基于惩罚二次推断函数的纵向数据罕见变异关联分析*

曹红艳[1]　沈　阳[1]　王菊平[1]　曾　平[2]　房瑞玲[1]　罗艳虹[1]　余红梅[1]　张岩波[1△]

【提　要】 **目的**　提出基于惩罚二次推断函数（penalized quadratic inference function，pQIF）的纵向数据罕见变异关联分析方法，以筛选出对人类复杂疾病有重要影响的遗传变异位点或基因，为人类复杂疾病相关基因的识别提供方法学支撑。**方法**　在pQIF框架下，结合加权合计检验（weighted sum statistic，WSS）罕见变异分析方法，对基因内所有变异进行加权相加，将新的基因得分变量作为自变量，构建pQIF纵向数据罕见变异关联分析方法。以连续变量为因变量，进行模拟研究，综合评价模型参数估计和基因筛选情况。进一步将其应用于GAW18真实数据分析，选取高血压相关的重要通路肾素-血管紧张素系统（renin-angiotensin system，RAS），以收缩压（systolic blood pressure，SBP）为因变量，识别高血压相关重要基因。**结果**　模拟研究表明，对于连续因变量，pQIF的参数估计精度显著优于未惩罚的QIF。pQIF以较高的正确选择率，极低的错误选择率，能够正确地选择出效应基因，且在不同作业相关矩阵下，pQIF的参数估计和变量选择具有稳健性和一致性。在RAS通路数据分析中，pQIF识别出*THOP*1基因和*ACE*基因。**结论**　构建了pQIF纵向数据罕见变异关联分析方法，为纵向二代测序数据分析提供方法支撑。

【关键词】 纵向数据　罕见变异关联分析　惩罚二次推断函数　加权合计检验
【中图分类号】 R195.1　　【文献标志码】 A　　**DOI**　10.11783/j.issn.1002-3674.2026.04.006

Longitudinal Data Analysis for Rare Variants Detection with Penalized Quadratic Inference Function

Cao Hongyan, Shen Yang, Wang Juping, et al (*Division of Health Statistics*, *Shanxi Provincial Key Laboratory of Biomedical Data and Statistics*, *Shanxi Medical University*, *Taiyuan* 030001)

【**Abstract**】 **Objective**　This study proposes a method for the rare variants association tests to genetic longitudinal studies under the penalized quadratic inference function (pQIF) framework, aiming to identify genetic variants or genes that have a significant impact on human complex diseases. **Methods**　Within the pQIF framework, the weighted sum statistic (WSS) is adopted to collapse multiple variants in a gene region to form a gene score, which is subsequently treated as a covariate in constructing a pQIF-based longitudinal rare variant association analysis method. Simulation studies with continuous outcomes are conducted to comprehensively evaluate parameter estimation performance and gene selection accuracy. The proposed method is further applied to real data from GAW18. Focusing on the renin - angiotensin system (RAS), a key pathway related to hypertension, systolic blood pressure (SBP) is used as the outcome to identify important genes associated with hypertension. **Results**　Compared with unpenalized QIF methods, pQIF methods achieve higher estimation accuracy and greater selection efficiency. Moreover, pQIF remains optimal even when the working correlation structure is misspecified. Within the renin-angiotensin system (RAS) pathway, the pQIF identifies *THOP*1 and *ACE* genes. **Conclusion**　Our pQIF method provides a general tool for the analysis of longitudinal next-generation sequencing data.

【**Key words**】 Longitudinal data; Rare variants association test; Penalized quadratic inference function; Weighted sum statistic

全基因组关联研究（genome-wide association studies，GWAS）通过对成千上万的单核苷酸多态性（single nucleotide polymorphism，SNP）关联分析，识别出了一系列与人类复杂性状和疾病相关的遗传变异位点，提升了人们对复杂疾病和遗传变异的认识[1-2]。截至2024年7月，可访问、可重复使用的GWAS数据包含625113个SNP位点，与超过15500个性状相关，对人类复杂疾病研究具有重要意义[3]。

常规的GWAS主要针对常见变异的关联分析，将最小等位频率（minor allele frequency，MAF）≥5%的SNP视为常见变异，MAF<5%的SNP为罕见变异。常见遗传变异位点仅能解释极小部分的复杂疾病的遗传度，即存在“遗传缺失”，罕见变异是遗传缺失的重要原因[4]。罕见变异在复杂疾病的发生、发展过程中起着举足轻重的作用[5]，罕见变异发生率极低，基于单个位点的关联分析方法统计效能过低[6]。因此，以基因为单位，联合分析一组罕见变异，从而增强变异的信号强度并实现数据降维，成为提高罕见变异关联分析统计效能的有效策略，如经典的基于合并思想的加权合计检验（weighted sum statistic，WSS）法[7]。

为了深入探究复杂疾病与遗传变异之间的关联，对同一研究对象在不同时间点对特定指标进行重复测量形成纵向数据。针对纵向数据模型，Qu等[8]在广义估计方程（generalized estimating equations，GEE）[9]

* 基金项目：国家自然科学基金（82473739；82273742）；山西省基础研究计划（202303021211130），山西省回国留学人员科研资助项目（2024-081）
1. 山西医科大学卫生统计教研室，生物医学数据与统计山西省重点实验室（030001）
2. 徐州医科大学公共卫生学院流行病与卫生统计学系
△ 通信作者：张岩波，E-mail：sxmuzyb@126.com

