Pure Mathematics 理论数学, 2024, 14(2), 817-827
Published Online February 2024 in Hans. https://www.hanspub.org/journal/pm
https://doi.org/10.12677/pm.2024.142079

Hans汉斯

基于函数型Cox比例风险模型的基因－基因交互效应分析及在眼病中的应用

郭诗雨[1]，李运明[1,2]，郑海涛[1*]

[1]西南交通大学数学学院统计系，四川 成都
[2]西部战区总医院医疗保障中心，四川 成都

收稿日期：2024年1月7日；录用日期：2024年1月29日；发布日期：2024年2月29日

摘　要

疾病关联性研究存在大量的基因与基因的交互效应(gene-gene interaction)和基因与环境因素的交互效应(gene-environment interaction)分析，以上交互效应对个体化诊疗具有极为重要的参考价值。针对基因与基因交互效应，本文提出了一种具有函数交互效应的Cox比例风险模型。该方法将基因的多个单核苷酸多态性(SNP)之间的交互效应进行函数化处理，大大降低了待估参数的维数。基因－基因的交互作用的假设检验采用似然比(LRT)检验统计量。经模拟研究表明，所提方法能够较好地控制第I类错误率，功效也比较高。实例分析表明，利用所提出的方法能够有效地检测出与老年黄斑变性和白内障(AREDS)相关联的基因－基因交互作用。

关键词

交互效应，函数数据分析，Cox比例风险模型，SNP

Analysis of Gene-Gene Interaction Effects Based on Functional Cox Model and Its Application in Ophthalmopathy

Shiyu Guo[1], Yunming Li[1,2], Haitao Zheng[1*]

[1]Department of Statistics, School of Mathematics, Southwest Jiaotong University, Chengdu Sichuan
[2]Medical Support Center, The General Hospital of Western Theater Command, PLA, Chengdu Sichuan

Received: Jan. 7th, 2024; accepted: Jan. 29th, 2024; published: Feb. 29th, 2024

*通讯作者。

文章引用: 郭诗雨, 李运明, 郑海涛. 基于函数型 Cox 比例风险模型的基因-基因交互效应分析及在眼病中的应用[J]. 理论数学, 2024, 14(2): 817-827. DOI: 10.12677/pm.2024.142079

Abstract

Gene-Gene interaction and Gene-Environment interaction have been widely used in disease association analysis. In particular, the interaction effect of personalized medicine has a very important research value. In this paper, a Cox proportional hazards model with functional interaction effect is proposed, which mainly studies Gene-Gene interaction effect. The interaction effects between multiple Single-nucleotide polymorphism of genes (SNPs) are functionally processed, which greatly reduces the dimension of the parameters to be estimated. The likelihood ratio (LRT) test was used to test for Gene-Gene interaction. A large number of simulation studies show that the proposed method can control the Type I error rate better, and the Power is also high. The proposed method can effectively detect gene-gene interactions associated with macular degeneration and cataract (AREDS).

Keywords

Interaction Effects, Functional Data Analysis, Cox Model, SNP

1. 引言

全基因组关联分析(Genome-wide Association study, GWAS) [1] [2] [3]是疾病关联分析中应用较为广泛的分析方法。GWAS 主要是研究单核苷酸多态性(Single-nucleotide polymorphisms, SNP)与疾病的关联分析。研究人员已经利用 GWAS 识别出很多与疾病或复杂性状相关的 SNP。但大多数 SNP 的效应可能很小，也不能完全解释疾病或复杂性状的遗传效应。

进一步的研究表明基因与环境因素的交互效应对复杂性状或疾病的解释也起着极为重要的作用。由于基因主效应对于复杂性状或疾病的解释比较有限，基因与基因的交互效应(gene-gene interaction, GGI)分析得到了越来越多的研究[4] [5]，基因之间的相互作用能够影响复杂性状的表现，包括一些常见疾病。理解这些交互效应对于疾病的发病机制和个体之间的差异至关重要。基因交互效应有助于揭示遗传学中的复杂性，帮助我们更好地理解基因对生物体特征和疾病易感性的贡献[6]。

多数疾病都是由多个基因和环境因素相互作用引起的。基因之间的复杂交互可以导致疾病的多样性和不同个体之间的差异。特定基因的单一变异可能并不足以显著增加疾病风险，但是当这些基因与其他基因相互作用时，可以导致更高的疾病风险。同时基因交互作用可以从网络生物学的角度来理解[7]，即基因以复杂的网络方式相互连接。这种网络的破坏或改变可能与疾病的发生和发展密切相关。不同个体之间的基因组组合各异，这意味着对于相同疾病，可能存在不同的遗传机制。总体而言，深入研究基因与基因之间的交互作用有助于揭示疾病的遗传基础，为个性化医疗和疾病治疗提供更深入的理解和方法。

在研究遗传变量之间的交互效应中，基因与基因交互效应的分析涉及多种方法，研究人员提出了几种方法[8] [9]来检测 SNP-SNP 的交互效应，通过这种关联分析，研究人员可以检测基因变异与疾病之间的关系。这可以通过单一基因关联(单基因研究)或考虑多个基因同时的关联来进行(多基因关联研究)[10]：如基于熵的统计、Logit 模型等[11] [12] [13]；其他技术同时也包括多因素降维(Multifactor dimensionality reduction, MDR)、BOOST、RRIntCC、GenEpi [14] [15] [16] [17]以及一些加速方法[18]。这些方

