主因子逼近方法在变量选择中的应用

许　健, 崔靓然, 李雅芝, 张祎璠
(湖南农业大学 信息科学技术学院, 湖南 长沙 410128)

摘　要: 当数据中变量个数远大于样本个数时, 变量之间的共线性问题变得尤其突出. 偏最小二乘方法作为一种潜变量方法, 将原始变量通过线性组合的方式转化为几个新的潜变量用于对响应变量的建模解释, 但变量之间复杂共线性的存在使得变量选择困难重重. 本文采用主因子近似方法分离出原始变量之间的共线性信息, 再进行变量选择. 模拟研究表明主因子逼近方法能有效地提高变量选择的精度.

关键词: 变量选择; 主因子近似; 偏最小二乘; 变量共线性
中图分类号: O213.9　　**文献标识码**: A　　**文章编号**: 1672-5298(2019)01-0008-05

Application of Principal Factor Approximation Method in Variable Selection

XU Jian, CUI Liangran, LI Yazhi, ZHANG Yifan
(School of Information Science and Technology, Hunan Agriculture University, Changsha 410128, China)

Abstract: The problem of variable collinearity between variable becomes particularly acute when variables are far more than samples in data. As a method of latent variables, partial least squares transform original variables into a few new factors by collinear combination, which can interpret response variable modeling. But, the complex sample data correlation structure makes variable selection become a tough task. In this paper, we introduced a principal component approximation (PFA) method to directly eliminate the effect of sample correlation on the observed values of the regression coefficients. Simulation studies were performed under three typical sample data correlation structures and the results showed that PFA and PLS performs comparably well.

Key words: variable selection; principal factor approximation; partial least squares; variable collinearity

技术手段的进步使得数据收集的难度大大降低, 变量个数 p 远大于样本个数 n 的数据(大 p 小 n 数据)的建模分析也越发常见. 多元分析模型如偏最小二乘回归、主成分回归等常用于这类数据的分析. 为了提升模型精度同时增加模型的解释性, 变量选择常常用于降低模型的复杂度. 大 p 小 n 情形下变量选择的一个难题是解释变量与响应变量之间的虚假相关性(spurious correlation)可能会非常高[1], 一个不重要的变量可能会因为与某个重要变量之间的高度相关性而显得与响应变量也高度相关[2]. 就模型预报而言, 这些变量的存在并不会总是有损于模型性能, Trygg 等人将这些变量所携带的信息称为正交信息(orthogonal variation)[3]. 现代仪器设备收集的数据, 比如光谱色谱等, 变量之间往往显示出很强的相关性, 这使得虚假相关性问题尤其突出.

容许变量之间存在强相关性是偏最小二乘回归方法流行的一个原因. 本质上, 偏最小二乘是基于潜变量的方法, 它使用几个因子来解释变量空间中的相关性结构. 许多变量选择方法是基于偏最小二乘的, 比如 VIP 方法(Variable Importance for the Projection)[4], UVE-PLS 方法(Unimportant Variable Elimination by PLS)[5, 6], BVS-PLS 方法(Backward Variable Selection method for PLS)[7]等. 尽管这些方法在实际数据的处理中表现出有效性, 但至少有对两个问题尚不清楚: (1)变量之间的相关性对变量选择到底存在什么影响; (2)偏最小二乘方法是如何应对共线性问题的.

从根本上来说, 大多数变量选择方法都是基于响应变量与各个解释变量之间的“回归系数”估计值的大小来决定变量的去留的, 比如模型:

收稿日期: 2018-11-28
基金项目: 湖南农业大学青年自然科学基金(16QN11)
作者简介: 许　健(1983−), 男, 湖南张家界人, 讲师. 主要研究方向: 应用统计、高维数据分析

$$y = x_1\beta_1 + x_2\beta_2 + \cdots + x_p\beta_p + \varepsilon\ . \tag{1}$$

其中$|\beta_j|$的大小反映了变量x_j的重要性. 模型(1)可以由不同的方法得到, 比如偏最小二乘回归法. 简单起见, 我们采用 Person 相关系数来计算β_j的估计值:

$$\hat{\beta}_j = \frac{\widehat{\mathrm{cov}}(x_j, y)}{\hat{\sigma}_{x_j}\hat{\sigma}_y}\ .$$

