A powerful and efficient multivariate approach for voxel-level connectome-wide association studies

Weikang Gong, Fan Cheng, Edmund T. Rolls, Chun-Yi Zac Lo, Chu-Chung Huang, Shih-Jen Tsai, Albert C. Yang, Ching-Po Lin, Jianfeng Feng

Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, 200031, China

University of Chinese Academy of Sciences, Beijing, 100049, China

Department of Computer Science, University of Warwick, Coventry, CV4 7AL, UK

Shanghai Center for Mathematical Sciences, Fudan University, Shanghai, 200033, China

Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai, 200433, China

Institute of Neuroscience, National Yang-Ming University, Taipei, Taiwan

Institute of Psychiatry, Taipei Veterans General Hospital, Taipei, Taiwan

ARTICLE INFO

Keywords:
Multivariate analysis
Structured kernel principal component regression
Association study
Functional connectivity

ABSTRACT

We describe an approach to multivariate analysis, termed structured kernel principal component regression (sKPCR), to identify associations in voxel-level connectomes using resting-state functional magnetic resonance imaging (rsfMRI) data. This powerful and computationally efficient multivariate method can identify voxel-phenotype associations based on the whole-brain connectivity pattern of voxels, and it can detect linear and non-linear signals in both volume-based and surface-based rsfMRI data. For each voxel, sKPCR first extracts low-dimensional signals from the spatially smoothed connectivities by structured kernel principal component analysis, and then tests the voxel-phenotype associations by an adaptive regression model. The method's power is derived from appropriately modelling the spatial structure of the data when performing dimension reduction, and then adaptively choosing an optimal dimension for association testing using the adaptive regression strategy. Simulations based on real connectome data have shown that sKPCR can accurately control the false-positive rate and that it is more powerful than many state-of-the-art approaches, such as the connectivity-wise generalized linear model (GLM) approach, multivariate distance matrix regression (MDMR), adaptive sum of powered score (aSPU) test, and least-square kernel machine (LSKM). Moreover, since sKPCR can reduce the computational cost of non-parametric permutation tests, its computation speed is much faster. To demonstrate the utility of sKPCR for real data analysis, we have also compared sKPCR with the above methods based on the identification of voxel-wise differences between schizophrenic patients and healthy controls in four independent rsfMRI datasets. The results showed that sKPCR had better between-sites reproducibility and a larger proportion of overlap with existing schizophrenia meta-analysis findings. Code for our approach can be downloaded from https://github.com/weikanggong/sKPCR.

1. Introduction

Functional connectivity analysis using resting-state functional magnetic resonance imaging (fMRI) data has become increasingly popular in the last few years (Smith et al., 2015; Finn et al., 2015), and the advances have led to many investigations of functional dysconnectivity between brain areas in neurodegenerative and psychiatric brain diseases (Gong and He, 2015; Romme et al., 2017). Voxel-based functional connectivity analysis has also recently emerged (Cheng et al., 2015, 2016; Rolls et al., 2018; Satterthwaite et al., 2015; Kaczkurkin et al., 2017). However,
designing methods to explore the associations between the whole-brain voxel-level connectome and phenotypes is a challenging task, and well-developed approaches are usually designed for parcellation-based or seed-based connectivity studies (Meskaldji et al., 2013; Bellec et al., 2015; Xia and He, 2017).

The most popular method for functional connectivity analysis is the massive univariate generalized linear model (GLM) approach. This approach uses a GLM to test the association between each voxel-voxel connectivity and the phenotype of interest, and then corrects for multiple comparison (Cheng et al., 2015, 2016) by such methods as Bonferroni correction, false-discovery rate (Benjamini and Hochberg, 1995) and random field theory (Gong et al., 2018), to locate the significant signals. The major advantage of this approach is that it can provide the exact location of the signals. However, the large number of hypothesis tests require a stringent multiple correction threshold which usually decrease the power. In addition, univariate approach only tests the linear relationships between connectivities and phenotypes. Important higher-order information, such as the co-contribution of a set of functional connectivities and the non-linear associations, is usually ignored by this method.

In recent years, many improvements over the univariate method have been proposed. These approaches usually adopt global association tests to achieve higher power. In other words, they test whether the signal is present somewhere in a set of functional connectivities rather than localizing it. We briefly review some of them here. First, in the brain-wide association study (BWAS) approach (Gong et al., 2018), the authors proposed to test whether the observed cluster size of the supra-threshold functional connectivities is larger than that by chance. The BWAS is a generalization of the traditional cluster-size inference approach (Friston et al., 1994) and popular network-based statistic (NBS) approach (Zalesky et al., 2010) to voxel-level connectivity studies. Second, the multivariate distance matrix regression (MDMR) (Shehzad et al., 2014) is a nonparametric multivariate approach, which directly tests the association between a phenotype of interest and a between-subject distance matrix estimated using the functional connectivity data. Third, in Pan et al. (2014); Kim et al. (2014, 2015), the authors proposed the adaptive sum of powered score (aSPU) test and its extensions. This approach first assigns a score to measure the association between a phenotype and an individual connection. It then combines all the individual scores into a summary statistic and uses a permutation test to phenotype and an individual connection. It then combines all the individual scores into a summary statistic and uses a permutation test to

For our approach can be downloaded from https://github.com/weikanggong/sKPCR.

2. Method

2.1. Structured kernel principal component analysis (sKPCA)

2.1.1. Background: principal component analysis and its probabilistic model

Principal component analysis (PCA) is one of the most popular dimension reduction and feature extraction approaches (Jolliffe, 2002). It is usually defined as the orthogonal projection of a data matrix $X \in \mathbb{R}^{p \times n}$ onto a low-dimensional space which maximizes the projection variance, where $n$ is the number of subjects and $p$ is the number of features. Specifically, suppose that $X = (x_1, x_2, \ldots, x_n)$ has been centered to mean zero by rows. Let $u_j \in \mathbb{R}^p$ be a $p$-dimensional vector that projects each data point $x_i$ to a scalar value $t_i = u_j x_i$, then $t_j^2 = (u_j x_i)^2$ is proportional to the variance of the projection, and PCA seeks such $u_j$ that maximizes it. As an optimization problem, this can be written as:

$$\max_{u_j} \sum_{i=1}^{n} u_j^T X u_j$$

subject to $u_j u_j = 1, j = 1, 2, \ldots, k$

$$u_j u_j = 0, \forall \ell, j$$

This constrained optimization problem can be solved by an eigenvalue decomposition of the sample covariance matrix $XX^T \in \mathbb{R}^{p \times p}$, in which the first $k$ principal components are exactly the first $k$ eigenvectors $U = (u_1, u_2, \ldots, u_k)$ of $XX^T$. When the number of features $p$ is larger than the number of subjects $n$, note that we can equivalently perform an eigenvalue decomposition on $XX^T \in \mathbb{R}^{n \times n}$. We can only extract a maximum of $\min(n, p)$ number of principal components.

