**Statistics and Application** 统计学与应用, 2021, 10(2), 293-299
Published Online April 2021 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2021.102029

Hans汉斯

# 不确定基因型的可加模型及变量选择

**钟思敏，徐　萍**

广西师范大学，广西 桂林
Email: zhongsimin@stu.gxnu.edu.cn

收稿日期：2021年4月1日；录用日期：2021年4月16日；发布日期：2021年4月28日

## 摘　要

**全基因组关联分析(GWAS)是研究复杂疾病相关位点的有效方法.在基因不确定情形下，传统方法利用基因填补方式估计基因概率，继而展开后续基因关联分析。我们对大样本基因考虑一个非参数可加模型对可加分量维数大而非零加性分量数目小的基因数据进行建模，其中加性分量利用B样条基函数的线性组合工具来近似拟合基因概率对性状表征的效应关系；选择非零分量是利用组Lasso惩罚来获得初始估计量。最后我们利用蒙特卡洛模拟证明，可加模型的组lasso方法在基因表达样本中的效果良好。**

## 关键词

**全基因组关联研究，基因型不确定，可加模型，B样条，变量选择**

# Additive Model and Variable Selection for Uncertain Genotypes

**Simin Zhong, Ping Xu**

Guangxi Normal University, Guilin Guangxi
Email: zhongsimin@stu.gxnu.edu.cn

Received: Apr. 1st, 2021; accepted: Apr. 16th, 2021; published: Apr. 28th, 2021

## Abstract

**Genome-wide association analysis (GWAS) is an effective method to study the associated loci of complex diseases. In the case of genetic uncertainty, the traditional method uses the gene filling method to estimate the gene probability, and then carries out the subsequent gene association analysis. We used a nonparametric additive model to model the data of large samples of genes with large additive component dimensions but small non-zero-additive component numbers. The additive component was used as a linear combination tool of B-spline basis function to approx-**

文章引用: 钟思敏, 徐萍. 不确定基因型的可加模型及变量选择[J]. 统计学与应用, 2021, 10(2): 293-299.
DOI: 10.12677/sa.2021.102029

**imate the effect relationship of gene probability on trait characterization. The group Lasso penalty was used to obtain the initial estimator for selecting the non-zero component. Finally, Monte Carlo simulation was used to demonstrate that the group Lasso method of the additive model performed well in gene expression samples.**

## Keywords

**Genome-Wide Association Study, Genotype Uncertainty, Additive Model, B-Spline, Variable Selection**



## 1. 引言

近年来，全基因组关联研究(GWAS) [1]被广泛关注，成功地应用于识别与复杂疾病相关的位点及农业畜牧业等重要经济性状相关的遗传变异基因。复杂疾病是环境和遗传变异因子的共同结果，一般涉及多个基因，而大部分疾病的相关基因位点未知，这就要求科学家开发合理的方法识别相关位点。GWAS就是寻找致病基因的方法之一，主要用于数量性状的分析，应用基因组中数以百万计的单核苷酸多态(SNP)，在全基因组水平上进行大样本分析找出影响复杂性状的基因变异位点。也就是说 GWAS 的主要目标是筛选出与特定疾病相关的基因、确定 SNP 与疾病表型的关联[2]。Klein *et al.* (2005)最初利用该方法找出老年黄斑变性的相关致病基因[3]，Sladek *et al.* (2007)确定 4 个与 2 型糖尿病有关的致病基因[4]。相关实例应用陆续开发，在风湿性关节炎[5]、食管癌[6]、心血管疾病[7]和前列腺癌[8]等疾病以及农业(芝麻等)畜牧业等重要经济作物中有了相应研究成果，GWAS 快速发展。

