ABSTRACT
Multiple-set canonical correlation analysis (mCCA) is a generalization of canonical correlation analysis (CCA) to three or more sets of variables. It aims to study the relationships between several sets of variables and it subsumes a number of interesting multivariate data analysis techniques as special cases. The quality and interpretability of the mCCA components are likely to be affected by the usefulness and relevance of each set of variables. Therefore, it is an important issue to identify each set of significant variables that are active in the relationships between sets. In this paper mCCA is extended to address the issue of variable set selection. Specifically a block sparse multiple set canonical correlation analysis (BSmCCA) algorithm is proposed to combine mCCA with $\ell_2$-norm type penalty in a unified framework. Within this framework sets of variables that are not necessarily relevant are removed. This makes BSmCCA a flexible method for analyzing for Multi-Subject functional magnetic resonance imaging (fMRI) data sets. The performances of the proposed BSmCCA algorithm are illustrated through block design paradigm finger tapping fMRI datasets.

Index Terms— Multiple-set canonical correlation analysis, fMRI, penalized rank one approximation, variable set selection.

1. INTRODUCTION
A number of different approaches have been proposed for the study of brain function through the analysis functional magnetic resonance imaging (fMRI) data sets. These approaches can be broadly divided into two main classes: model-based or data-driven. Model-based methods through the general linear model (GLM) [1] have been widely used. They use the a priori knowledge about the properties of the data; i.e.; the hemodynamic response function (HRF) and the experimental paradigm; i.e.; the stimulus function, to investigate the goodness-of-fit of the model and make inferences about regional brain activities. The use of these methods can however be limited for a number of reasons among them the absence of experimental paradigm; when studying resting state or naturalistic paradigms such as continuous listening or watching a movie.

Data-driven methods have also been successfully applied to fMRI data analysis. Among the justification for their suitability for fMRI data analysis is the minimization of the assumptions on the underlying structure of the problem. These methods mainly tries to decompose the observed data based on a factor model and a specific constraint. Different constraints have led to different data-driven methods. The maximum variance constraint has led to principal component analysis (PCA) [2], the independence constraint has led to spatial ICA (sICA, for the format of the data described above) and temporal ICA (tICA) [3] and the sparsity constraint has led to dictionary learning [4].

The maximum correlation constraint which leads to canonical correlation analysis (CCA) [5] has also been successfully used for fMRI data analysis. It has for example been used to find latent sources in single subject fMRI data by taking advantage of the spatial or temporal autocorrelation in the data [6] or improve the specificity and sensitivity of dictionary learning methods for fMRI by accounting for the autocorrelation in the fMRI signals [7, 8]. Its extension to multiple data sets, termed multiset canonical correlation analysis (mCCA) [9] has also successfully been used in association with other methods for the analysis of multiple fMRI data sets. It has for example been successfully used in conjunction with dictionary learning for multi-subject fMRI data analysis in [10] and in conjunction with ICA to maintain the correspondence of the source estimates across different subjects in [11, 12, 13]. When working with multi-subjects data sets, the general canonical components obtained using the standard mCCA involves all the individual subjects data sets as mCCA doesn’t perform group variable selection. This makes it difficult to interpret these components without using subjective judgment. To ease this drawback of mCCA, a new method called block sparse mCCA (BSmCCA) is introduced in this paper. This method can be seen as a generalization of sparse canonical correlation analysis to three or more sets of variables.

This work was supported by the Australian Research Council through Grant FT: 130101394
BSmCCA combines the power of multi-block data analysis of mCCA and the interpretability of group variables selection [14][15]. Considering that some data sets or group of variables may be irrelevant or useless, the objective of BSmCCA is to find linear combinations of block variables (block components) such that the block components that are assumed to be connected are highly correlated.

2. BACKGROUND: CCA AND MCCA

2.1. Canonical correlation analysis: CCA

CCA is a standard approach for studying the relationships between two sets of random variables. Given two random vectors $\mathbf{x} = (x_1, ..., x_p)^\top$ and $\mathbf{y} = (y_1, ..., y_q)^\top$ and suppose $n$ i.i.d samples of $\mathbf{x}$ and $\mathbf{y}$ denoted by $\mathbf{X} \in R^{n \times p}$ and $\mathbf{Y} \in R^{n \times q}$, respectively, have been collected. Assuming that the columns of both $\mathbf{X}$ and $\mathbf{Y}$ have been centered and scaled, CCA aims to find two directions of projection $a_1 \in R^p$ and $b_1 \in R^q$ so that

$$(a_1, b_1) = \arg \max_{a_1, b_1} a_1^\top \Sigma_{xy} b_1$$

subject to $a_1^\top \Sigma_{xx} a_1 = 1$ and $b_1^\top \Sigma_{yy} b_1 = 1$ (1)

