基于动态协方差建模的纵向数据特征筛选方法

陈欣悦
西南大学 数学与统计学院,重庆 400715

摘 要:为了使统计分析有效进行,特征筛选问题在超高维领域已被众多学者广泛研究;针对现存特征筛选方法不能灵活处理超高维纵向数据的组内相关性问题,提出一个基于动态协方差建模的迭代特征筛选方法,并称之为迭代的动态特征筛选方法;在每次迭代过程中,均使用修正的 Cholesky 分解代替静态协方差矩阵建模方法对纵向数据的组内协方差矩阵进行动态建模,获得灵活的组内协方差矩阵估计,然后将所得估计代入广义估计方程中,并基于广义估计方程特征筛选方法的思想建立特征筛选准则进行筛选,最后当迭代算法收敛时得到最终的筛选子模型;引入随机模拟和酵母细胞周期循环基因表达数据集对迭代的动态特征筛选方法和基于广义估计方程的特征筛选方法以及其他 2 个经典的独立特征筛选方法进行测试,结果表明:迭代的动态特征筛选方法不仅可以快速地筛选出重要协变量,而且还能够更加灵活地处理纵向数据的组内相关性,拥有更高的筛选精度。

关键词:超高维纵向数据;特征筛选;修正的 Cholesky 分解;广义估计方程;动态协方差建模
中图分类号:O212.1 **文献标识码**:A doi:10.16055/j.issn.1672-058X.2023.0004.010

Feature Selection for Longitudinal Data Based on Dynamic Covariance Modeling

CHEN Xinyue
School of Mathematics and Statistics, Southwest University, Chongqing 400715, China

Abstract: In order to make statistical analysis effective, feature screening has been widely studied by many scholars in the ultra-high dimensional field. Aiming at the problem that the existing feature screening methods cannot flexibly deal with the intra-group correlation of ultra-high dimensional longitudinal data, an iterative feature screening method based on dynamic covariance modeling was proposed. This method is called the iterative dynamic feature screening method. At each iteration, the modified Cholesky decomposition was used to replace the static covariance matrix modeling method to dynamically model the intra-group covariance matrices of longitudinal data to obtain the flexible estimators of them, and then these estimators were substituted into the generalized estimating equation (GEE) to establish the feature screening criteria for screening according to the idea of GEE-based screening procedure (GEES). Finally, the final submodel was obtained when the iterative algorithm converged. Random simulations and yeast cell-cycle gene expression dataset were introduced to test the iterative dynamic feature screening method, GEES and the other two classical independent feature screening methods. The results show that the iterative dynamic feature screening method can quickly screen out important covariates, can deal with the intra-group correlation of longitudinal data more flexibly, and has higher screening accuracy.

Keywords: ultra-high dimensional longitudinal data; feature screening; modified Cholesky decomposition; generalized estimating equations; dynamic covariance modelling

收稿日期:2022-03-30 **修回日期**:2022-04-21 **文章编号**:1672-058X(2023)04-0069-76
基金项目:国家自然科学基金项目资助(11801466);重庆市自然科学基金项目资助(CSTC2021JCYJ-MSXMX0502).
作者简介:陈欣悦(1999—),女,重庆巫溪人,硕士研究生,从事高维统计降维研究.
引用格式:陈欣悦. 基于动态协方差建模的纵向数据特征筛选方法[J]. 重庆工商大学学报(自然科学版),2023,40(4):69—76.
CHEN Xinyue. Feature selection for longitudinal data based on dynamic covariance modeling[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2023, 40(4): 69—76.

1 引 言

伴随数据挖掘技术的深化和推广,人们使用的数据在维数方面经历着前所未有的爆炸式发展。例如:在基因微阵列或蛋白质组学数据中,常用数以千计的密码子数据以及大分子蛋白质表达作为协变量,使得协变量的维数远大于样本容量,这类问题可以归为超高维统计研究范畴。按照文献[1]的定义,超高维数据就是协变量的维数 p_n 随着样本量 n 呈指数阶增长的数据,即存在常数 $v\in(0,1/2)$,使得 $\log(p_n)=n^v$。在研究超高维数据时,人们常常遵循稀疏性假设,即并非每一个协变量都是重要的、不可或缺的,通常只有小部分协变量会对响应变量产生影响。但在超高维场合下,面对计算便利性、统计准确性和算法稳定性等挑战,传统变量选择方法的效果大打折扣[2-3]。因此,改进原有的变量选择方法或者提出新的变量筛选方法来处理超高维问题成为当今统计的热门研究领域。

文献[1]率先提出了一种独立特征筛选方法(Sure Independence Screening,SIS),针对超高维线性模型,通过计算响应变量和各协变量间皮尔逊相关系数的绝对值,对其进行排序,仅选择绝对值较大的一部分协变量即重要协变量进入模型。结果表明:SIS 易于计算,能够实现快速筛选,并且在一定条件下,SIS 具有确定筛选性质,能够准确高效地将超高维数据降维到高维数据,使得传统变量选择方法可以有效使用。特征筛选方法的良好性质吸引了许多学者在更多模型假设下对其进行大量研究,例如 NIS[4]、DC-SIS[5]、QaSIS[6]。但是这些方法都是针对独立同分布数据提出的。目前,纵向数据作为最常见的复杂数据之一,在生物医学、金融经济、环境科学等领域均具有广泛的应用。相比独立同分布数据,纵向数据的一个显著特征就是不同个体之间的观测是相互独立的,而同一个体在不同时间点所得的观测是相关的。众所周知,好的协方差矩阵估计可以提升纵向数据的分析效率。针对超高维纵向数据,如果直接沿用独立同分布数据下提出的特征筛选方法,虽然可以实现快速降维,但是忽略数据中存在的组内相关性会导致很多有效信息的丢失,使得特征筛选的精度较低,这会影响最终的建模效果。因此,对于超高维纵向数据,如何充分考虑组内相关性以得到筛选精度更高的特征筛选方法是本文所探索的。

一种常用的静态协方差矩阵建模方法为指定静态工作相关矩阵(如一阶自回归 AR(1)、等相关 CS、滑动平均 MA 等)来得到相应的工作协方差矩阵,这是文献[7]在构建广义估计方程(Generalized Estimating Equation,GEE)时所提出的方法。近年来,应用工作相关矩阵提出了一批超高维纵向特征筛选方法,例如文献[8]在广义线性模型假设下提出的基于广义估计方程的特征筛选方法(GEE-Based Screening Procedure,GEES);文献[9]在时变系数模型假设下,提出了一个新的特征筛选方法;文献[10]在广义变系数模型假设下,提出了基于广义估计方程的非参数独立特征筛选方法。实验结果表明:这些方法可以对超高维纵向数据进行有效的快速降维,并且筛选效果都优于 SIS、NIS,这意味着纳入组内相关性可以很好地提高纵向数据特征筛选的准确性[9]。但是针对不同的应用场景,选择出合适的工作相关结构是困难的,当工作相关结构被错误指定时,同样会严重影响到特征筛选的精度,这在实际应用中有很大的局限性。

