EEG Subspace Representations and Feature Selection for Brain-Computer Interfaces

Charles W. Anderson
Department of Computer Science
Colorado State University
Fort Collins, CO 80523

Michael J. Kirby
Department of Mathematics
Colorado State University
Fort Collins, CO 80523

Abstract

Electroencephalogram (EEG) signals recorded from a persons scalp have been used to control binary cursor movements. Multiple choice paradigms will require more sophisticated protocols involving multiple mental tasks and signal representations that capture discriminatory characteristics of the EEG signals. In this study, six-channel EEG is recorded from a subject performing two mental tasks. The signals are transformed via the Karhunen-Loéve or maximum noise fraction transformations and classified by quadratic discriminant analysis. In addition, classification accuracy is tested for all subsets of the six EEG channels. Best results are approximately 90% correct when training and testing data are recorded on the same day and 75% correct when training and testing data are recorded on different days.

1. Introduction

In 1929 Berger showed that electrical signals can be recorded externally from the scalp of human subjects [8]. Many characteristics of these electroencephalogram (EEG) signals have been studied and related to brain abnormalities and mental activity, but it is still relatively unknown just how much information about the underlying mental activity can be gleaned from these signals.

If we can tap this source of information, a new means of communication becomes available for people who have difficulty or are unable to communicate via any other means. For example, people with amyotrophic lateral sclerosis (ALS) experience loss of much of their voluntary muscle control, while their autonomic and intellectual functions remain normal. This can result in a locked-in syndrome in which a person is unable to communicate to the outside world, as described by [7] in a poignant, personal account. Once the technology is developed for extracting the needed information from EEG signals, paralyzed persons will have a new alphabet with which they can compose commands to devices like a wheel chair.

In this paper, we study six-channel EEG signals recorded from one subject performing two mental tasks, a mental multiplication task and an imaginary letter writing task. Signals are represented either as a Karhunen-Loève (KL) transform or as a maximum noise fraction (MNF) transform of windowed EEG data. The transformed data is then classified by quadratic discriminant analysis. Additional variations include subsets of electrodes, number of modes, or eigenvectors, included, and the number of time-delayed values included in each sample. Also, training and testing data is taken either from the same day or from two different days. These experiments are extensions of work described by Kirby and Anderson [19] who compared EEG representations based on KL, MNF, and canonical correlation analysis.

The next section is a brief survey of the BCI field. Then new experiments and results are described in the following two sections, followed by a section summarizing the conclusions drawn from the experiments.

2. Survey of BCI Field

This section briefly surveys the history and recent developments in the BCI field. Wolpaw, et al., [22] have written a detailed review of recent approaches to BCI.

The foundation of current BCI work was formed in the 1970's and 80's. Galen and Ornstein [14] studied the asymmetry of the EEG in the two hemispheres and were able to distinguish two cognitive modes—the verbal (left hemisphere) and spatial (right hemisphere). A frequency analysis of hemispheric EEG asymmetries was conducted by Doyle, et al., [12] who reported that right-to-left ratios of energy in the alpha band were higher in verbal and arithmetic tasks than in spatial tasks. Kraft, et al., [20] studied hemispheric asymmetries in EEG recorded from children and found that Piagetian conservation and reading tasks shifted from greater right-hemispheric processing during encoding of information to greater left-hemispheric processing during retrieval and verbal/logic expression of the information. Using a collection of the mental tasks used in these early studies, Keirn and Aunon [18] showed that it is possible to distinguish between pairs of mental tasks using only the EEG to a high degree of accuracy.

One of the first demonstrations of a BCI system was by Farwell and Donchin [13] who based their system on the P300 component of event-related brain potentials, or ERPs. The P300 component is a positive change in the measured EEG signal that occurs about 300 milliseconds following the presentation of a type of stimulus that is anticipated, but occurs relatively rarely. It is typically detected in the "oddball paradigm" that, for Farwell and Donchin, consisted of sequentially highlighting the rows or columns of a table containing the alphabet. When the row or column containing the desired letter is highlighted, which for this setup is the odd or rare stimulus, the P300 component is detected. Recently, Donchin, et al., [11] duplicated the oddball approach to test the speed with which characters can be selected with a P300-based BCI. Using EEG recorded from five electrodes, they found that perfect letter-choice detection was possible by averaging EEG over 40 trials, but this would require 60 seconds to communicate each character. 7.8 letters/minute could be obtained with a 80% level of accuracy and 4.8 letters/minute could be obtained with a 90% level of accuracy. These results were for off-line analysis of the data.

