# Clustering Linear Discriminant Analysis for MEG-Based Brain Computer Interfaces

Jinyin Zhang, *Student Member, IEEE*, Gustavo Sudre, *Student Member, IEEE*, Xin Li, *Senior Member, IEEE*, Wei Wang, Douglas J. Weber, *Member, IEEE*, and Anto Bagic

***Abstract*—In this paper, we propose a clustering linear discriminant analysis algorithm (CLDA) to accurately decode hand movement directions from a small number of training trials for magnetoencephalography-based brain computer interfaces (BCIs). CLDA first applies a spectral clustering algorithm to automatically partition the BCI features into several groups where the within-group correlation is maximized and the between-group correlation is minimized. As such, the covariance matrix of all features can be approximated as a block diagonal matrix, thereby facilitating us to accurately extract the correlation information required by movement decoding from a small set of training data. The efficiency of the proposed CLDA algorithm is theoretically studied and an error bound is derived. Our experiment on movement decoding of five human subjects demonstrates that CLDA achieves superior decoding accuracy over other traditional approaches. The average accuracy of CLDA is 87% for single-trial movement decoding of four directions (i.e., up, down, left, and right).**

***Index Terms*—Brain–computer interface (BCI), linear discriminant analysis (LDA), magnetoencephalography (MEG), spectral clustering.**

## I. Introduction

A BRAIN–COMPUTER interface (BCI) provides a direct control pathway from brain to external devices [1]–[3]. It is a new communication option for those with neuromuscular impairments that prevent them from using conventional augmentative communication methods. Although applications for patients with severe motor disabilities have been the driving force of most BCI research, the potential of BCI for healthy users is also extensive, including applications such as computer games and home entertainment systems.

To develop a practical BCI system, efficiently recording brain activity and decoding users' intention are two important but challenging tasks. Magnetoencephalography (MEG) is a noninvasive modality that measures magnetic fields generated by electrical neural activity [4]. MEG records brain signals with high temporal resolution. It is a valuable technique complementary to other noninvasive recording modalities like electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) [5]. In the context of BCI, MEG is an important tool to train human subjects to modulate their neural activity for movement control [6]–[8]. In this paper, we focus on an MEG-based BCI system where human subjects perform overt or imagined movement of their wrists.

Once brain activity is recorded, various signal processing and machine learning algorithms, e.g., linear discriminant analysis (LDA) [3], [8], [10], [11], support vector machine (SVM) [6], [10], [12], common spatial pattern (CSP) [13]–[15], etc., can be applied to decode users' intention in a BCI system. If the feature space is high-dimensional and the training data are limited, the movement decoding algorithm must be carefully designed to prevent the decoder from over-fitting the training data. To address this over-fitting problem, feature selection [6], [12] and/or regularization [8], [10] are often applied. For instance, diagonal LDA (DLDA) [11] and regularized LDA (RLDA) [8], [10] have been used in several BCI systems. Both of them pose extra constraints on BCI features to address the aforementioned dimensionality issue. In particular, DLDA assumes mutual independence among all features so that their covariance matrix can be approximated as a diagonal matrix. On the other hand, RLDA applies a Bayesian inference where a simple prior with diagonal covariance matrix is assumed for all features. In other words, both DLDA and RLDA rely on the prior knowledge that all features are mutually independent. While these two methods have been successfully applied to many practical BCI problems, they may not guarantee high decoding accuracy if the underlying prior knowledge does not represent the actual correlation structure of features.

In this paper, we propose a clustering linear discriminant analysis (CLDA) algorithm for BCI movement decoding. Unlike DLDA or RLDA, CLDA utilizes a unique group structure to model the correlation information of BCI features. It partitions all features into several groups where the within-group correlation is maximized and the between-group correlation is minimized. As such, the covariance matrix of all features can be approximated as a block diagonal matrix, thereby facilitating us to accurately extract the correlation information required by movement decoding from a small set of training data. Note that the traditional DLDA method can be conceptually viewed as a special case of CLDA where each group only contains a single

Manuscript received September 22, 2010; revised December 20, 2010; accepted January 31, 2011. Date of publication February 22, 2011; date of current version June 08, 2011. This work was supported in part by the National Institutes of Health under Grant 5 UL1 RR024153, Grant KL2 RR024154, Grant 1R01EB007749, and Grant 1R21NS056136), in part by the National Science Foundation under Grant EEEC-0540865, in part by the Telemedicine and Advanced Technology Research Center under Grant W81XWH-07-1-0716, in part by the Craig H. Neilsen Foundation, and in part by the University of Pittsburgh.

J. Zhang and X. Li are with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: jinyinz@ece.cmu.edu; xinli@ece.cmu.edu).

G. Sudre is with the Program in Neural Computation (PNC), Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: gsudre@pobox.com).

W. Wang and D. J. Weber are with the Department of Physical Medicine and Rehabilitation and the Department of Bioengineering, University of Pittsburgh, Pittsburgh, PA 15213 USA (e-mail: wangw4@upmc.edu; djw50@pitt.edu).

A. Bagic is with the Department of Neurology, University of Pittsburgh, Pittsburgh, PA 15213 USA (e-mail: bagica@upmc.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNSRE.2011.2116125

feature and, hence, no within-group correlation is modeled. From this point of view, the proposed CLDA algorithm is a generalized version of DLDA. It aims to achieve improved decoding accuracy by accurately capturing the correlation information among all features. As will be demonstrated by the experimental results in Section IV, CLDA achieves superior decoding accuracy over other traditional approaches. The average accuracy of CLDA is 87% for single-trial movement decoding of four directions (i.e., up, down, left, and right).

An important contribution of this paper is to apply a spectral clustering algorithm [16]–[19] to automatically identify the underlying group structure of BCI features and assign each feature to the appropriate group. The spectral clustering method first represents the correlation information of BCI features in form of a similarity graph. Next, an optimal partition is constructed to split the graph into several sub-graphs (i.e., groups) based on its Laplacian matrix. The optimal number of groups is automatically determined by measuring the "quality" of the clustering results [18]. In this paper, the spectral clustering algorithm is used, since it is not sensitive to the error of the correlation model estimated from a small set of training data, as is demonstrated by both theoretical studies and application examples in the machine learning community [16]–[19].

In addition, several theoretical aspects of the proposed CLDA algorithm are further examined in order to explain the reason why CLDA outperforms other traditional decoding techniques. An error bound is derived to quantitatively assess the approximation accuracy of the block diagonal covariance matrix and its impact on the final decoding accuracy. It can be shown that the decoding error of CLDA is directly related to the condition number of a normalized covariance matrix Σ0. If the condition number of Σ0 is sufficiently small, the accuracy of CLDA is close to that of an optimal classifier. These results provide theoretical evidence to support the practical utility of the proposed CLDA method.

The remainder of this paper is organized as follows. In Section II, we briefly review the background of LDA, and then propose our CLDA algorithm in Section III. The efficiency of CLDA is demonstrated by a number of experimental examples in Section IV. Several theoretical and practical aspects of CLDA are further discussed in Section V. Finally, we conclude in Section VI.

## II. Background

In this section, we briefly review the background of LDA and two of its modified versions: DLDA and RLDA. Consider two sets of training data $\{\mathbf{x}_{n,1}; n = 1, 2, \ldots, N_1\}$ and $\{\mathbf{x}_{n,2}; n = 1, 2, \cdots, N_2\}$ corresponding to two classes, where $\mathbf{x}_{n,k} = [x_{1,n,k}, x_{2,n,k}, \ldots, x_{M,n,k}]^T$ is the feature vector of the $n$th trial from the $k$th class, $M$ stands for the number of features, and $N_1$ and $N_2$ represent the numbers of training samples for these two classes, respectively. The key idea of LDA is to find the optimal projection direction $\mathbf{p}_{\text{OPT}} \in R^M$ so that the between-class scatter is maximized and the within-class scatter is minimized [29]. Define the within-class scatter matrix $\mathbf{S}_W \in R^{M \times M}$ as

$$\mathbf{S}_W = \sum_{n=1}^{N_1} (\mathbf{x}_{n,1} - \boldsymbol{\mu}_1) \cdot (\mathbf{x}_{n,1} - \boldsymbol{\mu}_1)^T + \sum_{n=1}^{N_2} (\mathbf{x}_{n,2} - \boldsymbol{\mu}_2) \cdot (\mathbf{x}_{n,2} - \boldsymbol{\mu}_2)^T \tag{1}$$

where $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ stand for the mean of $\{\mathbf{x}_{n,1}; n = 1, 2, \ldots, N_1\}$ and $\{\mathbf{x}_{n,2}; n = 1, 2, \ldots, N_2\}$, respectively. If $\mathbf{S}_W$ (1) is nonsingular, $\mathbf{p}_{\text{OPT}}$ can be determined as [29]

$$\mathbf{p}_{\text{OPT}} \propto \mathbf{S}_W^{-1} \cdot (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2). \tag{2}$$

Once $\mathbf{p}_{\text{OPT}}$ is found, the following decision function can be constructed for two-class classification:

$$\mathbf{p}_{\text{OPT}}^T \cdot \left( \mathbf{x} - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2} \right) = \begin{cases} \geq 0 & \text{(First Class)} \\ < 0 & \text{(Second Class)} \end{cases}. \tag{3}$$

The aforementioned two-class LDA can be extended to multiple classes. More details of LDA can be found in [29].

