 定义为所有应选特征中, 被模型正确选出的比例的期望. 本文方法的目标是充分利用网络内聚性现象, 发现并选择与响应变量真实相关的特征, 在正选比例 *Power* 足够高的前提下, 将错选比例 *FDR* 控制在目标水平.

1.3 构造 NLKF 统计量

构造一个服从充分性和反对称性质的一般统计量 W_j. 对于任意 $S\subseteq\{1,\cdots,p\}$, 构造统计量 W_j 满足:

$$W_j([\boldsymbol{X}\ \tilde{\boldsymbol{X}}]_{swap(S)},\boldsymbol{y})=W_j([\boldsymbol{X}\ \tilde{\boldsymbol{X}}],\boldsymbol{y})\cdot\begin{cases}+1, & j\notin S\\ -1, & j\in S\end{cases} \tag{8}$$

其中, $[\boldsymbol{X}\ \tilde{\boldsymbol{X}}]_{swap(S)}$ 表示交换矩阵 $[\boldsymbol{X}\ \tilde{\boldsymbol{X}}]$ 的 $\boldsymbol{X}_j, j\in S$ 列和 $\tilde{\boldsymbol{X}}_j, j\in S$ 列.

式 (1) 的回归系数 $\hat{\boldsymbol{b}}=[\hat{\boldsymbol{\beta}}^{\mathrm{T}}\ \hat{\tilde{\boldsymbol{\beta}}}^{\mathrm{T}}]^{\mathrm{T}}\in R^{2p}$, 可构造满足上述条件的网络连接数据回归模型的NLKF方法统计量为:

$$W_j=|\hat{\boldsymbol{b}}_j|-|\hat{\boldsymbol{b}}_{j+p}|,\ j\in\{1,\cdots,p\} \tag{9}$$

1.4 计算统计量的阈值

将上述统计量 $W_j,\ j\in[1,p]$ 按照 $|W_j|$ 从大到小依次排列, 寻找能够识别相关特征的特征统计量的最小值. 预先指定错误发现率 q, 定义 NLKF 方法的阈值 T 如下:

$$T=\min\left\{t\in\mathcal{W}:\frac{\#\{j:W_j\leqslant -t\}}{\#\{j:W_j\geqslant t\}\vee 1}\leqslant q\right\} \tag{10}$$

其中, $\mathcal{W}=\{|W_j|:j=1,\cdots,p\}$. 因统计量 W_j 具有反对称性, 式 (10) 将 T 定义为使错误发现率控制在指定值 q 以内的最小统计量的值. 当上述集合是空集时, 取 $T=+\infty$.

为使 *FDR* 被更好地控制, 定义一个修正的阈值 T_+:

$$T_+=\min\left\{t\in\mathcal{W}:\frac{1+\#\{j:W_j\leqslant -t\}}{\#\{j:W_j\geqslant t\}\vee 1}\leqslant q\right\} \tag{11}$$

当上述集合是空集时, 取 $T_+=+\infty$. 将使用阈值 T_+ 进行特征选择的方法为NLKF$_+$ 方法.

由式 (10) 和式 (11) 知, NLKF$_+$方法比NLKF方法选出的集合更保守, 因为NLKF$_+$方法的边界值 T_+ 总是高于NLKF方法的边界值 T.

NLKF方法可选出特征集合:

$$\hat{S}=\{j:W_j\geqslant T\}$$

NLKF$_+$方法可选出特征集合:

$$\hat{S}_+=\{j:W_j\geqslant T_+\}$$

设置目标 *FDR* 水平为 q, NLKF (NLKF$_+$) 方法通过模仿原始特征的内部依赖构造与目标变量无关的 Network-Knockoff 特征. 使用式 (2) 中提出的网络内聚性惩罚函数, 求解网络内聚性回归模型式 (1), 得到模型参数的估计. 借助参数估计值依次计算 NLKF 统计量 W 及其对应的阈值 T (T_+), 进行网络连接数据的变量选择. 下文从理论和实验两个方面, 证明NLKF方法能够将网络连接数据变量选择方法的 *FDR* 控制在预设水平 q.

2 理论性质

本节从理论上证明 NLKF 方法对变量选择 *FDR* 的控制. 这一证明过程说明了本文方法的统计可行性, 为网络数据特征筛选的可信度提供保障.

2.1 错误发现率控制在目标水平

首先, 介绍两个关键引理.