Birbaumer, et al., [9, 10] have shown that subjects can learn to control slow cortical potentials (SCPs) that last from one-half second to several minutes. After several months of practice, the subjects are presented with a typing task in which a letter is selected by first choosing the first or second half of the alphabet, then the correct half of the remaining letters, and so on until the single letter is selected. The median number of trials for the selection of a letter is 28, for a rate of one letter every 2 minutes.

Babiloni, et al., [6] investigate a BCI for recognizing imagined right or left hand, middle finger movements. Two frequency ranges were extracted from EEG, one covering the alpha and beta bands and another broader range from 8 to 30 Hz. The best classification performance resulted from the surface Laplacian transformation and the wider frequency range. Averaging over five subjects and the two movements, the best classification accuracy was about 87%.

Studies performed by Wolpaw, et al., [23] provide details on the mu and beta rhythms that are often used in BCIs based on a variety of imagined movement tasks. They find that the most discriminatory frequencies and electrode locations change from day to day. In their approach, the classification method continues to adapt as subjects perform multiple trials to track changes in how subjects perform the movement imagery. Some subjects report that they use the imagery less and less with further training.

In summary, the state-of-the-art of BCI is that EEG patterns correlated with limb movement or imagined movement can be reliably classified given a long enough time window of data. A few studies note that with practice subjects report they no longer use the initial imagined movement to trigger the BCI system but that they simply think about moving the cursor. These are fascinating results, but the approaches followed are limited in that binary decisions are being produced.

A truly practical BCI must provide more than two commands to the user and must perform the classification faster than is possible with current techniques. This goal has driven our work with multiple mental tasks [4, 2, 5, 3]. We have shown that for some subjects, the correct mental task out of five possible tasks can be predicted with approximately a 72% accuracy, using artificial neural networks as classifiers and autoregressive models as representations of the EEG signals. The autoregressive model seems to capture information in EEG that is discriminatory between tasks; there is little significant difference in classification accuracy between neural networks, support vector machines, and linear discriminant analysis [15].

3. Methodology

3.1. EEG Data

For the experiments reported here, data from one subject performing two tasks was used. The data set consisted of four 10-second trials, two for each task, recorded on one day and another four trials, two for each task, recorded on the second day. The tasks were a multiplication task and a letter-writing task. For the multiplication task, the subject is given a non-trivial multiplication problem to solve, such as multiply 49 times 78. For the letter-writing task, the subject is instructed to mentally compose a letter to a friend or relative without vocalizing. For both tasks, the subjects are instructed to not make any overt movements.

EEG was recorded from six channels (C_3, C_4, P_3, P_4, O_1, O_2, referenced to left and right mastoid) at 250 Hz with 12 bits of accuracy and bandpass filtered at 0.1–100 Hz. Each 10-second trial results in six simultaneous time series, one for each of the six electrodes. Concatenating each series as the rows of a vector, each trial produces a 6x2500 matrix of voltage measurements.

3.2. EEG Data Representations

Two representations of the EEG data are studied here: a representation based on the Karhunen-Loéve transform, and a representation based on the Maximum Noise Fraction transform, referred to here as the KL and MNF representations, respectively. Variations are generated by augmenting each sample with 0, 1, or 2 time-delayed, or lagged, samples and by selecting subsets of electrodes for feature set evaluation.

Additional variations are generated by selecting one of the resulting basis vectors, referred to as modes.

The KL and MNF representations of the data matrix D for a particular trial are produced by the following procedure. Each data sample is augmented by extending each sample with $l-1$ lagged samples, so each sample represented by l values. The augmented data is segmented into 1/4-second windows of 64 time samples that overlap by 1/8 second, or 32 samples. Let $G_{k,l}$ be the resulting $64 \times 6l$ matrix of 64 consecutive samples (from sample k to $k+63$) of 6-channel data with each sample augmented by $l-1$ lags:

$$G_{k,l} = \Big([D]_{k:k+63,1:6} \mid [D]_{k+1:k+63+1,1:6} \mid \\ \ldots \mid [D]_{k+l-1:k+63+l-1,1:6} \Big)$$

The notation $[M]_{a:b,c:d}$ describes selecting rows $a,\ldots,b$ and columns $c,\ldots,d$ of matrix M. Below, a non-consecutive set of indices is used. For example, if $F = \{1,3\}$, then $[M]_{a:b,F}$ selects rows $a,\ldots,b$ and columns 1 and 3.