Probabilistic PCA reformulates PCA as the maximum likelihood solution of a probabilistic latent variable model (Tipping and Bishop, 1999; Bishop, 2006), which is closely related to factor analysis (Bartholowen and Knott, 1999; Bishop, 2006) and probabilistic canonical correlation analysis (Bach and Jordan, 2005). In probabilistic PCA, the generative model of the $i$-th data point $x_i$ is $\mathbb{R}^{p-1}$ of a data matrix $X \in \mathbb{R}^{p \times n}$ is
assumed to be a projection of latent variable $z_k \in \mathbb{R}^{k \times 1}$ using weight matrix $W \in \mathbb{R}^{k \times p}$ plus isotropic Gaussian noises $\varepsilon$ among features:

$$z_k \sim N(0, I) \quad \varepsilon \sim N(0, \sigma^2 I) \quad x_i = Wz_k + \varepsilon$$  \hspace{1cm} (2)

To estimate the model parameter $W, \sigma^2$ and latent variable matrix $Z = (z_1, z_2, \ldots, z_k) \in \mathbb{R}^{k \times n}$, Tipping and Bishop (1999) proposed to maximize the (complete-data) log likelihood with respect to the parameters: max $\log p(X, Z|W, \sigma^2)$. They showed that the solution of the above maximum likelihood problem can be obtained by using an expectation–maximization (EM) algorithm or in closed-form as:

$$W_{MLE} = U(\Lambda - \sigma_{MLE}^2 I)^{-1/2}$$

$$\sigma_{MLE}^2 = \frac{1}{p-k} \sum_{j=k+1}^{p} \lambda_j$$

$$Z_{MLE} = (W_{MLE} W_{MLE}^T + \sigma_{MLE}^2 I)^{-1} W_{MLE} X$$ \hspace{1cm} (3)

where $\Lambda \in \mathbb{R}^{k \times k}$ is a diagonal matrix, the diagonal elements of which are the first $k$ eigenvalues of sample covariance matrix $XX^T$, i.e., $\lambda_j, j = 1, 2, \ldots, p$, and $U \in \mathbb{R}^{k \times k}$ is the corresponding eigenvectors.

The equivalence between conventional PCA and probabilistic PCA can be seen from the closed-form solution, as demonstrated in Eq. (3). That is, when $k - p$, then $\sigma^2 \to 0$, the maximum-likelihood estimation (MLE) of latent variable $Z_{MLE} = (W_{MLE} W_{MLE}^T)^{-1} W_{MLE} X$, which is an orthogonal projection of the data onto the latent space, enabling recovery of the standard PCA model (Tipping and Bishop, 1999; Bishop, 2006).

From the probabilistic interpretation of PCA, we can see that the noise is assumed to be independent between features (and subjects). However, the assumption may break down when analyzing voxel-based functional connectivity data because the noise terms among the connectivities are spatially correlated. In addition, PCA can only perform linear dimension reduction and feature extraction, and many important non-linear factors may be missed by this method. Therefore, we propose a structured kernel principal component analysis (sKPCA) approach in the next section, which can model the spatial structure of the noise and extract both linear and non-linear features.

2.1.2. Structured PCA

We propose a framework for structured PCA (sPCA) in this section. It allows structured noise among features and subjects to be modelled. We first describe sPCA from a probabilistic perspective by modelling the noise as a multivariate Gaussian distribution, and then provide an efficient algorithm for estimating the principal components.

To accomplish this, we first introduce the matrix normal distribution $\mathcal{MN}_{n \times p}(M, Q, R)$, which is a generalization of the multivariate normal distribution to matrix-valued random variables $X \in \mathbb{R}^{p \times n}$, the probability density function of which is: $p(X; M, Q, R) = \exp\left[-\frac{1}{2} \log \det \left(2\pi Q^{-1}(X - M)^T R^{-1}(X - M)\right)\right] / \det(R)^{n/2} |Q|^{p/2}$, where $M \in \mathbb{R}^{p \times n}$ is the mean, and $R \in \mathbb{R}^{p \times p}$ and $Q \in \mathbb{C}^{n \times n}$ are the covariance matrix of the rows and columns of $X$. The connection between the matrix normal distribution and multivariate normal distribution is: $\text{vec}(X) \sim \mathcal{N}(\text{vec}(M), Q \otimes R)$, where $\otimes$ denotes the Kronecker product and vec denotes the vectorization $M$. 

Fig. 1. An overview of the structured kernel principal component regression (sKPCR) in a voxel-level connectome-wide association study. First, for each voxel and each subject, the whole-brain functional connectivity map is computed. Second, a dimension reduction technique, termed structured kernel principal component analysis (sKPCA), is applied to extract important features in this connectivity map, which utilizes the spatial information of functional connectivities. Third, an adaptive regression is fitted to test the association between a phenotype of interest and principal components of this voxel, which automatically selects the optimal number of components. Finally, voxel-wise multiple correction is performed to identify significant clusters.
We can represent probabilistic PCA using the matrix normal distribution as:

\[ X = WZ + E; \quad E \sim \mathcal{N}(0, \sigma^2 I) \]

(4)

Here, both the rows (e.g., features) and columns (e.g., subjects) of the noise matrix are assumed to be independent from each other. However, for neuroimaging data, noises among voxels/vertices are known to be smoothly correlated with each other. Therefore, our spCA model is a generalization of the probabilistic PCA model, which allows two-way dependence between noise terms:

\[ X = WZ + E; \quad E \sim \mathcal{N}(0, R, Q) \]

(5)

where \( W \) is the weight matrix and \( Z \) is the latent variable matrix in accordance with probabilistic PCA.