的基础上提出了二次推断函数(quadratic inference function, QIF)。Cho和Qu提出了惩罚QIF(penalized quadratic inference function, pQIF)[10],pQIF能够适用于自变量维数随样本量增长的情形且具有oracle性质。由于罕见变异极低的发生率,不能直接应用pQIF进行纵向数据罕见变异关联分析。

本文将WSS罕见变异分析方法同pQIF有机结合,构建pQIF纵向数据罕见变异关联分析方法。以连续变量为因变量,进行模拟研究,比较不同模型条件下pQIF参数估计和基因筛选情况。进一步将其应用于GAW18[11]真实数据分析,在高血压相关的重要通路肾素-血管紧张素系统(renin-angiotensin system, RAS)上,以收缩压(systolic blood pressure, SBP)为因变量,识别高血压关联基因。pQIF纵向数据罕见变异关联分析,将为复杂疾病基因筛选提供有效的方法支撑。

资料与方法

1. 研究资料

数据来源于GAW18真实血压表型数据,考虑142例独立个体的全基因组测序数据(whole genome sequencing data, WGS)及其相对应的前三次纵向收缩压SBP数据(第一次测量了133例,第二次89例,第三次92例),同时包括是否服用高血压药物、年龄、性别和吸烟状态等。三次测量的相关特征和SBP数据描述如表1,其中年龄在三次测量的缺失数分别为9例,53例和50例;吸烟状态在三次测量的缺失数分别为11例,61例和50例。为了校正服用高血压药物所带来的影响,若测量时服用了高血压药物,则相应的SBP增加10mm Hg[12]。对同一个体不同测量时点的协变量缺失,年龄的缺失采用在前一次年龄基础上加上平均间隔测量时间进行填补。吸烟状态的缺失采用最近距离填补法进行填补,充分利用纵向数据的自身对照设计:通过个体自身的邻近测量值进行填补,若三次吸烟状态全部缺失,则以GAW18数据中的吸烟分布比例(22.9%)填补第一次测量,再按照吸烟转为不吸烟的比例填补第二、三次的吸烟状况缺失。pQIF适用于因变量不平衡数据,因此,血压表型值的缺失无需填补。

表1 三次测量的数据描述

变量	特征	第一次测量	第二次测量	第三次测量
年龄/岁($\bar{X}\pm S$)		53.30±15.91	56.38±12.96	58.24±11.89
性别/n(%)	男	61(42.96)	61(42.96)	61(42.96)
	女	81(57.04)	81(57.04)	81(57.04)
吸烟/n(%)	吸烟	33(23.24)	14(17.28)	14(15.22)
	不吸烟	109(76.76)	67(82.72)	78(84.78)
SBP/mmHg ($\bar{X}\pm S$)		128.67±21.71	134.79±20.42	130.10±18.87

注:性别基于全部142例样本统计;其余变量为各次测量时的观测值,存在缺失。

2. 二次推断函数

纵向数据分析的模型可表示为:

$$E(Y_{it})=\mu_i, g(\mu_i)=X_i'\beta \quad (1)$$

其中,Y_{it}表示个体i在时间点t的观测结果,X_{it}表示相应的p_n维协变量向量,$Y_i=(Y_{i1},\cdots,Y_{it})^T$,$X_i=(X_{i1},\cdots,X_{it})^T$。Liang等[9]通过定义伪得分方程(quasi-score equations),只需指定作业相关矩阵,就能获得参数的一致性估计。GEE伪得分方程定义为[13]:

$$U=\sum_{i=1}^{N}D_i^TA_i^{-\frac{1}{2}}R_i^{-1}(\alpha)A_i^{-\frac{1}{2}}(Y_i-\mu_i)=0 \quad (2)$$

其中,A_i为对角矩阵,$A_i=\mathrm{diag}\{\mathrm{var}(Y_{it})\}$,$D_i=\partial\mu_i/\partial\beta_j$,$R_i(\alpha)$为$Y_i$的作业相关矩阵,为了简化估计方程,可将$R_i(\alpha)$设定为一阶自回归结构[autoregressive structure, AR(1)],独立型(independent structure, INDEP)和可交换相关结构(exchange structure, EXCH)等。

QIF[8]是在GEE基础上的改进,QIF的β参数估计值较GEE精确,特别是当作业相关矩阵结构错误设定时,QIF的参数估计和变量选择结果具有稳健性和一致性。QIF采用一组基础矩阵$M_1,\cdots,M_S$的线性组合对GEE的作业相关结构矩阵$R(\alpha)$的逆进行转换,即

$$R^{-1}(\alpha)\approx a_0I+a_1M_1+\cdots+a_SM_S \quad (3)$$

其中I为单位矩阵,$M_1,\cdots,M_S$为已知的对称矩阵,$a_0,\cdots,a_S$为未知常数。将式(3)代入到GEE的伪得分方程(2)中,即得如下函数

$$\sum_{i=1}^{N}D_i^TA_i^{-\frac{1}{2}}(a_0I+a_1M_1+\cdots+a_SM_S)A_i^{-\frac{1}{2}}(Y_i-\mu_i) \quad (4)$$

将式(4)伪似然函数进行转化,定义扩展伪似然函数:

$$\psi_i(\beta)=\begin{pmatrix}\dot{\mu}_i^TA_i^{-1}(Y_i-\mu_i)\\ \dot{\mu}_i^TA_i^{-1/2}M_1A_i^{-1/2}(Y_i-\mu_i)\\ \vdots\\ \dot{\mu}_i^TA_i^{-1/2}M_SA_i^{-1/2}(Y_i-\mu_i)\end{pmatrix} \quad (5)$$