在充分考虑组内相关性的基础上,要提升现存特征筛选方法的灵活性,关键在于建立良好的协方差模型,允许更一般的相关结构。一种灵活有效的协方差建模方法是由文献[11]提出的修正 Cholesky 分解,可以将协方差矩阵分解为相关部分和方差部分,然后再用回归方法分别对这两部分中的参数进行建模。该分解方法的优点:对分解后两部分中的参数没有限制;可以保证最终所得的协方差矩阵估计是正定的;建模过程中可以依赖于时间或空间位置,比使用传统静态协方差阵估计方法(即指定组内相关结构为 AR(1)、MA 或 CS)更加灵活。近年来,修正的 Cholesky 分解被广泛应用于纵向数据的研究工作中,一种同时对均值和协方差阵建模的新途径被提出:如文献[12]利用广义估计方程和修正的 Cholesky 分解在纵向线性模型下建立了纵向数据的动态均值协方差模型;文献[13]放松参数假设建立了纵向半参数联合均值协方差模型;文献[14,15]基于 Huber 函数和修正的 Cholesky 分解研究了纵向数据下的稳健参数估计与变量选择问题;相关研究还可见文献[16,17]。以上这些基于修正 Cholesky 分解的研究成果都是建立在低维或高维框架下的,在超高维特征筛选领域尚无相关的理论与应用研究成果。因此,本文将在超高维纵向数据下,结合修正的 Cholesky 分解和广义估计方程原理建立一个新的

基于动态协方差建模的迭代特征筛选方法,并称之为迭代的动态特征筛选方法(Iterative Dynamic Feature Screening Procedure,IDFS)。新方法采用 GEES 中构造特征筛选准则思想,在每次筛选时只需计算一次 p_n 维广义估计方程在原点的值,而无需分别计算 p_n 个协变量与响应变量的边际相关性,继承了 GEES 的快速筛选性,同时采用修正的 Cholesky 分解来改进静态协方差阵估计方法带来的局限性,对组内协方差阵进行更加灵活和准确的动态建模,使新方法拥有更高的灵活性和特征筛选精度。

2　预备知识

2.1　超高维纵向线性模型

假设有 n 个个体,$\boldsymbol{y}_i=(\boldsymbol{y}_{i1},\cdots,\boldsymbol{y}_{im_i})^{\mathrm{T}}$ 表示在时间 $t_{i1}<\cdots<t_{im_i}$ 内针对第 i 个个体收集到的响应变量的观测值,$\boldsymbol{x}_i=(\boldsymbol{x}_{i1},\cdots,\boldsymbol{x}_{imi})^{\mathrm{T}}$ 是相对应的 $m_i\times p_n$ 阶协变量观测矩阵,其中 $\boldsymbol{x}_{ij}=(\boldsymbol{x}_{ij1},\cdots,\boldsymbol{x}_{ijp_n})^{\mathrm{T}}$,则称 $\{\boldsymbol{y}_i,\boldsymbol{x}_i,i=1,2,\cdots,n\}$ 为纵向数据集。

假定协变量维数 p_n 为超高维,考虑如下超高维纵向线性模型:

$$\boldsymbol{y}_i=\beta_0\mathbf{1}_{m_i}+\boldsymbol{x}_i\boldsymbol{\beta}+\boldsymbol{\varepsilon}_i,i=1,2,\cdots,n \tag{1}$$

其中,$(\beta_0,\boldsymbol{\beta})^{\mathrm{T}}$ 是 (p_n+1) 维未知参数向量,$\mathbf{1}_{m_i}$ 是 m_i 维单位向量,$\boldsymbol{\varepsilon}_i=(\varepsilon_{i1},\cdots,\varepsilon_{im_i})^{\mathrm{T}}$ 是 m_i 维随机误差向量,并且 $E(\boldsymbol{\varepsilon}_i|\boldsymbol{x}_i)=0$,$\mathrm{Cov}(\boldsymbol{\varepsilon}_i|\boldsymbol{x}_i)=\boldsymbol{\Sigma}_i$,其中 $\boldsymbol{\Sigma}_i$ 是第 i 个个体的重复观测之间的协方差矩阵。为了便于描述,本文记 $\boldsymbol{\eta}=(\beta_0,\boldsymbol{\beta})^{\mathrm{T}}$,$\boldsymbol{S}_i(\boldsymbol{\eta})=(S_{i1}(\boldsymbol{\eta}),\cdots,S_{im_i}(\boldsymbol{\eta}))^{\mathrm{T}}$,其中 $S_{ij}(\boldsymbol{\eta})=\varepsilon_{ij}=\boldsymbol{y}_{ij}-\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}-\beta_0$。

2.2　广义估计方程原理

在模型式(1)的假设下,可得 $\boldsymbol{\eta}$ 的最优估计方程为

$$G(\boldsymbol{\eta})\triangleq\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\Pi}_i\boldsymbol{\Sigma}_i^{-1}\boldsymbol{S}_i(\boldsymbol{\eta})=0 \tag{2}$$

其中,$\boldsymbol{\Pi}_i=(\mathbf{1}_{m_i}\vdots\boldsymbol{x}_i)^{\mathrm{T}}$。然而在实际问题中,每个个体组内的真实协方差矩阵 $\boldsymbol{\Sigma}_i$ 通常是未知的,因此想要求解式(2),得到 $\boldsymbol{\eta}$ 的有效估计就需要合理估计 $\boldsymbol{\Sigma}_i$。于是,文献[7]构建了一个易估的工作协方差矩阵 $\boldsymbol{V}_i=\boldsymbol{B}_i^{1/2}\boldsymbol{R}_i(\boldsymbol{\alpha})\boldsymbol{B}_i^{1/2}$ 作为真实协方差矩阵 $\boldsymbol{\Sigma}_i$ 的替代值,从而建立了 $\boldsymbol{\eta}$ 的 GEE:

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\Pi}_i\boldsymbol{V}_i^{-1}\boldsymbol{S}_i(\boldsymbol{\eta})=0$$

其中,$\boldsymbol{R}_i(\boldsymbol{\alpha})$ 为工作相关矩阵,$\boldsymbol{\alpha}$ 为一个有限维的未知待估参数向量,$\boldsymbol{B}_i$ 为 $m_i\times m_i$ 阶对角矩阵,其第 j 个对角元素为 $\mathrm{Var}(\boldsymbol{y}_{ij}|\boldsymbol{x}_{ij})$。