where $\Sigma_{xx}$, $\Sigma_{yy}$ and $\Sigma_{xy}$ are covariance and cross-covariance matrices. The vectors $a_1$ and $b_1$ are called the first pair of canonical vectors while the variables $\mathbf{a}_1^\top \mathbf{X}$, $\mathbf{b}_1^\top \mathbf{Y}$ are called the first pair of canonical components and $\rho_1 = \mathbf{a}_1^\top \Sigma_{xy} \mathbf{b}_1$ is referred to as the first canonical correlation. Given $d \leq \min(rank(\mathbf{X}, \mathbf{Y}))$, $d$ of such canonical vectors $\mathbf{A} \in R^{p \times d}$ and $\mathbf{B} \in R^{q \times d}$ are obtained using a deflationary approach with the following additional constraints, first

$$\mathbf{A}^\top \Sigma_{xx} \mathbf{A} = \mathbf{B}^\top \Sigma_{yy} \mathbf{B} = I_d$$

which traduces the none correlation of the canonical components, these components are zero mean and unit variance. Second

$$\mathbf{A}^\top \Sigma_{xy} \mathbf{B} = \mathbf{R} = \text{diag}(\rho_1, ..., \rho_d)$$

which traduces the correlation of the corresponding canonical components only. The closed form solutions for $\mathbf{A}$ and $\mathbf{B}$ obtained as the solution of a constrained optimization problem using Lagrange multipliers results in the first $d$ eigenvectors of $\Sigma_{xx}^{-1} \Sigma_{xy} \Sigma_{yy}^{-1}$ and $\Sigma_{xy}^\top \Sigma_{xx}^{-1} \Sigma_{xy}$ respectively, CCA results in vectors $a$ and $b$ that are not sparse, and these vectors are not unique if $p$ and $q$ exceeds $n$. In applications where $p$ and $q$ are large, one possible interest is in finding projections vectors that leads to high correlation but that are also sparse [16][17][18]. This is realized by imposing a sparsity constraint to (1).

2.2. Multiset Canonical correlation analysis: mCCA

CCA between two sets of random variables can be extended to three or more sets of variables in different ways [9]. In an extension of the CCA principle where correlation between two canonical components is maximized, mCCA generalizes the CCA principle to optimize an objective function of the correlations between pairs of canonical components associated to the different sets of random variables with respect to the canonical vectors so that an overall correlation is maximized. Given $n$ realizations of $J$ sets of $p_j$ dimensional random vectors $\mathbf{x}_j$ so that $\mathbf{x}_j$ is of size $n \times p_j$, the canonical vectors $a_1, ..., a_J$ are determined as

$$a_1, ..., a_J = \arg \max_{a_1, ..., a_J} \sum_{j,k=1}^J g(a_j^\top \Sigma_{jk} a_k)$$

In the optimization (4), the definition of the function $g$ leads to specific methods defined in [9]. For example, the function $g(x) = x$ corresponds to the sum of correlation method (SUMCOR) and the function $g(x) = x^2$ corresponds to the sum of squared correlations method (SSQCOR). Furthermore, these functions are not constrained, but several natural choices for constraints under which to carry out the optimizations (4) can be taken, among them $a_1^\top a_1 = 1$, $\sum_{j=1}^J a_j^\top a_j = 1$ and $a_j^\top \Sigma_{ii} a_j = 1$. For the subsequent stages, the canonical vectors are further constrained to be uncorrelated to the ones generated in the previous stages. In all cases defined in [9], mCCA reduces to CCA when $J = 2$. Similar to the two sets case, the linear combinations $u_j = \mathbf{X} a_j$, $j = 1, ..., J$ represent the canonical correlation components associated to the canonical vectors $a_1, ..., a_J$. In what follows the focus is put on the SUMCOR method.

3. BLOCK SPARSE MCCA

3.1. mCCA as a rank one matrix approximation

In what follows all $\Sigma_{ij}$ are assumed to be full rank and the data considered are the whitened data sets $\mathbf{x}_j = \mathbf{x}_j \Sigma_j^{-1/2}$ so that there are no difference between the constraints $\sum_{i=1}^J a_i^\top a_i = 1$ and $\sum_{i=1}^J a_i^\top \Sigma_{ii} a_i = 1$. This constraint is the natural extension from the two sets case and is the only one considered in this paper. Let $\mathbf{X}$ denote the $n \times p$, where $p = \sum_{j=1}^J p_j$ row block matrix, $\mathbf{X} = [\mathbf{X}_1, ..., \mathbf{X}_J]$.