If there are a sufficient number of training samples, $\mathbf{S}_W$ in (1) is an accurate estimator of the covariance matrix and LDA yields the optimal projection direction $\mathbf{p}_{\text{OPT}}$ that maximizes classification accuracy. However, if only a small number of training samples are available for a high-dimensional feature space, it is extremely difficult to accurately estimate the covariance matrix required by LDA. To address this dimensionality issue, DLDA [11] and RLDA [8], [10] have been proposed. DLDA assumes mutual independence among all features, thereby forcing $\mathbf{S}_W$ to be diagonal. Alternatively, RLDA adds an additional regularization term to the estimator: $\mathbf{S}_W + \lambda \cdot \mathbf{I}$, where $\mathbf{I}$ is an identity matrix and $\lambda \geq 0$ is a regularization parameter that is typically determined by cross-validation. Both DLDA (with a diagonal within-class scatter matrix) and RLDA (using a diagonal covariance matrix to model the prior distribution for Bayesian inference) rely on the prior knowledge that all BCI features are mutually independent. While these two methods have been successfully applied to a broad range of practical applications, they cannot guarantee high decoding accuracy if the underlying prior knowledge does not represent the actual correlation structure of BCI features. In addition, as will be demonstrated by the experimental results in Section IV, simple feature selection (e.g., by using Fisher criterion [29]) does not lead to high decoding accuracy. These observations, therefore, motivate us to develop a new CLDA algorithm to achieve improved classification accuracy by carefully modeling the mutual correlation among all features.

## III. Clustering Linear Discriminant Analysis

The proposed CLDA algorithm relies on a unique group structure to extract the correlation information required for movement decoding. Namely, we assign each feature to the appropriate group so that the within-group correlation is maximized and the between-group correlation is minimized. Note that such a feature clustering task is not trivial, since the correlation information extracted from training data is likely to be inaccurate, especially if only a limited number of training samples are available for a high-dimensional feature space. In other words, the challenging issue here is how to develop a *robust* clustering scheme that is not sensitive to the error of the correlation model estimated from a small set of training data.

In this section, we propose to borrow the spectral clustering algorithm [16]–[19] from graph theory [27] to address the aforementioned challenge on feature clustering. Spectral clustering is one of the most important clustering techniques developed by the machine learning community. It first forms a similarity graph based on the mutual correlation of different

$$\mathbf{S}_W = \begin{bmatrix} \mathbf{S}_{W,1,1} & \mathbf{S}_{W,1,2} & 0 & \mathbf{S}_{W,1,4} \\ \mathbf{S}_{W,1,2} & \mathbf{S}_{W,2,2} & 0 & \mathbf{S}_{W,2,4} \\ 0 & 0 & \mathbf{S}_{W,3,3} & \mathbf{S}_{W,3,4} \\ \mathbf{S}_{W,1,4} & \mathbf{S}_{W,2,4} & \mathbf{S}_{W,3,4} & \mathbf{S}_{W,4,4} \end{bmatrix}$$

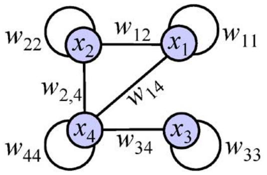

Fig. 1. Simple example of within-class scatter matrix for four BCI features $\{x_1, x_2, x_3, x_4\}$ and the corresponding similarity graph.

features. Next, the features are partitioned into several groups based on the Laplacian matrix of the similarity graph. Here, the spectral clustering algorithm is selected, because it can provide robust performance, even if the input data are noisy, as is demonstrated by both theoretical studies and application examples in the machine learning community [16]–[19]. Hence, the spectral clustering algorithm perfectly fits the need of our feature clustering problem. Based upon spectral clustering, a modified LDA algorithm (i.e., CLDA) is further proposed for BCI movement decoding using the grouped features. In what follows, we will describe the technical details of the algorithms and highlight their novelties.

### *A. Feature Clustering*

Given a set of BCI features $\{x_m; m = 1, 2, \ldots, M\}$, the goal of feature clustering is to partition all features into several groups such that the features in the same group are similar and different features in different groups are dissimilar to each other. In our application, correlation is the criterion to quantitatively measure the "similarity" between features. Namely, we want to maximize the within-group correlation and simultaneously minimize the between-group correlation.

To mathematically define the aforementioned feature clustering problem, we represent all features $\{x_m; m = 1, 2, \ldots, M\}$ in form of a weighted undirected graph $G = (X, E)$ that is referred to as *similarity graph* in [19]. In this graph $G$, each vertex represents a feature $x_m$ where $m \in \{1, 2, \ldots, M\}$. Two vertices $x_m$ and $x_n$ are connected by an edge $e_{mn}$, if and only if the correlation between these two features is non-zero. The weight $w_{mn}$ of $e_{mn}$ is equal to the correlation coefficient

$$w_{mn} = |\mathbf{S}_{W,m,n}| / (\sqrt{\mathbf{S}_{W,m,m}} \cdot \sqrt{\mathbf{S}_{W,n,n}}) \qquad (m, n = 1, 2, \cdots, M) \quad (4)$$

where $\mathbf{S}_{W,m,n}$ stands for the $(m, n)$th element of the within-class scatter matrix $\mathbf{S}_W$ in (1). For each vertex $x_m$, there is a self-edge $e_{mm}$ and the weight $w_{mm}$ is equal to 1. In (4), the correlation coefficient is defined by the within-class scatter matrix and is always nonnegative. Fig. 1 shows a simple example of within-class scatter matrix for four features $\{x_1, x_2, x_3, x_4\}$ and the corresponding similarity graph.

Based on the similarity graph $G = (X, E)$, we want to partition $G$ into several sub-graphs such that the edges between different sub-graphs have small weights (i.e., the corresponding features are weakly correlated) and the edges within the same sub-graph have large weights (i.e., the corresponding features are strongly correlated). Such a partition can be constructed by using the Laplacian matrix of $G$. In what follows, we will first define several important terminologies in graph theory [19], [27].

The *adjacency matrix* $\mathbf{W} \in R^{M \times M}$ of the similarity graph $G$ is defined as

$$\mathbf{W} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1M} \\ w_{12} & w_{22} & \cdots & w_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ w_{1M} & w_{2M} & \cdots & w_{MM} \end{bmatrix}. \quad (5)$$

Namely, the $(m, n)$th element of $\mathbf{W}$ is the weight $w_{mn}$ of the edge $e_{mn}$. If two vertices $x_m$ and $x_n$ are not connected, $w_{mn}$ is simply set to zero. Since the similarity graph $G$ is undirected, the adjacency matrix $\mathbf{W}$ is symmetric. Based on the adjacency matrix $\mathbf{W}$, the *degree* of a vertex $x_m$ is defined as

$$d_m = \sum_{n=1}^{M} w_{mn} \quad (m = 1, 2, \cdots, M). \quad (6)$$

Remember that the weight $w_{mn}$ is nonzero, if and only if the vertices $x_m$ and $x_n$ are connected. Hence, the degree $d_m$ in (6) is determined by all edges that are connected to $x_m$. The *degree matrix* $\mathbf{D} \in R^{M \times M}$ is defined as a diagonal matrix with $\{d_m; m = 1, 2, \ldots, M\}$ on its diagonal

$$\mathbf{D} = diag(d_1, d_2, \cdots, d_M). \quad (7)$$

Now, we are ready to define the *Laplacian matrix* $\mathbf{L} \in R^{M \times M}$ of the similarity graph $G$

$$\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2} \cdot \mathbf{W} \cdot \mathbf{D}^{-1/2}. \quad (8)$$

It can be shown that the Laplacian matrix $\mathbf{L}$ in (8) is positive semi-definite. All of its eigenvalues are within the interval [0, 1]. In particular, $\lambda = 0$ is one of its eigenvalues (i.e., the smallest eigenvalue) and the corresponding eigenvector is

$$\mathbf{v} = [\sqrt{d_1} \quad \sqrt{d_2} \quad \cdots \quad \sqrt{d_M}]^T. \quad (9)$$

More details on Laplacian matrix can be found in [19], [27].