实际上, 采用 Person 相关系数的文献并不少见[8, 9], 并且有人证明了这种简单方法也不乏一些优良性质. 然而, 变量之间的相关性会使得问题变得十分复杂. 由于变量之间存在相关性, 回归系数$(\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_p)$之间也由此产生相关性. 对于一个不重要的变量$x_j$, 观测到的回归系数$\hat{\beta}_j$可能并不接近于 0, 因为与$\hat{\beta}_j$相关的某些$\hat{\beta}_i$的取值较大, 类似地, 对于重要的变量, 观测到的回归系数也可能非常接近于 0. 因此, 在进行变量选择时最好要考虑变量之间的相关性结构. Leek 等人针对大规模假设检验问题提出了一个处理变量相关性的一般框架[10]. 他们指出相关性问题源于变量之间所包含的共有信息, 可以通过所谓的相依核(dependence kernel)来捕捉在高维数据中观测到的相关性结构信息, 并将此部分信息在后续的建模过程中剥离出来, 从而将一个相关性问题转化为一个独立性问题. 依照这一框架, 许多新方法被陆续提了出来, 其中范剑青等人提出了一种所谓的主因子近似方法(principal factor approximation, PFA)[11]来控制大规模假设检验问题中的错误发现率.

本文以 Person 相关系数为例来说明相关性结构是怎样影响变量选择的, 进一步采用 PFA 方法分离变量之间的相关性信息来改善变量选择的效果. 同时模拟了在三种典型的变量相关性结构下 PFA 方法的改进效果.

1 理论与方法

理论上, PFA 方法利用因子模型将相依的正态随机向量分解为因子与弱相关的正态随机误差之和[11]. 我们用 PFA 来分离出回归系数估计量中的相关性信息. 为了便于探讨相关系数的理论性质, 首先引入响应变量对各个解释变量的边缘线性回归模型.

1.1 边缘线性回归

假设对$(y, x_1, x_2, \cdots, x_p)$进行了$n$次观测, 观测数据保存在一个$n\times p$矩阵$\boldsymbol{X} = (x_{ij})$和一个$n\times 1$向量$\boldsymbol{y}$中.

考虑$\{y_i\}_{i=1}^n$对$\{x_{ij}\}_{i=1}^n$边缘线性回归:

$$(\alpha_j, \beta_j) = \arg\min_{a_j, b_j} \frac{1}{n}\sum_{i=1}^{n} E(y_i - a_j - b_j x_{ij})^2,\ j = 1, 2, \cdots, p. \tag{2}$$

如果$\beta_j \neq 0$, 则可以认为x_j与y之间存在线性关系, 且$|\beta_j|$越大, 则x_j对y的预报作用越重要. 由于$\{x_j\}_{j=1}^p$之间存在相关性, 故$\{\beta_j\}_{j=1}^p$的最小二乘估计量$\{\hat{\beta}_j\}_{j=1}^p$之间也具有相关性. 假设$r_{kj}$表示变量$x_k$与 x_j的相关系数, s_{jj}表示变量x_j的标准差, 给定$x_{i1}, x_{i2}, \cdots, x_{ip}$之后$y_i$的条件分布为 $N(\mu(x_{i1}, x_{i2}, \cdots, x_{ip}), \sigma^2)$, 则根据文[11], 给定$\boldsymbol{X} = (x_{ij})$之后$\{\hat{\beta}_j\}_{j=1}^p$的联合分布为

$$(\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_p) \sim N((\beta_1, \beta_2, \cdots, \beta_p)^T, \boldsymbol{\Sigma}),$$

其中矩阵$\boldsymbol{\Sigma}$的第 k 行第 j 列的元素$\boldsymbol{\Sigma}_{kj} = \dfrac{\sigma^2 r_{kj}}{n s_{kk} s_{jj}}$. 方差$\sigma^2$可以用 RCV 方法(refitted cross-validation)估计得出. 如果对数据进行标准化处理, 则边缘回归系数的最小二乘估计$\{\hat{\beta}_j\}_{j=1}^p$与 Pearson 相关系数$\{\dfrac{\widehat{\mathrm{cov}}(x_j, y)}{\hat{\sigma}_{x_j}\hat{\sigma}_y}\}_{j=1}^p$是相等的.