$\bar{\psi}_n(\beta)=\frac{1}{n}\sum_{i=1}^{n}\psi_i(\beta)$,函数(4)相当于式(5)元素的线性组合,构建QIF函数:

$$Q_n(\beta)=\bar{\psi}_n^T(\beta)\bar{C}_n^{-1}\bar{\psi}_n(\beta) \quad (6)$$

其中 $\bar{C}_n = 1/n \sum_{i=1}^{n} \psi_i(\beta)\psi_i^T(\beta)$，$\bar{C}_n$ 为 $\Omega = \mathrm{var}(\psi_i)$ 的一致估计。

$$\hat{\beta} = \mathrm{argmin}_{\beta} Q_n(\beta) \quad (7)$$

β 的估计可以通过最小化 $Q_n(\beta)$ 得到一致的估计值，同时还可构建出 $\hat{\beta}$ 的渐进正态分布，从而对 $\hat{\beta}$ 进行统计推断。

3. 惩罚二次推断函数

Cho 和 Qu（2013）提出了基于平滑剪切绝对偏差（smooth clipped absolute deviation，SCAD）惩罚[14]的 pQIF[10]，适用于自变量随样本含量增大的情况，不需要设定似然函数，对非零变量的识别率依概率收敛到 1。非凸 SCAD 惩罚函数如下：

$$p'_{\lambda_n}(\theta) = \lambda_n \left\{ I(\theta \leqslant \lambda_n) + \frac{(a\lambda_n - \theta)_+}{(a-1)\lambda_n} I(\theta > \lambda_n) \right\} \quad (8)$$

其中，λ 为调和参数，I 为指示函数，a 为预先选择的常数，函数“+”表示取结果为正的部分，$a > 2, \theta > 2$，SCAD 惩罚函数能同时选择和估计系数，并具有 oracle 性质。pQIF 在 QIF 的 $Q_n(\beta)$ 统计量上，增加了一项 SCAD 惩罚函数，即为：

$$U_n(\beta) = Q_n(\beta) + n \sum_{j=1}^{p} p_{\lambda_n}(|\beta_j|) \quad (9)$$

通过最小化 $U_n(\beta)$ 以得到 β 的估计值，由于 SCAD 惩罚函数是非凸的，可采用局部二次近似法最小化惩罚得分函数式（9）进行参数估计。

pQIF 在 QIF 目标函数 $Q_n(\beta)$ 基础上，构建类似于 BIC 的调和参数选择标准，定义为贝叶斯信息二次推断函数准则（Bayesian information quadratic inference function criterion，BIQIF）：

$$BIQIF_{\lambda_n} = Q_n(\hat{\beta}_{\lambda_n}) + df_{\lambda_n} \log(n) \quad (10)$$

其中，$\hat{\beta}_{\lambda_n}$ 为给定 λ_n 时，最小化惩罚得分函数 $U_n(\beta)$ 得到的边缘回归参数，df_{λ_n} 为 $\hat{\beta}_{\lambda_n}$ 中非零系数的个数。选择最小 BIQIF 相对应的 λ_n，为最优调和参数。

4. pQIF 纵向数据罕见变异关联分析

由于罕见变异极低的发生率，不能直接应用 pQIF 进行纵向数据罕见变异关联分析。本文将 WSS 罕见变异分析方法同 pQIF 有机结合，构建 pQIF 纵向数据罕见变异关联分析方法。

采用 WSS 罕见变异分析方法[7]，对基因内一组变异进行加权求和得到一个新的基因得分变量，假定基因内包含 J 个 SNP，在可加遗传模型下，个体 i 在位点 j 的等位基因数目为 G_{ij}，$G_{ij} = 0, 1, 2$，对于给定的基因，每个个体在该基因上的基因得分为：

$$C_i = \sum_{j=1}^{J} w_j G_{ij} \quad (11)$$

其中，w_j 为等位基因的样本标准差的倒数，如 $w_j = 1/\sqrt{p_j(1-p_j)}$，$p_j$ 为 SNP_j 的最小等位频率 MAF，对罕见变异赋予更大的权重。得到每个基因的基因得分后，将其引入到纵向数据分析模型中：

$$E(Y_{it}) = \mu_{it},\ g(\mu_{it}) = \alpha_0 + \sum_{k=1}^{K} \gamma_k E_{ikt} + \sum_{j=1}^{p_n} \beta_j C_{ij} \quad (12)$$

其中，E_{kt} 为协变量，C_{ij} 为第 i 个个体第 j 个基因的得分，不随时间的变化而变化。采用 pQIF 的方法对模型（13）进行参数估计和变量选择，即可实现基因筛选。

5. 模拟研究

基于 QIF 和 pQIF 纵向数据的稀有变异关联分析，探讨因变量为连续变量时的模型拟合效果。考察在三种不同协方差矩阵结构［AR（1）、EXCH、INDEP］下，QIF 和 pQIF 参数估计及变量选择的稳健性与一致性。