在求解的过程中,首先给定 $\boldsymbol{\eta}$ 的初始估计量 $\hat{\boldsymbol{\eta}}$,然后使用皮尔逊残差和矩方法,得到 $\hat{\boldsymbol{V}}_i=\hat{\boldsymbol{B}}_i^{1/2}\hat{\boldsymbol{R}}_i\hat{\boldsymbol{B}}_i^{1/2}$,最后再将 $\boldsymbol{\eta}$ 的 GEE 改写为下式进一步求解:

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\Pi}_i\hat{\boldsymbol{V}}_i^{-1}\boldsymbol{S}_i(\boldsymbol{\eta})=0$$

上式可能没有数值解,因此通常会利用 Newton-Raphson 迭代法或者 Fisher 得分迭代法来寻求近似解。

3　迭代的动态特征筛选方法

3.1　基于修正 Cholesky 分解的动态协方差建模

在超高维背景下,协变量的维数异常巨大,很难直接得到 $\boldsymbol{\Sigma}_i=\mathrm{Cov}(\boldsymbol{S}_i(\boldsymbol{\eta})|\boldsymbol{x}_i)$ 的估计。因此首先采用现存的特征筛选方法(如文献[1]所提方法)把模型式(1)降维到如下稀疏纵向线性模型:

$$\boldsymbol{y}_i=\beta_{0A}\mathbf{1}_{m_i}+\boldsymbol{x}_{iA}\boldsymbol{\beta}_A+\boldsymbol{\varepsilon}_{iA},i=1,2,\cdots,n \tag{3}$$

其中,A 是所选协变量的指标集,集合大小表示为 $d_n=|A|\leqslant p_n$,$\boldsymbol{x}_{iA}$ 是相对应的 $m_i\times d_n$ 阶协变量观测矩阵。为了便于描述,本文令 $\boldsymbol{\eta}_A=(\beta_{0A},\boldsymbol{\beta}_A)^{\mathrm{T}}$ 为 (d_n+1) 维未知参数向量,$\boldsymbol{S}_i(\boldsymbol{\eta}_A)=\boldsymbol{y}_i-\boldsymbol{x}_{iA}\boldsymbol{\beta}_A-\beta_{0A}\mathbf{1}_{m_i}$ 为 m_i 维随机误差向量,其中 $\boldsymbol{S}_i(\boldsymbol{\eta}_A)$ 的第 j 个元素记为 $S_{iA,j}$,然后将估计 $\boldsymbol{\Sigma}_i$ 转化为估计 $\boldsymbol{\Sigma}_{iA}=\mathrm{Cov}(\boldsymbol{S}_i(\boldsymbol{\eta}_A)|\boldsymbol{x}_{iA})$,而 $\boldsymbol{\Sigma}_{iA}$ 可理解为工作协方差矩阵。

接下来,通过修正 Cholesky 分解对 $\boldsymbol{\Sigma}_{iA}$ 进行建模。该建模方法不需要事先假定任何工作相关结构,比传统静态协方差估计方法更加灵活。利用修正的 Cholesky 分解,协方差矩阵的逆 $\boldsymbol{\Sigma}_{iA}^{-1}$ 可以分解为

$$\boldsymbol{\Sigma}_{iA}^{-1}=\boldsymbol{T}_i^{\mathrm{T}}\boldsymbol{D}_i^{-1}\boldsymbol{T}_i$$

其中,$\boldsymbol{T}_i$ 是主对角元素全为 1,第 (j,l) 个元素为 $-\psi_{ij,l}$ $(j>l)$ 的 $m_i\times m_i$ 维下三角矩阵,$\boldsymbol{D}_i$ 为 $m_i\times m_i$ 维对角矩阵,对角线元素为 $\sigma_{ij}^2=\mathrm{Var}(S_{iA,j})$。定义 $\boldsymbol{e}_{iA}=(e_{iA,1},\cdots,e_{iA,m_i})^{\mathrm{T}}=\boldsymbol{T}_i\boldsymbol{S}_i(\boldsymbol{\eta}_A)$,可得:

$$S_{iA,1}=e_{iA,1},S_{iA,k}=e_{iA,k}+\psi_{ik,1}S_{iA,1}+\cdots+\psi_{ik,k-1}S_{iA,k-1}$$

$$E(\boldsymbol{e}_{iA}|\boldsymbol{x}_{iA})=0,\mathrm{Cov}(\boldsymbol{e}_{iA}|\boldsymbol{x}_{iA})=\boldsymbol{D}_i$$

于是,要得到 $\boldsymbol{\Sigma}_{iA}$ 的估计,只需要估计广义自回归参数 $\psi_{ij,l}$ 和更新方差 σ_{ij}^2。根据文献[12]的建议,使用如下广义线性模型对其进行建模:

$$\psi_{ij,l}=\boldsymbol{w}_{ijl}^{\mathrm{T}}\boldsymbol{\gamma},\log(\sigma_{ij}^2)=\boldsymbol{z}_{ij}^{\mathrm{T}}\boldsymbol{\lambda}$$

其中,$\boldsymbol{\gamma}=(\gamma_1,\cdots,\gamma_q)^{\mathrm{T}}$,$\boldsymbol{\lambda}=(\lambda_1,\cdots,\lambda_d)^{\mathrm{T}}$ 都是未知的,

$\boldsymbol{w}_{ijl}=(1,(t_{ij}-t_{il}),\cdots,(t_{ij}-t_{il})^{q-1})^{\mathrm{T}}$, $\boldsymbol{z}_{ij}=(1,t_{ij},\cdots,t_{ij}^{d-1})^{\mathrm{T}}$。

进一步,只需估计 $\boldsymbol{\gamma}$ 和 $\boldsymbol{\lambda}$,可以通过求解如下广义估计方程来得到:

$$U_1(\boldsymbol{\gamma})=\sum_{i=1}^{n}\left(\frac{\partial \boldsymbol{e}_{iA}^{\mathrm{T}}}{\partial \boldsymbol{\gamma}}\right)\boldsymbol{D}_i^{-1}\boldsymbol{e}_{iA}=0 \tag{4}$$

$$U_2(\boldsymbol{\lambda})=\sum_{i=1}^{n}\boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{D}_i\boldsymbol{W}_i^{-1}(\boldsymbol{e}_{iA}^2-\boldsymbol{\sigma}_i^2)=0 \tag{5}$$