Let F be a subset of electrodes, i.e., F is an element of the power set of electrode indices, or $F \in \mathcal{P}\{1,\ldots,6\} - \emptyset$. This dependency on F is designated by an additional subscript on $G_{k,l,F}$, where

$$G_{k,l,F} = \Big([D]_{k:k+63,F} \mid [D]_{k+1:k+63+1,F} \mid \\ \ldots \mid [D]_{k+l-1:k+63+l-1,F} \Big)$$

The KL transform of the data matrix $G_{k,l,F}$ is given by the singular value decomposition $G_{k,l,F} = U_{k,l,F}\Sigma_{k,l,F}V_{k,l,F}^T$. $V_{k,l,F}$ is the $|F|l \times |F|l$ matrix of right-singular vectors, where $|F|$ denotes the size of set F. The basis defined by $V_{k,l,F}$ is equivalent to the basis composed of $V_{k,l,F}$ with any of its columns negated. To maintain consistency across multiple data matrices, a column is negated if its first component is negative.

The KL representation of a trial is defined by selecting a singular vector, or mode, from the columns of each $V_{k,l,F}$. Using the first singular vectors results in a KL representation consisting of the direction of maximum variance for each data window. The maximum variance directions may not result in the best classification, so representations consisting of each mode are tested.

The KL representation, $K_{l,m,F}$, of D using l lags, mode m and features F is given by

$$K_{l,m,F} = \Big([V_{1,l}]_{1:|F|l,m} \mid [V_{32,l}]_{1:|F|l,m} \mid \\ \cdots \mid [V_{2500-63,l}]_{1:|F|l,m} \Big)$$

for which each column now contains the KL transformed representation of a 1/4-second window of data.

The KL procedure produces an optimal set of data-dependent eigenvectors as the columns of V. Given a data matrix G, eigenvector i may be expressed

$$[V]_{*,i} = G\phi_i$$

where $*$ denotes all rows and the coefficient vectors ϕ_i reflect the weighting of the data required to make an optimal set of eigenvectors as defined by the KL transform. As a result, if the data itself consists of signal plus noise, then so will the optimal eigenvectors.

Anderle and Kirby [1] describe the maximum noise fraction transform developed to maximally filter out the noise component. Assume that a data matrix G is composed of signal P and an additive noise component Q, so $G = P+Q$. The subspace that best represents the data matrix P may be found by expressing the data dependence of an eigenvector as

$$\phi = G\psi \tag{1}$$
$$= P\psi + Q\psi. \tag{2}$$

In other words, each basis vector may be decomposed as

$$\phi = \phi_P + \phi_Q$$

where the component $\phi_P = P\psi$ represents the P data dependency and $\phi_Q = Q\psi$ represents the Q data dependency.

Now the *Q–fraction* of a basis vector ϕ is defined as

$$D(\phi) = \frac{\phi_Q^T \phi_Q}{\phi^T \phi}.$$

Maximizing the function $D(\phi)$ maximizes the representation of Q by the basis vector ϕ. This may now be rewritten as

$$D(\phi(\psi)) = \frac{\psi^T Q^T Q \psi}{\psi^T G^T G \psi}.$$

The maximization problem leads to a *symmetric definite generalized eigenproblem*

$$\alpha G^T G \psi = \beta Q^T Q \psi. \tag{3}$$

Thus, given a data matrix G and the solution matrix $\Psi = [\psi^{(1)}|\ldots|\psi^{(m)}]$ of the generalized singular vector problem, given in Equation (3), the orthonormal basis for $\mathbb{R}^m$ is given by $\Phi = [\phi^{(1)}|\ldots|\phi^{(m)}]$ where $\Phi = G\Psi$. The basis vectors are ordered by increasing noise fraction, so a truncation of the basis corresponds to noise filtering. If we include the complete basis consisting of m vectors, we may express the data without loss as $G = \Phi\Phi^T G$. Truncating columns of Φ filters the noise, i.e., the data may be decomposed as

$$\hat{G} = \Phi_{*,1:d} B_d,$$

where the smaller matrix

$$B_d = \Phi^T_{*,1:d}\hat{G}$$

consists of reconstruction coefficients and $d < m$.