In this model, the \( Q \) and \( R \) do not need to be estimated from the data; however, they do need to be prespecified based on the known topological structure of the data (see section 2.2 for details). Then, we can still use the MLE approach to estimate the \( W \) and \( Z \) in (5):

\[
\begin{align*}
\max_{W,Z} \log P(X,Z|W) &= \max_{W,Z} \left\{ -\frac{1}{2} \text{tr}\left[Q^{-1}(X - WZ)^T R^{-1}(X - WZ)\right] - \frac{1}{2} \text{tr}(ZZ^T) + \text{Const} \right\} \\
&= \max_{W,Z} \left\{ -\frac{1}{2} \text{tr}\left[(\tilde{R}\tilde{Q})^{-1/2}(\tilde{R}\tilde{Q})^{1/2}(X - WZ)\right] - \frac{1}{2} \text{tr}(ZZ^T) + \text{Const} \right\} \\
&= \max_{W,Z} \left\{ -\frac{1}{2} \text{tr}\left[(\tilde{X} - \tilde{W}Z)\tilde{X} - \tilde{W}Z\right] - \frac{1}{2} \text{tr}(ZZ^T) + \text{Const} \right\}
\end{align*}
\]

(6)

where we have decomposed \( Q^{-1} = \tilde{Q}\tilde{Q}^T \) and \( R^{-1} = \tilde{R}\tilde{R}^T \), and \( \tilde{X} = R\tilde{Q}, \tilde{W} = WR \) and \( \tilde{Z} = Z\tilde{Q} \). Therefore, the spCA problem (6) is equivalent to the standard PCA or probabilistic PCA problems using the ‘weighted’ data matrix \( \tilde{X} \) (Escoufier, 1977; Allen et al., 2014; Zhu et al., 2017), where the weights are learned from the external information, i.e., the spatial, temporal or population structures of data as:

\[
\begin{align*}
\max_{u_j \in \mathcal{S}_+^{n \times k}} u_j R^{-1/2}XQ^{-1}X'R^{-1/2}u_j \\
\text{subject to} \quad u_j R^{-1}u_j = 1, \quad j = 1, 2, \ldots, k \\
\quad u_j R^{-1}u_j = 0, \quad \forall \ j < j
\end{align*}
\]

(7)

Therefore, the principal components \( U \) (or \( W \)) can be obtained by a simple eigenvalue decomposition on the matrix \( R^{-1/2}XQ^{-1}X'R^{-1/2} \in \mathbb{R}^{p \times p} \) or when \( n < p \), on the matrix \( Q^{-1/2}X'R^{-1}XQ^{-1} \in \mathbb{R}^{n \times n} \).

2.1.3. Structured kernel PCA

The spCA model can be further generalized to perform non-linear dimension reduction by using kernel tricks specifically. Let \( \tilde{x}_i \) be the \( i \)-th ‘weighted sample’, i.e., the \( i \)-th column of \( \tilde{X} = R\tilde{Q} \), we first perform a non-linear mapping of the sample \( \tilde{x}_i \) to the high dimensional feature space as \( \Phi(\tilde{x}_i) = \Phi(x_i) \). Now, we assume that each \( \Phi(x_i) \) has been mean centered in the feature space and we will return to this point later. We can perform a PCA in the mapped high-dimensional feature space by maximizing the projected variance as:

\[
\max_{u_j \in \mathcal{S}_+^{n \times k}} \Phi(\tilde{x}_i)^T \Phi(\tilde{x}_i)u_j, \quad j = 1, 2, \ldots, n
\]

(8)

Similar to the kernel principal component analysis (Schölkopf et al., 1997), the optimization problem (8) can be solved by first performing a mean normalization of the kernel matrix \( K \in \mathbb{R}^{n \times n} \), where \( K_{ij} = \Phi(\tilde{x}_i)^T \Phi(\tilde{x}_i) \):

\[
K = K - LK - KL + LKL
\]

where \( L \) is a \( n \times n \) matrix in which each element takes the value \( 1/n \). Then, we solve the eigenvalue problem:

\[
n^{-1}\tilde{K}u_j = \lambda_j u_j, \quad j = 1, 2, \ldots, n
\]

and obtain \( n \) eigenvalues in a descending order as \( \lambda_1, \ldots, \lambda_n \) and the corresponding eigenvectors \( u_1, \ldots, u_n \). The \( k \)-th principal component is the \( k \)-th eigenvector \( u_k \).

Similar to most other kernel-based approaches, all the computations can be expressed in the form of the kernel matrix. When using the linear kernel, the skPCA is exactly the same as spCA. In addition, for many commonly used kernels, we do not even need to estimate \( Q \) and \( R \). For example, we can calculate \( X^+ = q^{-1}X'R^{-1}X'^{-1}X'^{-1}Q'^{-1} \), and the polynomial kernel can be calculated as \( K_{ij} = (\alpha x_i^T x_j + b)^p \), the sigmoid kernel can be calculated as \( K_{ij} = \exp(-||x_i - x_j||^2/\sigma^2) \), and the Gaussian kernel can be calculated as \( K_{ij} = \exp(-||x_i - x_j||^2/2\sigma^2) \).

2.2. The choice of skPCA parameters

Many methods have been developed to determine the number of principal components for conventional PCA, such as the ratio estimator (Lam and Yao, 2012; Li et al., 2017), the information criteria approaches (Bai and Ng, 2002, 2007), the distribution-based approach (Choi et al., 2014) or just by the amount of variance explained (e.g., 90%). Although these methods can be easily extended to the current skPCA framework, we have found that none of them works optimally in the subsequent association tests, which is our main goal in this paper. Therefore, we developed a novel adaptive regression approach in the next section, in order to address the problem of selecting the number of principal components in the association study.

Many possible choices are available for the covariance matrix (Allen et al., 2014; Ramsey, 2006), but we will introduce and compare just three in this paper. The first one is the Graph Laplacian (GL) operator, which has been widely used in Bayesian task-activation studies (Penny et al., 2005; Flandin and Penny, 2007; Sidén et al., 2017). It is also known as the inverse covariance operator (Allen et al., 2014; Ramsey, 2006).

