DOI:10.3969/j.issn.1005-202X.2020.05.010

医学影像物理

基于稀疏表示变量选择方法的影像遗传学数据分析

谢忠翔,武杰,项华中
上海理工大学医疗器械与食品学院, 上海 200093

【摘要】**目的:**采用影像遗传学研究方法探索精神分裂症的影像遗传学特征。**方法:**在传统稀疏回归模型的基础上,改进了其在不同范数条件下进行变量选择的能力,形成一种基于稀疏表示变量选择算法,并将该算法应用于208个受试者的41 236个功能磁共振成像数据和722 177个单核苷酸多态性数据的综合分析。通过对两类数据施加不同的权重因子,并使用不同的L_p(p=0、0.5、1)范数分别对模型进行求解,筛选出两类数据在不同条件下的显著特征。**结果:**基因DAOA和HTR2A在3种范数下均被筛选出。此外,在影像学数据方面,发现中央前回、枕上回、顶下缘角回、角回、内侧和旁扣带脑回、后扣带回脑区与精神分裂症相关,此发现与先前精神分裂症的临床医学研究结果一致。**结论:**基于稀疏表示变量选择方法应用于影像遗传学数据分析是一个有效可行的途径,为今后精神分裂症的影像遗传学研究提供了一种新的研究思路。

【关键词】精神分裂症;稀疏表示;变量选择方法;单核苷酸多态性;功能磁共振成像
【中图分类号】R318 【文献标志码】A 【文章编号】1005-202X(2020)05-0584-05

Sparse representation-based variable selection algorithm for analysis of imaging genetics data

XIE Zhongxiang, WU Jie, XIANG Huazhong
School of Medical Instrument and Food Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China

Abstract: Objective To explore the imaging genetic characteristics of schizophrenia using imaging genetics method. **Methods** A sparse representation-based variable selection algorithm with improved ability of variable selection under different norm conditions based on traditional sparse regression model is proposed. The proposed algorithm was applied for the comprehensive analysis of 41 236 functional magnetic resonance imaging data and 722 177 single nucleotide polymorphisms data of 208 subjects. By applying different weight factors to the two types of data and using different L_p (p=0, 0.5, 1) norms for solving the models, the significant features of the two types of data were extracted. **Results** DAOA and HTR2A genes were extracted under 3 different L_p norms. In addition, the results of imaging data suggested that precentral, occipital_sup, parietal_inf, angular, cingulum_mid, cingulum_post were associated with schizophrenia, which was consistent with previous clinical studies on schizophrenia. **Conclusion** Sparse representation-based variable selection algorithm is an effective and feasible approach for the analysis of image genetics data, providing a new direction for the image genetics study on schizophrenia.

Keywords: schizophrenia; sparse representation; variable selection algorithm; single nucleotide polymorphisms; functional magnetic resonance imaging

前 言

在信号恢复和重要成分识别等应用中,稀疏表示(包括压缩感知)受到了研究者的关注,该方法最主要的一个特点是:要恢复的信号或要检测的重要成分数量通常等于样本数[1-2]。然而,在基因组数据或生物医学成像数据分析中,样本数量通常远远少于变量数量。由于现有压缩感知方法不能直接应用于这些数据进行变量选择[3-4],因此,本研究在传统稀疏回归模型基础上,改进了它在不同范数条件下进行变量选择能力,形成一种基于稀疏表示变量选择算法(Sparse Representation-based Variable Selection Algorithm, SRVS),且结果表明,SRVS能更准确地选择精神分裂症影像遗传学特征,筛选有效性更高。

本研究利用基于稀疏回归模型的SRVS,改变L_p(p=0、0.5和1)范数,对208名受试者(96名精神分裂

【收稿日期】2019-12-17
【基金项目】国家自然科学基金(61605114);上海理工大学微创基金(YS30810175)
【作者简介】谢忠翔,在读硕士,研究方向:医学影像技术,E-mail: 2572237304@qq.com
【通信作者】武杰,博士,讲师,研究方向:医学影像技术,E-mail: jie-usst@163.com