随着测序数据的发展，多国协作的 HapMap 计划的完成，海量的基因型数据被用于 GWAS [9]，适用于大数据关联分析的算法也在不断更新。在大型样本的基因测序研究中，有时出现基因型不确定的情况[10]，在这种情况下，利用“填补基因型方法”确定每个样本对应的基因型[11] [12] [13]，继而展开后续的基因关联分析[14] [15] [16] [17]。对于特定的 SNP 基因分型，通常采用的“最大概率法”和“剂量法”。前者指定概率最大的基因型作为样本基因进行分析估计，后者是用线性组合的方式计算基因型概率，计算方法分别为：$\tilde{X}_{ij}=\left\{g:p_{ijg}=\max\left\{p_{ij0},p_{ij1},p_{ij2}\right\}\right\}$；$\bar{X}_{ij}=p_{ij1}+2p_{ij2}$。

传统的关联分析主要考虑单个基因型对疾病表征关系的关联分析[18]，但多个基因位点的基因型与特定疾病的关联分析及变量选择就很少[19]。Li Q. [20]、Loley [21]分别对多个基因位点对性状表征的传统线性回归模型、广义线性模型。在高维基因数据方面，成青[22]和刘璐[19]考虑高维数据中 $n$ 个样本的 $p$ 个基因位点的基因数据建立线性回归模型并结合 Lasso 罚函数进行变量选择。而这些方法没有考虑基因型的不确定，且认为基因对疾病表征关系是线性的。在遗传基因表达中，基因位点对疾病的效应关系可以是非线性的。对高维数据处理，一般传统的高维数据模型加罚函数可以初步对变量降维，在这方面已有大量的研究工作[23] [24] [25] [26]。这里我们将在考虑高维基因数据下，部分基因位点不确定，探索一种利用 B 样条拟合基因位点对基本性状响应的光滑效应关系，进而建立可加模型[27]，结合 grouplasso 罚函数对高维变量降维及非参估计以确定基因型与性状表征的关系。

本文的其余部分安排如下：在第 2 节中，我们介绍在基因不确定情形下的加性模型，我们将在第 3 节中详细介绍选择一致性证明结果；在第 4 节将我们的方法应用于模拟 GWAS 数据的分析。最后，我们

作一些结束语。

## 2. 符号与模型

### 2.1. B 样条曲线

在介绍我们的模型前，首先回顾 B 样条曲线概念。1972 年 Riesenfeld 等人[28]提出 B 样条曲线，它可以通过一组标准 B 样条基函数的线性组合来有效逼近上述非线性函数。假定 $a=\eta_0<\eta_1<\cdots<\eta_{K-1}<\eta_K=b$ 将区间 $[a,b]$ 分割为 $K$ 个子区间，记 $I_{Kt}=[\eta_t,\eta_{t+1}),t=0,\cdots,K-2$ 为前 $K-1$ 个子区间，对于第 $k$ 个子区间，其 $p$ 次样条基函数 $\phi_{k,p}$ 定义如下

$$\phi_{k,0}=\begin{cases}1,x\in[\eta_k,\eta_{k+1}]\\0,x\notin[\eta_k,\eta_{k+1}]\end{cases}$$

其递归公式为

$$\phi_{k,p}(x)=\frac{x-\eta_k}{\eta_{k+p}-\eta_k}\phi_{k,p-1}(x)+\frac{\eta_{k+p+1}-x}{\eta_{k+p+1}-\eta_{k+p}}\phi_{k+1,p-1}(x),k>0$$

假设 $S_n$ 为 $p\geq 1$ 次多项式样条空间，$\{\phi_k,k\leq d_n\}$ 是定义在 $S_n$ 上的一组标准基向量，则有

$$f_{nj}(x)=\sum_{k=1}^{d_n}\phi_k(x)\beta_{jk},1\leq j\leq p$$

其中 $d_n=K+p$，在适当的条件下，$f_{nj}(x)$ 可以和好的逼近 $f_j(x)$ [29]。记 $\boldsymbol{\beta}=\left(\boldsymbol{\beta}_1^{\mathrm{T}},\cdots,\boldsymbol{\beta}_p^{\mathrm{T}}\right)^{\mathrm{T}}$，$\boldsymbol{\beta}_j=\left(\beta_{j1},\cdots,\beta_{jd_n}\right)^{\mathrm{T}}$。