引理 1. 对于任意集合 $S\subset\{1,\cdots,p\}$:

$$[\boldsymbol{X}\ \tilde{\boldsymbol{X}}]_{swap(S)}^{\mathrm{T}}\boldsymbol{A}[\boldsymbol{X}\ \tilde{\boldsymbol{X}}]_{swap(S)}=[\boldsymbol{X}\ \tilde{\boldsymbol{X}}]^{\mathrm{T}}\boldsymbol{A}[\boldsymbol{X}\ \tilde{\boldsymbol{X}}]$$

即为, 对于任意 $j\in S$ 交换 X_j 和 $\tilde{\boldsymbol{X}}$, Gram 矩阵 $\boldsymbol{G}=[\boldsymbol{X}\ \tilde{\boldsymbol{X}}]^{\mathrm{T}}\boldsymbol{A}[\boldsymbol{X}\ \tilde{\boldsymbol{X}}]$ 不变. 其中 $\boldsymbol{A}=\boldsymbol{I}_n-(\boldsymbol{I}_n+\lambda\boldsymbol{L})^{-1}$, $\boldsymbol{L}$ 为网络连接数据的拉普拉斯矩阵.

引理 2. 记 $W_j=|\hat{b}_j|-|\hat{b}_{j+p}|,\ j\in\{1,\cdots,p\}$ 为式 (9) 中定义的NLKF统计量, 在已知 $(|W_1|,\cdots,|W_p|)$ 时, 与目标变量无关的变量对应的 NLKF 统计量 $W_j,\ j\in S_0^c$ 的正负性服从标准二项分布. 其中 S_0^c 为无关变量下标集.

引理 2 的证明见附录 A.

基于上述引理, 可给出本文关键的错误发现率控制定理及其证明如下.

定理 1. *FDR* 控制定理. 对于统计量 $W_j, j\in[1,p]$, 给定 q 为 *FDR* 的目标水平. 网络连接数据的变量选择方

法的阈值$T=\min\left\{t\in\mathcal{W}:\frac{\#\{j:W_j\leqslant -t\}}{\#\{j:W_j\geqslant t\}\vee 1}\leqslant q\right\}$, NLKF 挑选出的特征集$\hat{S}=\{j:W_j\geqslant T\}$,可以将 *FDR* 控制在目标水平, 即:

$$E\left[\frac{\#\{j:\beta_j=0 \text{ and } j\in\hat{S}\}}{\#\{j:j\in\hat{S}\}+1/q}\right]<q$$

其中, $\mathcal{W}=\{|W_j|:j=1,\cdots,p\}$. 对于一个保守的阈值, $T_+=\min\left\{t\in\mathcal{W}:\frac{1+\#\{j:W_j\leqslant -t\}}{\#\{j:W_j\geqslant t\}\vee 1}\leqslant q\right\}$, NLKF$_+$ 挑选出的特征集$\hat{S}_+=\{j:W_j\geqslant T_+\}$, 可以将 *FDR* 控制在目标水平, 即:

$$E\left[\frac{\#\{j:\beta_j=0 \text{ and } j\in\hat{S}\}}{\#\{j:j\in\hat{S}\}\vee 1}\right]<q$$

其中, $\mathcal{W}=\{|W_j|:j=1,\cdots,p\}$.

证明: 由引理 2 可知, 统计量W_j的符号是独立同分布的. 计算统计量W_j之后, 必须选择一个数据依赖的边界值: $T=\min\left\{t\in\mathcal{W}:\frac{\#\{j:W_j\leqslant -t\}}{\#\{j:W_j\geqslant t\}\vee 1}\leqslant q\right\}$,不失一般性地假定 $|W_1|\geqslant|W_2|,\cdots,|W_{p-1}|\geqslant|W_p|$, 寻找一个使得 $FDR\leqslant q$的最小边界值T. 从统计量的最小值$t=|W_p|$开始测试, 然后逐步测试$t=|W_{p-1}|$, 当测试到一个t值满足$FDR(t)\leqslant q$时停止. 通过可选停时定理, 上鞅在随机时间$t=T$的期望值受其在时间$t=0$的期望值限制: 令p_0为空特征的数量并写出 $Y=\#\{\beta_j=0 \text{ and } W_j\leqslant 0\}$, 其中"#"表示计数, 有:

$$\begin{aligned}&E\left[\frac{\#\{j:\beta_j=0 \text{ and } W_j\leqslant -T\}}{1+\#\{j:\beta_j=0 \textit{ and } W_j\geqslant T\}}\right]\\&\leqslant E\left[\frac{\#\{j:\beta_j=0 \text{ and } W_j\leqslant 0\}}{1+\#\{j:\beta_j=0 \text{ and } W_j\geqslant 0\}}\right]\\&=E\left[\frac{Y}{1+p_0-Y}\right]\leqslant 1\end{aligned}$$

由于$sign(W_j)\overset{\text{i.i.d.}}{\sim}\{\pm 1\}$中$j$为空特征, 因此 $Y=\#\{j:\beta_j=0 \text{ and } W_j\leqslant 0\}$为服从二项分布$(p_0,\frac{1}{2})$的随机变量.