Note that to compute the Ψ matrix (and ultimately the maximum signal basis vectors) it is necessary to estimate Q^TQ.[1] As suggested by [21], the covariance matrix of the noise may be estimated by shifting a time series and computing differences. For further details concerning the estimation of the covariance matrix see [16, 1].

The resulting MNF representation, $N_{l,m,F}$, for the time-lagged, window of data, $G_{k,l,F}$ becomes

$$N_{l,m,F} = \Big([\Phi_1]_{1:|F|(l+1),m} \mid [\Phi_{32}]_{1:|F|(l+1),m} \mid \cdots \mid [\Phi_{2500-63}]_{1:|F|(l+1),m} \Big)$$

for which each column now contains the MNF transform representation, Φ_k, of a 1/4-second window of data starting with sample k.

3.3. Classification

Quadratic discriminant analysis (QDA) [17] was used to classify the transformed EEG signals using the following procedure. Three ways of partitioning the data into training and testing sets were used. Recall that data from two trials of each task was recorded during two sessions on different days. One way of partitioning the data is to ignore the second day's data and just train and test the classifier on data from the same day. A second way of partitioning the data is to train the classifier on data from one day and test its classification accuracy on data from the second day. This would show the degree to which the same features are discriminatory across days. The third way is to train the classifier on data from both days, and test the resulting classifier on data from different trials from both days. Results using each partitioning method are shown in the next section.

We followed a validation procedure to determine the best mode and feature set for each number of lags. Every third sample in the training set was removed from the training set and added to a validation set of data. The reduced training set was used with QDA to calculate the classification rate of the validation data. The mode and feature set resulting in the maximum classification rate for each lag value was found. The training data represented with this mode and lag value was then used to classify the testing data. This classification rate on the testing data is the measure of performance used in the results reported in the next section.

4. Results

Detailed results are reported for the third way of partitioning the data—training and testing on data from both days. Figure 1 shows the results of variations in the number of electrodes, the number of lags, and the mode. The first graph shows the test classification accuracy for different numbers of electrodes. The three solid lines are for the MNF representation and for 1, 2, or 3 lagged samples. The dotted lines are for the KL representation. Each point is the mean of the top tenth of the runs for a particular number of electrodes and lagged values. The mean is over different combinations of electrodes and representation mode. The top tenth of the runs were identified by their validation set errors and the corresponding test set errors are averaged. This graph shows that the MNF representation results in more accurate classification rates than does the KL representation, and that there is little consistent effect on classification rate of the number of electrodes. The best classification rate is about 85%.

The second graph shows that for the MNF representation, the number of lagged values has no effect, but for the KL representation, two and three lags is better than one. The third graph shows that, for the MNF representation, the best number of features, or electrodes, decreases slightly as the number of lags increases, but for all lags it is between three and four electrodes. For the KL representation, as the number of lags increases from one to two, the best number of electrodes increases. This graph suggests that the MNF representation better captures relevant information in multi-lagged samples than does the KL representation. The

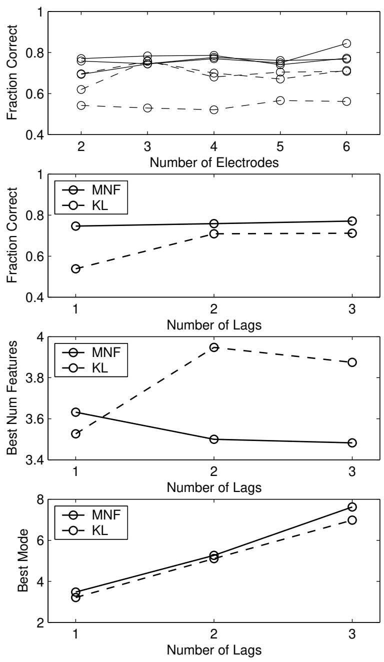

Figure 1: Results of training with data from both days and testing with data from both days. All results are the mean of the top tenth of the corresponding runs.

