一种用于影像遗传学关联分析的高阶统计量结构化稀疏算法

茹 锋 徐 锦 常 琪 阚丹会
(长安大学电子与控制工程学院 西安 710064)

摘 要 神经影像技术和分子遗传学的发展产生了大量的影像遗传学数据,极大地促进了复杂精神疾病的研究。但因为该数据的特征维度过高且相关性的度量都是假设数据服从高斯分布,所以传统的算法往往无法很好地解释两类数据之间的依赖关系。为了解决传统算法的问题,文中提出了一种对大量SNP和fMRI数据进行关联分析的方法,该方法通过构建稀疏的特征网络结构来指导fused lasso进行特征选择,与此同时,该方法利用高阶统计量提取出具有统计显著性的变量,从而识别出与精神疾病有关的生物标记物。实验结果表明,在模拟数据中所提算法得到的典型向量值的分布与实际数据中值的分布几乎一致且得到的相关系数与数据集中实际的相关系数最接近,所提算法的平均相关系数最高达到81%,比L1-SCCA提高了约20%,比FL-SCCA提高了约3%;在真实数据中,相比另外两种算法,所提算法可以找出更多的对精神分裂症有潜在影响的基因与脑区。实验结果证明:该算法可以在合理时间内有效识别出风险基因和异常脑区。

关键词 影像遗传学,关联分析,稀疏表示,特征选择,高阶统计量

中图法分类号 TP301.6 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.04.010

High Order Statistics Structured Sparse Algorithm for Image Genetic Association Analysis

RU Feng XU Jin CHANG Qi KAN Dan-hui
(School of Electronic Control, Chang'an University, Xi'an 710064, China)

Abstract The development of neuroimaging technology and molecular genetics has produced a large number of imaging genetic data, which has greatly promoted the study of complex mental diseases. However, because the dimensions of the data are too high and the correlation measure is based on the assumption that data obey Gaussian distribution, traditional algorithms often fail to explain the dependencies between two types of data. In order to solve the shortcomings of traditional algorithms, this paper proposed a method for correlation analysis of a large number of SNP and fMRI data. This method guides fused lasso to perform feature selection by constructing a network structure of features, and uses higher-order statistics to extract statistically significant variables. Thus, biomarkers associated with mental illness are identified. The experimental results show that the distribution of typical vector values obtained by the algorithm in simulation data are almost consistent with the real data, and the correlation coefficient obtained is the closest to the correlation coefficient in the real dataset. The average correlation coefficient of the proposed algorithm is up to 81%, which is about 20% higher than L1-SCCA and about 3% higher than FL-SCCA. Compared with the other two algorithms in real data, the proposed algorithm can find more genes and brain regions that have potential effects on schizophrenia. The experimental results show that the proposed algorithm can effectively identify risk genes and abnormal brain regions within a reasonable time.

Keywords Image genetics, Correlation analysis, Sparse representation, Feature selection, Higher-order statistics

1 引言

精神疾病的研究一直都是脑科学研究的重要部分,近年来受到了广泛关注。这类疾病大多是由各种遗传因素与外界环境相互作用引起的[1-2],作为一种高度可遗传的疾病,其诊断一直是医学界的一大难题[3]。脑区域异常和遗传变异是研究脑部精神疾病的重要标志。近年来,随着分子遗传学与神经影像技术的发展,影像遗传学结合脑部成像与遗传变异,极大地促进了复杂的精神疾病的研究。目前,从大量脑部图像数据和基因数据中找到了它们之间的关联信息,进而从关联信息中发现与精神分裂症等复杂疾病相关的生物标记物仍然存在很大的挑战。

到稿日期:2018-08-27 返修日期:2018-11-04 本文受西安市智慧高速公路信息融合与控制重点实验室(201805062ZD13CG46)资助。
茹 锋(1969-),男,博士,教授,主要研究方向为数据挖掘、模式识别,E-mail:35831406@qq.com(通信作者);徐 锦(1995-),女,硕士,主要研究方向为机器学习、模式识别;常 琪(1992-),女,硕士,主要研究方向为机器学习、模式识别;阚丹会(1993-),女,硕士,主要研究方向为机器学习、模式识别。

影像遗传学数据常采用精神病患者的影像和遗传数据，例如功能磁共振成像（functional Magnetic Resonance Imaging，fMRI），这种成像方式是基于血氧饱和度所依赖的对比测量脑部动态运动的一种成像；结构性磁共振成像（Structure Magnetic Resonance Imaging，sMRI），用于评估灰质、白质和脑脊液的体积和密度；弥散张量成像（Diffusion Tensor Imaging，DTI），用于追踪脑部白质束；磁共振波谱分析（Magnetic Resonance Spectroscopy，MRS），用来获取脑组织相关的生化信息；单核苷酸多态性（single nucleotide polymorphism，SNP），基因的一种表达形式。这些数据具有高维度和样本数量相对较少的特点，这些特点成为这类研究主要面对的难点问题。近些年，一些机器学习方法被用于寻找这两类数据的关联性，例如对 SNP 和 fMRI 数据同时降维，再度量这两类数据的相关性，如Parallel ICA[4]；使用 lasso，fused lasso 等以 L1 和 L2 范数为惩罚项的正则化方法进行稀疏的 sparse PLS[5]，sparse CCA[6-7] 等。

影像遗传学开始采用“大数据”技术，同时对全基因组和全脑影像数据进行建模，在全基因组和全脑水平上加入先验信息，将单核苷酸序列与表型功能相关的特征进行聚类[8]。由于 SNP 与脑体素的数量巨大，一些研究人员采用稀疏 PLS（Partial Least Squares）、稀疏 CCA（Canonical Correlation Analysis）、稀疏 RRR（Reduced-Rank Regression）、稀疏多规范相关分析（SMCCA）[9] 等多种方法来捕捉数据集之间的复杂关系。其中，稀疏 CCA 模型因可以发现双多变量之间的联系并选择出相关的特征而被广泛应用于影像基因学的研究中[10]。