$$\begin{aligned}FDR&=\mathbb{E}\left[\frac{\#\{j:\beta_j=0 \text{ and } W_j\geqslant T\}}{\#\{j:W_j\geqslant T\}\vee 1}\right]\\&=E\left[\frac{\#\{j:\beta_j=0 \text{ and } W_j\geqslant T\}}{1+\#j:\beta_j=0 \text{ and } W_j\leqslant -T}\cdot\frac{1+\#\{j:W_j\leqslant T\}}{\#\{j:W_j\geqslant T\}\vee 1}\right]\\&\leqslant E\left[\frac{\#\{j:\beta_j=0 \text{ and } W_j\geqslant T\}}{1+\#j:\beta_j=0 \text{ and } W_j\leqslant -T}\cdot q\right]\leqslant q\end{aligned}$$

定理 1 说明了本文方法能确保所选特征集合的 *FDR* 维持在预定的目标之内, 为变量选择的可靠性和网络数据分析的可重复性提供了保证. 由第 1 节可知, NLKF (NLKF$_+$) 方法中统计量的计算涉及原始变量回归系数与 Network-Knockoff 变量的回归系数共计$2\times p$个参数, 网络连接数据各节点独立效应的引入又为模型新增了 n 个未知参数. 为说明模型式 (1) 中解的存在性, 引入如下定理.

定理 2. 解的存在性. 对于网络内聚性回归模型, 见式 (1), 网络一致性惩罚的回归系数可求解为:

$$\hat{\boldsymbol{b}}=(\hat{\boldsymbol{\beta}}^{\mathrm{T}},\hat{\tilde{\boldsymbol{\beta}}}^{\mathrm{T}})^{\mathrm{T}}=[\boldsymbol{X},\tilde{\boldsymbol{X}}]^{\mathrm{T}}\boldsymbol{A}[\boldsymbol{X},\tilde{\boldsymbol{X}}]^{-1}[\boldsymbol{X},\tilde{\boldsymbol{X}}]^{\mathrm{T}}\boldsymbol{A}\boldsymbol{Y}$$

其中, $\boldsymbol{A}=\boldsymbol{I}_n-(\boldsymbol{I}_n+\lambda\boldsymbol{L})^{-1}$, 是一个对称正定阵.

证明: 求解模型式 (1), 最小化下方损失函数: $Loss(\boldsymbol{\alpha},\boldsymbol{\beta})=\|\boldsymbol{Y}-[\boldsymbol{X}\tilde{\boldsymbol{X}}]\boldsymbol{\beta}-\boldsymbol{\alpha}\|^2+\sum_{(u,v)\in E}(\alpha_u-\alpha_v)^2+\lambda\|\boldsymbol{\beta}\|_1$, 可得: $\hat{\boldsymbol{\theta}}=(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}})=(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}+\lambda\boldsymbol{M})^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Y}$, 其中, $\boldsymbol{Z}=(\boldsymbol{I}_n,\boldsymbol{X},\tilde{\boldsymbol{X}})$, $\boldsymbol{M}=\begin{bmatrix}\boldsymbol{L}&\boldsymbol{0}_{n\times 2p}\\\boldsymbol{0}_{2p\times n}&\boldsymbol{0}_{2p\times 2p}\end{bmatrix}$.

为保证解一定存在, 使用$\boldsymbol{L}+\gamma\boldsymbol{I}$代替$\boldsymbol{L}$, γ是一个很小的正的常数.

3 算法

在实际应用中, 我们经常遇到具有内部依赖性的网络数据, 这类数据的复杂性使得传统的变量选择方法难以有效处理. 为应对这一挑战, 本文第 1 节提出了一种新的方法, 即网络连接数据变量选择方法 (NLKF), 下文将详细阐述这一方法的具体实现流程.

为网络连接数据变量选择的NLKF (NLKF$_+$) 方法设计算法如算法 1.

算法 1. NLKF 算法

输入: $(\boldsymbol{X},\boldsymbol{y})$, $\boldsymbol{L}$, λ, q_{FDR}.

输出: $\hat{S}\subset\{1,\cdots,p\}$.

1) 给定协方差矩阵$\boldsymbol{\Sigma}$求精度矩阵$\boldsymbol{\Omega}=\boldsymbol{\Sigma}^{-1}$. 当协方差矩阵未知时, 用 GLasso 或 iSEE 估计协方差阵, 求$\hat{\boldsymbol{\Omega}}$.