采用 GAW18 真实测序数据，选取其中的 142 例独立个体，同时，采用 bootstrap 抽样，从 142 例个体中分别抽取不同样本含量的样本。从“GAW18−The answers”中提供的功能基因中选择 5 个基因：*TNN*、*MAP*4、*LEPR*、*FLT*3 *和 NRF*1，根据 WSS 法计算基因得分。噪声基因设置为包含 50 个 SNP，罕见变异与常见变异比例为 6∶4，在可加遗传模型下，SNP 的等位基因数目编码为{ 0, 1, 2 }。年龄缺失采用均值填补法进行填补，吸烟状态根据 GAW18 数据分布模拟生成。模拟三个不同时间点的连续纵向血压值 Y_{it}，生成 200 个样本。具体模型构造如下：

$$Y_{it} = \beta_0 + \beta_{age} age_{it} + \beta_{smoke} smoke_{it} + \sum_{j=1}^{p_n} \beta_j C_{ij} + \varepsilon_{it},\ t = 1, \cdots, 3 \quad (13)$$

其中，age_{it} 和 $smoke_{it}$ 分别为第 i 个个体在 t 时刻的年龄和吸烟状态，随机误差 $(\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3})^T$ 假定服从边缘均值为 0，边缘方差为 1 以及相关矩阵为参数 ρ 的 AR（1）结构。

（1）pQIF 已知相关结构时的模拟研究

假定已知 AR（1）为纵向数据协方差结构，则设定作业相关结构为 AR（1），考虑连续因变量时，不同模拟设置情况的 pQIF 拟合效果。考虑三种样本量：n = 142，n=200 和 n=500，其中 n=200 和 n=500 为 bootstrap 抽样样本。考虑 p=20 和 p=40 两种自变量维度，当 p=20 时，$\beta = (0.9, -0.7, -0.6, 0.5, 0, \cdots, 0)^T$，年龄和三个效应基因 *TNN*，*MAP*4，*NRF*1 对应非零系数；当 p=40 时，$\beta = (0.9, -0.7, -0.7, 0.6, 0.6, 0.5, 0, \cdots, 0)^T$，年龄和五个效应基因 *TNN*、*MAP*4、*LEPR*、*FLT*3 和 *NRF*1 对应非零系数。组内相关系数 ρ=0.4 和 ρ=0.8。

（2）pQIF 不同相关结构模拟研究

探讨连续因变量情况下不同作业相关结构时 pQIF 的稳健性。在三种不同协方差矩阵结构［AR

(1)、EXCH、INDEP]下，对于每一种协方差结构，分别拟合不同作业相关结构。以 n=300，p=20，ρ=0.5 为例，$\beta=(0.9,-0.7,-0.6,0.5,0,\cdots,0)^T$，年龄和效应基因对应非零系数。

(3)模型评价指标

计算总均方误差(total mean squared error, TMSE)和噪声变量均方误差(mean squared error for noisy variables, NMSE)：

$$TMSE=\frac{1}{200}\sum_{j=1}^{200}\left\|\hat{\beta}^{(j)}-\beta\right\|^2/p \qquad (14)$$

其中，p 为 β 的维度，$\hat{\beta}^{(j)}$ 为第 j 次迭代时 β 的估计系数，噪声变量的NMSE类推。进一步计算200次模拟结果的总误差平方中位数(total median of squared error, TMe)，以及噪声变量误差平方中位数(median of squared error for noisy variables, NMe)，以评价模型参数估计准确性。

同时，计算模型相对预测误差(relative prediction error, RPE)[15]及其标准误，以评价不同模型的预测精度。假定随机误差服从均值为0，边缘方差为1的多元正态分布，相应的RPE为：

$$RPE=E[(X^T\hat{\beta}-X^T\beta)^2] \qquad (15)$$

对每个变量分别计算选择率，包括非零系数变量的正确选择率(true positive rate, TPR)以及噪声变量的错误选择率(false positive rate, FPR)，以评价模型变量选择准确性。