症患者和112名健康对照者)的41 236个功能磁共振成像(function Magnetic Resonance Imaging, fMRI)体素和722 177个单核苷酸多态性(Single Nucleotide Polymorphisms, SNP)进行联合分析,以识别精神分裂症影像遗传学特征。精神分裂症是一种复杂疾病,它是由多种遗传因素(如基因调控的改变、mRNA和SNP的改变)和环境效应相互作用的结果[5-6]。近年来,许多研究集中于探索与精神分裂症相关的关键基因或SNP[7]。除了遗传学研究,fMRI还为精神分裂症的研究增加了一个维度,因为fMRI能够识别精神分裂症患者大脑区域的结构和功能异常。然而,在大多数研究中,fMRI和SNP都是分开单独分析[8]。本研究将结合fMRI和SNP,使用SRVS来选择精神分裂症的影像遗传学特征。

1 资料与方法

1.1 一般资料

本研究采用由Mind Clinical Imaging Consortium(MCIC)收集的数据,所选取的数据均符合《美国精神障碍诊断与统计手册》第4版(DISM-IV)精神分裂症诊断标准[9]。两种类型数据(41 236个fMRI体素和722 177个SNP)收集自208名受试者,包括96名精神分裂症患者和112名健康对照组。

1.2 研究方法

本研究将传统线性回归模型扩展到两类数据集的集成分析:

$$\boldsymbol{Y}=\left[\alpha_1\boldsymbol{X}_1,\alpha_2\boldsymbol{X}_2\right]\begin{bmatrix}\boldsymbol{\delta}_1\\ \boldsymbol{\delta}_2\end{bmatrix}+\varepsilon=\boldsymbol{X}\boldsymbol{\delta}+\varepsilon \tag{1}$$

其中,$\boldsymbol{Y}\in R^{m\times 1}$为观察向量,即受试者的表型,1表示健康人,0表示精神分裂症患者;$\boldsymbol{X}_1\in R^{m\times n_1}$和$\boldsymbol{X}_2\in R^{m\times n_2}$为两种数据类型的测量矩阵列标准化;$\boldsymbol{X}=\left[\alpha_1X_1,\ \alpha_2X_2\right]\in R^{m\times n}$;$\alpha_1$和$\alpha_2$为两种数据类型权重,且$\alpha_1+\alpha_2=1$,$\alpha_1,\ \alpha_2>0$;$\varepsilon\in R^{m\times 1}$为噪声引起的测量误差。该模型建立目的是基于已知的表型$\boldsymbol{Y}$和测量的数据矩阵$\boldsymbol{X}$来恢复未知的稀疏向量$\boldsymbol{\delta}=\begin{bmatrix}\boldsymbol{\delta}_1\\ \boldsymbol{\delta}_2\end{bmatrix}\in R^{n\times 1}$,其中$\boldsymbol{\delta}_1\in R^{n_1\times 1}$,$\boldsymbol{\delta}_2\in R^{n_2\times 1}$且$n_1+n_2=n$。

为了在具有少量非零项的稀疏向量$\boldsymbol{\delta}$(对应于少量的$\boldsymbol{X}$测量值)的情况下,对表型$\boldsymbol{Y}$进行最佳逼近[10],使用SRVS近似求解公式(1)给出的回归问题,并选取稀疏向量$\boldsymbol{\delta}$中非零项所对应测量矩阵$\boldsymbol{X}$中的列作为所要提取的特征。

SRVS算法流程如下:

(1)初始化$\boldsymbol{\delta}^{(0)}=0$;

(2)对于循环l,从fMRI和SNP测量矩阵$\boldsymbol{X}=\left\{x_1,\cdots,\ x_n\right\}\in R^{m\times n}$中随机选择$k$列,重组一个$m\times k$子矩阵,用$\boldsymbol{X}_l\in R^{m\times k}$表示;同时将所选列的索引向量表示为$\boldsymbol{I}_l\in\{1,\ 2,\ 3,\ \cdots\}$;