If the similarity graph $G$ can be exactly partitioned into $K$ sub-graphs $\{G_k; k = 1, 2, \ldots, K\}$ (i.e., there is no edge connecting any two sub-graphs) and all vertices are appropriately ordered (i.e., the vertices in the same sub-graph are grouped together), it is straightforward to verify that the Laplacian matrix $\mathbf{L}$ of the graph $G$ is block diagonal

$$\mathbf{L} = diag(\mathbf{L}_1, \mathbf{L}_2, \cdots, \mathbf{L}_K) \quad (10)$$

where $\mathbf{L}_k$ is the Laplacian matrix of the sub-graph $G_k$. Since $\mathbf{L}_k$ is a Laplacian matrix, $\lambda = 0$ is one of its eigenvalues and we represent the corresponding eigenvector as $\mathbf{v}_k$. Similar to (9), $\mathbf{v}_k$ can be determined by the degrees of the vertices in the sub-graph $G_k$. Based on these observations, we can conclude that the Laplacian matrix $\mathbf{L}$ in (10) has $K$ different eigenvectors that is associated with the same eigenvalue $\lambda = 0$

$$\mathbf{V} = \begin{bmatrix} \mathbf{v}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{v}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{v}_K \end{bmatrix} \quad (11)$$

where each column of the matrix $\mathbf{V} \in R^{M\times K}$ in (11) is an eigenvector of the Laplacian matrix $\mathbf{L}$, and the symbol $\mathbf{0}$ denotes a zero vector (i.e., all elements in $\mathbf{0}$ are zero).

Next, we normalize each row of the matrix $\mathbf{V}$ to unit length, resulting in

$$\tilde{\mathbf{V}} = \begin{bmatrix} \mathbf{1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1} \end{bmatrix} \tag{12}$$

where the symbol $\mathbf{1}$ denotes a vector in which all elements are one. Studying the matrix $\tilde{\mathbf{V}} \in R^{M\times K}$ in (12), one would notice that each row of $\tilde{\mathbf{V}}$ corresponds to a vertex $x_m$ of the similarity graph $G$ (i.e., the BCI feature $x_m$). We can conceptually consider the $m$th row of $\tilde{\mathbf{V}}$ as the coordinate of the $m$th feature $x_m$ in a "transformed" feature space $R^K$. For all features in the same sub-graph $G_k$, their coordinates are identical. Hence, we can use the "normalized" eigenvectors in (12) to partition all BCI features into $K$ groups (e.g., by applying K-means clustering [29]). Such a feature clustering scheme is based on spectral graph theory [27] and, hence, is referred to as spectral clustering in the literature [16]–[19].

While the above discussion covers the key idea of the proposed feature clustering based on spectral graph theory, there are three important implementation issues that should be further considered. First, our previous discussion assumes that all vertices of the similarity graph $G$ are appropriately ordered so that the Laplacian matrix $\mathbf{L}$ in (10) is block diagonal. In practice, this is never the case and the correct ordering of all vertices should be the output of the clustering algorithm. Note that if the ordering of the vertices is changed, the Laplacian matrix of the similarity graph $G$ can be written as a linear transformation of the block diagonal matrix $\mathbf{L}$ in (10): $\mathbf{P}\cdot\mathbf{L}\cdot\mathbf{P}^T$ where $\mathbf{P}$ is a permutation matrix (i.e., an identity matrix with its rows reordered [27]). Compared to the matrix $\mathbf{L}$, the matrix $\mathbf{P}\cdot\mathbf{L}\cdot\mathbf{P}^T$ has identical eigenvalues. The eigenvectors of $\mathbf{L}$ and $\mathbf{P}\cdot\mathbf{L}\cdot\mathbf{P}^T$ only differ by a simple permutation. In other words, the aforementioned properties for eigenvalues and eigenvectors still hold and, hence, the spectral clustering scheme can be applied to appropriately identify the subgraphs, even if all vertices in the similarity graph $G$ are arbitrarily ordered.

Second, the sub-graphs $\{G_k; k = 1, 2, \ldots, K\}$ are not necessarily disconnected in most practical applications, since the BCI features $x_m$ and $x_n$ in different sub-graphs can be weakly correlated and, hence, the weight $w_{mn}$ of the edge $e_{mn}$ is not exactly zero. In addition, since the weight $w_{mn}$ in (4) is calculated from the within-class scatter matrix $\mathbf{S}_W$ in (1) based on a set of training data, the estimation of $w_{mn}$ can be inaccurate, especially if the feature space is high-dimensional and only a small set of training data are available. In these cases, the Laplacian matrix $\mathbf{L}$ in (8) is not exactly block diagonal. However, the spectral clustering result is not sensitive to the small perturbation presented in the correlation model. In other words, even if the Laplacian matrix $\mathbf{L}$ deviates from the ideal (i.e., block diagonal) case, spectral clustering can still yield the correct clustering results, as is demonstrated by both theoretical studies and application examples in the machine learning community [16]–[19]. For instance, a detailed perturbation analysis of the invariant subspace (i.e., the subspace spanned by eigenvectors) has been studied in [19] for spectral clustering. It provides theoretical evidence that the spectral clustering method offers robust performance and, hence, perfectly fits the needs of our proposed feature clustering problem.

Third, the number of clusters (i.e., $K$) is not known in advance when the proposed feature clustering scheme is applied to BCI movement decoding. In other words, the optimal value of $K$ must be automatically determined as part of the clustering procedure. To this end, we borrow the concept of quality function $q(K)$ from [18]. Namely, $q(K)$ is defined to measure the quality of different clustering results with different $K$ values

$$q(K) = \sum_{k=1}^{K} \left\{ f(G_k, G_k)/f(G, G) - [f(G_k, G)/f(G, G)]^2 \right\} \tag{13}$$

where the function $f(A, B)$ measures the similarity between two sub-graphs $A$ and $B$ based on their weights

$$f(A, B) = \sum_{x_m \in A; x_n \in B} w_{mn}. \tag{14}$$

It has been empirically demonstrated by a broad range of simulated and real-world examples in [18] that a large $q(K)$ implies an improved clustering result. Hence, we can repeatedly perform feature clustering with different $K$ values and calculate the "quality" $q(K)$ in (13) for the clustering results. The optimal value of $K$ is then determined by finding the largest quality function $q(K)$.

**Algorithm 1: Feature Clustering**

```
1) Start from the training data $\{\mathbf{x}_{n,1}; n = 1, 2, \ldots, N_1\}$
   and $\{\mathbf{x}_{n,2}; n = 1, 2, \ldots, N_2\}$ corresponding to the
   BCI features $\{x_m; m = 1, 2, \ldots, M\}$ of two classes.
2) Calculate the within-class scatter matrix $\mathbf{S}_W \in R^{M\times M}$
   in (1). Construct the similarity graph $G$. Calculate the
   adjacency matrix $\mathbf{W} \in R^{M\times M}$ in (4), (5), the degree
   matrix $\mathbf{D} \in R^{M\times M}$ in (6), (7), and the Laplacian
   matrix $\mathbf{L} \in R^{M\times M}$ in (8).
For each $K \in \{1, 2, \ldots, M\}$
3) Find the $K$ smallest eigenvalues and the corresponding
   $K$ eigenvectors. Form the matrix $\mathbf{V} \in R^{M\times K}$ where
   each column is one of the $K$ eigenvectors.
4) Normalize each row of the matrix $\mathbf{V}$ to unit length,
   resulting in the matrix $\tilde{\mathbf{V}} \in R^{M\times K}$.
5) Consider each row of $\tilde{\mathbf{V}}$ as the coordinate of a BCI
   feature in the space $R^K$, and apply K-means clustering
   [29] to partition the features $\{x_m; m = 1, 2, \ldots, M\}$
   into $K$ groups.
6) Calculate the quality function $q(K)$ in (13).
End For
7) Find the optimal value $K_{\mathrm{OPT}}$ at which the quality
   function $q(K)$ reaches its maximum. Use the clustering
   result at $K_{\mathrm{OPT}}$ to partition all features into $K_{\mathrm{OPT}}$
   groups.
```

Algorithm 1 summarizes the major steps of the proposed feature clustering method. During the clustering procedure, the optimal number of clusters is automatically determined by evaluating the quality function in (13). Once the features are appropriately partitioned into $K$ groups, a modified LDA algorithm (i.e., CLDA) can be applied for movement decoding using the grouped features. The details of CLDA will be discussed in the next subsection.