其中,$\partial \boldsymbol{e}_{iA}^{\mathrm{T}}/\partial \boldsymbol{\gamma}$ 是一个 $q\times m_i$ 维矩阵,其第 1 列为零向量,第 $2\leqslant j\leqslant m_i$ 列为 $\partial \boldsymbol{e}_{iA}/\partial \boldsymbol{\gamma}=-\sum_{k=1}^{j-1}S_{iA,k}\,\boldsymbol{w}_{ijk}$, $\boldsymbol{Z}_i=(\boldsymbol{z}_{i1},\cdots,\boldsymbol{z}_{im_i})^{\mathrm{T}}$, $\boldsymbol{\sigma}_i^2=(\sigma_{i1}^2,\cdots,\sigma_{im_i}^2)^{\mathrm{T}}$。$\boldsymbol{W}_i=\boldsymbol{B}_i^{1/2}\boldsymbol{R}_i(\boldsymbol{\rho})\boldsymbol{B}_i^{1/2}$ 是 $\boldsymbol{e}_{iA}^2$ 的工作协方差矩阵,其中 $\boldsymbol{B}_i=2\mathbf{diag}\{\sigma_{i1}^4,\cdots,\sigma_{im_i}^4\}$,而 $R_i(\rho)$ 中则引入了一个新的参数 ρ 来代表 $e_{iA,j}^2$ 和 $e_{iA,k}^2$ 之间的相关性,但由于 ρ 的影响甚微(文献[14-16]),因此可假设 $\rho=0$, $\boldsymbol{W}_i=2\mathbf{diag}\{\sigma_{i1}^4,\cdots,\sigma_{im_i}^4\}$。

显然式(4)和式(5)不能直接获得数值解,因此,本文采用 Fisher 得分迭代算法来进行求解,在给定 $\boldsymbol{\eta}_A$ 和 $\boldsymbol{\lambda}$ 时,$\boldsymbol{\gamma}$ 的更新方程为

$$\hat{\boldsymbol{\gamma}}=\boldsymbol{\gamma}-\left\{\left[\sum_{i=1}^{n}\left(\frac{\partial \boldsymbol{e}_{iA}^{\mathrm{T}}}{\partial \boldsymbol{\gamma}}\right)\boldsymbol{D}_i^{-1}\left(\frac{\partial \boldsymbol{e}_{iA}^{\mathrm{T}}}{\partial \boldsymbol{\gamma}}\right)^{\mathrm{T}}\right]^{-1}\times\left[\sum_{i=1}^{n}\left(\frac{\partial \boldsymbol{e}_{iA}^{\mathrm{T}}}{\partial \boldsymbol{\gamma}}\right)\boldsymbol{D}_i^{-1}\boldsymbol{e}_{iA}\right]\right\} \tag{6}$$

在给定 $\boldsymbol{\eta}_A$ 和 $\boldsymbol{\gamma}$ 时,$\boldsymbol{\lambda}$ 的更新方程为

$$\hat{\boldsymbol{\lambda}}=\boldsymbol{\lambda}-\left\{\left[\sum_{i=1}^{n}\boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{D}_i\boldsymbol{W}_i^{-1}\boldsymbol{D}_i\boldsymbol{Z}_i\right]^{-1}\times\left[\sum_{i=1}^{n}\boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{D}_i\boldsymbol{W}_i^{-1}(\boldsymbol{e}_{iA}^2-\boldsymbol{\sigma}_i^2)\right]\right\} \tag{7}$$

最后可得 $\sum_{iA}$ 的估计为

$$\hat{\sum}_{iA}=\hat{\boldsymbol{T}}_i^{-1}\hat{\boldsymbol{D}}_i(\hat{\boldsymbol{T}}_i^{\mathrm{T}})^{-1},i=1,2,\cdots,n$$

其中,$\hat{\boldsymbol{D}}_i=\mathbf{diag}\{\exp(\boldsymbol{z}_{i1}^{\mathrm{T}}\hat{\boldsymbol{\lambda}}),\cdots,\exp(\boldsymbol{z}_{im_i}^{\mathrm{T}}\hat{\boldsymbol{\lambda}})\}$, $\hat{\boldsymbol{T}}_i$ 的主对角元素全为 1,第 (j,l) 个元素为 $-\hat{\psi}_{ij,l}=\boldsymbol{w}_{ijl}^{\mathrm{T}}\hat{\boldsymbol{\gamma}}(j>l)$。

3.2 迭代算法

基于估计函数 $\boldsymbol{G}(\boldsymbol{\eta})$, $\boldsymbol{\eta}_A$ 和 $\hat{\boldsymbol{\Sigma}}_{iA}$ 以及文献[8]所提方法,对于模型式(1),首先可以利用广义估计方程原理得到估计函数为

$$\begin{pmatrix}\hat{\boldsymbol{D}}_0(\boldsymbol{\eta})\\ \hat{\boldsymbol{D}}_1(\boldsymbol{\eta})\end{pmatrix}\triangleq\hat{\boldsymbol{D}}(\boldsymbol{\eta})=\begin{pmatrix}\frac{1}{n}\sum_{i=1}^{n}1_{\boldsymbol{m}_i}^{\mathrm{T}}\hat{\sum}_{iA}^{-1}\boldsymbol{S}_i(\boldsymbol{\eta})\\ \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i^{\mathrm{T}}\hat{\sum}_{iA}^{-1}\boldsymbol{S}_i(\boldsymbol{\eta})\end{pmatrix}$$

然后可以提出如下的一个特征筛选准则:

$$A'=\{1\leqslant j\leqslant p_n:|\hat{D}_{1j}(\boldsymbol{\eta}_A^*)|>\delta_n\}$$

其中,$\boldsymbol{\eta}_A^*=(\beta_{0A},\boldsymbol{0}^{\mathrm{T}})^{\mathrm{T}}$, $\hat{D}_{1j}(\boldsymbol{\eta}_A^*)$ 是 $\hat{\boldsymbol{D}}_1(\boldsymbol{\eta}_A^*)=(1/n)\sum_{i=1}^{n}\boldsymbol{x}_i^{\mathrm{T}}\times\hat{\sum}_{iA}^{-1}\boldsymbol{S}_i(\boldsymbol{\eta}_A^*)$ 的第 j 个元素,作为衡量第 j 个协变量相对重要程度的统计度量指标,δ_n 是一个事先给定的阈值。

注 1 在实际应用中,很难选择阈值 δ_n,因此通常使用如下一个等价的筛选准则:

$$A'=\{1\leqslant j\leqslant p_n:\text{前 }d_n\text{ 个较大的}|\hat{D}_{1j}(\boldsymbol{\eta}_A^*)|\}$$

其中,d_n 也是一个事先给定的阈值,规定了所选子模型的大小。

因此,本文将结合修正的 Cholesky 分解和广义估计方程原理建立一个基于动态协方差建模的迭代特征筛选算法,并称之为迭代的动态特征筛选方法,算法过程如下:

步骤 1 设 $m=1$,取 $d_n=[n/\log(n)]$。忽略重复测量之间的相关性,使用文献[1]提出的独立特征筛选方法得到初始子模型:$A^{(1)}=\{1\leqslant j\leqslant p_n:\text{前 }d_n\text{ 个较大的}\omega_j\}$,其中 $A^{(1)}$ 是所选协变量的指标集,ω_j 是响应变量和第 j 个协变量之间的皮尔逊相关系数,假设 $\boldsymbol{x}_{iA^{(1)}}$ 是相应的第 i 个个体的 $m_i\times d_n$ 维协变量观测矩阵。

步骤 2 对给定的指标集 $A^{(m)}$ $(m\geqslant 1)$,将式(3)中的指标集 A 替换为 $A^{(m)}$,应用现存惩罚估计方法(如自适应 LASSO)得到 $\boldsymbol{\eta}_{A^{(m)}}$ 的估计,记为 $\hat{\boldsymbol{\eta}}_{A^{(m)}}=(\hat{\beta}_{0A^{(m)}},\hat{\boldsymbol{\beta}}_{A^{(m)}}^{\mathrm{T}})^{\mathrm{T}}$;然后将式(4)和式(5)中的 $\boldsymbol{\eta}_A$ 替换为 $\hat{\boldsymbol{\eta}}_{A^{(m)}}$,根据式(6)和式(7)更新得到 $\hat{\boldsymbol{\gamma}}^{(m)}$ 和 $\hat{\boldsymbol{\lambda}}^{(m)}$,得到 $\hat{\sum}_{iA^{(m)}}=(\hat{\boldsymbol{T}}_i^{(m)})^{-1}\hat{\boldsymbol{D}}_i^{(m)}((\hat{\boldsymbol{T}}_i^{(m)})^{\mathrm{T}})^{-1}$,其中 $\hat{\boldsymbol{T}}_i^{(m)}$ 的主对角元素全为 1,第 (j,l) 个元素为 $-\hat{\psi}_{ij,l}^{(m)}=\boldsymbol{w}_{ijl}^{\mathrm{T}}\hat{\boldsymbol{\gamma}}^{(m)}(j>l)$, $\hat{\boldsymbol{D}}_i^{(m)}=\mathbf{diag}\{\exp(\boldsymbol{z}_{i1}^{\mathrm{T}}\hat{\boldsymbol{\lambda}}^{(m)}),\cdots,\exp(\boldsymbol{z}_{im_i}^{\mathrm{T}}\hat{\boldsymbol{\lambda}}^{(m)})\}$。

步骤 3 基于上一步得到的 $\hat{\sum}_{iA^{(m)}}$,更新所选协变量的指标集:

$$A^{(m+1)}=\{1\leqslant j\leqslant p_n:\text{前 }d_n\text{ 个较大的}|\hat{D}_{1j}(\hat{\boldsymbol{\eta}}_{A^{(m)}}^*)|\}$$

其中,$\hat{\boldsymbol{\eta}}_{A^{(m)}}^*=(\hat{\beta}_{0A^{(m)}},\boldsymbol{0}^{\mathrm{T}})^{\mathrm{T}}$, $\hat{D}_{1j}(\hat{\boldsymbol{\eta}}_{A^{(m)}}^*)$ 是 $\hat{\boldsymbol{D}}_1(\hat{\boldsymbol{\eta}}_{A^{(m)}}^*)=(1/n)\sum_{i=1}^{n}\boldsymbol{x}_i^{\mathrm{T}}\hat{\sum}_{iA^{(m)}}^{-1}\boldsymbol{S}_i(\hat{\boldsymbol{\eta}}_{A^{(m)}}^*)$ 的第 j 个元素,令 $m=m+1$。

步骤 4 重复迭代步骤 2 和步骤 3,当指标集

$A^{(m)}$ 和 $A^{(m+1)}$ 相等且元素顺序相同时或者迭代次数大于 10 时,停止迭代。

令 $\widehat{\boldsymbol{\eta}}_A=(\widehat{\beta}_{0A},\widehat{\boldsymbol{\beta}}_A^{\mathrm{T}})^{\mathrm{T}},\widehat{\boldsymbol{\gamma}},\widehat{\boldsymbol{\lambda}},\widehat{\Sigma}_{iA}$ 为算法收敛时各统计量的估计量,则此时所选协变量的指标集为

$$\widehat{A}=\{1\leqslant j\leqslant p_n:\text{前 } d_n \text{ 个较大的}|\widehat{D}_{1j}(\widehat{\boldsymbol{\eta}}_A^*)|\}$$

其中,$\widehat{\boldsymbol{\eta}}_A^*=(\widehat{\beta}_{0A},0^{\mathrm{T}})^{\mathrm{T}}$,$\widehat{D}_{1j}(\widehat{\boldsymbol{\eta}}_A^*)$ 是 $\widehat{\boldsymbol{D}}_1(\widehat{\boldsymbol{\eta}}_A^*)=(1/n)\times\sum_{i=1}^{n}\boldsymbol{x}^{\mathrm{T}}{}_i\widehat{\Sigma}_{iA}^{-1}\boldsymbol{S}_i(\widehat{\boldsymbol{\eta}}_A^*)$ 的第 j 个元素。

注 2　为了在保证算法精度的同时保证算法效率,本文设置迭代算法的终止条件有两个:一是迭代次数,设置最大迭代次数为 10 次;二是算法精度,当第 k 次所选子模型自变量的下标与第 $k-1$ 次所选子模型自变量的下标相同时停止迭代。结果表明该迭代算法在 10 次以内就能达到收敛(见表 1)。

4　随机模拟测试

假设真模型为

$$\boldsymbol{y}_{ij}=\boldsymbol{\beta}_0+\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}+\varepsilon_{ij};i=1,\cdots,n;j=1,\cdots,m_i$$

其中,$\beta_0=0.65$,$\boldsymbol{\beta}=(\beta_1,\cdots,\beta_{15},0,\cdots,0)^{\mathrm{T}}$ 是 p_n 维向量,$c_k\sim\mathrm{Binomal}(1,0.5)$,$\beta_k\sim(-1)^{c_k}\times U(0.5,1)$,$k=1,\cdots,15$,这表示前 15 个协变量是重要协变量;$\boldsymbol{x}_{ij}=(x_{ij1},\cdots,x_{ijp_n})^{\mathrm{T}}$,其中 x_{ijk} 服从标准正态分布;设样本量 $n=100$,协变量个数 $p_n=500$、10 000、20 000,每个个体的重复观测数量 m_i 是从 10—20 里随机抽取的一个数;随机误差向量 $\boldsymbol{\varepsilon}_i=(\varepsilon_{i1},\cdots,\varepsilon_{im_i})^{\mathrm{T}}$ 考虑以下两种情况:

情况 1　随机误差向量 $\boldsymbol{\varepsilon}_i$ 以 0.5 的概率来自多元正态分布 $\mathrm{N}(0,\Sigma_i^a)$,以 0.5 的概率来自多元正态分布 $\mathrm{N}(0,\Sigma_i^b)$,其中 $\Sigma_i^a=\boldsymbol{A}_i^{1/2}\boldsymbol{C}_i^a\boldsymbol{A}_i^{1/2}$ 且 $\boldsymbol{C}_i^a$ 为相关系数为 0.85 的 AR(1)相关结构,$\Sigma_i^b=\boldsymbol{A}_i^{1/2}\boldsymbol{C}_i^b\boldsymbol{A}_i^{1/2}$ 且 $\boldsymbol{C}_i^b$ 为相关系数为 0.85 的可交换相关结构,$\boldsymbol{A}_i=\mathbf{diag}(\exp(\boldsymbol{z}_{i1}^{\mathrm{T}}\boldsymbol{\lambda}),\cdots,\exp(\boldsymbol{z}_{im_i}^{\mathrm{T}}\boldsymbol{\lambda}))$,$\boldsymbol{z}_{ij}=(1,t_{ij},t_{ij}^2,t_{ij}^3)^{\mathrm{T}}$,$\boldsymbol{\lambda}=(0.2,-0.05,0.15,-0.1)^{\mathrm{T}}$,$t_{ij}\sim U(0,1)$ 是第 i 个个体的第 j 次观测时间。

情况 2　随机误差向量 $\boldsymbol{\varepsilon}_i\sim\mathrm{N}(0,\boldsymbol{\Lambda}_i)$,其中 $\boldsymbol{\Lambda}_i=\boldsymbol{\Delta}_i\boldsymbol{B}_i(\boldsymbol{\Delta}_i)^{\mathrm{T}}$,$\boldsymbol{B}_i$ 是一个 $m_i\times m_i$ 维对角矩阵,其第 j 个对角元素为 $1+2|\sin(2\pi\zeta_{ij})|$,$\zeta_{ij}\sim\mathrm{U}(0,2)$,$\boldsymbol{\Delta}_i$ 是一个单位下三角矩阵且第(j,l)个元素为 $\sin(\pi(t_{ij}-t_{il}))-(t_{ij}-t_{il})$。

为了评估本文所提方法 IDFS 的有效性,将其与下列 5 种特征筛选方法进行比较:超高维线性模型下的独立特征筛选方法(SIS);超高维稀疏可加模型下的非参数独立筛选方法(NIS);基于广义估计方程的特征筛选方法(GEES),包括独立结构下的 GEES,记为 GEES-IND;AR(1)相关结构下的 GEES,记为 GEES-AR(1);可交换相关结构下的 GEES,记为 GEES-CS。

然后使用以下两个准则来评价特征筛选的效果:MMS,即包含所有重要协变量的最小模型大小,该值越接近重要协变量的个数越好;P_a,即在给定模型大小 $d_n=[n/\log(n)]$ 下,T 次数值模拟后所有重要协变量都被选入最终子模型中的比例,该值越接近于 1 越好。

按照上述模拟的设定,将每种情况重复模拟 200 次。表 1 给出了情况 1 和情况 2 中 IDFS 的平均迭代次数,可以看到在两种情况下 IDFS 都很快收敛了,平均迭代次数不超过 3 次。

表 1　算法的平均迭代次数

Table 1　Average number of iterations of the algorithm

	p_n	平均迭代次数	标准差
情况 1	5 000	2.360	0.908
	10 000	2.300	0.481
	20 000	2.365	0.909
情况 2	5 000	2.485	0.558
	10 000	2.570	0.614
	20 000	2.510	0.549

表 2、表 3 分别给出了在情况 1 和情况 2 中使用上述 6 种方法进行 200 次数值模拟后 5%、25%、50%、75%、95%分位点下的 MMS 值和 P_a 值。由表 2 可知:在情况 1 下,IDFS 的 MMS 值更接近重要协变量个数 15,P_a 值更接近 1,因此 IDFS 的表现优于 SIS、NIS、GEES-IND、GEES-AR(1)、GEES-CS,具有更高的筛选精度。由表 3 可知:在情况 2 下,与情况 1 相比,各方法的筛选精度都有所下降,但 IDFS 的表现仍是最优的。这说明在纵向数据组内相关结构复杂的情况下,IDFS 更加灵活,能更充分地利用纵向数据组内相关性,达到更高的筛选精度。

综上所述,IDFS 可以快速地筛选出重要变量且比现存方法更灵活,具有更高的筛选精度。

表 2 情况 1 下 200 次模拟后的筛选效果

Table 2 Screening performance over 200 simulations for Case 1

p_n	特征筛选方法	MMS						P_a
		5%	25%	50%	75%	95%	100%	
5 000	SIS	15	15	15	15	15	18	0.995
	NIS	15	15	15	15	20	34	0.980
	GEES-IND	15	15	15	15	15	18	0.995
	GEES-AR(1)	15	15	15	15	18	77	0.980
	GEES-CS	15	15	15	15	15	18	0.995
	IDFS	15	15	15	15	15	16	0.995
10 000	SIS	15	15	15	15	15.05	19	0.985
	NIS	15	15	15	15	20.05	104	0.975
	GEES-IND	15	15	15	15	15.05	19	0.985
	GEES-AR(1)	15	15	15	15	19.05	79	0.970
	GEES-CS	15	15	15	15	15	17	0.995
	IDFS	15	15	15	15	15	17	0.995
20 000	SIS	15	15	15	15	15.05	20	0.990
	NIS	15	15	15	15	21.05	81	0.975
	GEES-IND	15	15	15	15	16	21	0.990
	GEES-AR(1)	15	15	15	15	19.05	260	0.965
	GEES-CS	15	15	15	15	15	16	0.990
	IDFS	15	15	15	15	15	16	0.990

表 3 情况 2 下 200 次模拟后的筛选效果

Table 3 Screening performance over 200 simulations for Case 2

p_n	特征筛选方法	MMS						P_a
		5%	25%	50%	75%	95%	100%	
5 000	SIS	15	15	15	17	62.05	348	0.940
	NIS	15	16	19	35	207.20	1 552	0.870
	GEES-IND	15	15	15	17	62	314	0.935
	GEES-AR(1)	15	15	15	17	57.20	315	0.920
	GEES-CS	15	15	15	17	48.05	342	0.925
	IDFS	15	15	15	16	24.10	109	0.945
10 000	SIS	15	15	15	18	94.50	1 165	0.940
	NIS	15	16	21	55.25	581.95	2 203	0.855
	GEES-IND	15	15	15	18	88.20	1 124	0.920
	GEES-AR(1)	15	15	15	19	91.10	1 002	0.915
	GEES-CS	15	15	15	18	76.15	1 128	0.930
	IDFS	15	15	15	17	54.25	766	0.945
20 000	SIS	15	15	16	20	185.05	1 343	0.915
	NIS	15	17	26.50	82.75	1 428.60	6 711	0.825
	GEES-IND	15	15	16	21.25	213.55	1 337	0.930
	GEES-AR(1)	15	15	16	22.25	174.05	926	0.925
	GEES-CS	15	15	16	21	193.30	1 199	0.910
	IDFS	15	15	15	17.25	88.15	3 447	0.945