(3)解决L_p最小化问题[11],找到最优稀疏解$\boldsymbol{\delta}_l\in R^{k\times 1}$:

$$\min\left\|\boldsymbol{\delta}_l\right\|_p\ \ s.t.\ \ \left\|\boldsymbol{Y}-\boldsymbol{X}_1\boldsymbol{\delta}_1\right\|_2\leqslant\varepsilon \tag{2}$$

(4)根据上一步得到的$\boldsymbol{\delta}_l$来更新$\boldsymbol{\delta}^{(l)}\in R^{n\times 1}$,$\boldsymbol{\delta}^{(l)}\left(I_l\right)=\boldsymbol{\delta}^{(l-1)}\left(I_l\right)+\boldsymbol{\delta}_l$;其中,$\boldsymbol{\delta}^{(l)}\left(I_l\right)$和$\boldsymbol{\delta}^{(l-1)}\left(I_l\right)$分别对应表示$\boldsymbol{\delta}^{(l)}$和$\boldsymbol{\delta}^{(l-1)}$中第$I_l$项;

(5)如果不满足停止准则,则更新$l=l+1$,并返回步骤2;否则,令$\boldsymbol{\delta}=\boldsymbol{\delta}^{(l)}/l$并循环终止,稀疏向量$\boldsymbol{\delta}$中非零项对应于数据测量矩阵$\boldsymbol{X}$中选择的列向量。

在步骤(3)中,目前已经有许多行之有效的方法来解决L_p最小化问题。例如,对于$p=1$时,可以用同伦算法来解决[12];$p=0$时可用正交匹配追踪算法[13];$p=0.5$时可用MFCOUSS算法[14]。

在步骤(5)中,设置了以下两个算法迭代停止准则:

(1)$\left\|\boldsymbol{\delta}^{(l)}/l-\boldsymbol{\delta}^{(l-1)}/(l-1)\right\|_2<\alpha$,其中,$\alpha$是预定义的阈值;

(2)$\boldsymbol{X}$中的每一对列向量被比较的概率应该大于$1-p_{\text{stop}}$。

当两个停止准则都满足时,迭代停止。

2 结 果

在本研究中,分别将α_1和α_2作为fMRI数据和SNP数据的权重因子,$\alpha_1+\alpha_2=1$。当权重因子$\alpha_1=1$或者$\alpha_2=1$时,相当于只选择了一种类型数据(fMRI数据或SNP数据)。通过对模型中fMRI数据和SNP数据各自权重的适当调整,提取具有显著意义的fMRI特征和SNP特征,同时将其与公认的45个精神分裂症基因[15-16]进行对比验证。

2.1 不同权重及范数下SNP特征提取

为选出最佳权重因子α,设置参数VarNum(即最终提取的变量数)为200,令$\alpha_1=0.20\sim0.85$,以0.05间隔取14次,在L_0范数模型下,选取精神分裂症影像遗传学特征。

不同权重下筛选的fMRI特征和SNP特征数量对比图见图1。由于两类数据的数据量差别巨大(41 236个fMRI数据,722 177个SNP数据),在L_0范数模型下,只有当fMRI权重$\alpha_1\geqslant0.45$时,才能提取到fMRI特征,而当SNP权重$\alpha_2\geqslant0.85$时,几乎提取不到fMRI特征。当权重$\alpha_1\geqslant0.85$或者$\alpha_1\leqslant0.45$时,只能提取到一类数据特征(fMRI特征或者SNP特征)。