### B. Discriminant Analysis

Once the BCI features $\{x_m; m = 1, 2, \ldots, M\}$ are partitioned into $K$ groups, all features are ordered according to their group assignment. Since the mutual correlation between different groups is almost zero, the within-class scatter matrix $\mathbf{S}_W \in R^{M \times M}$ can be approximated by a block diagonal form

$$\mathbf{S}_{WB} = diag(\mathbf{S}_{W,1}, \mathbf{S}_{W,2}, \cdots, \mathbf{S}_{W,K}) \qquad (15)$$

where $\mathbf{S}_{W,k}$ stands for the within-class scatter matrix for the features in the $k$th group. The key idea of CLDA is to estimate the block diagonal matrix $\mathbf{S}_{WB}$ in (15), and use it to replace $\mathbf{S}_W$ in (2) to calculate the optimal projection direction $\mathbf{p}_{\mathrm{OPT}}$. Since the block diagonal matrix $\mathbf{S}_{WB}$ is more constrained than the original within-class scatter matrix $\mathbf{S}_W, \mathbf{S}_{WB}$ is less sensitive to the dimensionality issue posed by high-dimensional feature space and small training data set. In other words, the proposed CLDA can efficiently approximate the within-class scatter matrix using a block diagonal form and then accurately estimate the optimal projection direction $\mathbf{p}_{\mathrm{OPT}}$, even if the feature space is high-dimensional and the training data are limited. This is the primary advantage of CLDA over LDA.

On the other hand, CLDA can be viewed as a direct extension of DLDA [11]. Unlike DLDA that approximates the within-class scatter matrix $\mathbf{S}_W$ by a diagonal matrix and completely ignores the correlation between different features, the proposed CLDA is able to automatically identify the critical correlation information (i.e., by partitioning all features into different groups) and then accurately extract the correlation (i.e., by estimating a block diagonal within-class scatter matrix) from a small set of training data. For this reason, CLDA can capture the mutual correlation of BCI features more accurately than DLDA. Hence, it is expected to achieve superior decoding accuracy over DLDA.

**Algorithm 2: Clustering Linear Discriminant Analysis**

```
1) Start from the training data $\{\mathbf{x}_{n,1}; n = 1, 2, \ldots, N_1\}$
   and $\{\mathbf{x}_{n,2}; n = 1, 2, \ldots, N_2\}$ corresponding to the
   BCI features $\{x_m; m = 1, 2, \ldots, M\}$ of two classes.
2) Apply Algorithm 1 to partition all features into $K_{\mathrm{OPT}}$
   groups.
3) Order all features according to their group assignment.
   Construct the block diagonal within-class scatter
   matrix $\mathbf{S}_{WB}$ in (15).
4) Replace $\mathbf{S}_W$ in (2) by $\mathbf{S}_{WB}$ to calculate the optimal
   projection direction $\mathbf{p}_{\mathrm{OPT}}$.
5) Create the two-class classifier for movement decoding
   based on the decision function in (3).
```

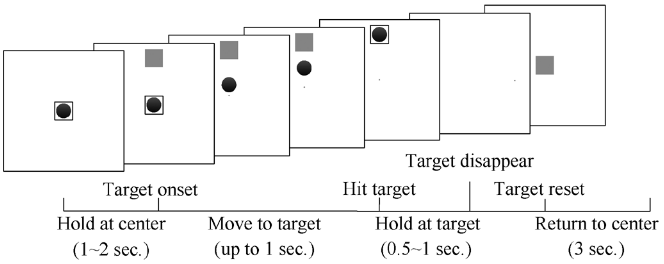

Fig. 2. Simplified diagram of the experimental setup for our MEG-based movement decoding. A human subject first holds the wrist at the center to start a trial. After a peripheral target onset, the subject moves (or imagines moving) the wrist to the target direction and holds that position until the peripheral target disappears. Next, the human subject waits for the target to reappear at the center and then moves (or imagines moving) the wrist back to the center.

Algorithm 2 summarizes the simplified flow of the proposed CLDA method for two classes. It should be noted that Algorithm 2 can be easily extended to multiple classes by following the standard flow of multiclass LDA [29]. Since the extension to multiclass CLDA is straightforward, we will not present its details in this paper.

The efficiency of CLDA will be demonstrated by several experimental examples in the next section. As will be shown in Section IV, CLDA results in substantially higher decoding accuracy than other traditional approaches, including LDA with feature selection, DLDA and RLDA. In addition, a theoretical study will be presented in Appendix to derive the error bound of the proposed block diagonal approximation. It, in turn, further explains why CLDA achieves superior accuracy for BCI movement decoding.

## IV. Experimental Results

In this section, CLDA is applied to MEG-based movement decoding and its performance is compared to other traditional decoding techniques on five human subjects. In what follows, we will describe the experimental setup and the movement decoding results in detail.

### A. Experimental Setup

In our experiment, five human subjects performed a four-target center-out task with their wrist holding an MEG-compatible joystick. During overt movements, subjects were instructed to move the cursor from the center target to one of the four locations (i.e., up, down, left, or right) by making wrist movements (i.e., radial deviation, ulnar deviation, flexion, and extension) while keeping the rest of the body in a relaxed position. A successful repetition was characterized by reaching one of the four peripheral targets within a prespecified time window after the onset of the target and holding the cursor position there without overshooting, as shown in Fig. 2. Only successful repetitions were used for our offline data analysis. During imagined movements, subjects were instructed to imagine making the wrist movements to one of the four targets displayed on the screen, while the cursor moved from the center to the target automatically. For both overt and imagined conditions, subjects were instructed to keep their gaze at the center of the screen, and only attend to the targets using their peripheral vision.

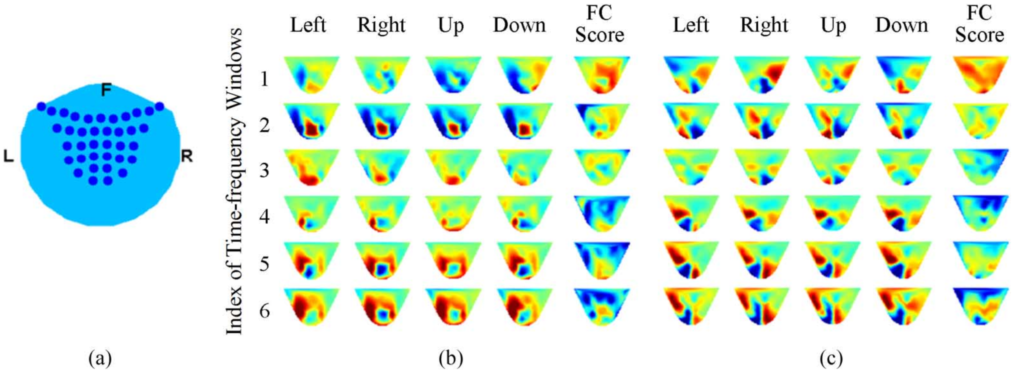

Fig. 3. (a) The 74 selected MEG channels located on top of the sensorimotor area (two gradiometers at each location), (b) the spatial distribution of wavelet coefficients from the first gradiometer, and (c) the spatial distribution of wavelet coefficients from the second gradiometer. In (a), the symbols "L," "R," and "F" represent left, right and front, respectively. In both (b) and (c), the first four columns correspond to the four movement directions, and the last column shows the scores calculated by Fisher criterion (FC). Each color map was calculated by averaging the wavelet coefficients over all trials of the same class. Each row corresponds to the wavelet coefficients associated with the same time-frequency window. Red color indicates large value and blue color indicates small value. All plots of wavelet coefficients share the same color scale, and all plots of FC scores share the same color scale.