5 实例分析与应用

转录因子通过与基因上的特定序列相结合,激活或者抑制基因的表达,在调控基因表达方面具有重要作用(详见文献[18])。本文将分析由文献[19]清洗过的酵母细胞周期基因表达数据集(可在 R 软件的 spls 包中获取),用于识别影响酵母细胞周期中基因表达的重要转录因子。该数据集包含 18 个时间点上 542 个酵母细胞的细胞周期调控基因的观测,响应变量 $\boldsymbol{y}_{ij}$ 为第 i 个基因在第 j 次观测的基因表达水平,协变量向量 $\boldsymbol{x}_{ij}=(x_{ii1},\cdots,x_{ii106})^{\mathrm{T}}$,其中 x_{iik} 为第 i 个基因与第 k 个转录因子的结合概率,$i=1,\cdots,542;j=1,\cdots,18;k=1,\cdots,106$。

为了说明该数据重复观测之间的确存在相关性,本文使用修正的 Cholesky 分解方法对其组内协方差矩阵进行建模,得到如表 4 所示的参数向量 $\boldsymbol{\gamma}$ 和 $\boldsymbol{\lambda}$ 的估计值,并通过分块 bootstrap 方法得到其标准差估计和 95%正态置信区间。由表 4 可得:在显著性水平为 5%的情况下,$\boldsymbol{\gamma}$ 和 $\boldsymbol{\lambda}$ 显著不为 0,说明该数据的重复观测之间存在显著的相关性。

因此,对该数据建立如下纵向线性回归模型:

$$\boldsymbol{y}_i=\beta_0 1_{18}+\boldsymbol{x}_i\boldsymbol{\beta}+\widetilde{\boldsymbol{x}}_i\boldsymbol{\alpha}+\boldsymbol{\varepsilon}_i;i=1,2,\cdots,542 \tag{8}$$

其中,$(\beta_0,\boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$ 和 $\boldsymbol{\alpha}$ 分别为 $p+1$ 维和 p 维未知参数向量($p=106$),$\boldsymbol{y}_i=(\boldsymbol{y}_{i1},\cdots,\boldsymbol{y}_{i18})^{\mathrm{T}}$,$\boldsymbol{x}_i=(\boldsymbol{x}_{i1},\cdots,\boldsymbol{x}_{i18})^{\mathrm{T}}$,$\widetilde{\boldsymbol{x}}_i=(\boldsymbol{x}_{i1}\boldsymbol{t}_{i1},\cdots,\boldsymbol{x}_{i18}\boldsymbol{t}_{i18})^{\mathrm{T}}$,$\boldsymbol{x}_{ij}=(\boldsymbol{x}_{ii1},\cdots,\boldsymbol{x}_{ii106})^{\mathrm{T}}$,$t_{ij}=(j-1)/17$,$1_{18}$ 是 18 阶单位向量,$\boldsymbol{\varepsilon}_i$ 是随机误差向量。此时协变量个数为 $2p$,为了便于描述,本文对数据集中的每一个协变量进行编号,例如第一个协变量即第一个转录因子 ABF1 的结合概率记为 ID=1,第二个协变量即第二个转录因子 ACE2 的结合概率记为 ID=2,以此类推,总共有 212 个 ID。

根据文献[20],已知 21 个被生物实验证明确实对细胞周期循环有显著影响的转录因子为 ABF1(ID=1)、ACE2(ID=2)、BAS1(ID=9)、CBF1(ID=11)、FKH1(ID=21)、FKH2(ID=22)、GCN4(ID=27)、GCR1(ID=28)、GCR2(ID=29)、LEU3(ID=46)、MBP1(ID=51)、MCM1(ID=52)、MET31(ID=53)、NDD1(ID=61)、REB1(ID=70)、SWI4(ID=93)、SWI5(ID=94)、SWI6(ID=95)、STB1(ID=88)、SKN7(ID=84)、STE12(ID=89)。

设置筛选的阈值 $d_n=[n/(2\log(n))]=43$,采用本文所提方法 IDFS 以及其他 5 种方法 SIS、NIS、GEES-IND、GEES-AR(1)、GEES-CS,得到如图1和表5所示的特征筛选结果。表 5 报告了 IDFS 和其他 5 种方法的筛选结果中,包含 21 个已知的重要转录因子的个数,将其作为性能比较指标,可知 IDFS 比其他 5 种方法表现更好,更具实用性。例如,IDFS 筛选出的排名前 5 的转录因子中包含 4 个已知的有显著影响的转录因子,GEES-AR(1)和 GEES-CS 方法选中的排名前 5 的转录因子中分别包含 2 个和 1 个已知的有显著影响的转录因子,而基于独立结构的 SIS、NIS、GEES-IND 此时已经失效了。

表 4　$\boldsymbol{\gamma}$ 和 $\boldsymbol{\lambda}$ 的估计值及其标准差估计和 95%正态置信区间

Table 4　The estimators and sample standard errors and 95% normal confidence intervals of $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$

		估计值	标准差	95%正态置信区间
$\boldsymbol{\gamma}$	γ_1	0.297	0.014	(0.270,0.323)
	γ_2	−2.996	0.086	(−3.165,−2.827)
	γ_3	6.211	0.212	(5.796,6.627)
	γ_4	−3.864	0.153	(−4.163,−3.565)
$\boldsymbol{\lambda}$	λ_1	−0.746	0.099	(−0.940,−0.553)
	λ_2	−4.301	0.806	(−5.881,−2.720)
	λ_3	4.804	1.728	(1.417,8.191)
	λ_4	−1.918	1.051	(−3.978,0.142)

表 5　6 种方法下前 5、前 15 和前 25 个被选中的协变量中包含 21 个已知重要转录因子的个数

Table 5　The first 5,15 and 25 selected covariates contain the numbers of 21 known important transcription factors for the six methods