To define the GL operator \( G \), we first define the feature-feature adjacency matrix \( A \) as a binary matrix such that \( a_{ij} = 1 \) if the spatial distance between feature \( i \) and \( j \) \((i \neq j)\) equals one (or feature \( i \) and \( j \) are spatial neighbours) and \( a_{ij} = 0 \) otherwise. Based on \( A \), we can define \( G \) as, for feature \( i \) and \( j \),

\[
G_{ij} = \frac{1}{\sum_{i' \neq i} a_{ij}} \quad \text{if} \quad i \neq j.
\]

The second one is the normalized Graph Laplacian (NGL) operator \( G^* \). Based on \( A \), it is defined as \( G_{ij}^* = 1 \) and \( G_{ij}^* = \frac{1}{\sqrt{\sum_{i' \neq i} a_{ij}}} \) if \( i \neq j \). The third one is called the Gaussian random field operator. It assumes that the noise covariance between two features is a functional of their spatial distance:

\[
\Sigma_{ij} = \exp\left(-\frac{||X_i - X_j||^2}{2\sigma^2}\right) \quad \text{where} \quad ||X_i - X_j||_2 \text{ represents the spatial Euclidian distance.}
\]
distance between feature \( i \) and feature \( j \) in the volume space, and \(||X_i - X_j|||\) can also be the geodesic distance in surface space. The \( \sigma \) can be specified based on the estimated Full Width at Half Maximum (FWHM) of the images using the relationship FWHM = \( 2\sqrt{2} \log 2\sigma \). We will show that sKPCR is very stable for selecting different covariance operators and that the GL operator has a slightly higher power. Therefore, it is used in all our analyses.

### 2.3. Structured kernel principal component regression (sKPCR) for identifying connectome-wide associations

#### 2.3.1. Overview

We extend sKPCA to structured kernel principal component regression (sKPCR) to identify connectome-wide associations. In our study, the individual-level brain functional network is estimated by the Fisher’s Z transformed Pearson correlation coefficient between every pair of voxels’ BOLD signal time series. Let \( n \) be the number of subjects in a study, and \( p \) be the number of voxels; as such, the total number of functional connectivities in each individual’s brain network is \( p(p - 1)/2 \), and there are \( p - 1 \) functional connectivities connecting a voxel to all other voxels across the whole brain. Let \( Y \in \mathbb{R}^{n \times 1} \) be the phenotype of interest of \( n \) subjects (e.g. disease status, clinical symptoms) and \( Z \in \mathbb{R}^{n \times q} \) be the nuisance covariates (e.g. age, gender, motion terms). Our aim is to test, for each voxel, whether the phenotype of interest \( Y \) is associated with the voxel’s whole-brain functional connectivity pattern \( X \in \mathbb{R}^{n \times (p-1)} \), conditioned on the nuisance covariates \( Z \). Since connectivity is of ultra-high dimensionality (e.g. for each voxel, there are 10^4 to 10^5 whole-brain functional connectivities, but only a few hundred samples), the basic idea of sKPCR is to (1) extract important low-dimensional features (principal components) in the data \( X \) by sKPCA, and then (2) test the association between the extracted principal components \( U = (u_1, …, u_k) \) and the phenotype of interest \( Y \).

#### 2.3.2. An adaptive regression model

To test associations, we propose a novel adaptive regression approach which can estimate a single \( p \)-value per voxel to summarize the overall significance of the association. Traditionally, we would manually select the top \( k \) principal components and then use a general linear model with \( F \) statistic for statistical testing. However, we found the pre specification of \( k \) to be very difficult, and the top principal components may not always explain the phenotype of interest. Therefore, we propose a new approach which is able to adaptively choose the optimal number of principal components to include in the model, one that is sufficiently robust to include noise components. The idea is similar to many other adaptive test approaches widely used in the neuroimaging (Kim et al., 2015) and genomics fields (Pan et al., 2014; Lee et al., 2012).

In detail, let \( n \) be the partial correlation between a principal component \( u_i \) and a phenotype \( Y \) conditioned on covariates \( Z \), then we define a score \( S_i \) to measure the overall correlations between \( k (k = 1, 2, …, K) \) extracted components and the phenotype as:

\[
S_i = \sum_{i=1}^{K} r_i^2
\]

We can get a score vector \( S = (S_1, S_2, …, S_K) \). Using a non-parametric permutation approach, we can get the \( p \)-value of each score by permuting the phenotype \( Y \) and recalculating the ‘null’ score vector \( M \) times. That is, let \( \hat{Y} \) be a randomly permuted phenotype in the \( j \)-th permutation. We first calculate the partial correlation between \( \hat{Y} \) and \( u_i \) conditioned on covariates \( Z \) to get the permuted coefficient \( \hat{r}_i \). Then, as above, we calculate the score as:

\[
\hat{S}_i = \sum_{i=1}^{K} (\hat{r}_i)^2
\]

With a total of \( M \) permutations, we can get a vector of null score \( \hat{S}_i = (\hat{S}_1, \hat{S}_2, …, \hat{S}_K) \). Therefore, the \( p \)-value of score \( S_i \) can be estimated non-parametrically as:

\[
p_i = \frac{\#(\hat{S}_i \geq S_i) + 1}{M + 1}
\]

where \( \#(A \geq a) \) denotes the number of times the elements in vector \( A \) is larger than a number \( a \). After the above steps, we can get the \( p \)-values of the \( K \) scores \( S_i \) denoting them as \( (p_1, p_2, …, p_K) \). The above steps can be computationally very efficient by using a simple matrix computation strategy.