虽然这些方法可以发现这两种数据之间的相关关系，但仍存在不足，主要表现在这些相关特征的度量都是基于高斯分布计算相关系数或是协方差的，但实际的基因和脑影像数据并不一定服从高斯分布，如果数据不是严格地服从高斯分布，那么仅考虑二阶统计量很难发现真正有意义的信息。

本文针对上述 SNP 与 fMRI 数据的特点和已有关联分析算法中存在的缺陷，提出一种新的基于 CCA 与 ICA 的探索大量基因与大量脑影像数据之间关联性的方法，以提取出两类数据中高度相关的特征并保证特征相互独立。一方面，通过数据驱动获得特征的网络结构并将之作为先验，以指导 fused lasso 进行特征选择；另一方面，利用高阶统计量提取出统计独立和非高斯的成分，并筛选出潜在的重要特征。使用病人的 SNP 和 fMRI 数据对本文方法进行测试。实验结果表明，本文方法可以在合理时间内有效找出与精神分裂症有关的基因和脑区。

2　数学模型

2.1　典型相关分析

典型相关分析（Canonical Correlation Analysis，CCA）[11] 是由著名学者 Hotelling 于 1936 年提出的一种经典的用于多元统计分析的方法。当前，该算法已被广泛地应用于模式识别与机器学习领域。这里，用 $\boldsymbol{X}\in\mathbb{R}^{n\times p}$ 和 $\boldsymbol{Y}\in\mathbb{R}^{n\times q}$ 分别表示 SNP 与 fMRI 数据，$\boldsymbol{X},\boldsymbol{Y}$ 均有 n 个样本，其中 $\boldsymbol{X}$ 有 p 维特征，$\boldsymbol{Y}$ 有 q 维特征。CCA 算法的目的是找到 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 各自特征之间的线性组合，即 $U=\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}$，$V=\boldsymbol{v}^{\mathrm{T}}\boldsymbol{Y}$。其中，$\boldsymbol{u}\in\mathbb{R}^{p\times 1}$，$\boldsymbol{v}\in\mathbb{R}^{q\times 1}$，用这两个综合变量 U 和 V 来表征两个数据集 $\boldsymbol{X},\boldsymbol{Y}$ 的整体相关关系，使得综合变量 U 和 V 之间的 Pearson 相关系数 $\rho_{U,V}$ 达到最大，公式如下：

$$\rho_{U,V}=Corr(U,V)=\frac{Cov(U,V)}{\sqrt{D_U}\sqrt{D_V}}=\frac{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{v}}{\sqrt{u^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{u}}\sqrt{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{v}}} \tag{1}$$

2.2　稀疏典型相关分析的两种形式

近年来，各种稀疏 CCA 算法被广泛应用于影像遗传学的研究，即通过在 CCA 的基础上加入惩罚项来设计合适的模型以达到稀疏效果。本文则是研究两个高维数据集 SNP 与 fMRI 之间的关联性，通过分配系数权重的大小来表示特征的重要程度，即不重要的特征系数的权重很小或为 0，重要特征的系数权重较大，从而从大量特征中找出与精神分裂症密切相关的潜在变量。Witten 等[12] 提出的稀疏 CCA 算法的统一表示形式如下：

$$\begin{aligned}&\max\ mize_{\boldsymbol{u},\boldsymbol{v}}\ \boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{v}\\&\text{subject to }\ \|\boldsymbol{u}\|^2\leqslant 1,\|\boldsymbol{v}\|^2\leqslant 1,P_1(\boldsymbol{u})\leqslant c_1,P_2(\boldsymbol{v})\leqslant c_2\end{aligned} \tag{2}$$

其中，P_1 和 P_2 是凸惩罚函数，通过选择合适的 P_1 和 P_2 使 $\boldsymbol{u}$ 和 $\boldsymbol{v}$ 达到稀疏，P_1 和 P_2 最常用的两种惩罚形式为：

1）P_1 是 L1（lasso）惩罚（L1-SCCA），即 $P_1(\boldsymbol{u})=\|\boldsymbol{u}\|_1$，通过选择合适的 c_1 来控制 $\boldsymbol{u}$ 的稀疏度，假设 $1\leqslant c_1\leqslant\sqrt{P_1}$。

2）P_1 是 fused lasso 惩罚（FL-SCCA）[13]，即 $P_1(\boldsymbol{u})=\sum_j|\boldsymbol{u}_j|+\sum_j|\boldsymbol{u}_j-\boldsymbol{u}_{j-1}|$，这个惩罚会产生稀疏且平滑的典型向量 $\boldsymbol{u}$，第一项 $\sum_j|\boldsymbol{u}_j|$ 使系数具有稀疏性，第二项 $\sum_j|\boldsymbol{u}_j-\boldsymbol{u}_{j-1}|$ 使系数间的差分具有稀疏性。FL-SCCA 是两个相邻系数之差，所以被用于特征具有自然顺序的数据集中。

2.3　稀疏 CCA-ICA 模型

由于基因与影像数据的特征维度远远高于样本数量，即 $p,q\gg n$，为了改善过拟合现象，我们采用基于正则化的稀疏表示方法，通过给高维矩阵乘以一个稀疏向量，将大部分特征置为零，仅保留主要的显著特征。本文为了找到 SNP 与 fMRI 中具有强相关性的重要特征，考虑基因的连锁不平衡、3D 影像的空间结构信息以及特征间的连接权重与正负相关性，利用特征的网络结构指导 fused lasso 进行特征选择[7,18]，构造两类数据的惩罚项，如式（3）所示：

$$\begin{aligned}P_1(\boldsymbol{u})&=\|\boldsymbol{u}\|_1+|\boldsymbol{u}|^{\mathrm{T}}L_1|\boldsymbol{u}|\\P_2(\boldsymbol{v})&=\|\boldsymbol{v}\|_1+|\boldsymbol{v}|^{\mathrm{T}}L_2|\boldsymbol{v}|\end{aligned} \tag{3}$$