2) 根据定义 1, 构造 Network-Knockoff 特征$\tilde{\boldsymbol{X}}$:

$$\begin{cases}\boldsymbol{C}=\boldsymbol{I}_p-\mathrm{diag}\{s\}\hat{\boldsymbol{\Omega}},\ \boldsymbol{B}=(2\mathrm{diag}\{s\}-\mathrm{diag}\{s\}\hat{\boldsymbol{\Omega}}\mathrm{diag}\{s\})^{1/2}\\\tilde{x}_i|x_i\sim N(\boldsymbol{C}x_i,\boldsymbol{B}^2),\ \tilde{\boldsymbol{X}}=(\tilde{x}_1,\cdots,\tilde{x}_n)\in R^{n\times p}\end{cases}$$

3) 引入网络内聚性惩罚函数, 即式 (2):

$$Loss(\boldsymbol{\alpha},\boldsymbol{\beta})=\|\boldsymbol{Y}-[\boldsymbol{X}\tilde{\boldsymbol{X}}]\boldsymbol{b}-\boldsymbol{\alpha}\|^2+\mu\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{\alpha}+\lambda\|\boldsymbol{b}\|_1$$

求解模型$\boldsymbol{Y}=\boldsymbol{\alpha}+\boldsymbol{X}\boldsymbol{\beta}+\tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}+\boldsymbol{\varepsilon}$得:

$$\hat{\boldsymbol{b}}=(\hat{\boldsymbol{\beta}}^{\mathrm{T}},\hat{\tilde{\boldsymbol{\beta}}}^{\mathrm{T}})^{\mathrm{T}}=[\boldsymbol{X},\tilde{\boldsymbol{X}}]^{\mathrm{T}}\boldsymbol{A}[\boldsymbol{X},\tilde{\boldsymbol{X}}]^{-1}[\boldsymbol{X},\tilde{\boldsymbol{X}}]^{\mathrm{T}}\boldsymbol{A}\boldsymbol{Y}$$

其中, $\boldsymbol{A}=\boldsymbol{I}_n-(\boldsymbol{I}_n+\lambda\boldsymbol{L})^{-1}$.

4) **for** j=1 **to** p **do**:

$$W_j=|\hat{\boldsymbol{b}}_j|-|\hat{\boldsymbol{b}}_{j+p}|$$

end

5) 求阈值T和T_+

6) **return** $\hat{S}=\{j:W_j\geqslant T\}$和$\hat{S}_+=\{j:W_j\geqslant T_+\}$.

当原始变量协方差结构已知时, 使用定义 1 构造原始特征的 Network-Knockoff 特征$\tilde{\boldsymbol{X}}$. 原始变量协方差未知时, 先估计原始变量的协方差, 再构造$\tilde{\boldsymbol{X}}$. 引入原始变量的 Network-Knockoff 特征后, 模型 (1) 可更直观地表达为$\boldsymbol{Y}=\boldsymbol{\alpha}+\boldsymbol{X\beta}+\tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}+\boldsymbol{\varepsilon}$, 其中$\boldsymbol{\alpha}$, $\boldsymbol{\beta}$和$\tilde{\boldsymbol{\beta}}$分别对应$n$, p, p个未知参数. 定理 2 说明了模型式 (1) 的解的存在性, 计算方法如算法 1 中的步骤 3). 基于模型求解结果, 可依次计算特征统计量W及阈值T (T_+), 进而挑选出目标变量的相关特征.

4 数值模拟

本节主要考察使用网络连接数据仿变量方法 (NLKF) 在模拟数据上的方法表现. 为了更直观地评估模型效果, 将本文方法与一般变量选择法 (network cohesion Lasso, NCL) 进行比较实验.

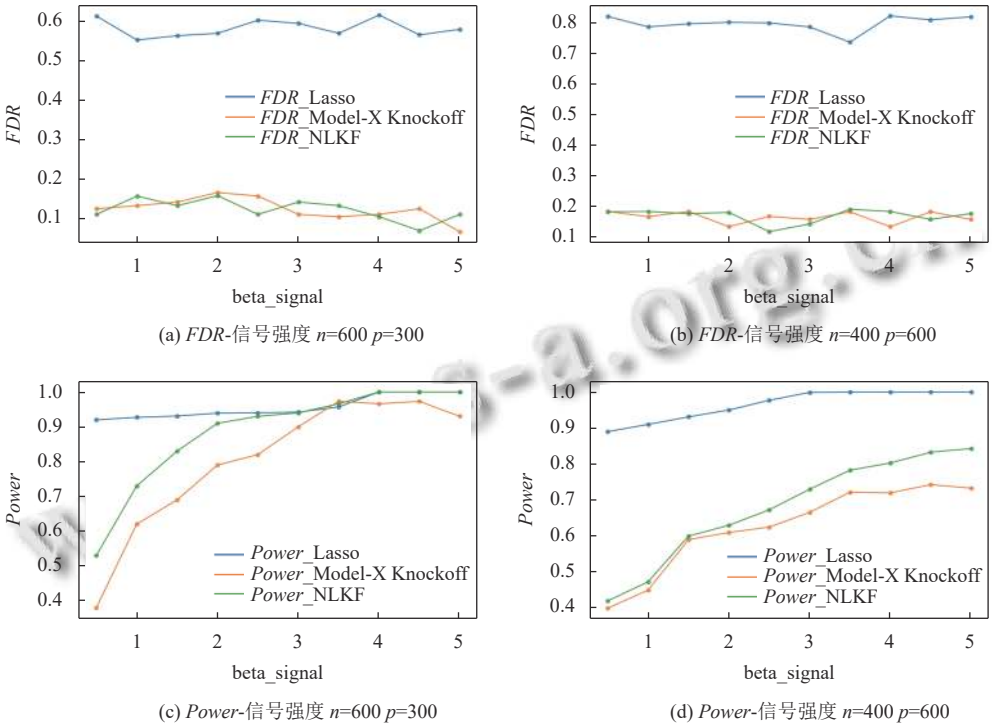
模拟数据的网络结构由随机图模型生成, 共有n个节点, p维属性, K个随机生成块. 各节点样本由参数为$(\pi_1,\cdots,\pi_K)$的多元正态分布生成, 并独立分配给K个图块. 各个图块的标签为c_i, $i=1,\cdots,n$. 图形的边E_{ij}, $1\leqslant i\leqslant j\leqslant n$, 为$P(E_{ij}=1)=B_{c_ic_j}$的 Bernoulli 随机变量. 设置$K=2$, $\pi_1=\pi_2=1/2$, 同一图块内部节点相互连接的概率$B_{kk}=0.8$, 不同图块之间节点相互连接的概率$B_{kl}=0.2$.