### 2.2. 基于基因不确定的可加模型

考虑等位基因为 a 和 A 的单核苷酸多态性。假设 A 是导致疾病的高风险等位基因。三种基因型分别表示为 aa、Aa、AA，对应编码为 0、1、2。定义 $X_{ij}$ 为第 $i$ 个个体第 $j$ 个点位的基因型，则对应的基因型概率定义为 $p_{ij}=P\left(X_{ij}=g\right),g=0,1,2$，$Y_i$ 为第 $i$ 个个体的性状观测值。这里基因不确定数据，我们的 $X_{ij}$ 可采取传统填补基因型方法：$\tilde{X}_{ij}=\left\{g:p_{ijg}=\max\left\{p_{ij0},p_{ij1},p_{ij2}\right\}\right\}$；$\bar{X}_{ij}=p_{ij1}+2p_{ij2}$。

假设我们的数据向量 $(\boldsymbol{X}_i,Y_i)$ 独立同分布服从 $(\boldsymbol{X},Y)$，$Y$ 是响应变量，$\boldsymbol{X}=\left(X_1,\cdots,X_p\right)^{\mathrm{T}}$ 是 $p$ 维基因数据，考虑如下非线性可加模型：

$$Y_i=\mu+\sum_{j=1}^{p}f_j\left(X_{ij}\right)+\varepsilon_i,\ i=1,\cdots,n;\ j=1,\cdots,p$$

其中 $\mu$ 为截距项，$f_j(*)$ 为未知光滑函数，$\varepsilon_i$ 为服从均值为 0 方差为 $\sigma^2$ 的随机误差。在高维稀疏性的假定下，这里的大部分 $f_j(*)\equiv 0,\ j=1,\cdots,p$，我们的目的是找出不恒等于 0 的部分。结合 B 样条拟合光滑函数的工具，表达式等价于：

$$Y_i\triangleq\mu+\sum_{j=1}^{p}\sum_{k=1}^{d_n}\phi_k\left(X_{ij}\right)\beta_{jk}+\varepsilon_i,\ \ i=1,\cdots,n;\ j=1,\cdots,p;k=1,\cdots,d_n$$

这里的待估参数 $\boldsymbol{\beta}=\left(\mu,\beta_{11},\cdots,\beta_{1d_n},\cdots,\beta_{p1},\cdots,\beta_{pd_n}\right)$。这里利用 Bspline 基函数线性组合工具，可视为其将非线性模型转化线性关系模型。

由于参数 $\boldsymbol{\beta}$ 的维数是 $pd_n+1$，利用 Bspline 线性组合近似非线性函数 $f_j(*)$ 时，$d_n$ 很大则系数的稀疏性是我们考虑的范围，因而这里用稀疏组 lasso 选择重要分量和非零系数，这一部分内容我们在下一小节详细介绍。

### 2.3. grouplasso 变量选择及参数估计

特定疾病对应与之相关的基因位点往往只有少部分，假定真实模型是稀疏的，也就是重要的协变量就是少数，只有小部分加性分量 $f_j$ 是非零的，记真实的非零分量指标集为 $M^*=\left\{j: f_j \neq 0\right\}$，这里指标集中的数目大小为 $s=\left\|M^*\right\|$。我们的目标是利用罚函数对高维基因数据降维，选择分量 $\hat{M}=\left\{j: \hat{\boldsymbol{\beta}} \neq \mathbf{0}\right\}$，这里 $\hat{s}=\left\|M^*\right\|$，且 $\hat{s} \geq s$。即当样本量增大时，重要变量选入模型的概率趋于 1，我们的模型具有选择一致性。

在高维基因数据的情况下，在模型上加 groupLasso 惩罚项有选择一致的特性，这里我们考虑惩罚最小二乘法：

$$S_n\left(\mu, \boldsymbol{\beta}_n\right)=\sum_{i=1}^n\left[y_i-\mu-\sum_{j=1}^p \sum_{k=1}^{d_n} \phi_k\left(x_{ij}\right) \beta_{jk}\right]^2+\lambda_n \sum_{j=1}^p\left\|\boldsymbol{\beta}_{nj}\right\|_2$$