	SIS	NIS	GEES-IND	GEES-AR(1)	GEES-CS	IDFS
前 5 个	0	0	0	2	1	4
前 15 个	0	0	0	3	4	7
前 25 个	0	0	0	4	7	8

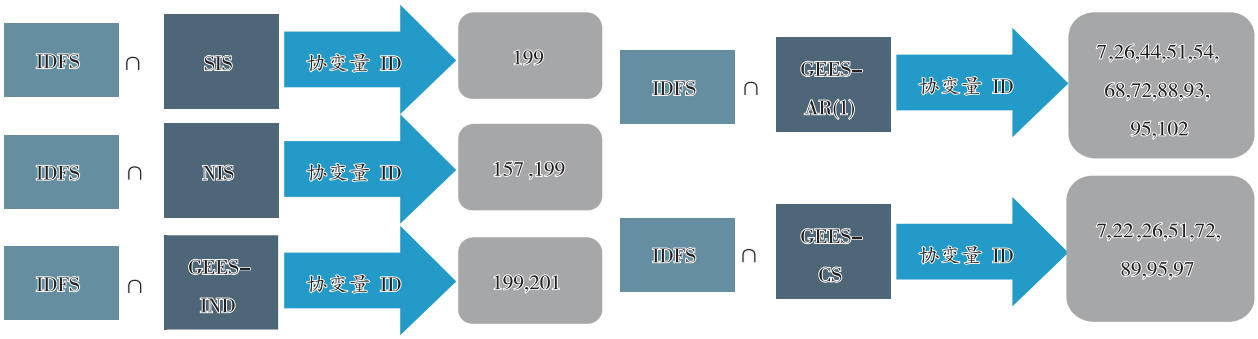

图 1　本文所提方法(IDFS)与现有方法(NIS, SIS, GEES-IND, GEES-AR(1), GEES-CS)在前 25 个所选协变量中相同的协变量 ID

注:符号“∩”代表了在两种方法下前 25 个所选协变量 ID 的交集。

Fig. 1　The ID of the same covariate in the first 25 selected covariates between the proposed method (IDFS) and the existing methods (NIS, SIS, GEES-IND, GEES-AR(1), GEES-CS)

Note: the symbol “∩” represents the intersection of the first 25 selected covariate IDs under the two methods.

6 结 论

基于静态工作相关结构提出的超高维特征筛选方法,当工作相关结构被错误指定时,特征筛选的精度会受到严重影响。因此,在超高维纵向线性模型下,提出一个迭代的动态特征筛选方法(IDFS),将广义估计方程原理和基于修正 Cholesky 分解的动态协方差建模方法结合起来,在每次迭代过程中,使用修正 Cholesky 分解得到良好的组内协方差阵估计,基于广义估计方程特征筛选方法的思想建立特征筛选准则,当迭代算法收敛时得到最终的子模型。随机模拟结果显示:IDFS 可以快速筛选出重要协变量,并且比现存方法具有更高的灵活性与筛选精度。最后,选取酵母细胞周期循环基因表达数据集作为实证研究对象,检验了新方法在寻找影响酵母循环基因表达的重要转录因子的能力,结果证明新方法相比于现存方法更具实用性。

参考文献(References):

[1] FAN J Q, LYU J C. Sure independence screening for ultra-high dimensional feature space (with discussion)[J]. Journal of the Royal Statistical Society Series B (Statistical Methodology), 2008, 70(5): 849—911.

[2] 牛勇, 李华鹏, 刘阳惠, 等. 超高维数据特征筛选方法综述[J]. 应用概率统计, 2021, 37(1): 69—110.
NIU Yong, LI Hua-peng, LIU Yang-hui, et al. Overview of feature screening methods for ultra-high dimensional data[J]. Chinese Journal of Applied Probability and Statistics, 2021, 37(1): 69—110.

[3] 黄文静, 邓丹, 杜杰琳, 等. 基于预测变量图结构的高维逻辑回归模型[J]. 重庆工商大学学报(自然科学版), 2021, 38(5): 107—113.
HUANG Wen-jing, DENG Dan, DU Jie-lin, et al. High-dimensional logic regression model based on graph structure of predictive variables[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2021, 38(5): 107—113.

[4] FAN J Q, FENG Y, SONG R. Nonparametric independence screening in sparse ultra-high dimensional additive models[J]. Journal of the American Statistical Association, 2011, 106(494): 544—557.

[5] LI R I, ZHONG W, ZHU L P. Feature screening via distance correlation learning[J]. Journal of the American Statistical Association, 2012, 107(499): 1129—1139.

[6] HE X M, WANG L, HONG H G. Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data[J]. Annals of Statistics, 2013, 41(1): 342—369.

[7] LIANG K Y, ZEGER S L. Longitudinal data analysis using generalized linear models[J]. Biometrika, 1986, 73(1): 13—22.

[8] XU P R, ZHU L X, LI Y. Ultrahigh dimensional time course feature selection[J]. Biometrics, 2014, 70(2): 356—365.

[9] CHU W H, LI R, REIMHERR M. Feature screening for time-varying coefficient models with ultrahigh dimensional longitudinal data[J]. The Annals of Applied Statistics, 2016, 10(2): 596—617.

[10] ZHANG S, ZHAO P X, LI G R, et al. Nonparametric independence screening for ultra-high dimensional generalized varying coefficient models with longitudinal data[J]. Journal of Multivariate Analysis, 2019, 171(3): 37—52.

[11] POURAHMADI M. Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation [J]. Biometrika, 1999, 86(3): 677—690.

[12] YE H J, PAN J X. Modelling of covariance structures in generalised estimating equations for longitudinal data [J]. Biometrika, 2006, 93(4): 927—941.

[13] LENG C L, ZHANG W P, PAN J X. Semiparametric mean-covariance regression analysis for longitudinal data [J]. Journal of the American Statistical Association, 2010, 105 (489): 181—193.

[14] ZHENG X Y, FUNG W K, ZHU Z Y. Robust estimation in joint mean-covariance regression model for longitudinal data [J]. Annals of the Institute of Statistical Mathematics, 2013, 65(4): 617—638.

[15] ZHENG X Y, FUNG W K, ZHU Z Y. Variable selection in robust joint mean and covariance model for longitudinal data analysis[J]. Statistica Sinica, 2014, 24(2): 515—531.

[16] LYU J, GUO C H, YANG H, et al. A moving average Cholesky factor model in covariance modeling for composite quantile regression with longitudinal data[J]. Computational Statistics & Data Analysis, 2017, 112(8): 129—144.

[17] LYU J, GUO C H, WU J B. Subject-wise empirical likelihood inference for robust joint mean-covariance model with longitudinal data[J]. Statistics and Its Interface, 2019, 12(4): 617—630.

[18] 冯琳璎. 转录因子结合的协同性及其对基因表达的调控[D]. 呼和浩特: 内蒙古大学, 2019.
FENG Lin-ying. Cooperativity of transcription factor binding and its regulation role on gene expression[D]. Hohhot: Inner Mongolia University, 2019.

[19] CHUN H, KELES S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2010, 72(1): 3—25.

[20] WANG L F, CHEN G, LI H Z. Group SCAD regression analysis for microarray time course gene expression data[J]. Biometrics, 2007, 23(12): 1486—1494.

责任编辑:李翠薇