各个节点X_i独立取样自正态分布$N(0,\Sigma_X)$, 构造一个系数向量$\beta\in\mathbb{R}^p$, 其中随机选出 15 个的位置放置均值为β_{signal}的 15 个信号值, 其他$p-k$项的值为 0. 独立节点效应α_i独立取样自均值与该节点所在图块相关的正态分布$N(\eta_{c_i},s^2)$. 其中, η_{c_i}表示该节点所在图块c_i对应的均值, 取$\eta_{c_1}=-1$, $\eta_{c_2}=1$. s表示每个图块中个节点的内聚性强度, 取$s=1$.

设置目标错误发现率 (*FDR*) 水平为 0.2. 两类方法 NLKF 和 NCL 分别借助 R 程序包“knockoff”和 “glmnet” 实现. 用 10 折交叉验证的方法调优参数.

4.1 数值模拟 1: 信号强度比较实验

为观察信号强度的大小对实验结果的影响, 设置β_{signal}的均值从 0.5 逐渐增加至 5, 分别在高维和低维进行比较试验, 实验结果如图 1 所示.



(a) *FDR*-信号强度 n=600 p=300　(b) *FDR*-信号强度 n=400 p=600

(c) *Power*-信号强度 n=600 p=300　(d) *Power*-信号强度 n=400 p=600

图 1　不同信号强度下, NLKF、Model-X Knockoff 与 Lasso 变量选择效果图

实验结果可知, NLKF 方法明显优于经典的 Model-X Knockoff 和 Lasso 变量选择方法. 低维弱信号强度情况下, NLKF 方法的 *Power* 比 Model-X Knockoff 提高 0.1, 且随着信号强度的增强, 二者的 *Power* 有明显提升, 且相比于 Lasso 方法, NLKF 方法能够将 *FDR* 控制在目标水平. 高维弱信号强度情况下, NLKF 方法表现与经典的 Model-X Knockoff 方法差别不大, 随着信号强度的增强, NLKF 方法的 *Power* 明显优于 Model-X

Knockoff 和 Lasso 变量选择方法.

4.2 数值模拟 2: NLKF、NLKF$_+$ 比较实验

设置信号强度均值为 3.5, $\Sigma_{X,ij}=0.5^{|i-j|}$, 数值模拟结果见表 1. 其中, $N_\epsilon(0,1)$表示误差项ϵ服从均值为 0, 方差为 1 的正态分布. 表 1 显示, 使用 NLKF 方法可以将 *FDR* 控制在目标水平 (20%) 以下, 并且 *Power* 值大于 NCL 方法. 相比于NLKF方法, 改进的NLKF$_+$对 *FDR* 的控制效果更优.

表 1 数值模拟结果 1 ($\Sigma_{X,ij}=0.5^{|i-j|}$)

参数	方法	*FDR*	*Power*
$n=400,\ p=150$ $N_\epsilon(0,1)$	NLKF	0.190	0.973
	NLKF$_+$	0.182	0.967
	NCL	0.802	1
$n=400,\ p=150$ $N_\epsilon(0,5)$	NLKF	0.183	0.978
	NLKF$_+$	0.174	0.967
	NCL	0.434	1
$n=400,\ p=600$ $N_\epsilon(0,1)$	NLKF	0.197	0.996
	NLKF$_+$	0.183	0.930
	NCL	0.567	1

4.3 数值模拟 3: 泛化能力测试

设置信号强度均值 3.5, $\Sigma_{X,ij}=0.5+I(i=j)\times 0.5$, 数值模拟结果见表 2. 表 2 显示, 在原始数据协方差为不同结构时, NLKF 方法仍能够在 *Power* 足够高的情况下, 将 *FDR* 控制在目标水平. 说明了本文方法具有一定的泛化能力.