记 $\boldsymbol{\beta}_n=\left(\boldsymbol{\beta}_{n1}^{\mathrm{T}}, \cdots, \boldsymbol{\beta}_{np}^{\mathrm{T}}\right)^{\mathrm{T}}=\left(\beta_{11}, \cdots, \beta_{1 d_n}, \cdots, \beta_{p 1}, \cdots, \beta_{p d_n}\right)^{\mathrm{T}}$、$1 \leq j \leq p$，$\lambda_n$ 是惩罚参数。

为了模型的可识别性，加入约束条件 $\sum_{i=1}^n \sum_{k=1}^{d_n} \phi_k\left(x_{ij}\right) \beta_{jk}=0, j=1, \cdots, p$。对 $S_n\left(\mu, \boldsymbol{\beta}_n\right)$ 极小化求参数：

$$\hat{\boldsymbol{\beta}}=\arg \min \left\{\sum_{i=1}^n\left(y_i-\mu-\sum_{j=1}^p \sum_{k=1}^{d_n} \phi_k\left(x_{ij}\right) \beta_{jk}\right)^2+\lambda_n \sum_{j=1}^p\left\|\boldsymbol{\beta}_{nj}\right\|_2\right\}$$

通过 GroupLasso 估计出每组参数的估计，根据稀疏性假定，其中只有少部分 $\left\|\hat{\boldsymbol{\beta}}_{nj}\right\|_2 \neq 0, j=1, \cdots, p$，等价于 $f_j(*)$ 不恒等于，也就是说对应的第 j 个基因位点是与疾病存在显著关系的。

## 3. 模拟结果

利用蒙特卡罗模拟数值，比较基因型数据 $X_i$ 采用剂量法与最大可能概率法在线性模型与非线性可加模型的表现。这里 $X_{ij}$ 中的三个概率( $p_{ij0}, p_{ij1}, p_{ij2}$ )通过将基因不确定率的情况考虑在内的三项分布生成。这里的是三项分布可以通过三种频率对模拟生成特定频率生成，在模拟设定基因确定率和不确定基因频率则为 $r$ 和 $1-r$。有不确定率的存在，一个基因位点会有三个基因型(aa、Aa、AA)的概率。基因型的检测频率，如表 1 所示。

**Table 1.** Frequency of genotype detection
**表 1.** 基因型的检测频率

| 基因位点 | 基因型 | | | 最大概率法 | 剂量法 |
|---|---|---|---|---|---|
| | 0 | 1 | 2 | | |
| $SNP_1$ | 0.853 | 0.127 | 0.020 | 0 | 0.167 |
| $SNP_2$ | 0.156 | 0.102 | 0.742 | 2 | 1.586 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| $SNP_p$ | 0.314 | 0.421 | 0.265 | 1 | 0.951 |

下面我们通过两个例子比较说明可加模型的 groupLasso 罚函数的变量选择的优良性。

例 1：从下面的模型中产生数据：

$$Y_i=\mu+\sum_{j=1}^p f_j\left(X_{ij}\right)+\varepsilon_i, i=1, \cdots, n$$

这里取 $\mu=2$，$f_1(x)=-x$；$f_2(x)=2x^2$；$f_3(x)=-3\log(x)$；$f_j(x) \equiv 0, 4 \leq j \leq 100=p$；这里非零分量个数 $s=3$，$p=100$，样本量 $n=1000$，随机误差服从标准正态分布。模型超参数 $\lambda$ 的选取采用十折

交叉验证进行筛选，数据按照 7:3 比例随机分为训练集与测试集。统计变量正确选择平均个数(TM)，错误选择平均个数(FM)以及测试集均方误差(TMSE)如表 2。