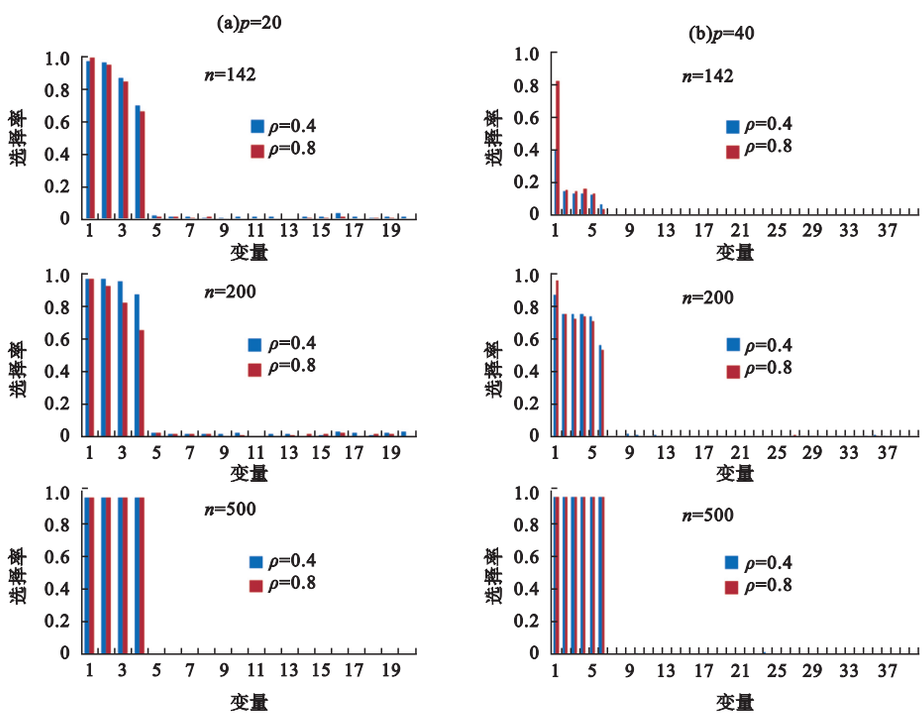
本文所有分析均通过R软件实现。

结　果

1. 模拟研究结果

(1)pQIF在已知相关结构时的模拟结果

图1为pQIF在不同模型条件下的变量选择结果，横坐标表示变量编号，横轴的编号1为年龄，当p=20时，横轴的编号2~4表示效应基因TNN、MAP4和NRF1，编号5~20为噪声变量；当p=40时，横轴的编号2~6表示五个效应基因TNN、MAP4、LEPR、FLT3和NRF1，编号7~40为噪声变量。由图1可知，pQIF对噪声变量的选择率(FPR)非常低，随着样本量的增大，FPR接近或等于0；pQIF对非零系数变量的选择率(TPR)受维度和样本量的影响，当n=142，p=40时，尽管200次模拟均能收敛，但是有关pQIF的模型选择指标BIQIF失效，将系数全惩罚为0，无法得到有意义的结果，随着样本含量的增大，TPR大幅度增加，由组内相关系数所引起的TPR差别减小；当样本量达到500时，不同p和ρ所对应的TPR以及FPR均达到理想状态。



注：(a)中的非零系数变量为横坐标变量编号1~4，编号5~20为噪声变量；(b)中非零系数变量为编号1~6，而编号7~40为噪声变量。非零系数变量对应的选择率为TPR，噪声变量对应的选择率为FPR。

图1　连续pQIF在不同模型条件下的变量选择结果

表2列出了QIF和pQIF在不同模型条件下的参数估计精度，从表2可知，除了 n=142，p=40之外，惩罚QIF的TMe、NMe以及RPE均显著小于未惩罚的QIF，随着样本量的增大，惩罚QIF的参数估计精度和oracle QIF相近或相等，pQIF NMe均为0，表明一般而言，pQIF参数估计精度显著地优于QIF。当样本量较

表 2　连续 QIF 和 pQIF 在不同模型条件下的参数估计精度

	n	方法	ρ=0. 4			ρ=0. 8		
			TMe	NMe	RPE(SE)	TMe	NMe	RPE(SE)
p=20	n=142	oQIF	0. 0007	–	0. 0129(0. 0007)	0. 0009	–	0. 0189(0. 0014)
		QIF	0. 0051	0. 0051	0. 0995(0. 0028)	0. 0078	0. 008	0. 1529(0. 0030)
		pQIF	0. 0016	0	0. 0295(0. 0029)	0. 0024	0	0. 0432(0. 0045)
	n=200	oQIF	0. 0005	–	0. 0095(0. 0008)	0. 0008	–	0. 0151(0. 0011)
		QIF	0. 0033	0. 0032	0. 0650(0. 0023)	0. 0005	0. 0048	0. 0982(0. 0030)
		pQIF	0. 0009	0	0. 0175(0. 0013)	0. 0012	0	0. 0221(0. 0019)
	n=500	oQIF	0. 0002	–	0. 0034(0. 0002)	0. 0003	–	0. 0049(0. 0004)
		QIF	0. 0011	0. 0011	0. 0231(0. 0008)	0. 0017	0. 0017	0. 0345(0. 0009)
		pQIF	0. 0002	0	0. 0038(0. 0003)	0. 0003	0	0. 0052(0. 0003)
p=40	n=142	oQIF	0. 0006	–	0. 0243(0. 0009)	0. 0009	–	0. 0332(0. 0020)
		QIF	0. 0092	0. 0092	0. 3654(0. 0116)	0. 0142	0. 0145	0. 5658(0. 0200)
		pQIF	0. 0690	0	2. 5863(0. 0779)	0. 0488	0	1. 8049(0. 0013)
	n=200	oQIF	0. 0004	–	0. 0134(0. 0008)	0. 0005	–	0. 0207(0. 0011)
		QIF	0. 0044	0. 0044	0. 1730(0. 0049)	0. 0065	0. 0066	0. 2603(0. 0070)
		pQIF	0. 0012	0	0. 0452(0. 0150)	0. 0017	0	0. 0602(0. 0419)
	n=500	oQIF	0. 0001	–	0. 0055(0. 0005)	0. 0002	–	0. 0086(0. 0004)
		QIF	0. 0013	0. 0013	0. 0503(0. 0009)	0. 0019	0. 0019	0. 0767(0. 0016)
		pQIF	0. 0002	0	0. 0072(0. 0004)	0. 0003	0	0. 0097(0. 0005)

注：TMe 为总误差平方中位数；NMe 为噪声变量误差平方中位数；oQIF= oracle QIF，oQIF 不存在噪声变量，无 NMe 值，用“–”表示。

小时，p=20 的参数估计精度显著地优于 p=40，随着样本量的增大，二者之间的差距缩小，当 n=500 时，p=20 和 p=40 的 TMe 相同，RPE 相近；当样本含量较小时，ρ=0. 4 的 TMe 和 RPE 均小于 ρ=0. 8，随着样本量的增大，ρ=0. 4 和 ρ=0. 8 的参数估计精度相近。

（2）pQIF 不同相关结构模拟结果

图 2 表示因变量为连续变量时，pQIF 在不同作业相关矩阵下的变量选择结果，由图可见，pQIF 在不同作业相关结构下，变量选择结果基本一致，表现出良好的稳健性和一致性。表 3 为 pQIF 在不同作业相关矩阵下的模型估计精度，针对同一真实的协方差结构，采用三种不同的作业相关矩阵进行拟合，其参数估计精度未见明显差异，进一步说明 pQIF 在不同作业相关结构矩阵下，其参数估计和变量选择的稳健性和一致性。