Proposition 1: Let the $p \times d$ matrix $\mathbf{A}$ be partitioned conformally with the partition of $\mathbf{X}$, that is, $\mathbf{A} = [\mathbf{A}_{1}, ..., \mathbf{A}_{J}]^\top$, where $\mathbf{A}_j$ is a $p_j \times d$ matrix. In SUMCOR, $\mathbf{A}$ comprises the first $d$ right singular vectors associated to the $d$ largest singular values of $\mathbf{X}$.

Proof: Under the SUMCOR objective and the constraint $\sum_{j=1}^J a_j^\top a_j = 1$, the first set of canonical vectors is obtained as the solution of the optimization

$$a_1, ..., a_J = \arg \max_{a_1, ..., a_J} \sum_{j,k=1}^J a_j^\top \Sigma_{jk} a_k - \lambda \left( \sum_{j=1}^J a_j^\top a_j - 1 \right)$$
By setting the derivative with respect to \( a_j \) to zero we get
\[
\sum_{k=1}^{J} \Sigma_{jk} a_k = \lambda a_j, \quad j = 1, \ldots, J
\]
or
\[
X^\top X a = \lambda a, \quad a = (a_1, \ldots, a_J).
\]
since \( \Sigma_{jk} = X_k^\top X_k \). Therefore \( a \) is the first eigenvector of \( X^\top X \) and thus the first left singular vector of \( X \). Adding the orthogonality constraint between subsequent sets of canonical vectors, we have
\[
A = \text{arg max}_A \text{tr} \left( A^\top X^\top X A \right)
\]
subject to the restriction \( A^\top A = I_d \). Furthermore we can easily see that the left singular vector is given by
\[
u = \sum_{j=1}^{J} X_j a_j
\]
and is also of unit norm as the whitened data sets are considered.

The SUMCOR problem can then be reformulated as the following rank one matrix approximation optimization problem
\[
\min_{u,a} \| X - u a^\top \|_F^2
\]
where \( \bar{a} = \lambda a \). The subsequent pairs \( (u_k, \lambda_k a_k), 1 < k \leq J \), provide best rank one approximations of the corresponding residual matrices or rank one matrix deflation. For example, \( \lambda_2 u_2 a_2^2 \) is the best rank one approximation of \( X - \lambda_1 u_1 a_1^\top \).

### 3.2. Block sparse mCCA via penalization rank one matrix approximation

The key idea of the proposed block sparse mCCA algorithm is the observation that the optimization (5) is a least squares problem. For fixed \( u \), the optimal \( \bar{a} \) is the least squares co-efficient vector of regressing the columns of \( X \) on \( u \). Introducing block sparsity on \( \bar{a} \) in such a context is the familiar group variable selection problem in least squares regression.

A group sparsity on \( \bar{a} \) has two main advantages. First, the canonical correlation components are more interpretable as the irrelevant or negligible \( \bar{a}_i \)’s will not appear in \( \bar{a} \). Second the canonical correlation component \( u \) will not lose much in terms of the sum of correlation it explains.

To achieve group sparsity on \( \bar{a} \), we propose to use the \( \ell_2 \)-norm type penalty [14][15] in (5) to promote shrinkage and sparsity of the groups \( \bar{a}_i \)’s
\[
\min_{u,a} \frac{1}{2} \| X - u \bar{a}^\top \|_F^2 + \alpha \sum_{i=1}^{J} \sqrt{\rho_i} \| \bar{a}_i \|_2
\]
where \( \alpha \) is the tuning parameter that can be adjusted using a model selection criterion [19, 20, 21, 22]. The minimization of (6) with respect to \( u \) and \( \bar{a} \) under the constraint \( \| u \|_2 = 1 \) can be obtained via an iterative algorithm. Consider first the problem of optimizing over \( u \) for a fixed \( \bar{a} \).

**Proposition 2**: For a fixed \( \bar{a} \), the \( u \) that minimizes (6) and verifies \( \| u \|_2 = 1 \) is \( u = X \bar{a} / \| X \bar{a} \| \).

**Proof**: From (6) we have
\[
u = \arg \min_u \frac{1}{2} \| X - u \bar{a}^\top \|_F^2 + \alpha \sum_{i=1}^{J} \sqrt{\rho_i} \| \bar{a}_i \|_2
\]
\[
\propto \arg \min_u -u^\top X \bar{a} + \frac{1}{2} u^\top u \bar{a}^\top \bar{a}
\]
(7)

Proposition 2 is obtained by deriving (7) with respect to \( u \) and imposing the norm one constraint.

Second the problem of optimizing over \( \bar{a} \) for fixed \( u \) is considered.

**Proposition 3**: For a fixed \( u \), the \( \bar{a} \) that minimizes (6) is obtained by iteratively applying
\[
\bar{a}_j = \left(1 - \frac{\alpha \sqrt{\rho_j}}{\| u^\top X_j \|}_2 \right) + u^\top X_j
\]
where \( (x)_+ = \max(x, 0) \) to \( \bar{a}_j, j = 1, \ldots, J \).