其中，L_1,L_2 分别是 SNP 与 fMRI 的 Laplacian 矩阵，有助于改善特征关联中非线性的影响。采用这种惩罚形式的主要优势在于：如果我们具有先验信息，例如有基因标记的通路信息，它们被同时选中的可能性会较高；在不同的疾病中基因和脑影像的标记发挥着不同的作用，也就是说，某些标记在某一疾病中是有意义的，而在其他疾病中是不相关的，因此在第一项中加入了 lasso 惩罚以保证模型的稀疏性。根据式（1）和式（3），得到目标函数为：

$$\begin{aligned}&\max_{\boldsymbol{u},\boldsymbol{v}}\ \boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{v}\\&\text{subject to }\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}=1,\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}=1,P_1(\boldsymbol{u})\leqslant c_1,P_2(\boldsymbol{v})\leqslant c_2\end{aligned} \tag{4}$$

其中，c_1，c_2 为常数，用于控制 $\boldsymbol{u}$，$\boldsymbol{v}$ 的稀疏程度。

另外，我们要求提取出的每类重要特征相互独立，负熵是信息论中衡量随机变量独立性的度量，被广泛应用于 ICA。负熵越大，表明变量之间相互独立的概率越大，非高斯性越强[14]。根据最大熵原理，我们采用一种基于期望的方法来近似估计负熵，文献[15]表明，这种近似值比使用传统的基于累积量的近似更精确，由于 SNP 与 fMRI 数据的复杂性，我们认为这两类数据均具有非高斯性，表达形式如下：

$$\max_{\boldsymbol{u}}[E(G(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}))-E(G(w_1))]^2 \tag{5}$$

$$\text{subject to } \boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}=1$$

$$\max_{\boldsymbol{v}}[E(G(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{Y}))-E(G(w_2))]^2 \tag{6}$$

$$\text{subject to } \boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}=1$$

其中，w_1，w_2 为标准的高斯变量，$G(\cdot)$ 表示非二次函数，通常选择如下两种形式[14]：

$$G_1(w)=\frac{1}{a_1}\log\cosh(a_1 w)$$
$$g_1(w)=\tanh(a_1 w) \tag{7}$$

$$G_2(w)=-\frac{1}{a_2}\exp\left(-\frac{a_2 w^2}{2}\right)$$
$$g_2(w)=w\exp\left(\frac{-a_2 w^2}{2}\right) \tag{8}$$

通常，$1\leqslant a_1\leqslant 2$，$a_2\approx 1$，$g(\cdot)$ 为 $G(\cdot)$ 的导函数。

为了封装上述 3 个最大化目标，最直观的方法是将三者结合成一个单一的目标函数并在各目标函数之间实现较好的折中。因此，我们采用了最简单的子目标加权线性和的方式，从而得到最终的优化目标：

$$\max_{\boldsymbol{u}_1,\boldsymbol{u}_2}\alpha[\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{v}]+\beta[E(G(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}))-E(G(w_1))]^2+\theta[E(G(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{Y}))-E(G(w_2))]^2 \tag{9}$$

$$\text{subject to } \boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}=1,\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}=1,P_1(\boldsymbol{u})\leqslant c_1,P_2(\boldsymbol{v})\leqslant c_2$$

其中，α，β，θ 分别是各子函数所占的权重，满足 $\alpha+\beta+\theta=1$，具体每个权重的选择将在 2.4 节介绍。

2.4 各子目标函数的权重选择

考虑各子函数可能具有不同的收敛速度，为各子函数分配合适的权重非常重要。我们将 α，β，θ 3 个权重分别分解为 3 个部分：$\alpha_{\mathrm{sig}}\cdot\alpha_{\mathrm{scale}}\cdot\alpha_{\mathrm{adj}}$，$\beta_{\mathrm{sig}}\cdot\beta_{\mathrm{scale}}\cdot\beta_{\mathrm{adj}}$，$\theta_{\mathrm{sig}}\cdot\theta_{\mathrm{scale}}\cdot\theta_{\mathrm{adj}}$，这里 α_{sig}，β_{sig}，θ_{sig} 为显著性因子，表示每个子目标的相对重要性，它们可以根据具体应用进行主观设置。α_{scale}，β_{scale}，θ_{scale} 为比例因子，定义如下：

$$\alpha_{\mathrm{scale}}=\frac{1}{|L_1(\boldsymbol{u},\boldsymbol{v})|},\beta_{\mathrm{scale}}=\frac{1}{|L_2(\boldsymbol{u})|},\theta_{\mathrm{scale}}=\frac{1}{|L_3(\boldsymbol{v})|} \tag{10}$$

其中，分母分别是对应子目标的 Lagrange 方程的模。α_{adj}，β_{adj}，θ_{adj} 为可调因子，用来平衡它们不同的收敛速度且在迭代的过程中实时更新。通常，梯度函数用来估计收敛速度，因此，可调因子被定义为如下形式：

$$\alpha_{\mathrm{adj}}=\frac{1}{(\|\nabla L_{1,\boldsymbol{u}}\|_2+\|\nabla L_{1,\boldsymbol{v}}\|_2)/2}$$
$$\beta_{\mathrm{adj}}=\frac{1}{\|\nabla L_{2,\boldsymbol{u}}\|_2} \tag{11}$$
$$\theta_{\mathrm{adj}}=\frac{1}{\|\nabla L_{3,\boldsymbol{v}}\|_2}$$