1.2 主因子近似法(PFA)

由于原始变量之间存在相关性, 因此 $(\hat{\beta}_1, \hat{\beta}_2,\cdots, \hat{\beta}_p)$ 的分量之间也彼此相关, 而 PFA 方法可对此进行校正. PFA 使用因子模型将 $(\hat{\beta}_1, \hat{\beta}_2,\cdots, \hat{\beta}_p)$ 之间的共有信息提取出来, 基于条件稀疏框架, 在共有信息剥离之后回归系数估计量之间应该是至多弱相关的. 因子载荷可通过主成分法计算. 首先, 对协方差矩阵 $\boldsymbol{\Sigma}$ 进行谱分解, 假设 $\boldsymbol{\Sigma}$ 的特征值从大到小依次为 $\lambda_1,\lambda_2,\cdots,\lambda_p$, 对应的标准正交特征向量为 $\boldsymbol{\xi}_1,\boldsymbol{\xi}_2,\cdots,\boldsymbol{\xi}_p$, 则

$$\boldsymbol{\Sigma}=\sum_{j=1}^{p}\lambda_j\boldsymbol{\xi}_j\boldsymbol{\xi}_j^T .$$

对整数 K, 令 $\boldsymbol{L}=(\sqrt{\lambda_1}\boldsymbol{\xi}_1,\sqrt{\lambda_2}\boldsymbol{\xi}_2,\cdots,\sqrt{\lambda_K}\boldsymbol{\xi}_K)$, $\boldsymbol{A}=\sum_{j=K+1}^{p}\lambda_j\boldsymbol{\xi}_j\boldsymbol{\xi}_j^T$, 则

$$\boldsymbol{\Sigma}=\boldsymbol{L}\boldsymbol{L}^T+\boldsymbol{A} .$$

$\hat{\beta}_j$ 可以写成

$$\hat{\beta}_j=\beta_j+\boldsymbol{b}_j^T\boldsymbol{W}+e_j, j=1,2,\cdots,p . \tag{3}$$

其中因子载荷 $\boldsymbol{b}_j=(b_{j1},b_{j2},\cdots,b_{jK})^T$ 是载荷矩阵 $\boldsymbol{L}$ 的第 j 行, 且

$$\boldsymbol{W}=(W_1,W_2,\cdots,W_K)^T\sim N_K(0,\boldsymbol{I}_K),\ (e_1,e_2,\cdots,e_p)^T\sim N(0,\boldsymbol{A}) .$$

此外, 假设 $W_1, W_2,\cdots,W_K$ 与 $e_1, e_2,\cdots,e_p$ 之间相互独立. 在式(3)中, $\{\beta_j=0\}$ 意味着相应的 $\{x_j\}_s$ 与响应变量 y 之间没有线性关系. 因子个数 K 的选择通常要让 $e_1, e_2,\cdots,e_p$ 之间至多弱相关. 实践中, 可以选择使得

$$\frac{\sqrt{\lambda_{k+1}^2+\cdots+\lambda_p^2}}{\lambda_1+\cdots+\lambda_p}<\varepsilon$$

成立的最小的 k, 其中 ε 是指定的一个较小的阈值, 比如 0.01.

正常情况下, 一般用 $\hat{\beta}_j=\beta_j+E_j$ 来估计 β_j, 其中 E_j 是获取 $\hat{\beta}_j$ 时的模型误差. 变量的相关性对 $\{\hat{\beta}_j\}_s$ 造成的干扰可以认为包含在 $\{E_j\}_s$ 中. 通过式(3), E_j 被分解为 $\boldsymbol{b}_j^T\boldsymbol{W}$ 与 e_j, 其中 $\boldsymbol{b}_j^T\boldsymbol{W}$ 解释了变量之间的相关性所造成的模型误差. 显然, 如果直接用 $\{\hat{\beta}_j\}_s$ 的大小来衡量 $\{x_j\}_s$ 的重要性, 准确性必然受到变量间相关性的影响. 根据式(3), $\hat{\beta}_j-\boldsymbol{b}_j^T\boldsymbol{W}\sim N(\beta_j,\frac{\sigma^2r_{jj}}{ns_{jj}^2}-\sum_{k=1}^{K}b_{jk}^2)$, 且

$$\hat{\beta}_j\sim N(\beta_j,\frac{\sigma^2r_{jj}}{ns_{jj}^2}),\ \frac{\sigma^2r_{jj}}{ns_{jj}^2}-\sum_{k=1}^{K}b_{jk}^2\le\frac{\sigma^2r_{jj}}{ns_{jj}^2} .$$