TABLE I
NUMBER OF SUCCESSFUL TRIALS PER CLASS OF EACH DATA SET

| Subject ID | SubA | SubB | SubC | SubD | SubE |
|---|---|---|---|---|---|
| Overt | 81 | 84 | 155 | 123 | 93 |
| Imagined | 75 | 179 | 129 | 126 | 98 |

During the experiment, MEG data were acquired by using a 306-channel whole-head MEG system (Elekta Neuromag) with 1 kHz sampling frequency. In addition, electrooculography (EOG) was used to monitor eye blinks and eye movements. Electromyography (EMG) of wrist flexor and extensor muscles was recorded to make sure that no movement happened during the imagined sessions. All trials with EOG or EMG contamination were rejected.

All five subjects performed both overt and imagined movements in the experiment, resulting in a total of ten data sets. The number of successful trials in each data set was not the same after rejecting contaminated repetitions. This number was adjusted for each data set among the four classes (i.e., up, down, left, and right) such that each class in the same data set had the same number of successful trials. In other words, we discarded the last few trials for the classes with more trials than the others. Table I summarizes the data set size for the aforementioned experimental setup.

### B. Data Preprocessing

The recorded MEG signals were processed by the signal space separation (SSS) method [20] to remove the interference signals due to magnetic impurities (e.g., sensor electronics, electrical activities from arm muscles, etc.). SSS also compensated the signal distortions caused by head movement. Next, a notch filter was applied to remove the 60 Hz power line interference. A linear approximation was then determined by least-squares fitting for each channel and each trial, and the linear trends were subtracted from the recorded MEG signals. In this study, although the MEG signals contained 306 channels, only 74 channels were used for decoding. These 74 channels correspond to the gradiometers located on top of the sensorimotor area, as shown in Fig. 3(a). They are expected to carry useful information about the motor activity in which we are interested.

There are many possible ways to extract features from brain signals, including wavelet coefficients [21], [22], power spectral density [3], [10], [23], autoregressive model [6], [7], [12], etc. Previous neuroscience research on MEG-based BCI demonstrated that significant power modulation of MEG activity was observed in three different frequency bands [8]: 1) $\leq$7 Hz (low-frequency band), 2) 62–87 Hz, and 3) 10–30 Hz. In [8], the authors further mentioned that movement directions could be inferred from the low-frequency band only, but not from the other two bands. In addition, the important neural activity that carries movement information was observed during a short time window [8]. For these reasons, we only considered the low-frequency band for the time window $t \in [0.2\ \mathrm{s}, 0.6\ \mathrm{s}]$, where $t = 0\ \mathrm{s}$ represented target onset. We applied discrete wavelet transform with second-order Symlet wavelet function [30] to decompose the MEG signals from each channel and each trial to multiple resolution levels. Six wavelet coefficients corresponding to the selected time-frequency window were used to represent the features for each channel. Here, each time-frequency window is around 60 ms in length and covers the low frequency band ($\leq$7 Hz). Since 74 channels are considered in total, the dimensionality of the feature space is: $M = 6 \times 74 = 444$. Each wavelet coefficient (i.e., the feature) is correlated to the signal energy in a specific time-frequency window of a given channel. Taking the overt case of SubC as an example, Fig. 3(b) and (c) shows the spatial distribution of the selected wavelet coefficients for four different classes.

### C. Feature Clustering

Given the MEG features extracted in the previous subsection, we applied Algorithm 1 to partition these features into several groups. The optimal number of groups (i.e., $K_{\mathrm{OPT}}$) was automatically determined by evaluating the clustering quality in Algorithm 1. Note that $K_{\mathrm{OPT}}$ can be different for different data sets. Fig. 4 shows the adjacency matrices [i.e., $\mathbf{W}$ in (5)] for all ten data sets. In Fig. 4, the MEG features are ordered based

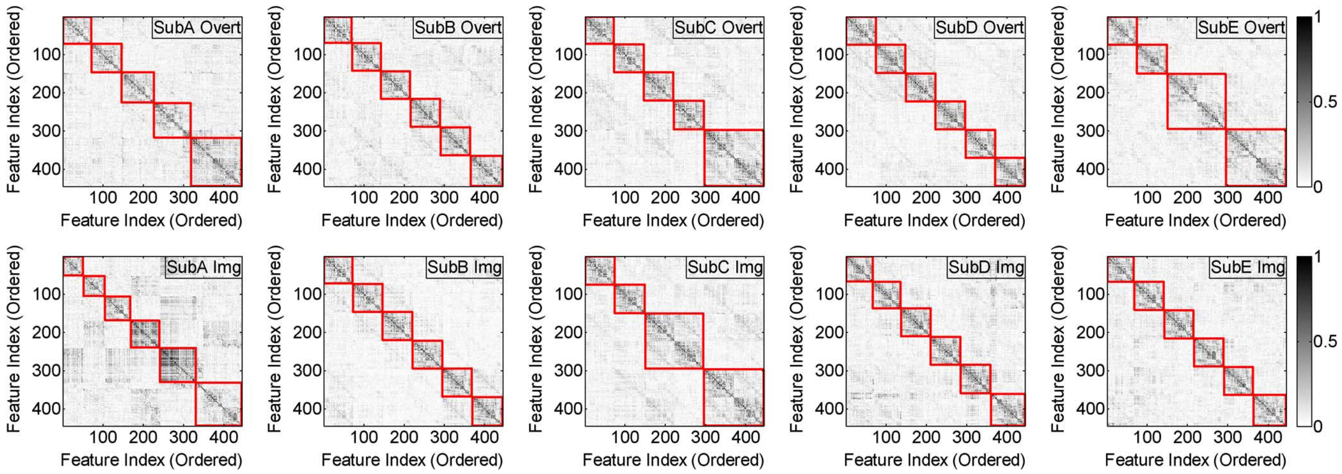

Fig. 4. The adjacency matrices are almost block diagonal for all ten data sets. The MEG features are ordered based on the feature clustering results where all features in the same cluster are grouped together.

on the feature clustering results where all features in the same cluster are grouped together.

Studying Fig. 4, we would have two important observations. First, once the features are appropriately ordered, the adjacency matrices are almost block diagonal. Remember that the adjacency matrix $\mathbf{W}$ in (5) contains the correlation coefficients for all features. Hence, a block diagonal $\mathbf{W}$ implies that different features in different groups are uncorrelated. In other words, a unique group structure of feature correlation exists for all data sets collected by our experiment. Second, the proposed feature clustering algorithm (i.e., Algorithm 1) successfully identified the appropriate feature groups for our data sets. Both the optimal number of groups and the appropriate group assignment for each feature were successfully found by Algorithm 1. Such a feature clustering scheme will eventually lead to superior decoding accuracy of the proposed CLDA algorithm, as will be discussed in detail in the next subsection.

### D. Movement Decoding

We implemented four different movement decoding algorithms for comparison purpose. For each decoding method, its accuracy was estimated by using leave-one-out cross-validation [29], where feature selection and/or feature clustering were repeatedly applied for each run within the cross-validation loop.

1) **FC-LDA**: Apply Fisher criterion [29] to select a set of important features for dimension reduction. Next, LDA is used for movement decoding. During the feature selection phase, the optimal number of required features is determined via an extra cross-validation step using the training data only.
2) **DLDA**: Assume mutual independence among all features and force the within-class scatter matrix $\mathbf{S}_W$ in (1) to be diagonal. Next, LDA is used for movement decoding based on the diagonal approximation of $\mathbf{S}_W$.
3) **RLDA**: An additional regularization term is added to the within-class scatter matrix: $\mathbf{S}_W + \lambda \cdot \mathbf{I}$. The regularization parameter $\lambda \geq 0$ is determined via an extra cross-validation step using the training data only. Next, LDA is used for movement decoding based on the regularized scatter matrix $\mathbf{S}_W + \lambda \cdot \mathbf{I}$.
4) **CLDA**: Algorithm 2 is applied for movement decoding.