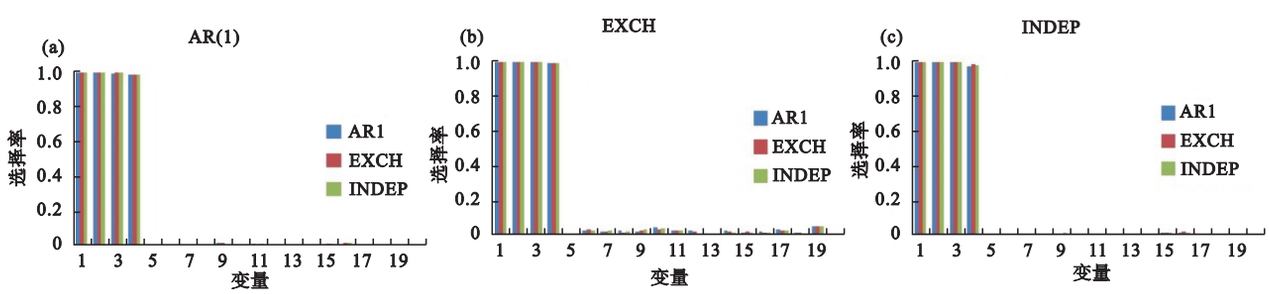


图 2　连续 pQIF 在不同作业相关矩阵下的变量选择结果

表 3　连续 pQIF 在不同的作业相关矩阵下的模型估计精度

真实相关结构	作业相关矩阵	NMSE	TMSE	RPE(SE)
AR(1)	AR(1)	0	0. 0009	0. 0090(0. 0007)
	EXCH	0	0. 0007	0. 0093(0. 0006)
	INDEP	0	0. 0007	0. 0088(0. 0007)
EXCH	AR(1)	0	0. 0009	0. 0096(0. 0006)
	EXCH	0	0. 0008	0. 0089(0. 0006)
	INDEP	0	0. 0008	0. 0096(0. 0008)
INDEP	AR(1)	0	0. 0005	0. 0060(0. 0006)
	EXCH	0	0. 0005	0. 0066(0. 0004)
	INDEP	0	0. 0005	0. 0064(0. 0006)

注：NMSE 表示噪声变量均方误差；TMSE 表示总均方误差；RPE(SE)为相对预测误差及其标准误。

2. 真实数据分析

应用 QIF 和 pQIF 分析 GAW18 真实的血压表型数据，以 SBP 为因变量，基于通路进行分析。以信号通路为单位，研究信号通路上的一系列基因对疾病的影响尤为重要，其中 RAS 是研究高血压遗传机制的重要通路之一，在血压调控中起着关键的作用。若 RAS 异常激活，血压会突然升高，很多药物通过抑制 RAS 来控制血压。因此，基于 RAS 进行纵向数据罕见变异关联分析，研究该通路内高血压相关基因。从 KEGG 通路数据库中选取 RAS 通路（http://www. kegg. jp/entry/map04614. html）[16]。

RAS 通路相关的 13 种蛋白质的编码基因分别为：

AGT、*REN*、*ACE*、*KLK*2、*KLK*1、*ANPEP*、*CPA*3、*MME*、*THOP*1、*NLN*、*PRCP*、*MRGPRD*、*AGTR*1 和 *LNPEP*，其中 *KLK*2、*KLK*1 共同编码 Tonin 蛋白质。以 SBP 为因变量，13 个编码基因为自变量，其他协变量包括年龄、性别和吸烟状况。表 4 表示在 RAS 通路中，QIF 和 pQIF 参数估计结果。pQIF 选择了年龄和基因 *ACE*、*THOP*1，吸烟和性别以及其他 12 个基因的系数均被惩罚为 0，而未惩罚的 QIF 却未达到稀疏性，因此无法实现变量选择的目的。

基因 *ACE* 包含 128 个 SNP，其中 95 个为罕见变异，33 个为常见变异；而 *THOP*1 基因包含 177 个 SNP，其中 139 个为罕见变异，38 个为常见变异。两个基因的 pQIF 系数均为正值，表明 *ACE* 和 *THOP*1 基因各自的综合效应为高血压的危险因素，对高血压的发生、发展将产生重要的影响。

表 4 QIF 和 pQIF 在 RAS 通路中的参数估计值

变量	QIF(SE)	pQIF
截距项	0.0904(0.0725)	–
年龄	0.6113(0.0747)	0.5436
性别	−0.0011(0.0741)	–
吸烟	0.1547(0.0544)	–
AGT	0.1327(0.0708)	–
REN	−0.1348(0.0809)	–
ACE	0.2247(0.0677)	0.0010
*KLK*2	0.0123(0.0615)	–
*KLK*1	−0.0069(0.0496)	–
ANPEP	0.0102(0.0612)	–
*CPA*3	0.1353(0.0699)	–
MME	0.0657(0.0658)	–
*THOP*1	0.1988(0.0484)	0.0599
NLN	0.0245(0.0440)	–
PRCP	0.0776(0.0325)	–
MRGPRD	0.0611(0.0770)	–
*AGTR*1	−0.1024(0.0666)	–
LNPEP	0.0555(0.0620)	–