Now, we define our test statistic as the smallest \( p \)-values in \( (p_1, p_2, …, p_K) \):

\[
T_{\text{sKPCR}} = \min(p_1, p_2, …, p_K)
\]

Note that \( T_{\text{sKPCR}} \) is a not a \( p \)-value, because its null distribution is no longer subjected to a uniform distribution. Therefore, its statistical significance should be estimated using non-parametric permutation again. However, it is interesting to note that we do not need to run another set of permutations, but rather simultaneously estimate the null distribution of \( T_{\text{sKPCR}} \) using the above permutations, which have been used to calculate \( (p_1, p_2, …, p_K) \). Specifically, the permutation empirical \( p \)-value of the \( k \)-th score in the \( j \)-th permutation is:

\[
\tilde{p}_k = \frac{\#(\hat{S}_k \geq \hat{S}_k) + 1}{M}
\]

Therefore, the most significant \( p \)-value across \( K \) scores in the \( j \)-th permutation is:

\[
\tilde{T}_{\text{sKPCR}} = \min(\tilde{p}_1, \tilde{p}_2, …, \tilde{p}_K)
\]

Thus, for all the \( M \) permutations, we get \( \tilde{T}_{\text{sKPCR}} = (\tilde{T}_{\text{sKPCR}}, \tilde{T}_{\text{sKPCR}}, …, \tilde{T}_{\text{sKPCR}}) \). Finally, the \( p \)-value of our test statistic \( T_{\text{sKPCR}} \) can be estimated as:

\[
p(T_{\text{sKPCR}}) = \frac{\#(\tilde{T}_{\text{sKPCR}} \geq \tilde{T}_{\text{sKPCR}}) + 1}{M + 1}
\]

Note also that this approach can provide an exact control of false-positive rate. However, for the general linear model approach, the \( p \)-value of a \( F \)-test or likelihood-ratio test may not provide a valid \( p \)-value when the number of components is comparable to the number of subjects (Sur et al., 2017).

#### 2.3.3. Multiple comparison correction

After getting the voxel-wise \( p \)-values, we can use a nonparametric permutation approach (Nichols and Holmes, 2002) or false-discovery rate method (Benjamini and Hochberg, 1995) to perform multiple comparison correction. For permutation-based approaches, we can still use the same set of permutations above to perform topological inference, including peak-level inference (Worsley et al., 1996), cluster-size inference (Friston et al., 1994), cluster-mass inference (Zhang et al., 2009), and threshold-free cluster enhancement (Smith and Nichols, 2009).

### 3. Evaluating sKPCR using simulation studies: false-positive rate, power and robustness

#### 3.1. Data

We use two resting-state fMRI datasets to evaluate different methods, including 281 subjects from the Southwest University (SWU) dataset in the International Data-sharing Initiative (IDNI, http://icon1000.projects.nitrc.org/indi/retro/southwestuni_qiu_index.html), and 150 subjects from the Human Connectome Project (HCP_REST1_LR, https://www.humanconnectome.org/). All subjects were healthy adults with
similar demographic information.

The data from SWU were preprocessed using a standard volume-based fMRI pipeline (code can be downloaded from https://github.com/weikanggong/Resting-state-fMRI-preprocessing). For each individual, the preprocessing steps include: slice timing correction (FSL slicetimer), motion correction (FSL mcflirt), spatial smoothing by a 3D Gaussian kernel (FWHM = 6 mm), despiking motion artifacts using the BrainVatoolbox (Patel et al., 2014), registering to 4 × 4 × 4 mm³ standard space by first aligning the functional image to the individual T1 structural image using boundary based registration (Greve and Fischl, 2009), and then to standard space using FSL’s linear and non-linear registration tool (FSL flirt and flipt), regressing out nuisance covariates including 12 head motion parameters (6 head motion parameters and their corresponding temporal derivatives), white matter signal, cerebrospinal fluid signal and global signal, band-pass filtering (0.01–0.1 Hz) using AFNI (3dfitproject). All the images were manually checked to ensure successful preprocessing. Finally, 14364 grey matter voxels located in each subject’s cerebrum were extracted for the subsequent analysis.

The data from HCP-S900 were preprocessed using the fMRIprep pipeline (Glasser et al., 2013; Smith et al., 2013). The basic steps included: correcting for spatial distortions caused by gradient nonlinearity, correcting for head motion by registration to the single band reference image, correcting for B₀ distortion, and registering to the T1w structural image. The global intensity was normalized. Then, independent component analysis (ICA) was run using MELODIC with automatic dimensionality estimation (Beckmann and Smith, 2004). These components were fed into FIX (Salimi-Khorshidi et al., 2014), which classified components into ‘good’ vs. ‘bad’. Bad components were removed from the data. From this resulting volume time-series, the data were mapped onto the standard 32k Conte69 cortical surface using the Multimodal Surface Matching approach (MSMAll pipeline (Robinson et al., 2014)). Finally, the Gaussian spatial smoothing was carried out on the cortical surface with a Full-Width at Half Maximum of 5 mm. In our analysis, BOLD time series of 32492 cortical vertices from each subject’s left cortical surface were used.

3.2. Type I error rate evaluation

To evaluate whether the proposed sKPCR approach could control the type I error rate, we evaluated whether it had a nominal false-positive rate when comparing the connectome of two groups of healthy subjects with similar demographic information (Eklund et al., 2016). If a method can provide a valid control of type I error rate, the observed false-positive rate will be around its nominal level (e.g. 0.05). Specifically, first, a voxel was randomly selected, and functional connectivities between it and all other voxels across the whole brain were estimated for every subject. Second, subjects were randomly divided into two groups, and sKPCR with 5 different types of kernel was then applied to test whether this voxel showed differences between the two randomly assigned groups. This step resulted in one p-value for sKPCR per kernel. Third, the above two steps were repeated for 1000 times, and the observed false-positive rate was estimated as the proportion of times the p-value was below 0.05. Some commonly used kernels we evaluated here were a linear kernel, polynomial kernel with degree 2,3,4,5 and a Gaussian kernel with σ parameter equaling the middle of the Euclidian distance among data points (Brown et al., 2000).

3.3. Comparing the statistical power of detecting linear signals with other methods

In this simulation, we considered methods for detecting linear signals. We compared the linear sKPCR with 5 other approaches, including a connectivity-wise general linear model (GLM) approach controlling the family-wise error rate (i.e., SPU(Inf) approach (Kim et al., 2014)); multivariate distance matrix regression (MDMR) (Shehzad et al., 2014); and adaptive sum of powered score (aSPU) and extensions (i.e., SPU(1), SPU(2), aSPU (Kim et al., 2014)). All of these approaches can produce voxel-wise p-value maps based on the rsfMRI data. A brief description of these methods and their parameter settings are shown in the Supplementary Material.