法有几个潜在的优点。首先，它通常含有更少的基因，可减少成对检验的数量。由于交互效应的存在，这些方法也适用于基因的主效应研究。此外，生物学的一些先验(例如，关于蛋白质与蛋白质交互效应(Protein-Protein Interaction Network, PPI)或已知的基因关联信息)也可以很容易地引入到研究中。同时使用机器学习算法，如随机森林、支持向量机、神经网络等，也可以来探索基因之间的复杂交互效应。不过，这些方法也存在一些挑战和不足：基因交互效应分析往往涉及大量基因组数据，导致维度灾难[19] [20]，使得分析变得更为复杂。在分析多个基因或变异时，需要考虑多重比较问题，以避免虚假的关联结果。机器学习等复杂模型的解释性可能较差，难以理解具体的基因之间交互的生物学意义。最近，研究人员提出了一种称为 AGGrGATOR [21]的方法，该方法在标记水平上结合 P 值检验基因交互效应的显著性，这是用于检测定量表型下的交互效应的策略[22]。也有研究人员提出了一种基于熵的非参数方法 GBIGM [23]检测交互效应。在生物学中，植物的性状[24]也受到许多基因与基因之间交互效应的影响。有研究表明 EYA4 基因和 GRHL2 基因之间的交互效应与噪声性耳聋(Noise-induced hearing loss, NIHL)的发生存在关联[25]。基因之间的交互效应得到了越来越多的研究验证。这些方法遇到的主要问题之一是多重检验的校正会导致的高阶效应或成对检验的显著性受到影响。

同时近年来，越来越多的高频观测数据以函数曲线的形式出现，比如每分钟股票价格数据、汇率数据、气温数据、小时 PM2.5 数据、光谱数据等，这些数据常用的分析方法是函数型数据分析(Functional Data Analysis, FDA) [26]。大数据时代传统的结构化数据也从简单的点数据，扩展到区间数据、符号数据和函数型数据等[27]。函数型数据分析已经成为统计分析中越来越重要的研究方向。函数回归分析也得到了越来越广泛的研究和应用。

从很多已有的研究结果中可以看到，与疾病相关联的基因数目一般都比较大，若进一步考虑交互效应将使得模型待估参数的个数变得非常高，这给统计分析带来了很大的挑战。另外有些主效应和交互效应比较弱，检测比较困难。综合来看，要解决该问题需要综合运用多种方法来克服挑战，如何降维以及如何处理交互效应，在这里本文利用函数回归模型对主效应和交互效应进行整体考虑，基因组数据通常是高维度的，处理这些大规模的遗传信息需要发展更有效的统计和计算方法。模型中的函数效应可通过常用的基展开进行估计，这也可以有效地降低的降低模型待估参数的维数，也更有利于检测出弱效应。本文将基因位点(loci)连续化[28]，把基因的主效应和基因与基因的交互效应进行函数化处理，这种扩展的 Cox 模型对于生存分析中考虑基因与基因之间的交互作用是一种常见的方法。当考虑基因与基因的交互作用时，Cox 比例风险模型可以包含这些交互项。交互作用项评估基因之间的相互作用对生存时间的影响是否显著。基于此，我们建立了函数型交互效应 Cox 比例风险模型。

2. 模型介绍

2.1. Cox 比例风险模型

在这里我们知道 Cox 比例风险模型的基本形式为：

$$h(t,X)=h_0(t)\mathrm{e}^{\beta_1x_1+\beta_2x_2+\cdots+\beta_mx_m}$$

在上式中，$x_1,\cdots,x_m$ 表示自变量，$\beta_1,\beta_2,\cdots,\beta_m$ 为自变量的偏回归系数。$h_0(t)$ 表示 t 时刻的基准风险率函数。

2.2. 函数交互效应 Cox 比例风险模型

假设有 m 个 SNP，对应的基因位点为 $0\le u_1\le u_2\le\cdots\le u_m$，在具体分析中可以将基因位点进行归一化处理。

令 T_i 表示个体 i 的生存时间，C_i 表示其右截尾时间。设 $y_i=\min\left(T_i,C_i\right)$ 是观察到的时间，$\delta_i=1\left(y_i=T_i\right)$。此外，设 $X_i=\left(X_i\left(u_1\right),\cdots,X_i\left(u_m\right)\right)^{\mathrm{T}}$ 表示 m 个 SNP 的基因型，其中 $X_i\left(u_j\right)\left(=0,1,2\right)$ 是 u_j 位置的次等位基因(Minor Allele)的数目，并且 $Z_i=\left(Z_{i1},\cdots,Z_{ic}\right)^{\mathrm{T}}$ 表示协变量向量。

对于上面的数据，常规的 Cox 比例风险模型为

$$\lambda_i\left(Z_i,X_i\right)=\lambda_0\exp\left\{Z_i'\alpha+\sum_{j=1}^{m}X_i\left(u_j\right)\beta_j+\sum_{j=1}^{m-1}\sum_{k>j}^{m}X_i\left(u_j\right)X_i\left(u_k\right)\beta_{jk}\right\}$$