如果将误差项 E_j 中由变量共线性所导致的那一部分信息去除, 回归系数估计的信噪比就会得到提升. 因此, 如果能得到因子 $\boldsymbol{W}=(W_1,W_2,\cdots,W_k)^T$ 的真实值, 则可以将 $\mathrm{adj}(\hat{\beta}_j)=\hat{\beta}_j-\boldsymbol{b}_j^T\boldsymbol{W}$ 作为回归系数的估计量来衡量变量的重要性. 而因子 $\boldsymbol{W}$ 可以用 LASSO 方法估计:

$$\arg\min_{\boldsymbol{W},\beta}\sum_{j=1}^{p}(\hat{\beta}_j-\beta_j-\boldsymbol{b}_j^T\boldsymbol{W})^2+\sum_{j=1}^{p}p_\lambda(|\beta_j|) .$$

其中 $p_\lambda(\cdot)$ 是 L_1 惩罚函数.

2 模拟研究

2.1 模拟方法

假设有线性模型(1), 考虑一个高维稀疏问题, 即尽管变量个数 p 很大, 但真实与 y 有关的变量的个数 p_1 相对很小, 本模拟中假设系数不为 0 的变量的个数 $p_1=20$ 且相应的 $\beta_j=2$. 样本数据 $(x_1,x_2,\cdots,x_p)^T$

产生自多元正态分布 $N(\mu,\boldsymbol{\Sigma})$, 误差项 $\varepsilon\sim N(0,1)$. 考虑三种典型结构的协方差矩阵 $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma}^{(1)}=\boldsymbol{I}_{p\times p},\ \boldsymbol{\Sigma}_{ij}^{(2)}=\rho^{\theta|i-j|},\ \boldsymbol{\Sigma}_{ij}^{(3)}=\begin{cases}1, & i=j,\\ \rho, & i\neq j.\end{cases}$$

其中 $0<\rho<1,\ 0<\theta\leqslant 1,\ i,j=1,2,\cdots,p$.

$\boldsymbol{\Sigma}=\boldsymbol{\Sigma}^{(1)}$ 意味着变量间相互独立; $\boldsymbol{\Sigma}^{(2)}$ 是一个“带状”矩阵, 强调临近变量相关, 而相距较远的变量则不相关; $\boldsymbol{\Sigma}^{(3)}$ 则表示几乎所有变量都是相关的. 在设定好 $\boldsymbol{\Sigma}=\boldsymbol{\Sigma}^{(k)}$ $(k=1,2,3)$ 的取值后, 按照以下步骤进行模拟:

(1) 产生解释变量 X 数据. 从 p 维正态分布中生成 $n=100$ 个正态随机向量, 均值 $\mu=0_{p\times 1}$;

(2) 生成响应变量 y. 在系数向量 $(\beta_1,\beta_2,\cdots,\beta_p)$ 中随机选择 20 个分量, 设定其等于 2, 其余 $p-20$ 个系数全为 0. 根据模型 $y=x_1\beta_1+x_2\beta_2+\cdots+x_p\beta_p+\varepsilon$ 来生成响应变量;

(3) 模型校正.

2.2 模拟结果评价方法

本文根据回归系数绝对值的大小来衡量变量的重要性, 但不讨论阈值的选择问题, 因此采用 ROC 曲线(受试者工作特征曲线)来综合展示变量选择方法的性能.

2.3 模拟结果与讨论

2.3.1 $\boldsymbol{\Sigma}$ 为单位矩阵情况

当 $\boldsymbol{\Sigma}$ 为单位矩阵时, 变量之间相互独立, 此时可以预期 PFA 方法不能提升变量选择的精度. 为了使模拟结果更加可靠, 将 2.1 中的模拟实验重复 100 次, 在每一次模拟中, 针对设定好的一系列的阈值, 计算变量选择过程的敏感度(Sensitivity)与专一度(Specificity), 作出 ROC 曲线. 图 1 展示的是 100 次试验的平均 ROC 曲线. 显然, 三种方法的变量选择效果几乎相当, 其中 PLS 方法的表现要稍微好一点, 那是因为 PLS 是同时对所有变量计算出回归系数的, 更能兼顾变量间的相对重要性.