**Proof**: From (6) we have
\[
\bar{a}_j = \arg \min_{\bar{a}_j} \frac{1}{2} \| X - u \bar{a}^\top \|_F^2 + \alpha \sum_{i=1}^{J} \sqrt{\rho_i} \| \bar{a}_i \|_2
\]
\[
\propto \arg \min_{\bar{a}_j} -u^\top X \sum_{i=1, i \neq j}^{J} X_i \bar{a}_i - u^\top X_j \bar{a}_j + \alpha \sqrt{\rho_j} \| \bar{a}_j \|_2
\]
\[
+ \frac{1}{2} \| u \|_2^2 \sum_{i=1, i \neq j}^{J} \bar{a}_i^\top \bar{a}_i + \frac{1}{2} \| u \|_2^2 \bar{a}_j^\top \bar{a}_j
\]
\[
+ \alpha \sum_{i=1, i \neq j}^{J} \sqrt{\rho_i} \| \bar{a}_i \|_2.
\]
(9)

By deriving (9) with respect to \( \bar{a}_j \), and accounting for the norm one of \( u \) we easily obtain the necessary and sufficient condition for \( \bar{a} \) to be a solution (6)
\[
\bar{a}_j + \alpha \sqrt{\rho_i} \| \bar{a}_j \|_2 = u^\top X_j \quad \forall \bar{a}_j \neq 0
\]
(10)
\[
\| u^\top X_j \|_2 \leq \alpha \sqrt{\rho_i} \quad \forall \bar{a}_j = 0
\]
(11)

It can easily be verified that the solution to expressions (10) and (11) is (8).

The above derivations leads to the following iterative procedure for minimizing (6).
BSmCCA Algorithm:
1. Initialize: apply the SVD to $X$ to obtain an initial $u$ and $\tilde{a} = \lambda a$ as indicated in (5).
2. Update $\tilde{a}$: While $\| \tilde{a}_{iter} - \tilde{a}_{(iter-1)} \| > \epsilon$ iterate (8) for $j = 1 : J$
3. Update $u$ using $u = X\tilde{a} / \| X\tilde{a} \|$. 
4. Repeat 2 and 3 until convergence.
5. Standardize $\tilde{a} = \tilde{a} / \| \tilde{a} \|$ to obtain $\lambda$.

Setting $\alpha = 0$ in the BSmCCA above algorithm reduces step 2 to $\tilde{a}$ to the minimizer of (5) and therefore the normal SUMCOR algorithm. The computation cost of each step of this algorithm is $O(np)$. The BSmCCA algorithm is particularly suited when $p > n$.

4. EXPERIMENTAL RESULTS

For experimental verification of our method, we have used three subject tMRI datasets from a block-paradigm right finger tapping (RFT) task from [4]. During the image acquisition, subjects were asked to perform right finger tapping task for 15 sec followed by a 72 sec resting period. This task was repeated 4 times. For details of image acquisition process, reader is referred to [4].

Preprocessing steps were done using SPM-12 in Matlab which included head motion correction, coregistration, normalization to standard MNI template, resampled to $2 \times 2 \times 2$ mm$^3$ voxels and spatial smoothing was performed using a $8 \times 8 \times 8$ mm$^3$ full-width at half-maximum (FWHM) Gaussian kernel. To improve SNR and the low frequency scanner drifts, we used a discrete cosine transform (DCT) with a Gaussian kernel. To improve SNR and improve the low frequency scan.

After the BSmCCA, we combine the canonical vectors of all the data-sets. This is then used to project all the data and construct a mean group activation map that represents the group of patients. From the figure 1, it can be observed that the activation maps contain strong activation in the motor cortex area as also seen in [23, 24, 25]. Thus, the BSmCCA,

- Successfully removes the noisy patient from the analysis, by reducing it’s canonical vector to zero.
- Constructs a mean activation map that represents the group of patients, excluding noise.

5. CONCLUSION

A new multiple-set canonical correlation analysis (mCCA) method is proposed in this paper. The proposed method is based on the SUMCOR method for mCCA and allows the identification of significant sets of variable that are active in the relationships between sets of variables. The algorithm is derived by establishing the link between the set of canonical vectors and rank one matrix approximation. The selection of the most significant sets of variables is obtained via penalized rank one matrix approximation where the sum of $\ell_2$-norm of variables sets penalty is used to shrink the irrelevant sets of variables to zero. While only the SUMCOR methods is addressed in this paper, the proposed method can easily be extended to the SSQCOR method. Note that the proposed algorithm can also be seen as a block sparse principal component analysis (PCA) algorithm.

6. REFERENCES