其中，$\nabla L_{1,\boldsymbol{u}}$，$\nabla L_{1,\boldsymbol{v}}$，$\nabla L_{2,\boldsymbol{u}}$，$\nabla L_{3,\boldsymbol{v}}$ 分别为对应子目标的 Lagrange 方程对 $\boldsymbol{u}$ 和 $\boldsymbol{v}$ 的一阶偏导。从以上的定义中可以看到，α_{adj}，β_{adj}，θ_{adj} 会在迭代的过程中在线调整，如果子目标变化较快，那么梯度的范数就会较大，相应地可调因子就会变小，这就使得所有的子目标并行改变。当结果接近最优时，梯度趋近于 0，因此，我们设置 α，β，θ 这 3 个权重如下：

$$\begin{cases}\alpha=\alpha_{\mathrm{sig}}\cdot\alpha_{\mathrm{scale}}\cdot\alpha_{\mathrm{adj}}, & \alpha_{\mathrm{adj}}\leqslant TH\\ \beta=\beta_{\mathrm{sig}}\cdot\beta_{\mathrm{scale}}\cdot\beta_{\mathrm{adj}}, & \beta_{\mathrm{adj}}\leqslant TH\\ \theta=\theta_{\mathrm{sig}}\cdot\theta_{\mathrm{scale}}\cdot\theta_{\mathrm{adj}}, & \theta_{\mathrm{adj}}\leqslant TH\\ \alpha=\alpha_{\mathrm{sig}}\cdot\alpha_{\mathrm{scale}}, & \alpha_{\mathrm{adj}}>TH\\ \beta=\beta_{\mathrm{sig}}\cdot\beta_{\mathrm{scale}}, & \beta_{\mathrm{adj}}>TH\\ \theta=\theta_{\mathrm{sig}}\cdot\theta_{\mathrm{scale}}, & \theta_{\mathrm{adj}}>TH\end{cases} \tag{12}$$

其中，TH 为预先设置的阈值，当迭代靠近最优解时，相应的梯度接近 0，此时权重变化为式(12)的后 3 个式子。

3 模型求解

3.1 求解方法

基于这个优化问题，主要的目标是使各个子目标函数同时达到最优，鉴于每个子目标函数的复杂性，需要依据各子函数所占权重 α，β，θ 来实现全局最优。为求解这个优化问题，我们构造如下的 Lagrange 方程：

$$\begin{aligned}L(\boldsymbol{u},\boldsymbol{v},\Gamma)=&\alpha[\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{v}]+\beta[E(G(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}))-E(G(w_1))]^2+\\&\theta[E(G(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{Y}))-E(G(w_2))]^2+\frac{\lambda_1}{2}|\boldsymbol{u}|^{\mathrm{T}}L_1|\boldsymbol{u}|+\\&\gamma_1\|\boldsymbol{u}\|_1+\frac{\mu_1}{2}\|\boldsymbol{u}\|_2^2+\frac{\lambda_2}{2}|\boldsymbol{v}|^{\mathrm{T}}L_2|\boldsymbol{v}|+\\&\gamma_2\|\boldsymbol{v}\|_1+\frac{\mu_2}{2}\|\boldsymbol{v}\|_2^2\end{aligned} \tag{13}$$

其中，$\Gamma=\{\lambda,\gamma,\mu\}\geqslant 0$ 为 Lagrange 乘子，通过交叉验证选择。

我们将上面的目标函数 L 分别对 $\boldsymbol{u}$，$\boldsymbol{v}$ 求偏导，并令其等于 0，即：

$$\frac{\partial L}{\partial\boldsymbol{u}}=0,\frac{\partial L}{\partial\boldsymbol{v}}=0 \tag{14}$$

整理得到：

$$2\beta[E(G(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}))-E(G(w_1))]E(Xg(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}))+(\mu_1 I+\lambda_1 L_1+\gamma_1\boldsymbol{D}_1)\boldsymbol{u}=-\alpha\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{v} \tag{15}$$

$$2\theta[E(G(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{Y}))-E(G(w_2))]E(Yg(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{Y}))+(\mu_2 I+\lambda_2 L_2+\gamma_2\boldsymbol{D}_2)\boldsymbol{v}=-\alpha\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{u} \tag{16}$$

其中，$\boldsymbol{D}_1$ 是对角元素为 $\frac{1}{2|u_i|}(i\in[1,p])$ 的对角矩阵，$\boldsymbol{D}_2$ 是对角元素为 $\frac{1}{2|v_j|}(j\in[1,q])$ 的对角矩阵。

由于 $\boldsymbol{u}$，$\boldsymbol{v}$ 都是未知的，使用交替最小二乘法不断更新 $\boldsymbol{u}$，$\boldsymbol{v}$，在每次的迭代过程中，首先固定 $\boldsymbol{v}$，根据式(15)来计算 $\boldsymbol{u}$，然后固定 $\boldsymbol{u}$，根据式(16)来计算，如此重复进行，直到算法收敛，使用 $\max\{\delta|\delta\in(u^{t+1}-u^t)\}<\varepsilon$ 和 $\max\{\delta|\delta\in(v^{t+1}-v^t)\}<\varepsilon$ 作为停止条件，本文中 ε 设为 10^{-5}。本文将在加入 fused lasso 惩罚的基础上结合结构约束，在先验信息不准确或不可得到时采用本文方法对精神分裂症患者的遗传基因与脑影像数据做关联分析，通过构造特征间图网络的惩罚项，并使用高

阶统计量从每类数据中分离出具有统计独立性的变量，来保证每一类中的特征相互独立。在求解的过程中多次迭代更新对角阵 $\boldsymbol{D}$ 和典型向量 $\boldsymbol{u}$，$\boldsymbol{v}$，找出 SNP 与 fMRI 之间的关联特征，进而发现与疾病有潜在关系的异常生物标记物。

3.2　算法流程

算法的主要计算步骤是交替最小二乘法，其过程可概括为：根据第 t 次估计出的 v^t，计算第 $t+1$ 次的 u^{t+1}，而 u^{t+1} 又被当作下一次迭代的已知条件加入到 v^{t+1} 的计算中，并分别验证相邻两次 $\boldsymbol{u}$，$\boldsymbol{v}$ 的差是否在误差范围内。

根据以上描述，本文算法的完整步骤可归纳如下：

1）初始化。在算法第一次迭代之前分别对 $\boldsymbol{u}$，$\boldsymbol{v}$ 和对角阵 $\boldsymbol{D}$ 进行初始化。

2）求解 $\boldsymbol{u}$。由前一次计算得到的 fMRI 数据的典型向量 $\boldsymbol{v}$，根据式(15)求解 SNP 数据的典型向量 $\boldsymbol{u}$。