表 2 数值模拟结果 2 ($\Sigma_{X,ij}=0.5+I(i=j)\times 0.5$)

参数	方法	*FDR*	*Power*
$n=400,\ p=150$ $N_\epsilon(0,1)$	NLKF	0.222	0.908
	NLKF$_+$	0.168	0.897
	NCL	0.593	1
$n=400,\ p=150$ $N_\epsilon(0,5)$	NLKF	0.221	0.900
	NLKF$_+$	0.171	0.893
	NCL	0.637	1
$n=400,\ p=600$ $N_\epsilon(0,5)$	NLKF	0.210	0.927
	NLKF$_+$	0.189	0.873
	NCL	0.634	1

3 个数值模拟分别从不同角度验证本文方法在数据计算上的可行性、实用性和可靠性. 从模拟结果来看, 本文提出的NLKF (NLKF$_+$) 方法能够在较高的统计功效 (*Power*) 下, 将错误发现率 (*FDR*) 控制在目标水平, 显著优于不使用仿变量的传统 Lasso 变量选择方法 (NCL).

5 实际数据分析

使用沪深两市股票的因子数据及股票所属申万一级行业构建的网络关系做实证分析. 因子是指在股票市场中, 能够解释股票收益率的一些特定因素. 常见的因子包括市场风险溢酬因子、市值因子、账面市值比因子等. 使用因子来构建投资组合的方法有很多种, 其中最常见的方法是基于因子收益率的方法. 具体来说, 首先根据投资需求制定投资目标. 然后从因子池中选择合适因子、计算单因子对个股收益率贡献、进而基于各个股票的因子表现来构建股票投资组合. 最后根据自己的需求和风险偏好, 适时调整每个因子在投资组合中的权重以适应市场变化. 其中, 能否从因子库中选出对股票收益率具有更好解释性的因子直接决定了后续投资的表现.

本文使用 2022 年 1–12 月沪深两市的 4 000 只 A 股 ($n=4000$) 的 200 个因子 ($p=200$) 作为实证数据集. 旨在从中选出对股票收益率有影响的重要因子. 计算过程中的协方差矩阵通过GLasso的方法估计, 参数λ用 5 折交叉验证调节. 因子频率有日度、周度和月度, 此处选取月底日度数值, 并将空缺值用当月日均值填补. 根据申万二级行业构建股票标的之间的网络关系. 设置目标错误发现率水平$q=0.2$. 本文参数λ和μ使用交叉验证使均方误差 (MSE) 最小的方法调整. 最终特征选择结果记录见表 3.

对相同数据集使用 Lasso 变量选择方法可选出 54 个对当前收益率具有较强解释性的因子, 详情见附录 B.

对比 Lasso 变量选择方法选出的 54 个特征, NLKF (NLKF$_+$) 方法选出的特征数量有明显降低. 参考第 4 节数值模拟可知, Lasso 变量选择方法具有较高的统计功效, 即 Lasso 方法能够有效地挑出与目标变量相关的特征. 但 Lasso 方法在选出相关特征的同时也选出较多无关特征, 表现为较高的错误发现率. 实证结果表明, 与传统的变量选择方法 Lasso 相比, 本文提出的NLKF和NLKF$_+$方法在高统计功效的前提下, 错误发现率明显降低.

2022 年 12 月, 食品饮料、酒店航空等消费行业呈现了较为明显的复苏态势. 从分析结果中可以看到, 短期反转、季节反转因子、波动性因子等都是影响 A 股股票收益率的关键因素. 并且, 公司本身的经营状况和资产质地也是影响公司价值的重要指标, 因此流动资产比总资产、资产周转率等指标也被视为重要变量被挑选出来.

表 3 实证结果记录表

NLKF			$NLKF_+$	
1.流动资产/总资产	13.基于波段切割的动量	25.不可任意支配总应计费用	1.流动资产/总资产	13. 基于波段切割的动量
2.资产周转率	14.超买/卖因子(DDI)	26.调整后每股收益	2.资产周转率	14.超买/卖因子(DDI)
3.营业费用率	15.超买/卖因子(DBCD)	27.每股营业收入	3.营业费用率	15.超买/卖因子(DBCD)
4.股票周转率	16.超买/卖因子(DDI)	28.经营性现金流量率	4.股票周转率	16.超买/卖因子(DDI)
5.贸易周转率	17.趋势因子(VHF)	29.动量排序	5.贸易周转率	17.趋势因子(VHF)
6. 偏移量	18.波动因子(CVI)	30.股权集中度	6. 偏移量	18.波动因子(CVI)
7.前K月动量	19.总量因子(CHO)	31.弗拉齐尼·佩德森贝塔	7.前K月动量	19.总量因子(CHO)
8.短期反转	20.波动因子(RI)		8.短期反转	20.波动因子(RI)
9.季节反转	21.销售额比价值比率		9.季节反转	21.销售额比价值比率
10.时间序列动量	22.波动性因子		10.时间序列动量	22.波动性因子
11.总量因子(VO)	23.市场beta		11.总量因子(VO)	23.市场beta
12.反转因子(PSY)	24.系统偏度		12.反转因子(PSY)	24.系统偏度