当 m 较小时，上面的模型可以用于交互效应和主效应分析；当 m 很大时，则该模型可能会因为参数过多而无法估计，即使能够估计模型参数，其功效也会较低。

为了有效降低模型参数个数和提高功效，基于 Zhang et al. (2021) [28]，本文提出如下函数型 Cox 比例风险模型：

$$\lambda_i\left(Z_i,X_i\right)=\lambda_0\exp\left\{Z_i'\alpha+\int_0^1X_i\left(u\right)\beta\left(u\right)\mathrm{d}u+\iint_{D:0\le u<v\le1}X_i\left(u\right)X_i\left(v\right)\beta\left(u,v\right)\mathrm{d}u\mathrm{d}v\right\}\tag{1}$$

其中 λ_0 是基线风险函数，$\beta\left(u\right)$ 是关于 u 的主效应函数，$0\le u<v\le1$，$\beta\left(u,v\right)$ 是关于 u、v 的交互效应函数。

在模型(1)中，假设主效应函数 $\beta\left(u\right)$ 是光滑的，即 $\beta\left(u\right)$ 是关于 u 的连续函数。它可以用 B 样条函数展开。假设 $\beta\left(u\right)$ 由一系列 K_β 基函数 $\psi_1\left(u\right),\cdots,\psi_{K_\beta}\left(u\right)$ 展开为

$\beta\left(u\right)=\left(\psi_1\left(u\right),\cdots,\psi_{K_\beta}\left(u\right)\right)\left(\beta_1,\cdots,\beta_{K_\beta}\right)'=\psi\left(u\right)'\beta$，其中 $\beta=\left(\beta_1,\cdots,\beta_{K_\beta}\right)'$ 是系数的 $K_\beta\times1$ 向量，

$\psi\left(u\right)=\left(\psi_1\left(u\right),\cdots,\psi_{K_\beta}\left(u\right)\right)'$。在这里我们利用 B 样条基：$\psi_k\left(u\right)=B_k\left(u\right),k=1,\cdots,K_\beta$ 来展开主效应函数。

设 $\varphi\left(u\right),k=1,\cdots,K$ 是一系列 K 基函数，如 B 样条基函数。设 Φ 表示含有 $\varphi_K\left(u_j\right)$ 的 $m\times K$ 矩阵，$\varphi\left(u\right)=\left(\varphi_1\left(u\right),\cdots,\varphi_K\left(u\right)\right)'$。利用已有结果[24] [25] [26]，$X_i\left(u\right)$ 可由下式进行估计：

$$\hat{X}_i\left(u\right)=\left(x_i\left(u_1\right),\cdots,x_i\left(u_m\right)\right)\Phi\left[\Phi'\Phi\right]^{-1}\phi\left(u\right)\tag{2}$$

对于交互作用，根据对称性，我们有

$$\iint_{D:0\le u\le v\le1}X_i\left(u\right)X_i\left(v\right)\beta\left(u,v\right)\mathrm{d}u\mathrm{d}v=\frac{1}{2}\iint_{\substack{D:0\le u\le1\\0\le v\le1}}X_i\left(u\right)X_i\left(v\right)\beta\left(u,v\right)\mathrm{d}u\mathrm{d}v$$

对于交互效应函数，我们利用张量积[29]对 $\beta\left(u,v\right)$ 进行展开。

张量积使一维 P 样条自然地延伸到二维，假设除了 x 之外，还有第二个解释变量 v。对于 $i=1,\cdots,m$，有数据三元组 $\left(v_i,x_i,y_i\right)$。寻找一个光滑的曲面 $\left(v,x\right)$，它可以给出数据一个很好的近似。设 B，$m\times L$ 是沿 x 的 B 样条基，B，$m\times K$ 是沿 v 的 B 样条基。构成 KL 张量积 $T_{kl}\left(v,x\right)=B_k\left(v\right)B_l\left(x\right),k=1,\cdots,K;l=1,\cdots,L$。设 $A=\left[kl\right]$ 是系数的 $K\times L$ 矩阵。然后，对于给定的 A，我们得到 $\left(v,x\right)$ 的拟合值是：

$$\mu\left(v,x\right)=\sum_k\sum_l\breve{B}_k\left(v\right)B_l\left(x\right)\alpha_{kl}$$

因此在这里我们设 $\gamma=\left[\gamma_{kl}\right]$ 是系数的 $K\times L$ 矩阵，则 $\beta\left(u,v\right)=\sum_k\sum_l\gamma_{kl}\psi_k\left(u\right)\psi_l\left(v\right)$。令

$$\overline{a_i}=\hat{X}_i\left(u\right)=\left(x_i\left(u_1\right),\cdots,x_i\left(u_m\right)\right)\Phi\left[\Phi'\Phi\right]^{-1}\phi\left(u\right)\tag{3}$$