注：SE 为估计系数的标准误；“–”表示变量系数惩罚为 0。

讨 论

本文将 WSS 罕见变异分析和 pQIF 纵向数据分析相结合，提出了 pQIF 纵向数据罕见变异关联分析方法。惩罚的 QIF 的参数估计精度均优于未惩罚的 QIF，pQIF 的错误选择率非常低，对不同作业相关矩阵具有稳健性和一致性。

在 RAS 信号通路分析中，pQIF 筛选出 *THOP*1 和 *ACE* 基因。其中 *THOP*1 基因编码的酶可促进血管紧张素Ⅰ转化为血管紧张素(1−7)[17]，而血管紧张素(1−7)通过 MAS 受体(血管紧张素 1−7 特异性受体)发挥血管保护作用，可抑制血管平滑肌增殖、改善内皮功能，从而对血压具有调节作用[18]。相反，*ACE* 基因编码的血管紧张素转换酶可促进血管紧张素Ⅰ转化为血管紧张素Ⅱ，血管紧张素Ⅱ通过 AT1 受体(血管紧张素Ⅱ受体 1 型)引起血管收缩和钠水潴留，从而升高血压[19]。因此，*THOP*1 和 *ACE* 基因的表达变化可能通过调节血管紧张素(1−7)/血管紧张素Ⅱ平衡，从而影响血管功能和血压水平。

pQIF 纵向数据罕见变异分析仍存在不足。首先，我们采用 WSS 的方法对基因内的一组变异进行集成，假定基因内所有位点对疾病的效应方向相同，一旦违背了这一基本的假设，将导致基因效应相互抵消，造成效能损失[20]。为了克服效应抵消的问题，可采用 Han 和 Pan 提出的自适应检验法(aSum)[21]。本文提供了如何将罕见变异分析方法和 pQIF 有机结合的一种思路，基于 aSum 的 pQIF 关联分析不难实现。其次，本文的 pQIF 模型中，每个自变量的系数不随时间的变化而变化，因此无法捕获基因随时间变化的动态效应，今后的研究中，我们将在惩罚 QIF 的框架下，研究非参变系数模型[22]。

综上所述，本文将 WSS 罕见变异分析方法和 pQIF 相结合，提出了 pQIF 纵向数据罕见变异关联分析方法，具有较好的参数估计准确性，对不同作业相关矩阵具有稳健性和一致性。pQIF 纵向数据罕见变异关联分析方法为纵向二代测序数据分析提供了统计学方法支撑。

参 考 文 献

[1] Abdellaoui A, Yengo L, Verweij KJH, et al. 15 years of GWAS discovery: realizing the promise[J]. The American Journal of Human Genetics, 2023, 110(2): 179-194.

[2] Uffelmann E, Huang QQ, Munung NS, et al. Genome-wide association studies[J]. Nature Reviews Methods Primers, 2021, 1(1):59.

[3] Cerezo M, Sollis E, Ji Y, et al. The NHGRI-EBI GWAS Catalog: standards for reusability, sustainability and diversity[J]. Nucleic Acids Research, 2025, 53(D1): D998-D1005.

[4] Young AI. Solving the missing heritability problem[J]. PLoS Genetics, 2019, 15(6):e1008222.

[5] 夏德润, 沃红梅, 顾珩, 等. 测序数据罕见遗传变异关联性分析方法评价[J]. 中国卫生统计, 2020, 37(5):649-653.

[6] Hecker J, Townes FW, Kachroo P, et al. A unifying framework for rare variant association testing in family-based designs, including higher criticism approaches, SKATs, and burden tests[J]. Bioinformatics, 2020, 36(22-23):5432-5438.

[7] Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic[J]. PLoS Genet, 2009, 5(2): e1000384.

[8] Qu A, Lindsay BG, Li B. Improving generalised estimating equations using quadratic inference functions[J]. Biometrika, 2000, 87(4): 823-836.

[9] Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models[J]. Biometrika, 1986, 73(1): 13-22.

（下转第 524 页）

层义诊和免费体检活动，提升居民对基层医疗服务的认知水平。加强基层慢性病健康管理，利用联合诊疗、对口支援、技能培训等方式，形成疾控中心、基层医疗机构和上级医院分级协同的慢性病诊疗机制，为患者提供精准、连续、全程的医疗服务[18]。扩展慢性病用药目录，规范医疗服务项目价格，门诊统筹报销比例进一步向基层倾斜，提升基层医疗服务的性价比，有效减轻居民经济负担。

4. 创新与不足

本研究利用离散选择实验方法，选取影响医疗服务利用的6个关键属性及其水平，探究小病假设下重庆市居民医疗服务选择的偏好及异质性，综合考量各区县人均综合卫生资源当量、经济水平及地理位置、常住居民年龄分布进行分层抽样，利用混合logit、潜类别logit模型进行同步分析，保证了研究人群的代表性和结果的可靠性。因研究方法对属性数量的限制，可能导致研究结果在一定程度上具有局限性。

参 考 文 献

[1] 重庆市卫生健康委员会. 重庆市医改办举办2023年三明医改经验专题培训班[EB/OL]. (2023-09-22)[2025-06-25]. https://wsjkw.cq.gov.cn/zwxx/bmdt/zyhd/202309/t20230922_12369054.html.

[2] 赵美英,苗艳青. 新中国70年基层卫生发展回顾与展望[J]. 中国卫生政策研究,2019,12(11):10-15.

[3] 郑蕾. 分级诊疗制度建设的影响因素剖析及建议[J]. 中国卫生经济,2019,38(9):12-15.