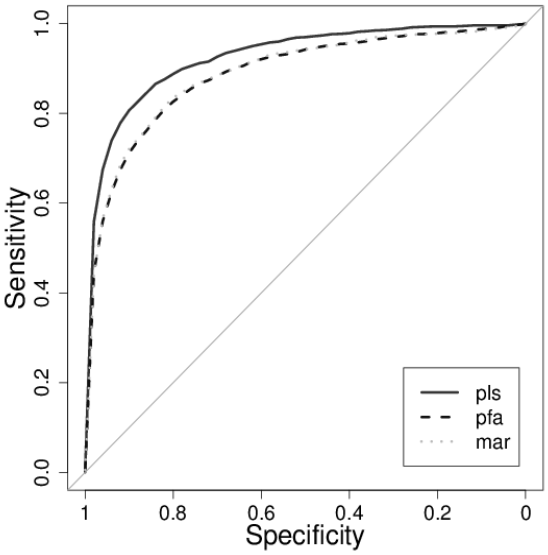


图 1　$\boldsymbol{\Sigma}=\boldsymbol{\Sigma}^{(1)}$ 时 100 次重复试验的平均 ROC 曲线

2.3.2 $\boldsymbol{\Sigma}$ 为带状矩阵情况

当 $\boldsymbol{\Sigma}$ 为带状矩阵时, 分两种情况来讨论变量选择的效果: (1)真实的重要变量聚集为几个连续片段; (2)真实的重要变量完全随机地散落在 p 个变量位置上. 图 2 中展示的是当真实的重要变量聚集为 4 个连续片段时, 分别用三种方法计算出的回归系数的绝对值. 重要变量的回归系数在图中刚好位于“山峰”位置, 同时与重要变量临近的变量其回归系数的绝对值也被“拉大”, 这种现象对于边缘线性回归方法(mar.reg)尤其明显, 而 PFA 以及 PLS 方法则对该现象有明显的改善. 当真实的重要变量的位置随机分布分布时, 则三种方法的表现都差强人意, 图 3(b)展示的是此时的 ROC 曲线, PFA 方法虽然有一定的改进作用, 但十分有限, 而图 3(a)中的所反映的重要变量为连续片段时, PFA 方法的改进作用则十分明显.

2.3.3 $\boldsymbol{\Sigma}$ 为稠密矩阵

当 $\boldsymbol{\Sigma}=\boldsymbol{\Sigma}^{(3)}$ 时, 所有的变量几乎都彼此相关, 图 3(c)展示了此时的 ROC 曲线. 由于没有考虑到变量之间的相互关系, 基于边缘回归系数的变量选择方法的效果非常差, 几乎等同于随机猜测变量的重要性, 而 PFA 以及 PLS 方法则有明显改善. 尤其是 PFA 方法, 在将回归系数中的共有信息剥离之后, 变量选择的精度在 PLS 的基础上有进一步的提高.