模型(1)可利用上面的基展开表示为

$$\lambda_i\left(Z_i,X_i\right)=\lambda_0\exp\left\{Z_i'\alpha+W_i'\cdot\beta+W_i''\cdot Vec\left(\gamma\right)\right\}\tag{4}$$

其中

$$Vec(\gamma)=\left(\alpha_{12},\alpha_{13},\cdots,\alpha_{1K_\beta},\alpha_{23},\alpha_{24},\cdots,\alpha_{2K_\beta},\cdots,\alpha_{34},\alpha_{35},\cdots,\alpha_{3K},\cdots,\alpha_{K_\beta K_\beta}\right)^{\mathrm{T}}$$

$$W_i'=\left(x_i(u_1),\cdots,x_i(u_m)\right)\Phi\left[\Phi'\Phi\right]^{-1}\int_0^1\phi(u)\psi'(u)\mathrm{d}u$$

$$W_i''=\left(w_{i,12},\cdots,w_{i,1K_\beta},w_{i,23},w_{i,2K_\beta},\cdots,w_{i,K_\beta K_\beta}\right),$$

$$w_{i,kl}=\overline{a_i}\iint_{\substack{D:0\le u\le 1\\0\le v\le 1}}\begin{pmatrix}\phi_1(u)\phi_1(v) & \cdots & \phi_1(u)\phi_K(v)\\ \vdots & \ddots & \vdots\\ \phi_K(u)\phi_1(v) & \cdots & \phi_K(u)\phi_K(v)\end{pmatrix}\cdot\psi_k(u)\psi_l(v)\mathrm{d}u\mathrm{d}v\cdot\overline{a_i}^{\mathrm{T}}$$

3. 模拟研究

3.1. 模拟设置

在下面的模拟中，显著性水平为 0.05、0.01 和 0.001；考虑三种删失情况：(a) $C_i\sim U(0,10)$，(b) $C_i\sim U(0,5)$，(c) $C_i\sim U(0,3)$。基因型是从 6 kb 和 9 kb 亚区的 SNP 中选择的，在下面模拟中的模型含有 100 个 SNP。样本量为 $n=2000$ 和 $n=2500$，每种设置组合下，重复进行 4000 次模拟。

模型中的 SNP 主效应 $|\beta_j|=c\left|\log_{10}(MAF_j)\right|$，其中 MAF_j (Minor Allele Frequency，次等位基因频率)是第 j 个 SNP 的次等位基因频率。罕见变异(Rare Variant)通常是指 MAF ≤ 0.05 的 SNP。在模拟中，考虑 10%的变异在 6 kb 和 9 kb 区域被选择作为关联变异，对于关联变异考虑两种组合：1、10%的常见变异(common variant)和 90%的罕见变异；2、100%的罕见变异。对于 6 kb 亚区，$c=\log(70)/k$；对于 9 kb 来说，$c=\log(70)/(2k)$。常数 k 和遗传效应大小随着区域大小的增加而减小。对于第一种变异组合，在 6 kb 亚区，$k=5.5$；在 9 kb 亚区，$k=6$；对于所有 SNP 为罕见变异的情况，在 6 kb 亚区，$k=1.25$；在 9 kb 亚区，$k=1.5$。

3.2. 第 I 类错误率

生存时间由如下模型生成：

$$T_i=\sqrt{-\frac{4\log U_i}{\exp 0.005(Z_{i1}-50)+0.05Z_{i2}+\beta_1x_i(u_1)+\cdots+\beta_qx_i(u_q)}}$$

其中 U_i 是均匀分布的随机变量 $U(0,1)$，Z_i 是从正态分布 $N(50,5^2)$ 到模型的连续协变量。基于所设置的模型和模拟参数，对每个产生的数据利用提出的交互效应模型(3)进行分析，并利用 LRT 检验累积交互效应是否显著。

在附表 1 中(详见附录)，所有变异(常见和罕见)都用于在零假设下生成基因和生存时间数据。显著性水平 α 分别取值为 0.05，0.01，0.001。总的来说，所提出的模型能够较好的控制了第 I 类错误率。这两个模型在 6 kb 和 9 kb 的区域尺寸以及 2000 和 2500 的样本量下都是稳定的，结果非常相似。模拟结果表明所提出的模型在区域大小、检查方案、名义水平均是稳定的和有效的。

3.3. 功效模拟分析

生存时间由如下模型产生：

$$T_i=\sqrt{-\frac{4\log U_i}{\exp 0.005(Z_{i1}-50)+0.05Z_{i2}+\beta_1x_i(u_1)+\cdots+\beta_qx_i(u_q)+\sum_{0\le u_k<v_l\le 1}X_i(u_k)X_i(v_l)\beta(u_k,v_l)}}$$

在这里 $\beta(u_k, v_l) = c_I \left|\log_{10}(MAF_k * MAF_l)\right|$，$c_I$ 对于 6 kb 亚区，$c_I = \log(70)/k_I$；对于 9 kb 来说，$c_I = \log(70)/(2k_I)$。当罕见变异占比为 90%，常见变异占比为 10%时，在 6 kb 亚区，$k_I = 8$；在 9 kb 亚区，$k_I = 8.5$；对于所有 SNP 为罕见变异的情况，在 6 kb 亚区，$k_I = 2.75$；在 9 kb 亚区，$k_I = 4$。