[1] Alternatively, one could choose to estimate the matrix P^TP.

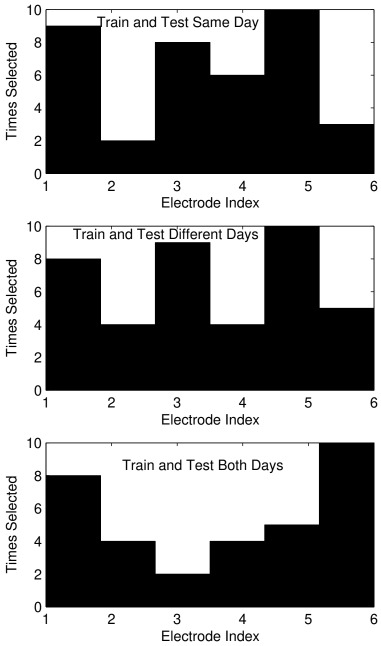

Figure 2: Counts of number of times each electrode appeared in the top tenth of the results.

fourth graph shows that the first mode is not the best mode to use for classification, and as the number of lags increases, higher modes are best.

The above analysis and similar analyses for the other two ways of partitioning the data results in the following classification accuracies. When training and testing with data from one day, the accuracy is approximately 90% correct. When training with data from one day and testing with data from the second day, the accuracy is reduced to about 75% correct. When training and testing with data from both days, the accuracy is about 85% as reported above.

To better understand which electrodes are most useful in classifying this data, the top tenth of the MNF representation results were collected and the number of times each electrode appeared in the feature subset was counted. The counts are shown in Figure 2. It appears that electrodes 1, 3, and 5, or C_3, P_3, O_1, are most useful when the training data consists of data only from the first day. These electrodes are all on the left side of the scalp. When training and testing with data on both days, the pattern of electrodes changes significantly, with electrodes 1 and 6, or C_3 and O_2, most often selected.

5. Conclusions

Experiments showed, for this data set of EEG signals recorded during performance of two mental tasks, a maximum noise fraction (MNF) transform resulted in a representation that is more accurately classified with quadratic discriminant analysis than did the Karhunen-Loève (KL) transform. Training and testing with data on the same day resulted in about 90% accuracy, while training with data from one day and testing with data from another resulted in about 75% accuracy. An exhaustive search of subsets of electrodes showed that combinations of less than four of the six electrodes produced the best classification accuracies. Results also show that the best feature subset varied significantly when data recorded on multiple days versus a single day was used to train the classifier.

The significance of these findings is that training with data from one day may result in discriminations based on information present only during that recording session, due to aspects unique to that recording session such as environmental noise, electrode cap placement, and state of the subject. Such effects will result in poor generalization performance. Information found to discriminate data on multiple days is more likely to correspond to signal characteristics related to the underlying cognitive processes, therefore to result in better generalization performance.

Note that the the MNF and KL representations are used to represent 64-sample windows. When using all six electrodes and 2 lagged samples, a single window is represented by MNF or KL modes that contain only 18 components. This is a vast reduction in dimensionality compared to the $64 \times 2 \times 6 = 768$ components that make up the initial data window. Using a subset of the electrodes results in better classification and an even greater reduction in dimensionality.

Acknowledgments

This work was funded by Grant No. 0208958 from the National Science Foundation.

References

[1] M. Anderle and M. Kirby. An application of the maximum noise fraction method to filtering noisy time-series. *Mathematics in Signal Processing V, Oxford University Press*, 2002.

[2] C. W. Anderson. Effects of variations in neural network topology and output averaging on the discrimination of mental tasks from spontaneous electroencephalogram. *Journal of Intelligent Systems*, 7(1-2):165–190, 1997.

[3] C. W. Anderson and D. A. Peterson. Recent advances in EEG signal analysis and classification. In R. Dybowski and V. Gant, editors, *Clinical Applications of*

Artificial Neural Network, chapter 8, pages 175–191. Cambridge University Press, UK, 2001.