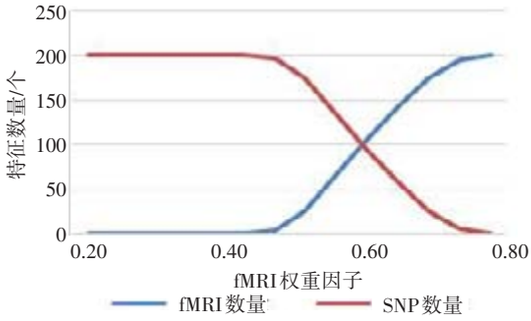


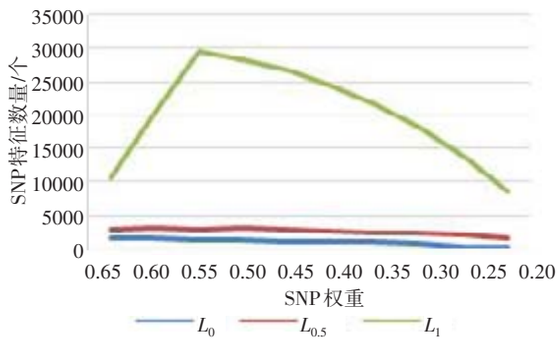
图1 不同权重下筛选的fMRI特征和SNP特征数量对比图

Fig.1 Comparison diagram of the number of fMRI features and SNP features extracted under different weights

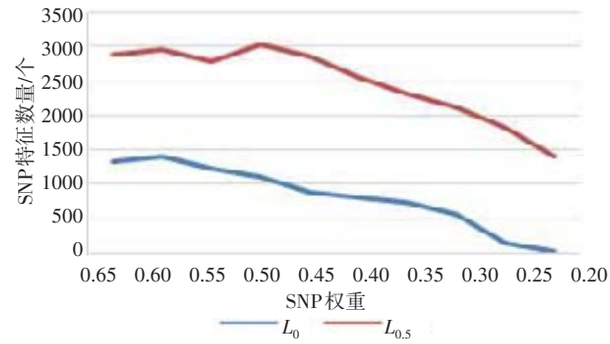
根据图1显示的结果，选取SNP权重范围为0.20~0.65，以0.5为间隔取10组，分别在3组L_p（p = 0、0.5、1）范数模型下提取SNP特征，结果见图2。、

由图2a能够发现，L_1范数下筛选出的SNP特征数量要远远超过L_0范数和$L_{0.5}$范数下筛选出的SNP特征数量，且当SNP权重为0.55时，L_1范数下筛选出的SNP特征数量达到一个峰值。在图2b中，可以发现，在L_0范数和$L_{0.5}$范数下，当SNP权重分别为0.6、0.5时，筛选出的SNP特征数量达到最大值。

设定权重α_1:α_2 = 0.5:0.5，在3种L_p模型（p =



a：L_0、$L_{0.5}$和L_1范数下SNP特征数量



b：L_0和$L_{0.5}$范数下SNP特征数量

图2 不同权重下L_p（p = 0、0.5、1）范数模型中筛选的SNP特征量折线图

Fig.2 Polyline graphs of SNP features extracted in L_p (p = 0, 0.5, 1) norm model under different weights

0、0.5、1）中进行处理提取特征，提取到的SNP特征分别为1 100、3 034、28 064个，找出对应的基因数据，并将其与公认的精神分裂症的45个易感基因进行对比[17]。

选取出精神分裂症相关的SNP数据，认为在多个模型下重复出现的数据更具有显著意义，因此选取至少被两种模型选中的SNP数据，结果见表1。

表1 3种L_p模型中筛选的SNP特征

Tab.1 SNP features extracted in 3 L_p models

L_0模型		$L_{0.5}$模型		L_1模型	
基因	SNP	基因	SNP	基因	SNP
HTR2A	rs6561354	HTR2A	rs6561354	HTR2A	rs6561354
DAOA	rs477122	DAOA	rs477122	DAOA	rs477122
			rs2256289		rs2256289
			rs1549059		rs1549059
			rs1345506		rs1345506
GABRB2	rs2617503	RELN	rs7783216	RELN	rs7783216
	rs1816071				
		DISC1	rs9658966	GABRB2	rs1816071
		GABRB2	rs1816071		

在该条件下，找出3种模型所共同提取出的941个SNP特征，并将其与之前学术界公认的精神分裂症的45个易感基因进行对比，发现有DAOA、HTR2A[18]和GABRB2属于这45个易感基因。