当样本大小为 2000 时，功率如图 A、B 所示，6 kb 和 9 kb 分别代表不同的区域，当样本大小为 2500

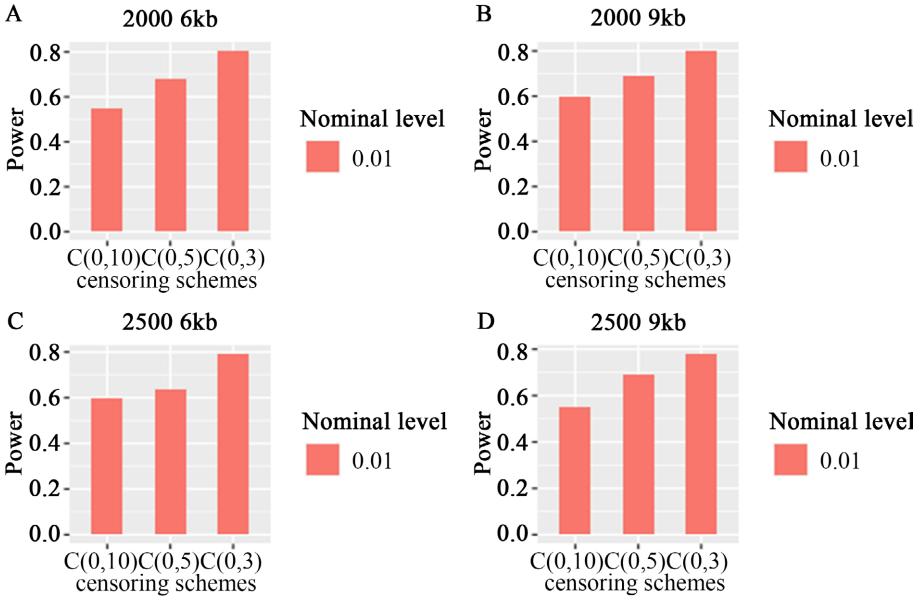

Figure 1. When the sample size was 2000 and 2500, and the region size was 6 KB and 9 kb, the LRT statistic was a potential function at α = 0.01, where some variants were common and others were rare. The Order of B-spline basis is 4

图 1. 当样本量为 2000 和 2500，区域大小为 6 kb 和 9 kb 时，LRT 统计量在 α = 0.01 时的势函数，其中一些变异是常见的，其余变异是罕见的。B 样条基的阶为 4

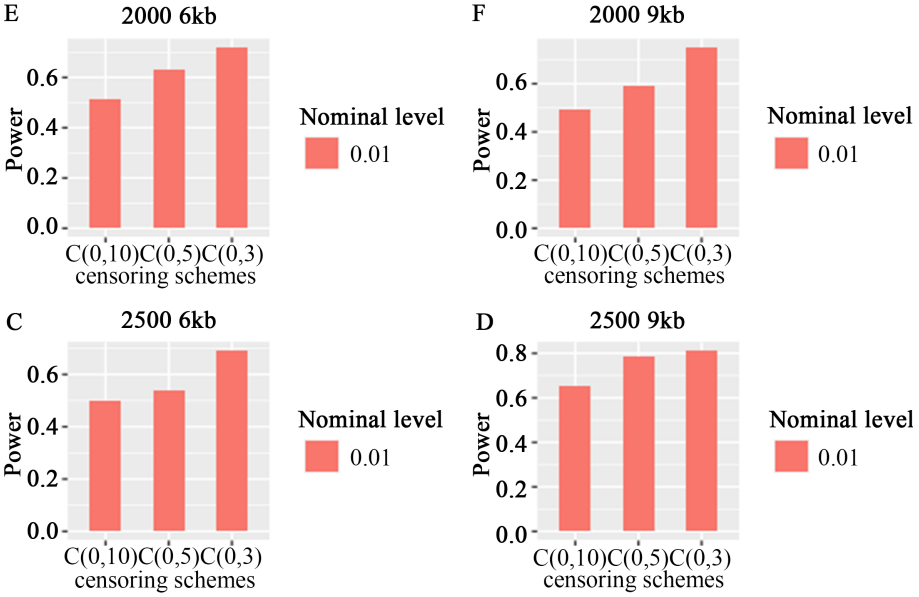

Figure 2. When the sample size was 2000 and 2500, and the region size was 6 KB and 9 kb, the LRT statistic was a potential function at α = 0.01, and all variants were rare

图 2. 当样本量为 2000 和 2500，区域大小为 6 kb 和 9 kb 时，LRT 统计量在 α = 0.01 时的势函数，所有变异都是罕见的

时，功率如图 C、D 所示，6 kb 和 9 kb 分别代表不同的区域；在图 A-D 中，常见变异和罕见变异占比分别为 90%和 10%。在图 E-H 中，所有 SNP 为罕见变异。

图 1 四个图的每个图中，比较了三种不同删失率、不同样本容量、不同区域的功效，所提出的 Cox 比例风险模型均能比较有效的检测出累积交互效应，功效随着样本容量的增加及删失率的减少而增加。图 2 中分析了变异全部为罕见变异的情况，从图中可以看到与图 1 类似的功效表现。函数交互效应 Cox 比例风险模型 LRT 统计量是非常稳定的，因为它们不强烈依赖于基因型数据是否平滑。