In our simulation, first, one voxel was randomly selected, and functional connectivities between it and all other voxels across the whole brain were estimated for every subject, and they were normalized to zero mean unit variance. Second, signals were then randomly added to a subset of functional connectivities (proportion of null functional connectivity 𝜌). For linear signals, we simulated a phenotype of interest y which was linearly correlated with some functional connectivities x. This was achieved by first simulating new functional connectivities as $x_{new} = \gamma y + x$, and then normalizing them again to zero mean unit variance. The signal-to-noise ratio $\gamma$, which was defined as the ratio of signal variance and noise variance, varied from 0 (no signal) to 0.25 in our simulation. Each method was then applied to test whether the overall connectivity pattern was associated with the signal y. Third, the above two steps were repeated 1000 times, and the empirical power was estimated by the proportion of times the p-value was below 0.05.

3.4. Comparing the statistical power of detecting nonlinear signals with other methods

For non-linear signals, we compared our approach with 2 other methods, including kernel principal component regression (KPCR) (Schölkopf et al., 1997) and least-squares kernel machine (LSKM) (Liu et al., 2007; Ge et al., 2012). A brief description of these methods and their parameter settings are shown in the Supplementary Material.

This simulation is similar to the above one. The only difference is the way of generating the nonlinear signals. This was achieved by simulating new functional connectivities $x_{new}$ as $x_{new} = \gamma y^{d} + x$ where d is the polynomial degree. Theoretically, therefore, methods using a polynomial kernel with corresponding degree should achieve the highest power.

3.5. The robustness of sKPCR when the kernel is misspecified

To evaluate whether sKPCR is robust when the kernel was misspecified, we also used sKPCR, KPCR and LSKM with linear kernels for signal detection. The simulation procedures were exactly the same as those described in Section 3.4, but we used the wrong kernels to detect signals.

3.6. The robustness of sKPCR when the number of principal components is misspecified

As illustrated in section 2.3, the proposed adaptive regression approach, which was used to detect association after sKPCA dimension reduction, was robust to the misspecification of number of principle components. To demonstrate this, we conducted a simulation study to compare it with the traditional general linear model approach.

We assumed a total of 200 subjects and we totally extracted 100 principal components. As the components are not correlated with each other, we first simulated independent Gaussian white noise data, which formed a 200 × 100 matrix X. The first 10 components were assumed to be correlated with a phenotype y, and the 11-th to 100-th components were assumed to be noise components. Therefore, we added $\gamma y$ to the first 10 components, where $\gamma$ could be treated as the effect size. In our experiment, $\gamma = 0.03, 0.06, 0.09, 0.12, 0.15$. We tested the association between the first i columns of X and y using either the adaptive regression or the linear regression model with the F statistic. The above procedures were repeated 1000 times for different effect sizes and numbers of components. We compared the power of the two methods when the number of principal components were misspecified by adding more noise components to the model.
3.7. The robustness of sKPCR under different covariance operators

Following the same simulation procedures as those detailed in section 3.3, we compared the power of linear sKPCR when using different covariance operators. The covariance operators we tested included GL, NGL, Gaussian with \( \sigma^2 = 2, 4, 6 \) voxels (Gaussian (2), Gaussian (4), Gaussian (6)).

4. Evaluating sKPCR in real data: brain-wide associations of the schizophrenic connectome

4.1. Data

Four resting-state fMRI datasets were used here: Taiwan, Centers of Biomedical Research Excellence (COBRE) (http://fcon_1000.projects.nitrc.org/indi/retro/cobre.html), BrainGluSch (http://schizconnect.org/), and NMorphCH (http://schizconnect.org/). All of them are datasets with schizophrenia patients and matched healthy controls. The demographic information is shown in Table 1. Resting-state fMRI data were preprocessed using the same pipeline as the SWU dataset (code can be downloaded from https://github.com/weikanggong/Resting-state-fMRI-preprocessing). Finally, 18757 voxels located in each subject’s cerebral regions were extracted for the subsequent analysis.

4.2. Between-sites reproducibility

We applied sKPCR with a linear kernel to identify voxels with significantly altered connectivities in each schizophrenia dataset separately. At the same time, five other methods, including the univariate approach, MDMR, SPU(1), SPU(2) and aSPU, were used for comparison. To compare the between-sites reproducibility of each method, they were applied to analyse the four schizophrenia datasets separately, and the Dice Coefficient (DC) between the resulting voxel-wise p-value maps of two sites was calculated. DC is defined as

\[
DC = \frac{2|E_i \cap E_j|}{|E_i| + |E_j|}
\]

where \(|E_i|\) is the number of significant voxels in the \(i\)-th experiment. We will mainly report the DC with \( p = 0.001 \) as the voxel-wise significance threshold, as it is widely used (Woo et al., 2014). We expect that a better approach has a larger overlap.

4.3. Evidence from the literature: overlaps with existing meta-analysis findings

We also evaluated whether the above results were consistent with existing meta-analysis findings reported in the Neurosynth database (Yarkoni et al., 2011). We searched for the term ‘schizophrenia’ on the Neurosynth website (http://neurosynth.org/analyses/terms/schizophrenia/), and downloaded the default forward inference map (FDR<0.01). We calculated the DC between the p-value map of each approach in each dataset and the Neurosynth schizophrenia forward inference map. We will mainly report the DC with \( p = 0.001 \) as the voxel-wise significance threshold, as it is widely used (Woo et al., 2014).

5. Results

5.1. Evaluating sKPCR using simulation studies

5.1.1. Type I error rate evaluation

We first evaluate whether sKPCR could control the type I error rate using different kernels in different types of data. We simulated a case-control study in the absence of any group difference (see Section 3.2), and sKPCR was applied to detect signals. The results show that the proposed approach can control the false-positive rate appropriately using a wide range of kernels (linear, polynomial and Gaussian) in both volume-based fMRI data compared to volume-based fMRI data (Fig. 2), because the observed false-positive rates are similar to their theoretical nominal level at 0.05.

5.1.2. Comparing the statistical power of detecting linear signals with other methods

Fig. 3 shows the results of comparing the power of different methods when the true signal is linear, i.e. some functional connectivities are linearly correlated with a simulated phenotype of interest, in volume-based and surface-based fMRI data (Fig. 2), because the observed false-positive rates are similar to their theoretical nominal level at 0.05.