[4] 黎相麟. 分级诊疗视角下武汉市居民医疗服务选择偏好及支付意愿研究[D]. 武汉:华中科技大学,2021.

[5] 张研,唐文熙,孙晓伟. 居民自由就医对我国卫生服务体系整合的影响[J]. 中国卫生事业管理,2024,31(9):678-680.

[6] Ryan M, Mcintosh E, Dean T. Trade-offs between location and waiting times in the provision of health care: the case of elective surgery on the Isle of Wight [J]. Journal of Public Health Medicine, 2008, 29(26): 3503-3513.

[7] Green C, Gerard K. Exploring the social value of health-care interventions:a stated preference discrete choice experiment [J]. Health Econ, 2009, 18(8): 951-976.

[8] 王翠连,赵林海,顾源远. 安徽省城镇老人对家庭医生服务的选择偏好:基于离散选择实验[J]. 现代预防医学,2021,48(7):1211-1214.

[9] 曹晓婧,戴力辉,王燕. 基于混合logit模型的我国农村居民就医选择的影响因素分析[J]. 卫生软科学,2022,36(3):82-86.

[10] 李程洪,陈景润,向晨. 武汉市某医院患者对互联网医院的选择偏好研究[J]. 医学与社会,2023,36(10):79-84.

[11] 张娇. 供需视角下家庭医生签约服务偏好研究:基于离散选择实验[D]. 济南:山东大学,2021.

[12] 巩田田. 我国6月~5岁儿童父母疫苗接种偏好分析:基于离散选择实验[D]. 济南:山东大学,2021.

[13] 蒋明珠,熊巨洋,张晗. 离散选择实验在居民基层医疗服务利用中的应用及偏好异质性分析[J]. 中国卫生经济,2019,38(4):55-59.

[14] 李明,张枫艳,王素梅. 某院互联网医院云药房运行模式和成效分析[J]. 中国处方药,2024,22(1):52-54.

[15] 王婕妤. 全市基层卫生健康工作会议绘制今年“施工图”群众在基层就近就医看好病[N]. 南京日报,2024-04-05 (A02).

[16] 中华人民共和国中央人民政府. 卫生健康委 发展改革委 财政部 人力资源社会保障部 医保局关于印发开展促进诊所发展试点意见的通知[EB/OL]. (2019-04-28)[2025-06-05]. https://www.gov.cn/gongbao/content/2019/content_5425334.htm.

[17] 徐晓丹,吴文强. 我国城乡中老年群体基层就医的影响因素分析:基于CHARLS数据的实证研究[J]. 中国卫生政策研究,2016,9(4):23-30.

[18] 王曼菲,陈馨仪,王安石. 基层医疗卫生机构的慢性病管理:经验、问题和建议: 基于案例分析[J]. 卫生经济研究,2022,39(3):46-49.

（责任编辑：张　悦）

（上接第519页）

[10] Cho H, Qu A. Model selection for correlated data with diverging number of parameters[J]. Statistica Sinica, 2013, 23: 901-927.

[11] Almasy L, Dyer TD, Peralta JM, et al. Data for Genetic Analysis Workshop 18: human whole genome sequence, blood pressure, and simulated phenotypes in extended pedigrees[J]. BMC Proc, 2014, 8 (Suppl 1): S2.

[12] Tobin MD, Sheehan NA, Scurrah KJ, et al. Adjusting for treatment effects in studies of quantitative traits: antihypertensive therapy and systolic blood pressure[J]. Statistics in Medicine, 2005, 24(19): 2911-2935.

[13] 曹红艳,曾平,李治,等. 惩罚广义估计方程在纵向数据基因关联分析中的应用[J]. 中国卫生统计,2017,34(4):534-537.

[14] Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties[J]. Journal of the American Statistical Association, 2001, 96(456): 1348-1360.

[15] Zou H. The adaptive lasso and its oracle properties[J]. Journal of the American Statistical Association, 2006, 101(476): 1418-1429.

[16] Kanehisa M, Furumichi M, Tanabe M, et al. KEGG: new perspectives on genomes, pathways, diseases and drugs[J]. Nucleic Acids Res, 2017, 45(D1): D353-D361.

[17] Pereira MG, Souza LL, Becari C, et al. Angiotensin Ⅱ-independent angiotensin-(1-7) formation in rat hippocampus: involvement of thimet oligopeptidase[J]. Hypertension, 2013, 62(5): 879-885.

[18] Zhang F, Li S, Song J, et al. Angiotensin-(1-7) regulates angiotensin Ⅱ-induced matrix metalloproteinase-8 in vascular smooth muscle cells[J]. Atherosclerosis, 2017, 261: 90-98.

[19] Chen H, Peng J, Wang T, et al. Counter-regulatory renin-angiotensin system in hypertension: Review and update in the era of COVID-19 pandemic[J]. Biochemical Pharmacology, 2023, 208: 115370.

[20] Lee S, Abecasis GR, Boehnke M, et al. Rare-variant association analysis: study designs and statistical tests[J]. Am J Hum Genet, 2014, 95(1): 5-23.

[21] Han F, Pan W. A data-adaptive sum test for disease association with multiple common or rare variants[J]. Hum Hered, 2010, 70(1): 42-54.

[22] Ma SJ, Yang LJ, Romero R, et al. Varying coefficient model for gene-environment interaction: a non-linear look[J]. Bioinformatics, 2011, 27(15): 2119-2126.

（责任编辑：林美艺）