4. 案例分析

在 AREDS (Age-Related Eye Disease Study 年龄相关眼病研究)中的应用

利用所提出的函数交互效应 Cox 比例风险模型来分析 AREDS [24]数据(年龄相关眼病研究组，1999)。AREDS 是一项临床试验，旨在了解影响老年黄斑变性和白内障的风险因素，这两种疾病是导致老年人视力丧失的主要原因。共有 2911 人被纳入这项分析，其中 1650 人为男性，1261 人为女性。2911 人的平均年龄为 68.65 岁，标准差为 4.92 岁。本文仅研究左眼的视力数据，左眼的删失率为 76%。在分析中，我们将年龄和性别作为协变量进行调整，且每个个体都有长期的表型数据。根据 Seddon 等人在 2007 的研究中表明在 AMD (age-related macular degeneration 老年黄斑变性)疾病研究中有两个基因区域，CFH 和 ARMS2，与 AMD 的风险及其进展相关。这里只研究 CFH 基因区的交互效应。在 CFH 基因区域有 103 个罕见变异，59 个常见变异，CFH 基因的常见和罕见变异都会影响 AMD 的进展。在这里我们在样条函数的处理中除了使用 B 样条之外还使用了傅立叶样条，此外我们还利用 Logistic 模型[30]对交互作用的显著性进行分析，结果见表 1，显示这个基因区内的 SNP 之间的交互作用对老年黄斑变性风险有显著影响，对于结果来说，我们提出的带有 B 样条的 Cox 模型效果更为显著，Logistic 模型的显著性明显不如我们所提出的模型。

Table 1. Interaction within CFH gene fragments
表 1. CFH 基因片段内的交互作用

变体类型	基因	SNP 数量	Cox 模型 LRT 统计的 P 值(B-Spline)	Cox 模型 LRT 统计的 P 值(Fourier-Spline)	Logistic 模型 LRT 统计的 P 值(B-Spline)	Logistic 模型 LRT 统计的 P 值(Fourier-Spline)
全部变异	CFH	162	$7.12 \times e^{-12}$	$3.74 \times e^{-9}$	0.0051598	0.0054327
	CFH × CFH		$9.23 \times e^{-13}$	$4.14 \times e^{-11}$	0.0433671	0.0465722
常见变异	CFH	59	$<2 \times e^{-16}$	0.00145679	0.0342618	0.0378653
	CFH × CFH		$5.49 \times e^{-9}$	$3.17 \times e^{-7}$	0.056437	0.052315
罕见变异	CFH	103	$8.18 \times e^{-16}$	$4.36 \times e^{-10}$	0.065725	0.065438
	CFH × CFH		$7.12 \times e^{-11}$	$5.25 \times e^{-9}$	0.076231	0.065438

注：罕见变异定义为 MAF ≤ 0.05 的变异，常见变异定义为 MAF > 0.05 的变异。CFH × CFH 表示 CHF 基因片段的 SNP 之间的交互效应。

5. 总结

本文在 Bingsong Zhang 等研究基础上对基因－基因交互效应的检测上进行了拓展。与疾病相关联的 SNP 数量比较大时，它们之间的交互效应维数将会极大的增加，有些交互效应也比较微弱，这使得交互

效应的分析变得较为困难。本文在基因主效应函数化的基础上把基因－基因交互效应也进行函数化，通过基展开，可有效降低待估参数的维数，也可对交互效应进行有效检测。在前面模拟中，SNP 的个数为100，两两交互效应的数量高达 4950，待估参数的个数远大于观测值的数量。通过将主效应和交互效应进行函数化，极大地降低了待估参数的个数。从模拟结果可以看出第 I 类错误率在各种情形下都得到了很好的控制。此外，在功效的模拟研究分析中，所提出方法可有效地检测出基因－基因交互效应。第四部分将所提出的函数型 Cox 模型应用在了眼部疾病的数据中，分析结果表明我们的模型验证了 CFH 基因会影响老年黄斑变性的进展，其基因片段内的 SNP 之间的交互效应也会影响老年黄斑变性的进展，不论它是常见变异还是罕见变异。这显示了加入交互效应的模型的有效性。

本文的研究对于基因－基因交互效应的分析主要检验的是累积交互效应，但是部分交互效应可能会不显著，也不应该放入模型进行分析，否则会降低检验的效率。另外，高阶交互效应、基因－环境因子等还没有进行考虑，这些均可作为未来研究的内容。

基金项目

本文由中央高校基本科研业务经费(SWJTU，2682021ZTPY078)；国家自然科学基金面上项目(51578471)资助。

参考文献

[1] Randall, C.J., George, W.N., Jennifer, L.T., *et al.* (2010) Accounting for Multiple Comparisons in Agenome-Wide Association Study (GWAS). *BMC Genomics*, **11**, Article No. 724. https://doi.org/10.1186/1471-2164-11-724

[2] Tam, V., Patel, N., Turcotte, M., *et al.* (2019) Benefits and Limitations of Genome-Wide Association Studies. *Nature Reviews Genetics*, **20**, 467-484. https://doi.org/10.1038/s41576-019-0127-1

[3] Beck, T., Hastings, R., Gollapudi, S., *et al.* (2014) GWAS Central: A Comprehensive Resource for the Comparison and Interrogation of Genome-Wide Association Studies. *European Journal of Human Genetics*, **22**, 949-952. https://doi.org/10.1038/ejhg.2013.274

[4] Emily, M. (2018) A Survey of Statistical Methods for Gene-Gene Interaction in Case-Control Genome-Wide Association Studies. *Journal de la Societe Française de Statistique*, **159**, 27-67.