[4] C. W. Anderson and Z. Sijerčić. Classification of EEG signals from four subjects during five mental tasks. In A. B. Bulsari, S. Kallio, and D. Tsaptsinos, editors, *Solving Engineering Problems with Neural Networks: Proceedings of the Conference on Engineering Applications in Neural Networks (EANN'96)*, pages 407–414. Systeemitekniikan seura ry, printed by Åbo Akademis Tryckeri, Turkey, 1996.

[5] C. W. Anderson, E. A. Stolz, and S. Shamsunder. Multivariate autoregressive models for classification of spontaneous electroencephalogram during mental tasks. *IEEE Transactions on Biomedical Engineering*, 45(3):277–286, 1998.

[6] F. Babiloni, F. Cincotti, L. Lazzarini, J. Millán, J. Mouri no, M. Varsta, J. Heikkonen, L. Bianchi, and M. G. Marciani. EEG recognition of imagined hand and foot movements through signal space projection. *IEEE Transaction on Rehabilitation Engineering*, 8(2):186–188, June 2000.

[7] J.-D. Bauby and J. Leggatt. *The Diving Bell and the Butterfly*. Vintage Books, New York, 1998.

[8] H. Berger. Uber das elektrenkephalogramm des menschen. *Archiv. für Psychiatrie*, 87:527, 1929.

[9] N. Birbaumer. Slow cortical potentials: Plasticity, operant control, and behavioral effects. *Neuroscientist*, 5(2):74–78, 1999.

[10] N. Birbaumer, A. K ubler, N. Ghanayim, T. Hinterberger, J. Perelmouter, J. Kaiser, I. Iversen, B. Kotchoubey, N. Neumann, and H. Flor. The thought translation device (TTD) for completely paralyzed patients. *IEEE Transactions on Rehabilitation Engineering*, 8(2):190–193, June 2000.

[11] E. Donchin, K. M. Spencer, and R. Wijesinghe. The mental prosthesis: Assessing the speed of a P300-based brain-computer interface. *IEEE Transactions on Rehabilitation Engineering*, 3(2):174–179, June 2000.

[12] J. C. Doyle, R. Ornstein, and D. Galin. Lateral specialization of cognitive mode: II EEG frequency analysis. *Psychophysiology*, 11(5):567–578, 1974.

[13] L. A. Farwell and E. Donchin. Talking off the top of your head: Toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and Clinical Neurophysiology*, 70:510–523, 1988.

[14] D. Galin and R. E. Ornstein. Hemispheric specialization and the duality of consciousness. *Human Behavior and Brain Function*, 1973.

[15] D. Garrett, D. A. Peterson, C. W. Anderson, and M. H. Thaut. Comparison of linear and nonlinear methods for eeg signal classification. *IEEE Transactions on Neural Systems and Rehabilitative Engineering*, to appear, 2003.

[16] A. A. Green, M. Berman, P. Switzer, and M. D. Craig. A transformation for ordering multispectral data in terms of image quality with implications for noise removal. *IEEE Transactions on Geoscience and Remote Sensing*, 26(1):65–74, January 1988.

[17] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, 2001.

[18] Z. A. Keirn and J. I. Aunon. A new mode of communication between man and his surroundings. *IEEE Transactions on Biomedical Engineering*, 37(12):1209–1214, 1990.

[19] M. Kirby and C. Anderson. Geometric analysis for the characterization of nonstationary time-series. In Ehud Kaplan, Jerry Marsden, and Katepalli R. Sreenivasan, editors, *Springer Applied Mathematical Sciences Series*. Springer-Verlag, 2003.

[20] R. H. Kraft, O. R. Mitchell, M. L. Languis, and G. H. Wheatley. Hemispheric asymmetry during six- to eight-year old performance of piagetian conservation and reading tasks. *Neuropsychologia*, 22:637–643, 1984.

[21] P. Switzer and A. Green. Min/max autocorrelation factors for multivariate spatial imagery. Technical report, Stanford University, Department of Statistics, 1984.

[22] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan. Brain-computer interfaces for communication and control. *Clinical Neurophysiology*, 113:767–791, 2002.

[23] J. R. Wolpaw, D. J. McFarland, and T. M. Vaughan. Brain-computer interface research at the wadsworth center. *IEEE Transactions on Rehabilitation Engineering*, 3(2):222–226, June 2000.