First, we consider a simple two-class decoding problem where the movement direction is either left or right. Fig. 5(a) shows the accuracy of the aforementioned four decoding algorithms. Note that CLDA outperforms the other three methods for all data sets. The decoding accuracy of CLDA is above 90% for all overt cases and it is above 80% for all imagined cases. The average decoding accuracy of CLDA is 97.3% and 94.5% for overt and imagined cases, respectively.

Next, we consider the four-class decoding problem where the movement direction can be up, down, left or right. The accuracy of four-class movement decoding is shown in Fig. 5(b). Similar to the two-class case, CLDA offers the best decoding accuracy in all test cases. The average decoding accuracy of CLDA is 90.2% and 83.7% for overt and imagined cases, respectively.

As previously mentioned, both DLDA (with a diagonal within-class scatter matrix) and RLDA (using a diagonal covariance matrix to model the prior distribution for Bayesian inference) rely on the prior knowledge that all features are mutually independent. Unlike DLDA or RDLA, CLDA is able to automatically identify the critical correlation structure (i.e., by partitioning all features into different groups) and then accurately extract the correlation (i.e., by estimating a block diagonal within-class scatter matrix) of all features. It, in turn, results in improved decoding accuracy over DLDA and RLDA.

On the other hand, FC-LDA applies Fisher criterion to select a small subset of important features. Such a feature selection method can be pessimistic, when there are a large number of important features and the training data are limited. In other words, given a small set of training data, FC-LDA has to filter out many useful features in order to reduce the dimensionality of the feature space and avoid over-fitting. These useful features carry the information that is needed for movement decoding. Since they are simply ignored by FC-LDA, the resulting decoding error becomes large. This is the primary reason why FC-LDA is less accurate than CLDA in our experiment.

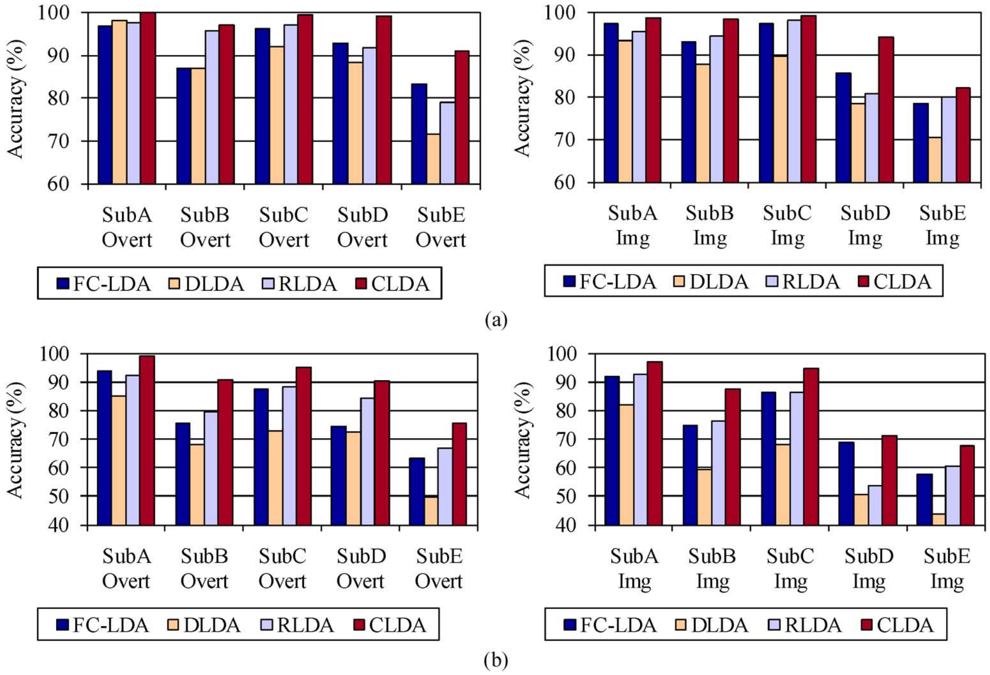

Fig. 5. Movement decoding accuracy estimated by using leave-one-out cross-validation: (a) decoding results of two movement directions (left and right), and (b) decoding results of four movement directions (up, down, left and right).

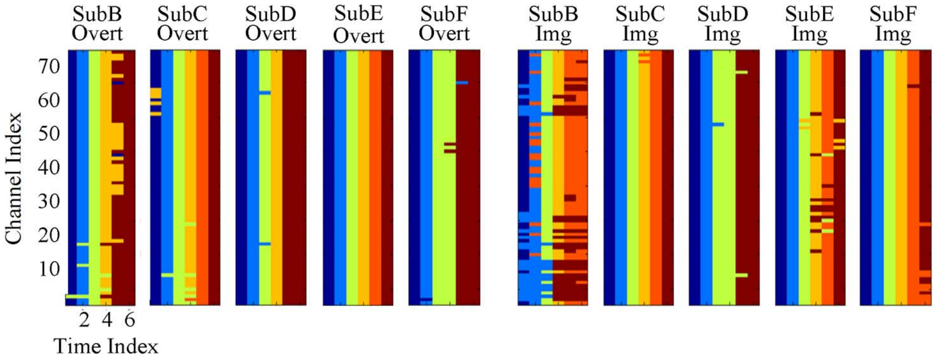

Fig. 6. Feature groups determined by the spectral clustering algorithm. The horizontal axes represent the indexes of wavelet coefficients. They can also be considered as the indexes of time windows, since the corresponding wavelet basis functions are associated with the same low-frequency band and have local support in different time windows. The vertical axes represent the indexes of 74 selected channels (i.e., the gradiometers on top of the sensorimotor area). Different colors indicate different feature groups.

## V. Discussions

In this section, we aim to explain why the correlation of our MEG features shows a unique group structure. Fig. 6 plots the group assignment of each feature that is determined by the proposed feature clustering algorithm (i.e., Algorithm 1). In our experiment, each feature (i.e., each wavelet coefficient) is associated with a particular time window of a particular channel, because the corresponding wavelet basis functions are in the same low-frequency band and have local support in different time windows.

Studying Fig. 6, we notice that most features of the same group are in the same time window or adjacent time windows, but from different channels. Remember that our proposed feature clustering algorithm attempts to maximize the within-group correlation and minimize the between-group correlation. It, in turn, implies that the MEG features from the same time window but different channels are strongly correlated, while the features from different time windows are weakly correlated. In other words, our MEG data sets present a strong spatial correlation but a weak temporal correlation.

Similar observations of strong spatial correlation for MEG data have been reported in many other applications [4]. From the physics point of view, MEG signals are generated by the primary current sources inside the brain. Based on Maxwell's equations, the magnetic field created by the same current source can propagate to multiple MEG sensors (i.e., the superconducting quantum interference devices) corresponding to different channels at different locations [4]. This is an important reason why a strong spatial correlation has been observed for MEG measurement data.

On the other hand, to explain the weak temporal correlation, we consider the following model for the time-domain signal $s_i(t)$ of the $i$th channel

$$s_i(t) = v_i(t) + n_t(t) \tag{16}$$

where $v_i(t)$ stands for the phase-locked evoked response associated with the stimulus and $n_i(t)$ denotes the ongoing activity. The linear model in (16) has been used in several previous studies to analyze the measured neural signals in time domain [24], [25]. The evoked response $v_i(t)$ in (16) can be estimated by averaging $s_i(t)$ over all trials corresponding to the same stimulus [25]. Once $v_i(t)$ is estimated and subtracted from $s_i(t)$, the resulting residue $n_i(t)$ does not contain event-related information and is often modeled as white noise [24], [25]. In other words, while $s_i(t)$ is a colored signal due to the evoked response $v_i(t), n_i(t)$ is white once $v_i(t)$ is removed from $s_i(t)$. It, hence, explains the weak temporal correlation that we observe, since the within-class scatter matrix $\mathbf{S}_W$ in (1) is mainly determined by $n_i(t)$.