**Table 2.** Results of variable selection
**表 2.** 变量选择结果

| | | $d_n$ | TM | FM | TMSE |
|---|---|---|---|---|---|
| 最大值法 | 可加模型 | 4 | 2.31 (0.51) | 1.24 (1.98) | 1.68 (0.13) |
| | | 6 | 2.43 (0.52) | 1.58 (2.37) | 1.56 (0.12) |
| | | 8 | 2.45 (0.52) | 1.45 (2.33) | 1.54 (0.12) |
| | | 10 | 2.41 (0.53) | 1.67 (2.63) | 1.57 (0.13) |
| | 线性模型 | - | 1.38 (0.60) | 0.70 (1.03) | 1.74 (0.14) |
| 剂量法 | 可加模型 | 4 | 2.91 (0.29) | 5.17 (5.16) | 1.42 (0.12) |
| | | 6 | 2.95 (0.22) | 5.22 (5.45) | 1.31 (0.10) |
| | | 8 | 2.96 (0.20) | 5.38 (5.73) | 1.32 (0.11) |
| | | 10 | 2.87 (0.34) | 5.43 (5.42) | 1.32 (0.11) |
| | 线性模型 | - | 1.92 (0.42) | 0.73 (1.24) | 1.58 (0.12) |

根据表 1 结果可以知道，在模型一样的情况下，剂量法相较于最大概率法能更多的识别出重要变量以及拥有更小的均方误差，与此同时识别变量错误的情况也偏多；当然线性模型应用范围比较局限，在存在非线性部分时，其变现结果很差。

例 2：从下面的模型中产生数据：

$$Y_i = \mu + \sum_{j=1}^{p} f_j\left(X_{ij}\right) + \varepsilon_i,\ i = 1,\cdots,n$$

这里取 $f_1(x) = 5x$， $f_2(x) = 3(2x-1)^2$； $f_3(x) = 4\sin(2\pi x)/(2-\sin(2\pi x))$；
$f_4(x) = 6\left(0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3\sin(2\pi x)^2 + 0.4\cos(2\pi x)^3 + 0.5\sin(2\pi x)^3\right)$； $f_5 = \cdots = f_p = 0$。这里非零分量个数 $s = 4$， $p = 1000$ 和三种不同的样本容量：n=100，200 和 1000，随机误差服从标准正态分布。这里的数据生成模型和 Huang 的一样，但是我们这里基于基因不确定情形。针对这个模型比较 GroupLasso 和 GroupSCAD 惩罚的优越性，得到下面表 3。

**Table 3.** Results of variable selection
**表 3.** 变量选择结果

| | | $d_n$ | TM | FM | TMSE |
|---|---|---|---|---|---|
| 最大值法 | Group Lasso | 4 | 3.13 (0.89) | 1.71 (2.03) | 2.03 (0.22) |
| | | 6 | 3.04 (0.82) | 1.67 (2.41) | 2.01 (0.21) |
| | | 8 | 3.14 (0.82) | 1.62 (2.53) | 1.97 (0.21) |
| | | 10 | 3.34 (0.80) | 1.81 (2.43) | 2.10 (0.25) |
| | Group SCAD | 4 | 3.21 (0.82) | 3.53 (3.97) | 1.99 (0.22) |
| | | 6 | 3.23 (0.82) | 3.63 (4.11) | 1.97 (0.22) |
| | | 8 | 3.41 (0.71) | 3.71 (4.14) | 1.95 (0.20) |
| | | 10 | 3.40 (0.75) | 3.59 (4.13) | 2.03 (0.21) |

**Continued**

| | | | | | |
|---|---|---|---|---|---|
| 剂量法 | Group Lasso | 4 | 3.68 (0.42) | 5.15 (4.21) | 1.89 (0.17) |
| | | 6 | 3.77 (0.37) | 5.57 (4.29) | 1.72 (0.16) |
| | | 8 | 3.83 (0.32) | 5.42 (4.24) | 1.71 (0.15) |
| | | 10 | 3.81 (0.33) | 5.09 (4.79) | 1.79 (0.19) |
| | Group SCAD | 4 | 3.64 (0.52) | 7.51 (3.03) | 1.63 (0.22) |
| | | 6 | 3.79 (0.47) | 8.32 (4.14) | 1.57 (0.17) |
| | | 8 | 3.77 (0.46) | 7.62 (3.02) | 1.58 (0.17) |
| | | 10 | 3.72 (0.53) | 7.61 (3.00) | 1.71 (0.20) |