3）求解 $\boldsymbol{v}$。根据式(16)，由前一次计算得到的 $\boldsymbol{u}$ 求解 $\boldsymbol{v}$。

4）判断收敛条件。$\max\{\delta|\delta\in(u^{t+1}-u^t)\}<\varepsilon$ 和 $\max\{\delta|\delta\in(v^{t+1}-v^t)\}<\varepsilon$，$\varepsilon$ 取 10^{-5}。

5）重复步骤 2）－步骤(4)，直到算法收敛。

3.3　算法的性能分析

从 3.2 节中的算法流程中可以看出，算法主要包括两个过程：迭代求解典型向量 $\boldsymbol{u}$，迭代求解典型向量 $\boldsymbol{v}$。采用交替最小二乘法求解，已知数据 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$，$\boldsymbol{u}$ 和 $\boldsymbol{v}$ 未知，通过初始化赋值后交替通过一个来更新另一个，通过设定合适的停止条件与最大迭代次数，可以很容易地求出 $\boldsymbol{u}$，$\boldsymbol{v}$。另外，对于 $\boldsymbol{u}$ 与 $\boldsymbol{v}$ 的求解，算法分别进行它们各自的迭代过程，而不是只使用一个大的迭代设置，这样直到两者均满足条件算法才会终止，保证了算法的有效性。

算法需要分别计算矩阵 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 各自的方差和协方差，其中 SNP 与 fMRI 数据的维度比较高，输出变量会随着输入数据集的规模呈线性变化，空间复杂度为 $O(n)$，其余计算均是简单的矩阵的加法、乘法与求逆运算。本文算法是由 Matlab 语言实现，在 Windows 7 平台下运行，在合理的时间内可以使算法达到收敛。另外，本文算法将每个特征看作图中的一个顶点，将特征之间的关联系数作为边的权重，对两类数据分别构建其网络图，这样在空间结构上有关联的特征会靠得更近，更有利于进行关联特征选择。

4　实验分析

4.1　实验数据

为了测试本文算法的有效性，我们选择了两个 L1-SCCA 算法与 FL-SCCA 算法与本文算法作比较。实验分别在 4 个模拟数据集和 79 个精神分类症病人的基因影像数据上进行。3 种算法均使用 Matlab 语言在 Windows 系统下实现，测试环境为 2.60 GHz CPU 和 128 GB 内存。

模拟实验数据集生成方法参考文献[16]。每个数据集均包含有 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$，$\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的真实权重系数 $\boldsymbol{u}$ 和 $\boldsymbol{v}$ 以及相关系数。其中，n 表示样本的数量，p 表示 $\boldsymbol{X}$ 的特征维度，q 表示 $\boldsymbol{Y}$ 的特征维度。我们遵循特征维度大、样本小的特点，通过分配系数权重 u，v 的大小来表示特征的重要程度，即不重要的特征系数权重很小或为 0，重要特征的系数权重较大，从而从大量特征中找出潜在的重要变量。生成模拟数据集的方法如下：1）根据预先设计的数据集的结构分别产生 u 与 v；2）生成重要变量 ζ 服从分布 $\zeta\sim N(0,\sigma_\zeta^2)$；3）生成数据集 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$，$\boldsymbol{X}$ 满足 $x=\zeta\boldsymbol{u}+\varepsilon_x$，$\boldsymbol{Y}$ 满足 $y=\zeta\boldsymbol{v}+\varepsilon_y$，其中，$\varepsilon_x$，$\varepsilon_y$ 为随机噪声向量，满足 ε_x，$\varepsilon_y\sim N(0,\sigma_\varepsilon^2 I)$，$u\in\mathbb{R}^p$，$v\in\mathbb{R}^q$。4 个模拟数据集的详情如表 1 所列。

表 1　模拟数据集

Table 1　Simulation dataset

数据集	n	p	q	相关系数
$Data_1$	80	100	120	0.6214
$Data_2$	100	250	600	0.8384
$Data_3$	100	250	600	0.7525
$Data_4$	100	500	900	0.6542

4.2　参数设置

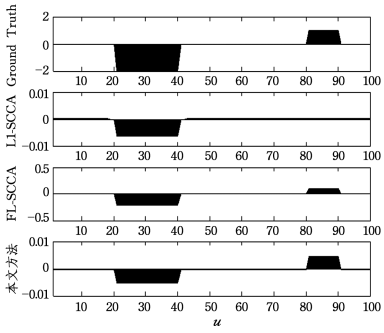
由式(15)和式(16)可知，有 6 个参数需要被调整，采用交叉验证的方式选择最优参数。我们将需要调整的参数限制在 $[10^{-2},10^{-1},10^0,10^1,10^2]$ 中，避免了盲目选择参数的缺陷，所有的参数都经过 5 折交叉验证产生，即：

$$CV(\lambda,\beta,\alpha)=\frac{1}{5}\sum_{j=1}^{5}Corr(\boldsymbol{X}_{\text{test}}\boldsymbol{u}_{\text{train}},\boldsymbol{Y}_{\text{test}}\boldsymbol{v}_{\text{train}}) \tag{17}$$

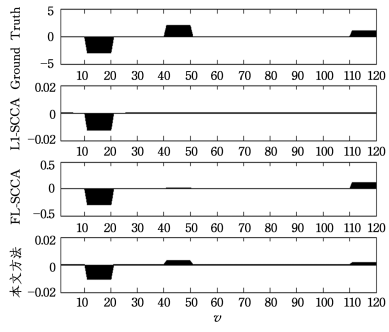
其中，$\boldsymbol{X}_{\text{test}}$，$\boldsymbol{Y}_{\text{test}}$ 为输入的测试集，$\boldsymbol{u}_{\text{trian}}$，$\boldsymbol{u}_{\text{train}}$ 为由训练集得到的典型向量，我们选择 $\arg\max CV(\lambda,\beta,\alpha)$ 作为最优参数。

4.3　模拟数据上的结果

模拟数据的评价标准为找到与真实相关系数最接近的一组 u，v，为了尽可能地减小训练集与测试集的差异对结果造成的不良影响，我们选择 5 次实验中训练集与测试集所得的相关系数之差 Δ_{corr} 最小的一组 u，v 作为最终的结果，如图 1 所示。