2.2 不同权重及范数下fMRI特征提取

通过比较不同权重下筛选的fMRI和SNP数据

的特征量，可以明显发现，对于fMRI数据，可以令其α_1 = 0.5~0.85，以0.05的间隔取8次，在范数模型p = 0时，选取与精神分裂症相关的影像遗传学特征。

在L_0范数模型中，fMRI的权重因子数值越大，所提取到的fMRI特征数量也就越多，尤其是当权重因子超过0.75时，提取到的fMRI特征数量有一个飞跃性的增长（图3）。

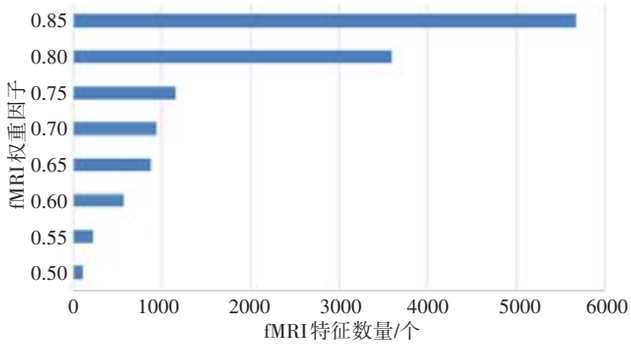


图3 不同权重下提取的fMRI特征数量

Fig.3 Number of fMRI features extracted under different weights

通过调整权重来提取fMRI特征，从而寻找合适的权重分配。设定权重因子α_1:α_2 = 0.5:0.5，在3种L_p模型（p = 0、0.5、1）中进行处理提取特征，筛选到的fMRI特征数量分别为118、77、1 973个。

在L_1范数下，提取1 973个fMRI特征，其中有些fMRI特征属于AAL脑模板116个脑区中的同一个脑区，接下来对此进行分区。在L_1范数下，权重因子α_1:α_2 = 0.5:0.5时，所提取到的fMRI特征总量为1 973个，远远超过L_0范数下的118个和$L_{0.5}$范数下的77个（图4）。其中，以96号（Cerebelum_3_R）和92号（Cerebelum_Crus1_R）脑区提取到的fMRI特征数量最多，而这两个脑区皆为Cerebellum_Superior，属于小脑，这与前人已知的小脑是受精神分裂症影响最为严重的脑区之一相一致。

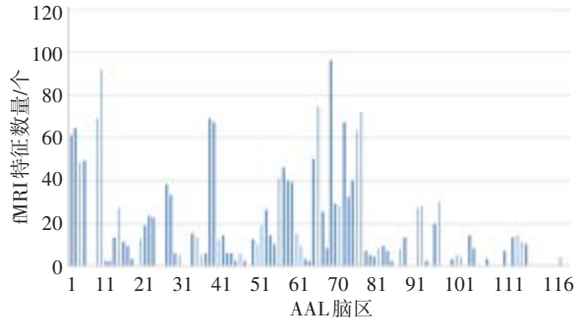


图4 L_1范数下不同脑区中提取到的fMRI特征数

Fig.4 Number of fMRI features extracted from different brain regions under L_1 norm

由于不同脑区结构大小不同，不能简单地通过某一脑区提取到的体素数量来评价它对于精神分裂症的相关性大小。因此，本研究提出通过对某一脑区提取出的体素数量占该脑区总体素数量的百分比来表示该脑区受精神分裂症的影响程度大小。

表2仅显示了提取出的体素占所属脑区总体素百分比前十的脑区。这里有一点需要注意的是，不能简单地从百分比的大小来判定某一脑区与精神分裂症相关性的大小，只能表示该脑区受精神分裂症的影响程度更大。比如，顶下缘角回（左）13.97%大于中央前回（右）8.72%，不能说中央前回脑区与精神分裂症的相关性就不如顶下缘角回脑区的，只能从一定程度上说明顶下缘角回脑区受精神分裂症的影响程度要大于中央前回脑区的。这其中，中央前回、枕上回、顶下缘角回和角回早已经有研究学者证实与精神分裂症相关[19-20]。内侧和旁扣带脑回、后扣带回则是与记忆、行为与情感有关，这与精神分裂症在临床上显示的情感和行为等方面的障碍相一致。