通过对表 3 两种变量选择方法进行比较，采用最大概率法时，GroupLasso 方法与 GroupSCAD 方法变量选择的效果差距不大；从错选变量(FM)个数看，Group Lasso 方法倾向于筛选出更少的伪变量，特别在用剂量法时，Group Lasso 错选变量个数明显小于 Group SCAD 方法。

## 4. 结束语

复杂疾病与基因位点通常存在非线性关系，全基因组测序建立复杂疾病与基因位点之间关系是一种有效途径，但是全基因组的准确测序也是一个耗时耗力的庞大工程。在部分基因位点基因型不确定下，采用“最大概率法”或“剂量法”来填补基因数据，GroupLasso 以及 GroupSCAD 方法对非线性可加模型进行重要基因位点识别是一种有效途径。

本文在基因不确定情形下，基于 GWAS 创新一种基因表达方法，不同于传统的基因填补估计概率方法，结合了 B 样条工具拟合加性分量建立非参数可加模型，考虑了大样本基因的环境下进行组 Lasso 变量选择以获得初始估计；最后利用数值模拟对比证明了我们方法的优良性。在基因测序中，我们非参加性模型更灵活，对模型的假设要求更低，适用范围更广。

## 参考文献


[1] Bush, W.S. and Moore, J.H. (2012) Chapter 11: Genome-Wide Association Studies. *PLOS Computational Biology*, **8**, e1002822. https://doi.org/10.1371/journal.pcbi.1002822

[2] 张学军. 复杂疾病的遗传学研究策略[J]. 安徽医科大学学报, 2007(3): 237-240.

[3] Klein, R.J., Zeiss, C., *et al*. (2005) Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science*, **308**, 385-388. https://doi.org/10.1126/science.1109557

[4] Sladek, R., Rocheleau, G., Rung, J., *et al*. (2007) A Genome-Wide Association Study Identifies Novel Risk Loci for Type 2 Diabetes. *Nature*, **445**, 881-885. https://doi.org/10.1038/nature05616

[5] Tamiya, G., Shinya, M., *et al*. (2005) Whole Genome Association Study of Rheumatoid Arthritis Using 27039 Microsatellites. *Human Molecular Genetics*, **14**, 2305-2321. https://doi.org/10.1093/hmg/ddi234

[6] Hu, N., Wang, C., Hu, Y., Yang, H.H., *et al*. (2005) Genome-Wide Association Study in Esophageal Cancer Using GeneChip Mapping 10K Array. *Cancer Research*, **65**, 2542-2546. https://doi.org/10.1158/0008-5472.CAN-04-3247

[7] Samani, N.J., Erdmann, J., Hall, A.S., *et al*. (2007) Genomewide Association Analysis of Coronary Artery Disease. *New England Journal of Medicine*, **357**, 443-453. https://doi.org/10.1056/NEJMoa072366

[8] Conti, D.V., Darst, B.F., Moss, L.C., *et al*. (2021) Trans-Ancestry Genome-Wide Association Meta-Analysis of Prostate Cancer Identifies New Susceptibility Loci and Informs Genetic Risk Prediction. *Nature Genetics*, **53**, 65-75. https://doi.org/10.1038/s41588-020-00748-0

[9] International HapMap Consortium (2005) A Haplotype Map of the Human Genome. *Nature*, **437**, 1299-1320. https://doi.org/10.1038/nature04226

[10] Lin, D., Hu, Y. and Huang, B. (2008) Simple and Efficient Analysis of Disease Association with Missing Genotype

Data. *The American Journal of Human Genetics*, **82**, 444-452. https://doi.org/10.1016/j.ajhg.2007.11.004

[11] Howie, B.N., Donnelly, P. and Marchini, J. (2009) A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genetics*, **5**, e1000529. https://doi.org/10.1371/journal.pgen.1000529