(a)



(b)

图 1　不同算法对典型向量 $\boldsymbol{u}$ 和 $\boldsymbol{v}$ 的估计

Fig. 1　Estimation of typical vector $\boldsymbol{u}$ and $\boldsymbol{v}$ by different algorithms

图 1 分别显示了典型向量 $\boldsymbol{u},\boldsymbol{v}$ 的真实值与估计值的分布，横坐标表示数据集的特征索引，纵坐标表示 $\boldsymbol{u},\boldsymbol{v}$ 的值，即特征权重值越大，表示该特征越重要。第一行为数据集中真实存在的 $\boldsymbol{u},\boldsymbol{v}$ 值，第 2－4 行分别为 L1-SCCA，FL-SCCA 以及本文方法对 $\boldsymbol{u},\boldsymbol{v}$ 的估计值。可以直观看出，本文算法明显优于另外两个算法。

为了更准确地说明本文算法的有效性，将表 1 中的 4 个模拟数据集均按照8∶2的比例划分训练集与测试集，进行 5 折交叉验证，每一折估计出的相关系数与其对应的平均值如表 2 所列。其中，NAN 表示用此方法计算典型向量 $\boldsymbol{u},\boldsymbol{v}$ 时失败，加黑数值为 5 次实验的平均值。可以明显看出，对于训练集，本文算法在 $Data_2$，$Data_3$，$Data_4$ 上得到的平均相关系数均显著大于其他两种算法得到的平均相关系数；对于测试集，本文算法也明显优于 L1-SCC，FL-SCCA 两种算法。一般来讲，测试集的结果比训练集的结果更能体现算法的有效性。

表 2 4 个模拟数据集上的 5 折交叉验证结果

Table 2 5-fold cross-validation results on 4 simulated datasets

Datasets/Methods		L1-SCCA					mean	FL-SCCA					mean	SC-SCCA					mean
Training Results	$Data_1$	0.54	0.62	0.67	0.58	0.60	**0.60**	0.55	0.57	0.65	0.66	0.66	**0.62**	0.63	0.67	0.47	0.64	0.65	**0.61**
	$Data_2$	NAN	NAN	NAN	NAN	NAN	NAN	0.82	0.80	0.76	0.70	0.83	**0.78**	0.81	0.81	0.83	0.80	0.79	**0.81**
	$Data_3$	0.64	0.61	0.56	0.43	0.69	**0.57**	0.67	0.64	0.75	0.17	0.71	**0.59**	0.77	0.76	0.70	0.78	0.75	**0.75**
	$Data_4$	0.32	0.41	0.50	0.47	0.61	**0.46**	0.66	0.67	0.28	0.43	0.57	**0.52**	0.64	0.65	0.65	0.65	0.66	**0.65**
Testing Results	$Data_1$	0.64	0.57	0.36	0.51	0.68	**0.55**	0.80	0.81	0.42	0.42	0.37	**0.56**	0.63	0.38	0.62	0.77	0.47	**0.57**
	$Data_2$	NAN	NAN	NAN	NAN	NAN	NAN	0.72	0.83	0.92	0.65	0.40	**0.71**	0.79	0.81	0.66	0.60	0.55	**0.68**
	$Data_3$	0.65	0.53	0.47	0.25	0.54	**0.49**	0.73	0.45	0.62	0.16	0.79	**0.55**	0.65	0.76	0.86	0.62	0.79	**0.74**
	$Data_4$	0.47	0.32	0.54	0.43	0.59	**0.47**	0.54	0.65	0.38	0.47	0.58	**0.52**	0.70	0.65	0.62	0.62	0.65	**0.65**

表 3 列出了利用 5 次实验中训练集与测试集所得的相关系数之差最小的一组典型向量 $\boldsymbol{u},\boldsymbol{v}$ 求得的相关系数，其中加黑数值为 3 种算法中最接近真实相关系数的值。如果考虑数据集真实的相关系数，可以看到，本文算法和 FL-SCCA 算法具有较小的平均评估误差，而且与真实相关系数更接近。也就是说，相比 L1-SCCA，本文算法与 FL-SCCA 在训练结果上更精确。另外，本文算法具有最小的评估误差，且在 $Data_3$，$Data_4$ 上求得的相关系数误差为 0。

表 3 完整数据集上的相关系数结果

Table 3 Correlation coefficient results on complete dataset

Methods/DataSets	$Data_1$	$Data_2$	$Data_3$	$Data_4$	Avg Error
True cc	0.62	0.84	0.75	0.65	—
L1-SCCA	0.58 (－0.04)	0.54(－0.30)	0.47(－0.28)	0.52(－0.13)	0.19
FL-SCCA	**0.62**(0.00)	0.77(－0.07)	0.64(－0.11)	**0.65**(0.00)	0.05
SC-SCCA	**0.63**(＋0.01)	**0.80**(－0.04)	**0.75**(0.00)	**0.65**(0.00)	0.01

图 2 更直观地展示了 3 种算法在不同数据集上的实验结果，前 3 列代表 3 种不同的方法，第 4 列表示数据集实际的相关系数，可以看出本文算法与 FL-SCCA 方法明显优于 L1-SCCA 方法，尤其在 $Data_2$－$Data_4$ 上。另外，在 $Data_2$ 和 $Data_3$ 上，本文算法优于 FL-SCCA 方法。