表2 L_1范数下提取的影像学特征

Tab.2 Imaging features extracted under L_1 norm

AAL脑区编号	脑区名称	占百分比/%
61	顶下缘角回（左）	13.97
36	后扣带回（右）	12.39
64	缘上回（右）	12.20
34	内侧和旁扣带脑回（右）	11.38
50	枕上回（右）	11.30
33	内侧和旁扣带脑回（左）	11.29
49	枕上回（左）	10.76
65	角回（左）	9.64
35	后扣带回（左）	9.16
2	中央前回（右）	8.72

3 讨 论

本研究的目的是整合分析与精神分裂症相关的fMRI数据和SNP数据，运用SRVS将选出的fMRI特征与已知的精神分裂症相关脑区进行对比验证，同时将选出的基因与之前普遍认同的精神分裂症的45个易感基因进行对比，结果证实了本研究提出的方法具有一定的可靠性。

在表1中，可以发现$L_{0.5}$模型下提取的精神分裂症的易感基因的数目最多，这说明p=0.5是此方法的

最佳约束范数。进行权重分析时，在$\alpha_1 \leqslant 0.45$时，选取到的影像遗传学特征几乎均为SNP特征，这是因为原始数据中fMRI数据量远远少于SNP数据量，而权重的失衡，又进一步使得这一比例失调。由于在该范围权重下提取到的所有变量全部为SNP数据，因此所选取的变量也可以看作是单变量分析的结果。

其中基因HTR2A、DAOA不仅在$\alpha_2 = 0.20\sim0.65$这10种权重下均被找到，还是在L_p范数在p取0、0.5、1的3种情况下的重叠基因，说明这两个基因对精神分裂症的研究有着不可忽视的重要意义，此前已经有研究表明这两个基因与精神分裂症有显著关联。

研究结果表明了SRVS表现优异，不仅找出了包括DAOA、HTR2A、DISC1在内的精神分裂症的易感基因，也找出了中央前回、枕上回、顶下缘角回、角回、内侧和旁扣带脑回、后扣带回等脑区，而这些脑区已被多名研究学者指出可能与精神分裂症相关。

把SRVS应用于影像遗传学数据分析是一个有效可行的途径，为今后精神分裂症的影像遗传学研究提供了新的研究思路，推测也可将其用于确定抑郁症等复杂精神类疾病的易感基因。

【参考文献】

[1] GRIBONVAL R, NIELSEN M. Sparse decompositions in unions of bases[J]. Trans Inf Theory, 2003, 49(12): 3320-3325.

[2] 郭建玉, 卜晓波, 韩彦龙. 精神分裂症相关基因遗传学研究进展[J]. 牡丹江医学院学报, 2016, 37(4): 116-118.
GUO J Y, BU X B, HAN Y L. Research progresses in genetics related to schizophrenia[J]. Journal of MuDanJiang Medical University, 2016, 37(4): 116-118.

[3] ZHOU S Y, SUZUKI M, TAKAHASHI T, et al. Parietal lobe volume deficits in schizophrenia spectrum disorders[J]. Schizophr Res, 2007, 89(3): 35-48.

[4] SCHERER L J, MCPHERSON J D, WASMUTH J J, et al. Human dopa decarboxylase: localization to human chromosome 7p11 and characterization of hepatic cDNAs[J]. Genomics, 2018, 13(2): 469-471.

[5] TROPP J A, GILBER A C, MUTHUKRISHNAN S, et al. Improved sparse approximation over quasi incoherent dictionaries[J]. Image Process, 2003, 12(1): 137-140.