[12] Li, Y., Willer, C.J., Ding, J., Scheet, P. and Abecasis, G.R. (2010) MaCH: Using Sequence and Genotype Data to Estimate Haplotypes and Unobserved Genotypes. *Genetic Epidemiology*, **34**, 816-834. https://doi.org/10.1002/gepi.20533

[13] Browning, B. and Browning, S. (2009) A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. *The American Journal of Human Genetics*, **84**, 210-223. https://doi.org/10.1016/j.ajhg.2009.01.005

[14] Zheng, J., Li, Y., Abecasis, G.R. and Scheet, P. (2011) A Comparison of Approaches to Account for Uncertainty in Analysis of Imputed Genotypes. *Genetic Epidemiology*, **35**, 102-110. https://doi.org/10.1002/gepi.20552

[15] Acar, E.F. and Sun, L. (2013) A Generalized Kruskal-Wallis Test Incorporating Group Uncertainty with Application to Genetic Association Studies. *Biom*, **69**, 427-435. https://doi.org/10.1111/biom.12006

[16] Ding, J. and Li, H. (2017) Comparison of Robust Tests for Genetic Association Analysis Incorporating Uncertain Genotype. *Communications in Statistics—Simulation and Computation*, **46**, 3436-3443.

[17] 黄蕊. 二阶段关联分析在基因型不确定情形的应用[D]: [硕士学位论文]. 桂林: 广西师范大学, 2018.

[18] Zheng, G. and Chen, Z. (2005) Comparison of Maximum Statistics for Hypothesis Testing When a Nuisance Parameter Is Present Only under the Alternative. *Biometrics*, **61**, 254-258. https://doi.org/10.1111/j.0006-341X.2005.030531.x

[19] 刘璐. 引入基因型线性模型的变量选择[D]: [硕士学位论文]. 桂林: 广西师范大学, 2019.

[20] Li, Q.Z., Xiong, W.J., Chen, J.B., Zheng, G., Li, Z.H., Mills, J.L. and Liu, A.Y. (2014) A Robust Test for Quantitative Trait Analysis with Model Uncertainty in Genetic Association Studies. *Statistics and Its Interface*, **7**, 61-68. https://doi.org/10.4310/SII.2014.v7.n1.a7

[21] Loley, C., König, I., Hothorn, L., *et al*. (2013) A Unifying Framework for Robust Association Testing, Estimation, and Genetic Model Selection Using the Generalized Linear Model. *European Journal of Human Genetics*, **21**, 1442-1448. https://doi.org/10.1038/ejhg.2013.62

[22] 成青. 高维基因数据中的变量选择[D]: [硕士学位论文]. 成都: 西南交通大学, 2014.

[23] Frank, I.E. and Friedman, J.H. (1993) A Statistical View of Some Chemometrics Regression Tools (with Discussion). *Technometrics*, **35**, 109-148. https://doi.org/10.1080/00401706.1993.10485033

[24] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*: *Series B* (*Methodological*), **58**, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[25] Fan, J. and Li, R. (2001) Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association*, **96**, 1348-1360. https://doi.org/10.1198/016214501753382273

[26] Huang, J., Horowitz, J.L. and Ma, S.G. (2008) Asymptotic Properties of Bridge Estimators in Sparse High-Dimensional Regression Models. *The Annals of Statistics*, **36**, 587-613. https://doi.org/10.1214/009053607000000875

[27] Stone, C.J. (1985) Additive Regression and Other Nonparametric Models. *The Annals of Statistics*, **13**, 689-705. https://doi.org/10.1214/aos/1176349548

[28] Gordon, W.J. and Riesenfeld, R.F. (1974) B-Spline Curves and Surfaces. In: *Computer Aided Geometric Design*, Academic Press, Cambridge, 95-126. https://doi.org/10.1016/B978-0-12-079050-0.50011-4

[29] Huang, J., Horowitz, J.L. and Wei, F. (2010) Variable Selection in Nonparametric Additive Models. *The Annals of Statistics*, **38**, 2282-2313. https://doi.org/10.1214/09-AOS781