[5] Emily, M., Sounac, N., *et al.* (2020) Gene-Based Methods to Detect Gene-Gene Interaction in R: The GeneGeneInteR Package. *Journal of Statistical Software*, **95**, 1-32. https://doi.org/10.18637/jss.v095.i12

[6] Chen, J.J., Song, Y.J., Li, Y., *et al.* (2023) A Trans-Omics Assessment of Gene—Gene Interaction in Early-Stage NSCLC. *Molecular Oncology*, **17**, 173-187. https://doi.org/10.1002/1878-0261.13345

[7] Yang, T.L., Guo, Y., *et al.* (2013) Gene-Gene Interaction between RBMS3 and ZNF516 Influences Bone Mineral Density. *Journal of Bone and Mineral Research*, **28**, 828-837. https://doi.org/10.1002/jbmr.1788

[8] Ritchie, M.D. and Van Steen, K. (2018) The Search for Gene-Gene Interactions in Genome-Wide Association Studies: Challenges in Abundance of Methods Practical Considerations and Biological Interpretation. *Annals of Translational Medicine*, **6**, Article 157. https://doi.org/10.21037/atm.2018.04.05

[9] Dong, C., Chu, X., Wang, Y., *et al.* (2008) Exploration of Gene-Gene Interaction Effects Using Entropy-Based Methods. *European Journal of Human Genetics*, **16**, 229-235. https://doi.org/10.1038/sj.ejhg.5201921

[10] Wu, W.K.K., Sun, R., Zuo, T., *et al.* (2018) A Novel Susceptibility Locus in MST1 and Gene-Gene Interaction Network for Crohn's Disease in the Chinese Population. *Journal of Cellular and Molecular Medicine*, **22**, 2368-2377. https://doi.org/10.1111/jcmm.13530

[11] Lin, H.H., Mueller-Nurasyid, M., *et al.* (2016) Gene-Gene Interaction Analyses for Atrial Fibrillation. *Scientific Reports*, **6**, Article No. 35371.

[12] Emily, M. (2012) IndOR: A New Statistical Procedure to Test for SNP-SNP Epistasis in Genome-Wide Association Studies. *Statistics in Medicine*, **31**, 2359-2373. https://doi.org/10.1002/sim.5364

[13] Ritchie, M., Hahn, L.W. and Moore, J.H. (2003) Power of Multifactor Dimensionality Reduction for Detecting Gene-Gene Interactions in the Presence of Genotyping Error, Missing Data, Phenocopy, and GeneticHeterogeneity. *Genetic Epidemiology*, **24**, 150-157. https://doi.org/10.1002/gepi.10218

[14] Wan, X., Yang, C., Yang, Q., *et al.* (2010) BOOST: A Fast Approach to Detecting Gene-Gene Interactions in Genome-Wide Case-ControlStudies. *The American Journal of Human Genetics*, **87**, 325-340. https://doi.org/10.1016/j.ajhg.2010.07.021

[15] Zhang, S., Jiang, W., Ma, R.C., *et al.* (2019) Region-Based Interaction Detection in Genome-Wide Case-Control Studies. *BMC Medical Genomics*, **12**, Article No. 133. https://doi.org/10.1186/s12920-019-0583-7

[16] Chang, Y.C., Wu, J.T., Hong, M.Y., *et al.* (2020) GenEpi: Gene-Based Epistasis Discovery Using Machine Learning. *BMC Bioinformatics*, **21**, Article No. 68. https://doi.org/10.1186/s12859-020-3368-2

[17] Nobre, R., Ilic, A., Santander, J., *et al.* (2021) Retargeting Tensor Accelerators for Epistasis Detection. *IEEE Transactions on Parallel and Distributed Systems*, **32**, 2160-2174. https://doi.org/10.1109/TPDS.2021.3060322

[18] Emily, M. (2016) AGGrEGATOr: A Gene-Based GEne-Gene InterActTiOn Test for Case-Control Association Studies. *Statistical Applications in Genetics and Molecular Biology*, **15**, 151-171. https://doi.org/10.1515/sagmb-2015-0074

[19] Liu, D., Wang, M., Yuan, Y., *et al*. (2019) Gene-Gene Interaction Among Cell Adhesion Genesand Risk of Nonsyndromic Cleft Lip with or without Cleft Palate in Chinese Case-Parent Trios. *Molecular Genetics & Genomic Medicine*, **7**, e00872. https://doi.org/10.1002/mgg3.872