## VI. Conclusion

In this paper, we proposed a new CLDA algorithm to decode movement directions from MEG signals recorded for human subjects. The proposed CLDA method can be conceptually viewed as a generalized extension of the traditional DLDA algorithm. It applies a spectral clustering algorithm to partition all BCI features into several groups where the within-group correlation is maximized and the between-group correlation is minimized. As such, the covariance matrix can be approximated as a block diagonal form that can be accurately estimated from a small set of training data. The efficiency of CLDA is studied by both theoretical analyses and practical examples. Our MEG-based movement decoding demonstrates that the average accuracy of CLDA is 87% for single-trial movement decoding of four directions (i.e., up, down, left, and right). Such high decoding accuracy implies that MEG can be used to provide accurate two-dimensional control for BCI and the proposed CLDA algorithm is a critical technique to make MEG-based BCI of practical utility. In addition, even though we focus on offline data analysis in this paper, the proposed CLDA algorithm can be extended to online BCI systems. As an important aspect of our future research, we will further study the feature correlation for other BCI applications and apply the proposed CLDA technique to those cases.

## Appendix
## Block Diagonal Approximation

In this section, we will theoretically analyze the quality of the proposed block diagonal approximation. Towards this goal, we consider a simple two-class movement decoding problem with the following assumptions.

1) The prior probability for each class (i.e., the probability for each class to occur) is identical.
2) The probability distributions of MEG features for both classes are multivariate Gaussian. These distributions have different mean values (denoted as $\mathbf{m}_1$ and $\mathbf{m}_2$ respectively), but share the same covariance matrix $\mathbf{\Sigma}$.

These two assumptions are not necessarily valid for all movement decoding problems; however, they define a simple classification problem for which we can show many insights on the proposed block diagonal approximation.

Given the aforementioned problem, it is straightforward to verify that the minimal decoding error (i.e., the probability of misclassification) of an optimal classifier is [26]

$$e_{\mathrm{OPT}} = 1 - \varphi(\rho_{\mathrm{OPT}}) \tag{17}$$

where $\varphi(\bullet)$ denotes the cumulative distribution function of standard Gaussian distribution (i.e., zero mean and unit variance), and $\rho_{\mathrm{OPT}}$ is defined as

$$\rho_{\mathrm{OPT}} = \sqrt{(\mathbf{m}_2 - \mathbf{m}_1)^T \cdot \mathbf{\Sigma}^{-1} \cdot (\mathbf{m}_2 - \mathbf{m}_1)}/2. \tag{18}$$

Studying (17), (18), we would have two important observations. First, if the difference of $\mathbf{m}_1$ and $\mathbf{m}_2$ increases, $\rho_{\mathrm{OPT}}$ in (18) increases and, hence the decoding error in (17) decreases. Second, the decoding error also decreases, if the variance of the MEG features (measured by the covariance matrix $\mathbf{\Sigma}$) decreases. These two observations are consistent with our intuition. Namely, the decoding error is small, if the two classes are substantially different (measured by the difference of $\mathbf{m}_1$ and $\mathbf{m}_2$) or the trial-to-trial variation is small (measured by the covariance matrix $\mathbf{\Sigma}$).

On the other hand, if the covariance matrix $\mathbf{\Sigma}$ is approximated by its block diagonal form $\mathbf{\Sigma}_B$, the decoding error of the proposed CLDA algorithm becomes

$$e_B = 1 - \varphi(\rho_B) \tag{19}$$

where

$$\rho_B = \frac{1}{2} \cdot \frac{(\mathbf{m}_2 - \mathbf{m}_1)^T \cdot \mathbf{\Sigma}_B^{-1} \cdot (\mathbf{m}_2 - \mathbf{m}_1)}{\sqrt{(\mathbf{m}_2 - \mathbf{m}_1)^T \cdot \mathbf{\Sigma}_B^{-1} \cdot \mathbf{\Sigma} \cdot \mathbf{\Sigma}_B^{-1} \cdot (\mathbf{m}_2 - \mathbf{m}_1)}}. \tag{20}$$

Equation (19), (20) can be derived by directly following the results in [26]. To study the difference between $e_{\mathrm{OPT}}$ in (17) and $e_B$ in (19), we further define

$$r = \frac{\rho_B}{\rho_{\mathrm{OPT}}} = \frac{\boldsymbol{\delta}_0^T \boldsymbol{\delta}_0}{\sqrt{\left(\boldsymbol{\delta}_0^T \mathbf{\Sigma}_0 \boldsymbol{\delta}_0\right) \cdot \left(\boldsymbol{\delta}_0^T \mathbf{\Sigma}_0^{-1} \boldsymbol{\delta}_0\right)}} \tag{21}$$

where

$$\boldsymbol{\delta}_0 = \mathbf{\Sigma}_B^{-1/2} \cdot (\mathbf{m}_2 - \mathbf{m}_1) \tag{22}$$

$$\mathbf{\Sigma}_0 = \mathbf{\Sigma}_B^{-1/2} \cdot \mathbf{\Sigma} \cdot \mathbf{\Sigma}_B^{-1/2}. \tag{23}$$

The value of $r$ in (21) indicates the decoding accuracy of CLDA, as compared to an optimal classifier. If $r$ is close to 1, the accuracy of CLDA is close to the maximal accuracy that can be achieved by the optimal classifier.

Based on the Kantorovich inequality [26], [28], we can obtain a lower bound of $r$

$$r \geq 2 \cdot \sqrt{\xi_0}/(1 + \xi_0) \tag{24}$$

where $\xi_0$ represents the condition number of the matrix $\mathbf{\Sigma}_0$ in (23) based on $\mathrm{L}_2$ norm. Combining (19), (21) and (24), we can derive the upper bound of the decoding error for CLDA

$$e_{\mathrm{OPT}} \leq e_B \leq e_{UP} = 1 - \varphi\left(\frac{2 \cdot \sqrt{\xi_0} \cdot \rho_{\mathrm{OPT}}}{1 + \xi_0}\right). \tag{25}$$

Note that the accuracy of CLDA strongly depends on $\xi_0$. If $\xi_0$ is sufficiently small, $e_{UP}$ is close to $e_{\mathrm{OPT}}$. Namely, the accuracy of CLDA is close to that of an optimal classifier. In the extreme case, if the covariance matrix $\mathbf{\Sigma}$ is exactly block diagonal, we have $\mathbf{\Sigma}_B = \mathbf{\Sigma}$, $\mathbf{\Sigma}_0 = \mathbf{I}$ and $\xi_0 = 1$. Therefore, CLDA is equivalent to the optimal classifier and both of them yield the same decoding accuracy. This result provides theoretical guidelines to quantitatively assess the quality of the proposed block diagonal approximation for CLDA.

## REFERENCES

[1] J. Vidal, "Toward direct brain-computer communication," *Annu. Rev. Biophys. Bioeng.*, vol. 2, pp. 157–180, 1973.

[2] J. Wolpaw, N. Birbaumer, W. Heetderks, D. McFarland, P. Peckham, G. Schalk, E. Donchin, L. Quatrano, C. Robinson, and T. Vaughan, "Brain-computer interface technology: A review of the first international meeting," *IEEE Trans. Rehab. Eng.*, vol. 8, no. 2, pp. 164–173, Jun. 2000.

[3] G. Pfurtscheller, C. Neuper, C. Guger, W. Harkam, H. Ramoser, A. Schlogl, B. Obermaier, and M. Pregenzer, "Current trends in Graz brain-computer interface (BCI) research," *IEEE Trans. Rehab. Eng.*, vol. 8, no. 2, pp. 216–219, Jun. 2000.

[4] M. Hamalainen, R. Hari, R. Ilmoniemi, J. Knuutila, and O. Lounasmaa, "Magnetoencephalography—Theory, instrumentation, and applications to noninvasive studies of the working human brain," *Rev. Modern Phys.*, vol. 65, no. 2, pp. 413–497, 1993.

[5] B. He and Z. Liu, "Multimodal functional neuroimaging: Integrating functional MRI and EEG/MEG," *IEEE Rev. Biomed. Eng.*, vol. 1, pp. 23–40, 2008.

[6] T. Lal, M. Schroder, N. Hill, H. Preissl, T. Hinterberger, J. Mellinger, M. Bogdan, W. Rosenstiel, T. Hofmann, N. Birbaumer, and B. Scholkopf, "A brain computer interface with online feedback based on magnetoencephalography," in *Proc. Int. Conf. Mach. Learn.*, 2005, pp. 465–472.

[7] J. Mellinger, G. Schalk, C. Braun, H. Preissl, W. Rosenstiel, N. Birbaumer, and A. Kuebler, "An MEG-based brain-computer interface (BCI)," *NeuroImage*, vol. 36, pp. 581–593, Jul. 2007.