For both volume-based and surface-based data, it can be clearly seen that the proposed sKPCR method with a linear kernel always has the highest power in different situations (different signal-to-noise ratios and proportions of non-null connectivities). The univariate approach and aSPU have similar power in different situations. The performance of MDMR and SPU(2) are similar to that of aSPU and the univariate method when the number of non-null functional connectivities is large (e.g., 20% non-null). However, the power of MDMR and SPU(2) decreases dramatically when only a few non-null connectivities exist (e.g., 1% non-null). SPU(1) performs the worst in these simulations. In addition, the power of sKPCR displays a larger gap with other approaches in surface-based fMRI data compared to volume-based fMRI data.

To get a more intuitive understanding of why sKPCR had a better performance, we simulated a case-control study with some of the functional connectivities in one group having a higher mean than another group with the same strategies as the above simulation. We applied sKPCA with a linear kernel and PCA to the simulated data and extracted the top 4 principal components. For each method, we plotted each pair of principal components in a 2D figure and used different colors to distinguish the two groups. As can be clearly observed from Fig. 5, sKPCA (top row) has much better performance because the case and control groups are better separated than with PCA dimension reduction (bottom row).

Table 1

Demographic information of subjects used in simulations (Southwest University and Human Connectome Project datasets) and real data analysis (4 schizophrenia datasets: COBRE, Taiwan, NMorphCH and BrainGluSch).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Group</th>
<th># Subjects</th>
<th>Age (mean ± std)</th>
<th>Gender (M/F)</th>
<th>mean FD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Southwest</td>
<td>University</td>
<td>281</td>
<td>19.7 ± 0.85</td>
<td>0/281</td>
<td>0.09 ± 0.03</td>
</tr>
<tr>
<td>HCP</td>
<td>Control</td>
<td>150</td>
<td>28.8 ± 3.71</td>
<td>64/86</td>
<td>0.08 ± 0.02</td>
</tr>
<tr>
<td>COBRE</td>
<td>Control</td>
<td>72</td>
<td>35.6 ± 11.7</td>
<td>50/22</td>
<td>0.21 ± 0.11</td>
</tr>
<tr>
<td></td>
<td>Patient</td>
<td>58</td>
<td>36.7 ± 13.5</td>
<td>49/9</td>
<td>0.24 ± 0.11</td>
</tr>
<tr>
<td></td>
<td>Statistic</td>
<td>0.01</td>
<td>0.01</td>
<td></td>
<td>0.07</td>
</tr>
<tr>
<td>Taiwan</td>
<td>Control</td>
<td>136</td>
<td>44.1 ± 12.0</td>
<td>57/79</td>
<td>0.11 ± 0.05</td>
</tr>
<tr>
<td></td>
<td>Patient</td>
<td>123</td>
<td>44.0 ± 11.3</td>
<td>51/72</td>
<td>0.10 ± 0.07</td>
</tr>
<tr>
<td></td>
<td>Statistic</td>
<td>0.79</td>
<td>1</td>
<td>0.07</td>
<td>0.66</td>
</tr>
<tr>
<td>NMorphCH</td>
<td>Control</td>
<td>39</td>
<td>30.6 ± 8.1</td>
<td>19/20</td>
<td>0.13 ± 0.07</td>
</tr>
<tr>
<td></td>
<td>Patient</td>
<td>42</td>
<td>32.8 ± 6.9</td>
<td>30/12</td>
<td>0.18 ± 0.01</td>
</tr>
<tr>
<td></td>
<td>Statistic</td>
<td>0.20</td>
<td>0.04</td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td>BrainGluSch</td>
<td>Control</td>
<td>76</td>
<td>38.7 ± 12.4</td>
<td>49/27</td>
<td>0.23 ± 0.11</td>
</tr>
<tr>
<td></td>
<td>Patient</td>
<td>60</td>
<td>34.5 ± 15.6</td>
<td>55/5</td>
<td>0.21 ± 0.11</td>
</tr>
<tr>
<td></td>
<td>Statistic</td>
<td>0.06</td>
<td>1.9e-4</td>
<td></td>
<td>0.54</td>
</tr>
</tbody>
</table>
Comparing the statistical power of detecting nonlinear signals with other methods

Fig. 6 shows the results of comparing the power of different methods when the true signal is nonlinear, i.e. a subset of functional connectivities are nonlinearly correlated with a phenotype of interest in volume-based fMRI data (see Section 3.4). Similarly, Fig. 7 reports the results of the same simulation using surface-based fMRI data.

For both volume-based and surface-based data, it can be clearly seen that the proposed sKPCR method with a polynomial or linear kernel always has the highest power in different situations (different signal-to-noise ratios and polynomial degrees). When the polynomial degree is an even number, sKPCR with a corresponding polynomial kernel outperforms all other approaches, but when the polynomial degree is an odd number, sKPCR with a corresponding polynomial kernel and linear kernel have the similar high power. This is because the polynomial signals with odd degree are quite similar to linear signals (e.g., consider $y = x^3 + \varepsilon$ and $y = x + \varepsilon$).

5.1.4. The performance of sKPCR when the kernel is misspecified

Again, when the true signals are nonlinear, we can see from Figs. 6...
and 7 that methods using a linear kernel usually have decreased power compared with the corresponding nonlinear kernel. This highlights the importance of specifying correct kernels in the analysis to achieve optimal power. However, this does not mean that our method is sensitive to the choice of kernels. In practice, we can run sKPCR with different kernels, and for each kernel, we will get a voxel-wise p-value map which reflects the strength of different types of association signals. In addition, although the power of sKPCR with a linear kernel decreases when the true signals are nonlinear, we can see that it is still higher than many other approaches, even with the correct nonlinear kernels. This may result from its effective modelling of the spatial structure of the data.