[6] DONOHO D L, ELAD M. Maximal sparsity representation *via* L1 minimization[J]. Science, 2015, 100(3): 2197-2202.

[7] TANG W, CAO H, DUAN J, et al. A compressed sensing based approach for subtyping of leukemia from gene expression data[J]. J Bioinf Computat Biol, 2011, 9(5): 631-645.

[8] CAO H, DUAN J, LIN D, et al. Sparse representation based clustering for integrated analysis of gene copy number variation and gene expression data[J]. IJCA, 2017, 19(2): 86-89.

[9] 李功迎, 宋思佳, 曹龙飞. 精神障碍诊断与统计手册第5版解读[J]. 中华诊断学电子杂志, 2014, 2(4): 310-312.
LI G Y, SONG S J, CAO L F. Interpretation of the fifth edition of diagnostic and statistical manual of mental disorders[J]. Chinese Journal of Diagnostics (Electronic Edition), 2014, 2(4): 310-312.

[10] WANG C D, BUCK M A, FRASER C M. Site-directed mutagenesis of alpha 2A-adrenergic receptors: identification of amino acids involved in ligand binding and receptor activation by agonists[J]. Mol Pharmacol, 2008, 40(2): 168-179.

[11] CAO H B, DENG H W, LI M, et al. Classification of multicolor fluorescence *in-situ* hybridization (M-FISH) images with sparse representation[C]. Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine, 2012: 125-129.

[12] CLARK D A, MATA I, KERWIN R W, et al. No association between ADRA2A polymorphisms and schizophrenia[J]. Am J Med Genet B Neuropsychiatr Genet, 2017, 144(1): 341-343.

[13] 芮国胜, 王林, 田文飚. 一种基于基追踪压缩感知信号重构的改进算法[J]. 电子测量技术, 2010, 33(4): 38-41.
RUI G S, WANG L, TIAN W B. Improved algorithm based basis pursuit for compressive sensing reconstruction[J]. Electronic Measurement Technology, 2010, 33(4): 38-41.

[14] LE STRAT Y, RAMOZ N, GORWOOD P. The role of genes involved in neuroplasticity and neurogenesis in the observation of a gene-environment interaction (GxE) in schizophreni[J]. Curr Mol Med, 2018, 9(4): 506-518.

[15] 李大伟, 顾鸣敏. 精神分裂症主要易感基因的研究进展[J]. 国际遗传学杂志, 2012, 35(5): 283-289.
LI D W, GU M M. Research progress in studies on main susceptibility genes of schizophrenia[J]. International Journal of Genetics, 2012, 35 (5): 283-289.

[16] LI Y Q, CICHOCKI A, AMARI S. Analysis of sparse representation and blind source separation[J]. Neural Comput, 2015, 16(3): 1193-1234.

[17] 魏凤仙, 武杰, 杨叶, 等. 基于组稀疏典型相关分析方法的影像遗传学方法在精神分裂症中的应用[J]. 中国医学影像技术, 2019, 35(2): 277-281.
WEI F X, WU J, YANG Y, et al. Application of imaging genetics method based on group sparse canonical correlation analysis in schizophrenia[J]. Chinese Journal of Medical Imaging Technology, 2019, 35(2): 277-281.

[18] 阿周存, 李合, 马志敏, 等. 5-HTR2A基因102T/C多态性与精神分裂症的相关性研究[J]. 大理学院学报, 2004, 3(5): 5-6.
A Z C, LI H, MA Z M, et al. Study on association of 102T/C polymorphism of 5HTR2A gene with schizophrenia in Chinese Han population[J]. Journal of Dali College, 2004, 3(5): 5-6.

[19] CAI T T, WANG L. Orthogonal matching pursuit for sparse signal recovery[J]. Inf Theory, 2016, 57(7): 1-26.

[20] LI Y, NAMBURI P, YU Z, et al. Voxel selection in fMRI data analysis based on sparse representation[J]. Trans Biomed, 2009, 56(10): 2439-2450.

（编辑：谭斯允）