6 结论与展望

2023 年 2 月, 《数字中国建设整体布局规划》明确提出, 到 2025 年基本形成横向打通、纵向贯通、协调有力的一体化格局, 数字中国建设取得重要进展. 到 2035 年, 数字化发展水平进入世界前列, 数字中国建设取得重大成就, 同时要把中国数字化建设方面的技术和资源大量输出到国外, 为推动全球数字化经济发展做出重要贡献. “数据”作为新时代的“关键能源”, 将成为新时代的重要生产力和发展引擎.

本文将网络连接数据的融合惩罚和 Knockoff 思想结合, 提出了针对网络结构数据变量选择的NLKF和$NLKF_+$方法. 从理论水平证明了这个方法对 *FDR* 的控制能力, 并从模拟和实证两个方面验证了模型的有效性. NLKF ($NLKF_+$) 方法适用于协方差矩阵未知的场景, 在处理复杂、高维且具有关联结构的数据方面表现出色, 能够满足当前网络连接数据分析的需求. 从 *FDR* 角度出发的数据分析, 在高维的、相互关联的数据建模上也将有更多、更广泛的应用.

参考文献

1 IDC. Data age 2025: The evolution of data to life-critical don’t focus on big data; focus on the data that’s big. https://www.seagate.com/www-content/our-story/trends/files/Seagate-WP-DataAge2025-March-2017.pdf. [2023-04-20].

2 Sengupta S. Statistical network analysis: Past, present, and future. arXiv:2311.00122, 2023.

3 Saqr M, Alamro A. The role of social network analysis as a learning analytics tool in online problem based learning. BMC Medical Education, 2019, 19(1): 160. [doi: 10.1186/s12909-019-1599-6]

4 Barabási AL, Oltvai ZN. Network biology: Understanding the cell’s functional organization. Nature Reviews Genetics, 2004, 5(2): 101–113. [doi: 10.1038/nrg1272]

5 Allen F, Babus A. Networks in finance. SSRN Electronic Journal, 2008, 6(1): 383–419. [doi: 10.2139/ssrn.1094883]

6 Girvan M, Newman MEJ. Community structure in social and biological networks. Proceedings of the National Academy of Sciences of the United States of America, 2002, 99(12): 7821–7826. [doi: 10.1073/pnas.122653799]

7 Bramoullé Y, Djebbari H, Fortin B. Identification of peer effects through social networks. Journal of Econometrics, 2009, 150(1): 41–55. [doi: 10.1016/j.jeconom.2008.12.021]

8 Lee LF. Identification and estimation of econometric models with group interactions, contextual factors and fixed effects. Journal of Econometrics, 2007, 140(2): 333–374. [doi: 10.1016/j.jeconom.2006.07.001]

9 Stephanie RL, Friedman JH. Variable fusion: A new adaptive signal regression method. Technical Report, Pittsburgh: Department of Statistics, Carnegie Mellon University Pittsburgh, 1997.

10 Tibshirani R, Saunders M, Rosset S, *et al*. Sparsity and smoothness via the fused Lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology, 2005, 67(1): 91–108. [doi: 10.1111/j.1467-9868.2005.00490.x]

11 Li TX, Levina E, Zhu J. Prediction models for network-linked data. The Annals of Applied Statistics, 2019, 13(1): 132–164. [doi: 10.1214/18-AOAS1205]

12 Guyon I, Elisseeff A. An introduction to variable and feature selection. The Journal of Machine Learning Research, 2003, 3: 1157–1182.

13 Brown PF, Cocke J, Della Pietra SA, *et al*. A statistical approach to machine translation. Computational Linguistics, 1990, 16(2): 79–85.

14 LeCun Y, Bottou L, Bengio Y, *et al*. Gradient-based learning

applied to document recognition. Proceedings of the IEEE, 1998, 86(11): 2278–2324. [doi: 10.1109/5.726791]

15 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778.