5.1.5. The robustness of sKPCR when the number of principal components is misspecified

Fig. 8 shows the results of a power comparison between adaptive regression and the general linear model in association testing with different numbers of noise components and signal-to-noise ratios (see Section 3.6). It can be seen that adaptive regression is robust to the misspecification of components because its power does not change much, even when an increasing number of noise components are added to the model. However, the power of the general linear model decreased dramatically when an increasing number of noise components are included in the model.
5.1.6. The stability of sKPCR under different covariance operators

Fig. 9 shows that the power of sKPCR using different covariance operators under different signal-to-noise ratios and different proportions of null functional connectivities (see Section 3.7). We find that its power is similar across three different covariance operators and operator parameters. The GL operator has a slightly higher power, so we used it throughout our simulations and real data analysis.
Fig. 8. A simulated example to demonstrate the power benefits of adaptive regression compared to the traditional general linear model $F$ test. In the absence of a signal (no PCs correlate with phenotypes of interest, top left figure), both approaches can control the type one error rate accurately, i.e., power $= \alpha \approx 5\%$. When signals exist and the number of noise PCs increases, however, the power of the adaptive regression approach is stable, which means that it is sufficiently robust with respect to the selection of the number of PCs, while the power of the traditional general linear model decreases dramatically. When the number of noise PCs is small, it should be noted that adaptive regression has a somewhat lower power than the traditional general linear model.

sKPCR power using different covariance operators

Fig. 9. The power of linear sKPCR with different covariance operators under different signal-to-noise ratio (0–0.25) and proportions of null functional connectivities (0.8–0.99). The covariance operators evaluated here are: Graph Laplacian (GL), Normalized Graph Laplacian (NGL), and Gaussian with $\sigma^2 = 2, 4, 6$ (Gaussian (2), Gaussian (4), Gaussian (6)).
5.2. Evaluating $sKPCR$ in real data: brain-wide associations of the schizophrenic connectome

5.2.1. Between-sites reproducibility

Table 2 shows the results of comparing the between-sites reproducibility of different approaches (see Section 4.2). In five of the six comparisons, our $sKPCR$ approach with a linear kernel achieved the highest between-sites reproducibility values, as measured by DC. Only the univariate approach outperformed our method in one comparison. The voxel-wise p-value map of $sKPCR$ in four datasets is shown in Figs. S1–S4, and the corresponding results with global signal regression are shown in Figs. S5–S9, which were very similar to the results without global signal regression.

5.2.2. Literature evidence: overlaps with existing meta-analysis findings

Table 3 shows the results of comparing the overlap of the findings with different methods with existing meta-analysis findings in the Neurosynth database (see Section 4.3). Our $sKPCR$ method with a linear kernel has the highest overlap with existing findings in the literature. The schizophrenia meta-analysis map is shown in Supplementary Fig. S5.

5.2.3. Computation time

Finally, we compare the computation time of $sKPCR$, MDMR, and $aSPU$ in the above analyses. All the analyses were implemented in MATLAB 2016b using 20 cores on a Linux workstation with Intel Xeon E5-2660 v3 (2.60 GHz) CPU and 128 GB memory. Table 4 shows that our method is the most efficient.

6. Discussion

The $sKPCR$ described in this paper is a powerful and efficient multivariate approach for voxel-level connectome-wide association studies. It can identify voxels, the overall connectivity pattern of which, as summarised by $sKPCA$ as low-dimensional features, correlates with the phenotypes of interest. The idea behind $sKPCR$ simply involves reducing the dimensionality of the connectivity features and then performing association studies. However, we went further and carefully refined these two steps, aiming to extract more information from the fMRI data. Specifically, $sKPCR$ models the spatial noise structure in the dimension reduction step and automatically selects an optimal number of principal components in the association testing steps. In our simulation, we demonstrated that $sKPCR$ usually had the highest power in both volume-based and surface-based fMRI data for detecting both linear and nonlinear signals. In real data analysis, we showed that $sKPCR$ usually had better between-sites reproducibility, larger overlap with existing findings, and faster computation speed.

A voxel/vertex identified by this approach can be interpreted as ‘there may exist one or more functional connectivities which connect it that are associated with the phenotype of interest’. To know the associated connections, a subsequent seed-based analysis can be performed. That is, we can extract a seed time series by averaging the voxel/vertex’s time series within a significant cluster and test the associations between the seed connectivity map and a phenotype of interest. However, no significant individual connections in the seed-based analysis may be found because our approach can detect more than simple linear association signals. For example, consider a scenario in which many of the connections only have small effect sizes. In addition, as our approach can produce a voxel-wise statistical map, but not a connectivity-wise result, it can be directly compared with results of other analyses, such as task-activation studies, voxel-based morphometry (VBM) analysis, and Neurosynth meta-analysis results, even though our results do not reflect the direction of the association.

Additional areas can be refined. First, the method currently can only analyse binary and continuous phenotype variables. However, it could be extended to analyse categorical and multivariate phenotypes. Second, a sparse version of $sKPCA$, which allows only a subset of functional connectivities related to a principal component, may further improve the performance of dimension reduction, just like sparse PCA (Witten et al., 2009) improves PCA. Third, with the larger size of the available datasets, such as HCP and the UK-Biobank, an online version of $sKPCA$ should be an important extension because it is not currently possible to fit thousands of high-resolution fMRI data into memory. However, $sKPCR$ could be equipped to analyse big datasets by borrowing ideas from the online/group PCA approach (Smith et al., 2014) and other related variants (Monti and Hyvärinen, 2018; Chen et al., 2015a). Fourth, it would also be very interesting to extend $sKPCR$ to infer causal relationships (Tran and Blei, 2017). Finally, the current $sKPCR$ method is designed for single-site studies. However, combining the $sKPCR$ results from multiple imaging sites is an important extension for the future. Possible methods include conventional meta-analysis methods and the model-based site-effect adjustment methods, such as ComBat (Johnson et al., 2007).

Acknowledgements

Weikang Gong would like to thank Professor Stephen M. Smith for valuable discussions and suggestions. Jianfeng Feng is supported by the National High Technology Research and Development Program of China (No. 2015AA020507), the Key Program of National Natural Science Foundation of China (No. 91230201), the International (Regional) Collaborative and Exchange Program of National Natural Science (No. 91583103), and the New Century Excellent Talents in University (No. NCET-11-0881).
Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuroimage.2018.12.032.

References

Bai, Jushan, Ng, Serena, 2002. Determining the number of factors in approximate factor models. Econometrica 70 (1), 191–221.

640