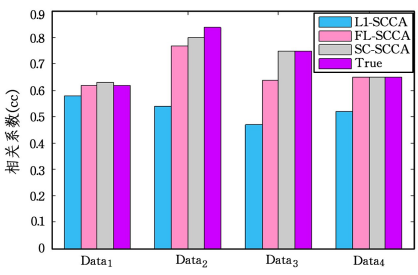


图 2 每种方法在各数据集上相关系数的比较

Fig. 2 Comparison of correlation coefficients of each method on each dataset

4.4 真实数据上的结果

我们在相同的实验设置下，比较 3 种算法分别在真实的基因影像数据集上的 5 折交叉验证结果，表 4 列出了每种算法在训练集与测试集上每折的相关系数与其对应的平均值。从中可以看出 3 种算法均存在微弱的过拟合问题，这是由于 SNP 与 fMRI 数据的维度较高。在训练集上，虽然本文算法得到的相关系数小于 L1-SCCA 算法得到的相关系数，但 L1-SCCA 算法在测试集上计算出的相关系数相差较大，过拟合问题也较严重。另外，3 种算法得到的相关系数均较低，这是由基因与脑影像这两种完全不同类的数据本身所决定的，参数的微小变化都会使实验结果有较大的差异，且所选参数在重复实验中变化很大。

由最大化 SNP 与 fMRI 两者的相关性求出的典型向量自动选择出与精神分裂症有关的重要基因和脑区。表 5－表 7 分别列出了 3 种方法得到的风险基因。表 8 列出了 3 种方法得到的异常脑区。

表 4 基因影像数据 5 折交叉验证的相关系数

Table 4 Correlation coefficient of 5-fold cross-validation of genetic image data

方法	训练结果					均值	测试结果					均值
L1-SCCA	0.42	0.38	0.40	0.32	0.39	**0.40**	0.30	0.35	0.41	0.28	0.37	**0.34**
FL-SCCA	0.36	0.41	0.49	0.29	0.37	**0.38**	0.19	0.42	0.36	0.38	0.40	**0.35**
SC-SCCA	0.35	0.40	0.39	0.43	0.38	**0.39**	0.29	0.42	0.37	0.41	0.38	**0.37**

表 5　L1-SCCA 算法得到的基因

Table 5　Gene obtained by L1-SCCA algorithm

特征索引	SNP ID	基因名称	染色体号
576429	rs12427675	CSNK1A1L[18]	13
576427	rs1555639	CSNK1A1L[18]	13
341736	rs10156115	C7orf16	7
772704	rs132966	**PLA2G6**	22
772706	rs132975	**PLA2G6**	22
450682	rs7033245	NOTCH1[18]	9
84891	rs10183370	**B3GNT2**	2
307508	rs9452354	EPHA7[18]	6

表 6　FL-SCCA 算法得到的基因

Table 6　Gene obtained by FL-SCCA algorithm

特征索引	SNP ID	基因名称	染色体号
307063	rs1334628	MAP3K7	6
368066	rs12705191	ORC5L	7
235444	rs7705425	PRLR	5
235393	rs4538595	AGXT2[18]	5
307101	rs1391506	MAP3K7	6
186132	rs12512830	SLIT2[18]	4
772703	rs4376	**PLA2G6**	22
368074	rs17586018	**CNTNAP2**	7
307248	rs958847	MAP3K7	6

表 7　本文方法得到的基因

Table 7　Gene obtained by the proposed method

特征索引	SNP ID	基因名称	染色体号
186132	rs12512830	SLIT2[18]	4
772703	rs4376	**PLA2G6**	22
368988	rs2888583	ZNF767[19]	7
368424	rs17824995	**CNTNAP2**	7
368568	rs4526286	**CNTNAP2**	7
341733	rs11772988	C7orf16	7
450682	rs7033245	NOTCH1[18,20]	9
235391	rs7717823	AGXT2[18]	5
307370	rs10455181	MAP3K7	6
235393	rs4538595	AGXT2[18]	5
175426	rs10433485	ST6GAL1	3
235399	rs163907	AGXT2[18]	5
175457	rs270144	RPL39L	3
84892	rs6545946	**B3GNT2**	2
368178	rs6945085	**CNTNAP2**	7
772696	rs5750542	**PLA2G6**	22

表 5－表 7 中加黑的基因 B3GNT2，CNTNAP2，PLA2G6 为已经公布的与精神分裂症有潜在关系的基因[21-22]，分别位于 2 号、7 号和 22 号染色体上。从表中可以看出，L1-SCCA 算法选择出了来自 6 条基因的 8 个 SNP 位点，成功找出 B3GNT2 和 PLA2G6 两条基因；FL-SCCA 算法选择出了来自 7 条基因的 9 个 SNP 位点，成功找出 PLA2G6 和 CNTNAP2 两条基因；本文算法选出了来自 11 条基因的 16 个 SNP 位点，成功找出 B3GNT2，CNTNAP2，PLA2G6 这 3 条基因。MAP3K7 基因是 FL-SCCA 与本文算法共同选择出来的，C7orf16 基因是 L1-SCCA 与本文算法共同选择出来的。另外，对比 3 种算法选择出的基因可知，基因 CSNK1A1L，NOTCH1，EPHA7，AGXT2，SLIT2 在文献[18]中也被选择出来，这说明了本文算法的有效性。

表 8　与 SNP 相关的脑区

Table 8　Brain region associated with SNP

方法	体素个数	脑区编号	脑区名称	中文名称
L1-SCCA	20	30	Insula_R[18]	脑岛
	20	14	Frontal_Inf_Tri_R	三角部额下回
	3	16	Frontal_Inf_Orb_R	眶部额下回
FL-SCCA	9	30	Insula_R[18]	脑岛
本文方法	5	56	Fusiform_R[9]	梭状回
	3	85	Temporal_Mid_L[9,23]	颞中回
	1	90	Temporal_Inf_R[9]	颞下回
	2	7	Frontal_Mid_L[9,23]	额中回
	6	86	Temporal_Mid_R[9]	颞中回
	2	30	Insula_R[18]	脑岛
	3	8	Frontal_Mid_R[9,23]	额中回
	3	89	Temporal_Inf_L[9]	颞下回
	2	38	Hippocampus_R[9,24]	海马
	2	13	Frontal_Inf_Tri_L[9]	三角部额下回
	1	37	Hippocampus_L[9,24]	海马
	1	9	Frontal_Mid_Orb_L	眶部额中回