[8] S. Waldert, H. Preissl, E. Demandt, B. Christoph, B. Niels, A. Aertsen, and C. Mehring, "Hand movement direction decoded from MEG and EEG," *J. Neurosci.*, vol. 28, no. 4, pp. 1000–1008, Jan. 2008.

[9] W. Wang, G. Sudre, R. Kass, J. Collinger, A. Degenhart, A. Bagic, and D. Weber, "Decoding and cortical source localization for intended movement direction with MEG," *J. Neurophysiol.*, vol. 104, pp. 2451–2461, Aug. 2010.

[10] P. Shenoy, K. Miller, J. Ojemann, and R. Rao, "Generalized features for electrocorticographic BCIs," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 1, pp. 273–280, Jan. 2008.

[11] J. Kronegg, G. Chanel, and S. Voloshynovskiy, "EEG-based synchronized brain-computer interfaces: A model for optimizing the number of mental tasks," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 15, no. 1, pp. 50–58, Mar. 2007.

[12] T. Lal, M. Schroder, T. Hinterberger, J. Weston, M. Bogdan, N. Birbaumer, and B. Scholkopf, "Support vector channel selection in BCI," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 1003–1010, Jun. 2004.

[13] H. Ramoser, J. Mueller-Gerking, and G. Pfurtscheller, "Optimal spatial filtering of single trial EEG during imagined hand movement," *IEEE Trans. Rehabil. Eng.*, vol. 8, no. 4, pp. 441–446, Dec. 2000.

[14] G. Blancharda and B. Blankertz, "BCI competition 2003—Data set IIa: Spatial patterns of self-controlled brain rhythm modulations," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 1062–1066, Jun. 2004.

[15] M. Grosse-Wentrup and M. Buss, "Multiclass common spatial patterns and information theoretic feature extraction," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 8, pp. 1991–2000, Aug. 2008.

[16] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, pp. 849–856.

[17] F. Bach and M. Jordan, "Learning spectral clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, pp. 305–312.

[18] M. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E*, vol. 69, no. 026113, 2004.

[19] U. Luxburg, "A tutorial on spectral clustering," *Stat. Comput.*, vol. 17, no. 4, pp. 395–416, 2007.

[20] S. Taulu, J. Simola, and M. Kajola, "Applications of the signal space separation method," *IEEE Trans. Signal Process.*, vol. 53, no. 9, pp. 3359–3372, Sep. 2005.

[21] I. Clark, R. Biscay, M. Echeverria, and T. Virues, "Multiresolution decomposition of non-stationary EEG signals: A preliminary study," *Comput. Biol. Med.*, vol. 25, no. 4, pp. 373–382, Jul. 1995.

[22] B. Graimann, J. Huggins, S. Levine, and G. Pfurtscheller, "Toward a direct brain interface based on human subdural recordings and wavelet-packet analysis," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 954–962, Jun. 2004.

[23] E. Leuthardt, G. Schalk, J. Wolpaw, J. Ojemann, and D. Moran, "A brain-computer interface using electrocorticographic signals in humans," *J. Neural Eng.*, vol. 1, no. 2, pp. 63–71, 2004.

[24] P. Karjalainen, J. Kaipio, A. Koistinen, and M. Vauhkonen, "Subspace regularization method for the single-trial estimation of evoked potentials," *IEEE Trans. Biomed. Eng.*, vol. 46, no. 7, pp. 849–860, Jul. 1999.

[25] W. Truccolo, M. Ding, K. Knuth, R. Nakamura, and S. Bressler, "Trial-to-trial variability of cortical evoked responses: Implications for the analysis of functional connectivity," *Clin. Neurophysiol.*, vol. 113, pp. 206–226, 2002.

[26] P. Bickel and E. Levina, "Some theory for fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations," *Bernoulli*, vol. 10, no. 6, pp. 989–1010, 2004.

[27] F. Chung, *Spectral Graph Theory*. Providence, RI: Am. Math. Soc., 1997.

[28] D. Luenberger, *Linear and Nonlinear Programming*. Reading, MA: Addison-Wesley, 1984.

[29] C. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.

[30] D. Percival and A. Walden, *Wavelet Methods for Time Series Anaysis*. Cambridge, U.K.: Cambridge Univ. Press, 2006.

**Jinyin Zhang** (S'11) received the B.S. and M.S. degrees in computer science from Beijing University of Astronautics and Aeronautics, Beijing, China, in 2002 and 2005, respectively. She is currently working toward the Ph.D. degree in electrical and computer engineering at Carnegie Mellon University, Pittsburgh, PA, advised by Dr. X. Li. Her research interests include signal processing, machine learning and optimization with specific interests in their application to neural signal processing.

**Gustavo Sudre** (S'06) received the B.S. degree in computer science at the University of Kansas in 2006, where he performed research on artificial intelligence, data mining, and knowledge representation. He received the M.S. degree in bioengineering at the University of Pittsburgh, Pittsburgh, PA, in 2008, where he was advised by Dr. D. Weber and used MEG to investigate the influence of somatosensory input in motor cortex activity. He is currently working toward the Ph.D. degree in neural computation at Carnegie Mellon University, Pittsburgh, PA, advised by Dr. T. Michell, and combines brain imaging data with machine learning techniques to investigate knowledge representation in the brain.

**Xin Li** (M'01–SM'06) received the B.S. and M.S. degrees in electronics engineering from Fudan University, Shanghai, China, in 1998 and 2001, respectively, and the Ph.D. degree in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA, in 2005,

He is currently an Assistant Professor in the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA. In 2005, he co-founded Xigmix Inc. to commercialize his Ph.D. research, and served as the Chief Technical Officer until the company was acquired in 2007. Since 2009, he has been appointed as the Assistant Director for FCRP Focus Research Center for Circuit & System Solutions (C2S2). His research interests include computer-aided design, neural signal processing, and power system analysis and design.

**Wei Wang** received the M.D. degree from Peking University Health Science Center, Beijing, China, in 1999, the M.Sc. degree in biomedical engineering from University of Tennessee Health Science Center, Memphis, TN, in 2002, and the Ph.D. degree in biomedical engineering from Washington University, St. Louis, MO, in 2006.

Prior to joining the University of Pittsburgh in 2007, he served as a Senior Scientist at St. Jude Medical, Inc., Sylmar, CA. He is currently an Assistant Professor in the Department of Physical Medicine and Rehabilitation with a secondary appointment in the Department of Bioengineering and the Clinical and Translational Sciences Institute at the University of Pittsburgh, Pittsburgh, PA. His current research interests include neural engineering, motor neuroprosthetics, brain–computer interface, ECoG, MEG, rehabilitation of movement disorders, and motor system neurophysiology.

**Douglas J. Weber** (M'94) received the B.S. degree in biomedical engineering from the Milwaukee School of Engineering in Milwaukee, WI, in 1994 and the M.S. and Ph.D. degrees in bioengineering from Arizona State University, Tempe, AZ, in 2000 and 2001, respectively.

He is currently an Assistant Professor in the Department of Physical Medicine and Rehabilitation, University of Pittsburgh, Pittsburgh, PA. He is also a faculty member in the Department of Bioengineering and the Center for the Neural Basis of Cognition. Previously, he was a Postdoctoral Fellow and then an Assistant Professor in the Centre for Neuroscience at the University of Alberta, Edmonton, AB, Canada. His primary research area is neural engineering. Specific research interests include functional electrical stimulation, activity-based neuromotor rehabilitation, neural coding, and neural control of prosthetic devices.

**Anto Bagic** is a Neurologist at Georgetown University, Washington, DC, sub-specialized clinical neurophysiology, epilepsy, sleep disorders, and magnetoencephalography, National Institutes of Health (NIH), Bethesda, MD. His clinical practice is focused on epilepsy and similar paroxysmal disorders. He is an Associate Professor of Neurology, Chief of Epilepsy Division, Director of the University of Pittsburgh Comprehensive Epilepsy Center (UPCEC), Founding Director of the 1st Center for Advanced Brain Magnetic Source Imaging (CABMSI), Pittsburgh, PA, Director of the UPMC MEG Epilepsy Program and Chief Scientific Advisor for MEG Research at University of Pittsburgh, Pittsburgh, PA. His research interests include applications of MEG in studying fundamental neural processes as they pertain to neurological disorders.