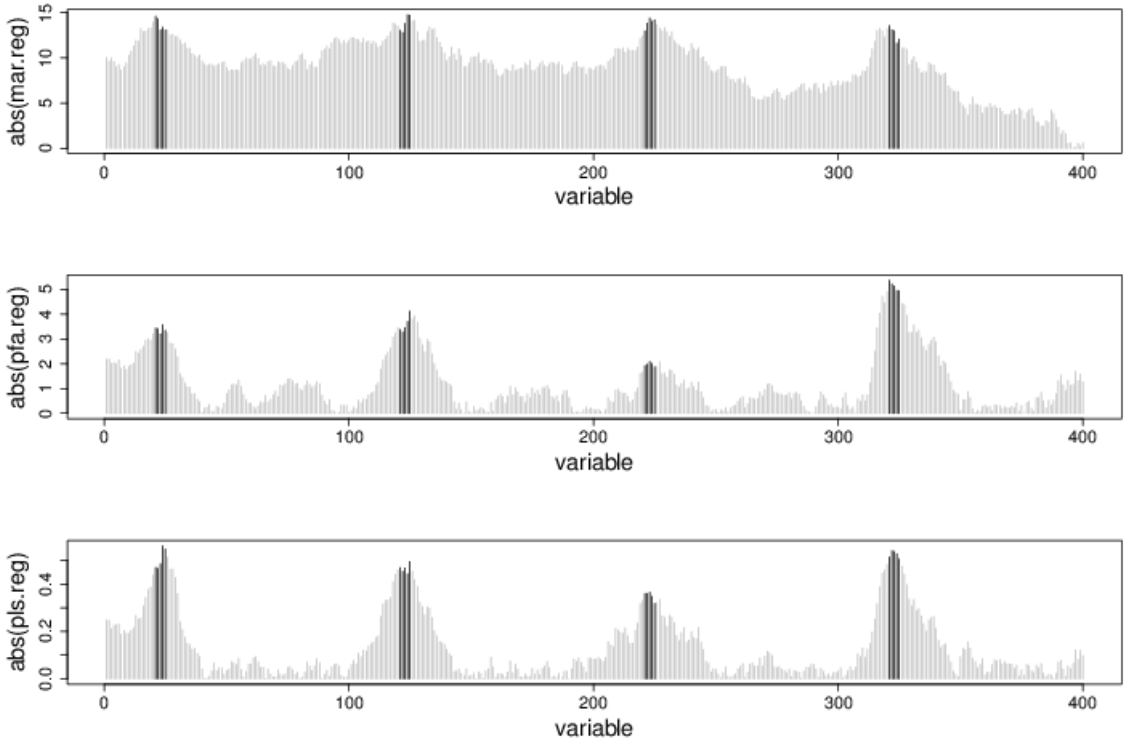
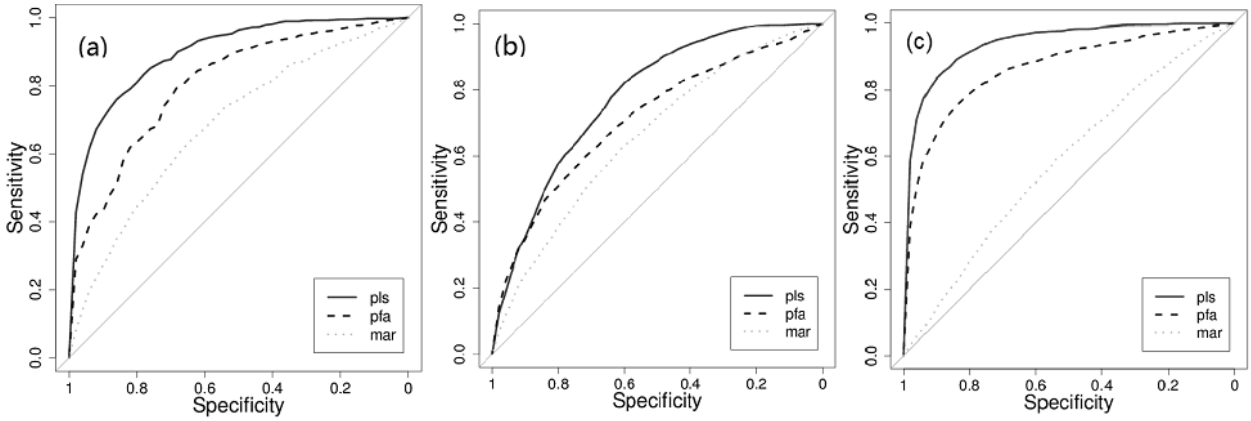


图 2 $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{(2)}$ 时用三种方法计算出的回归系数的绝对值大小, 从上往下依次是边缘回归方法(mar.reg), PFA 方法(pfa.reg)以及 PLS 方法(pls.reg)



(a) $\boldsymbol{\Sigma}=\boldsymbol{\Sigma}^{(2)}$ 且重要变量聚集为 4 个连续片段 (b) $\boldsymbol{\Sigma}=\boldsymbol{\Sigma}^{(2)}$ 且重要变量随机分布 (c) $\boldsymbol{\Sigma}=\boldsymbol{\Sigma}^{(3)}$ 且重要变量随机分布

图 3 $\boldsymbol{\Sigma}$ 为带状矩阵时的变量选择效果

3 结论

模拟研究充分表明变量之间的相关性会造成回归系数之间的相关性, 而回归系数之间的这种相关性将使得变量的相对重要性不能准确地通过系数的大小来反映. 本文使用 PFA 方法来提取变量之间的相关性信息从而对回归系数的估计量做出修正. 本质上, PFA 和 PLS 方法都属于因子模型方法, PLS 是有监督的因子模型, PFA 属于无监督的因子模型. 但 PFA 作为模块化的方法更加容易与别的变量选择方法有机结合.

(下转第 59 页)