表 8 显示了 3 种算法选择出的相关脑区及对应编号，其中 30 号脑区 Insula_R 是 3 种算法共同选择出来的。可以看出，L1-SCCA 与 FL-SCCA 算法只选择出了很少的可能存在异常的脑区，而本文算法选择出了 12 个脑区。根据对精神分裂症患者的功能脑影像的研究[25]，健康人和精神分裂症患者之间主要的差异位于 Hippocampus，Frontal，Temporal 和 Insula 脑叶中，L1-SCCA 算法定位了 Insula 和 Frontal 脑叶，FL-SCCA 算法只定位出了 Insula 脑叶，而本文算法定位出了 Hippocampus，Frontal，Temporal 和 Insula 4 个脑叶位置，这说明本文算法较另外两种算法具有很大的优势，并且这几个脑区在文献[9，18，23]中也被找到。

结束语　本文提出了一种在大量的基因和脑影像数据之间找到关联信息的方法，利用特征的网络结构和高阶统计量，通过交替最小二乘法选择出潜在的重要变量，进而发现与精神分裂症有关的生物标记物。使用模拟数据和精神分裂症患者的基因影像数据进行实验，本文方法可以在合理时间内有效识别风险基因和异常脑区。另外，如何融合多种影像数据和基因数据也是今后研究的重点之一。

参 考 文 献

[1] RIPKE S，NEALE B M，CORVIN A，et al. Biological insights from 108 schizophrenia-associated genetic loci[J]. Nature，2014，511(7510)：421.

[2] LIU J，CALHOUN V D. A review of multivariate analyses in imaging genetics [J]. Frontiers in Neuroinformatics，2014，8(29)：1-11.

[3] EDITION F. Diagnostic and statistical manual of mental disorders[M]. Arlington：American Psychiatric Publishing，2013.

[4] LIU J，PEARLSON G，WINDEMUTH A，et al. Combining fMRI and SNP data to investigate connections between brain function and genetics using parallel ICA[J]. Human Brain Mapping，2009，30(1)：241-255.

[5] LE FLOCH É，GUILLEMOT V，FROUIN V，et al. Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse

Partial Least Squares[J]. Neuroimage,2012,63(1):11-24.

[6] CHI E C,ALLEN G I,ZHOU H,et al. Imaging genetics via sparse canonical correlation analysis[C]// 2013 IEEE 10th International Symposium on Biomedical Imaging (ISBI). IEEE,2013:740-743.

[7] DU L,HUANG H,YAN J,et al. Structured sparse canonical correlation analysis for brain imaging genetics: an improved GraphNet method[J]. Bioinformatics,2016,32(10):1544-1551.

[8] BOGDAN R,SALMERON B J,CAREY C E,et al. Imaging Genetics and Genomics in Psychiatry: A Critical Review of Progress and Potential[J]. Biological Psychiatry,2017,82(3):165-175.

[9] HU W,LIN D,CAO S,et al. Adaptive sparse multiple canonical correlation analysis with application to imaging (epi) genomics study of schizophrenia[J]. IEEE Transactions on Biomedical Engineering,2018,65(2):390-399.

[10] DU L,HUANG H,YAN J,et al. Structured sparse CCA for brain imaging genetics via graph OSCAR[J]. BMC Systems Biology,2016,10(3):68-77.

[11] HOTELLING H. Relations Between Two Sets of Variates[J]. Biometrika,1936,28(3/4):321-377.

[12] WITTEN D M,TIBSHIRANI R J. Extensions of Sparse Canonical Correlation Analysis with Applications to Genomic Data [J]. Statistical Applications in Genetics & Molecular Biology,2009,8(1):1-27.

[13] TIBSHIRANI R,SAUNDERS M,ROSSET S,et al. Sparsity and smoothness via the fused lasso[J]. Journal of the Royal Statistical Society,2010,67(1):91-108.

[14] HYVÄRINEN A. Fast and Robust Fixed-Point Algorithms for Independent Component Analysis[J]. IEEE Transactions on Neural Networks,1999,10(3):626-634.

[15] HYVÄRINEN A. New Approximations of Differential Entropy for Independent Component Analysis and Projection Pursuit[J]. Advances in Neural Information Processing Systems,1997,10:273-279.

[16] CHEN X,LIU H. An Efficient Optimization Algorithm for Structured Sparse CCA, with Applications to eQTL Mapping [J]. Statistics in Biosciences,2012,4(1):3-26.

[17] HASTIE T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis [J]. Biostatistics,2009,10(3):515-534.

[18] FANG J,LIN D,SCHULZ C,et al. Joint sparse canonical correlation analysis for detecting differential imaging genetics modules[J]. Bioinformatics,2016,32(22):3480-3488.

[19] HU W,LIN D,CALHOUN V D,et al. Integration of SNPs-FMRI-methylation data with sparse multi-CCA for schizophrenia study[C]// Engineering in Medicine & Biology Society. IEEE,2016.

[20] CAO H,LIN D,DUAN J,et al. Biomarker Identification for Diagnosis of Schizophrenia with Integrated Analysis of fMRI and SNPs[C]// IEEE International Conference on Bioinformatics and Biomedicine. 2012:223-228.

[21] LAW M H,COTTON R G,BERGER G E. The role of phospholipases A2 in schizophrenia[J]. Molecular Psychiatry, 2006, 11(6):547-556.

[22] SANDERS A R,DUAN J,DRIGALENKO E I,et al. Transcriptome study of differential expression in schizophrenia[J]. Human Molecular Genetics,2013,22(24):5001-5014.

[23] CAO H,DUAN J,LIN D,et al. Integrating fMRI and SNP data for biomarker identification for schizophrenia with a sparse representation based variable selection method[J]. Bmc Medical Genomics,2013,6 (3):1-8.

[24] OZDEMIR H,ERTUGRUL A,BASAR K,et al. Differential effects of antipsychotics on hippocampal presynaptic protein expressions and recognition memory in a schizophrenia model in mice[J]. Progress in neuro-psychopharmacology & biological psychiatry,2012,39(1):62-68.

[25] KIRCHER T T,THIENEL R. Functional brain imaging of symptoms and cognition in schizophrenia[J]. Progress in Brain Research,2005,150(2):299-308.