16 Cheng HT, Koc L, Harmsen J, *et al*. Wide & deep learning for recommender systems. Proceedings of the 1st Workshop on Deep Learning for Recommender Systems. Boston: ACM, 2016. 7–10. [doi: 10.1145/2988450.2988454]

17 Wang MC, Wright J, Brownlee A, *et al*. A comparison of approaches to stepwise regression on variables sensitivities in building simulation and analysis. Energy and Buildings, 2016, 127: 313–326. [doi: 10.1016/j.enbuild.2016.05.065]

18 Tibshirani R. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological), 1996, 58(1): 267–288. [doi: 10.1111/j.2517-6161.1996.tb02080.x]

19 Hoerl AE, Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 1970, 12(1): 55–67. [doi: 10.1080/00401706.1970.10488634]

20 Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B: Statistical Methodology, 2005, 67(2): 301–320. [doi: 10.1111/j.1467-9868.2005.00503.x]

21 Ziegel ER. The elements of statistical learning. Technometrics, 2003, 45(3): 267–268. [doi: 10.1198/tech.2003.s770]

22 Deng L, Yu D. Deep learning: Methods and applications. Foundations and Trends® in Signal Processing, 2014, 7(3–4): 197–387. [doi: 10.1561/2000000039]

23 Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological), 1995, 57(1): 289–300. [doi: 10.1111/j.2517-6161.1995.tb02031.x]

24 Barber RF, Candès EJ. Controlling the false discovery rate via Knockoffs. The Annals of Statistics, 2015, 43(5): 2055–2085. [doi: 10.1214/15-AOS1337]

25 Candès E, Fan YY, Janson L, *et al*. Panning for gold: 'Model-X' Knockoffs for high dimensional controlled variable selection. Journal of the Royal Statistical Society Series B: Statistical Methodology, 2018, 80(3): 551–577. [doi: 10.1111/rssb.12265]

26 Fan YY, Demirkaya E, Li GR, *et al*. RANK: Large-scale inference with graphical nonlinear Knockoffs. Journal of the American Statistical Association, 2020, 115(529): 362–379. [doi: 10.1080/01621459.2018.1546589]

附录 A. 引理 2 的证明

对任意$S \subset S_0^c$是无关变量的子集, 由于:

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$$

可知:

$$[\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{swap(S)}^{\mathrm{T}}\boldsymbol{Ay} \sim \mathcal{N}([\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{swap(S)}^{\mathrm{T}}\boldsymbol{AX\beta}, \sigma^2[\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{swap(S)}^{\mathrm{T}}\boldsymbol{A}[\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{swap(S)})$$

即得:

$$[\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{swap(S)}^{\mathrm{T}}\boldsymbol{Ay} \overset{d}{=} [\boldsymbol{X}, \tilde{\boldsymbol{X}}]^{\mathrm{T}}\boldsymbol{Ay}$$

$$W_{swap(S)} = f([\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{swap(S)}^{\mathrm{T}}\boldsymbol{A}[\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{swap(S)}, [\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{swap(S)}^{\mathrm{T}}\boldsymbol{Ay}) \overset{d}{=} f([\boldsymbol{X}, \tilde{\boldsymbol{X}}]^{\mathrm{T}}\boldsymbol{A}[\boldsymbol{X}, \tilde{\boldsymbol{X}}], [\boldsymbol{X}, \tilde{\boldsymbol{X}}]^{\mathrm{T}}\boldsymbol{Ay}) = W$$

引理得证.

附录 B. 实证分析由 Lasso 方法选出的 54 个因子

1. 流动资产比总资产, 2. 流动负债比总负债, 3. 长期负债比率, 4. 不可任意支配总应计费用, 5. 财务质量因子, 6. 应收账款周转天数, 7. 现金周转周期, 8. 每股经营现金, 9. 资产经营现金流, 10. 营业费用率, 11. 股本回报率, 12. 总营业费用, 13. 经营性现金流量率, 14. 研发费用比市场权益, 15. 研发增长年份, 16. 股票发行量, 17. 换手率调整后成交量, 18. 条件流行性冲击, 19. 基于波段切割的动量, 20. 前 K 月动量, 21. 收入惊喜, 22. 短期反转, 23. 历史价格变动速度, 24. 价差, 25. 标准化意外收益, 26. 剩余资本利得, 27. 季节性逆转, 28. 季节性, 29. 时间序列动量, 30. 趋势动量, 31. 顶级股东分散率, 32. 股权集中度, 33. 技术指标动量/反转因子 (PSY), 34. 技术指标动量/反转因子 (EMV), 35. 偏移量 (BIAS), 36. 技术指标超买/超卖因子 (DBCD), 37. 技术指标 (DDI), 38. 技术因子超买/卖因子 (ROC), 39. 技术指标趋势因子 (VHF), 40. 技术指标波动因子 (CV), 41. 技术指标波动因子 (CVI), 42. 技术指标波动因子 (RI), 43. 技术指标总量因子 (CHO), 44. 技术指标总量因子 (KVO), 45. 技术指标总量因子 (VO), 46. 账面市值比, 47. EBITDA 比企业价值, 48. 销售额比企业价值, 49. 销售额比价格, 50. 有形资产变化, 51. 市场 beta, 52. 系统偏度, 53. 弗拉齐尼·佩德森贝塔, 54. 尾部风险.

(校对责编: 孙君艳)