[20] Hall, M.A., Verma, S.S., Wallace, J., *et al.* (2015) Biology-Driven Gene-Gene Interaction Analysis of Age-Related Cataract in the EMERGE Network. *Genetic Epidemiology*, **39**, 376-384. https://doi.org/10.1002/gepi.21902

[21] Ma, L., Clark, A.G. and Keinan, A. (2016) Gene-Based Testing of Interactions in Association Studies of Quantitative Traits. *PLOS Genetics*, **9**, e1003321. https://doi.org/10.1371/journal.pgen.1003321

[22] Li, J., Huang, D., Guo, M., *et al.* (2015) A Gene-Based Information Gain Method for Detecting Gene-Gene Interactions in Case-Control Studies. *European Journal of Human Genetics*, **23**, 1566-1572. https://doi.org/10.1038/ejhg.2015.16

[23] 杨青龙, 刘媛, 刘妍岩, 邹珺. 中等稀疏条件下基因交互作用的两步 Bayes 方法[J]. 中国科学(数学), 2021, 51(2): 393-410.

[24] 杨秋月, 王菁菁, 徐相蓉, 何丽华, 余善法, 陈国顺, 吴辉. 三种遗传性耳聋基因交互作用与高频听力损失易感性的关系[J]. 中国工业医学杂志, 2018, 31(2): 83-86, 93.

[25] Ramsay, J.O. and Silverman, B.W. (2005) Functional Data Analysis. Springer, New York. https://doi.org/10.1007/b98888

[26] 汪寿阳, 洪永淼, 霍红, 方颖, 陈海强. 大数据时代下计量经济学若干重要发展方向[J]. 中国科学基金, 2019, 33(4): 386-393.

[27] Zhang, B., Chiu, C., Yuan, F., *et al.* (2021) Gene-Based Analysis of Bi-Variate Survival Traits via Functional Regressions with Applications to Eye Diseases. *Genetic Epidemiology*, **45**, 455-470. https://doi.org/10.1002/gepi.22381

[28] Age-Related Eye Disease Study Research Group (1999) The Age-Related Eye Disease Study (AREDS): Design Implication AREDS Report No. 1. *Controlled Clinical Trials*, **20**, 573-600. https://doi.org/10.1016/S0197-2456(99)00031-8

[29] De, B.C. (2001) A Practical Guide to Splines (Applied Mathematical Sciences, 27). Springer, New York.

[30] Ferraty, F. and Romain, Y. (2010) The Oxford Handbook of Functional Data Analysis. Oxford University Press, Oxford.

附 录

Appendix Table 1. Type I error rates for two sample sizes
附表 1. 两种样本容量下的第 I 类错误率

样本量	区间量	删失分布	名义水平	情景 1	情景 2
2000	6 kb	$U(0,10)$	0.05	0.0503	0.0512
			0.01	0.0101	0.0132
			0.001	0.0011	0.0017
			0.0001	0.0000	0.0000
		$U(0,5)$	0.05	0.0506	0.0511
			0.01	0.0104	0.0101
			0.001	0.0015	0.0011
			0.0001	0.0000	0.0000
		$U(0,3)$	0.05	0.0504	0.0501
			0.01	0.012	0.012
			0.001	0.0014	0.0009
			0.0001	0.0000	0.0000
	9 kb	$U(0,10)$	0.05	0.0487	0.0507
			0.01	0.0089	0.0112
			0.001	0.0014	0.0013
			0.0001	0.0000	0.0000
		$U(0,5)$	0.05	0.0517	0.0509
			0.01	0.0117	0.0102
			0.001	0.0013	0.0008
			0.0001	0.0000	0.0000
		$U(0,3)$	0.05	0.0537	0.0492
			0.01	0.0112	0.0141
			0.001	0.0017	0.0013
			0.0001	0.0000	0.0000
2500	6 kb	$U(0,10)$	0.05	0.0455	0.0517
			0.01	0.01	0.014
			0.001	0.0008	0.0015
			0.0001	0.0000	0.0000
		$U(0,5)$	0.05	0.0528	0.0519
			0.01	0.0105	0.0112
			0.001	9e-04	0.0017
			0.0001	0.0000	0.0000

续表

2500	6 kb	$U(0,3)$	0.05	0.0493	0.0544
			0.01	0.0114	0.0129
			0.001	0.0014	0.00079
			0.0001	0.0000	0.0000
		$U(0,10)$	0.05	0.0551	0.0565
			0.01	0.0108	0.0101
			0.001	0.0012	0.0017
			0.0001	0.0000	0.0000
	9 kb	$U(0,5)$	0.05	0.0573	0.0462
			0.01	0.0112	0.0103
			0.001	0.0009	0.0012
			0.0001	0.0000	0.0000
		$U(0,3)$	0.05	0.0476	0.0511
			0.01	0.0113	0.0112
			0.001	0.0019	0.0016
			0.0001	0.0000	0.0000

(当区域大小为 6 和 9 kb 时，名义水平 α = 0.05，0.01，0.001，0.0001 的 Cox FR LRT 统计量的第 I 类错误率，情景 1 表示有一些变体是常见的，其余的是罕见的，情景二表示所有变异都是罕见的。)