MIXED EFFECTS MODELS FOR EEG EVOKED RESPONSE DETECTION

Yonghong Huang, Deniz Erdogmus, Misha Pavel

Oregon Health and Science University
OGI School of Science & Engineering
huang@csee.ogi.edu, derdogmus@ieee.org,
pavel@bme.ogi.edu

Santosh Mathan

Honeywell Laboratories
Human Centered Systems Group
Santosh.Mathan@honeywell.com

ABSTRACT

Human brain signals associated with perceptual processes have been shown to be useful for visual target image search. For the purpose of online training, we develop a novel mixed effects evoked response detector, which is capable of combining individual random effects and population fixed effects, for the analysis of neural signatures associated with targets. To avoid numerical problems in high dimensional matrix computations, we develop equivalent dimension reduced expressions for the mixed models. We construct the mixed effects evoked response model using principal component analysis to provide bases for the population model and linear discriminant analysis (LDA) to provide bases for the individual models. In addition, the LDA is adopted for Elecroencephalography channel dimensionality reduction. Data collected at different time and experimental conditions from two subjects performing image search tasks are utilized to assess the quality of the models. We also compare the proposed model with the support vector machine (SVM). The results demonstrate that the mixed models approach the SVM and provide reliable inference on cross session evaluation for the single-trial evoked response detection.

1. INTRODUCTION

Human brain signals associated with perceptual processes have been shown to be useful for visual target image search. Elecroencephalography (EEG) has been widely used for detecting cognitive disorders and other diseases. A brain's electrical response present in EEG signals to a stimulus, related to aspects of cognitive processing is referred to as an event-related potential (ERP) [1]. The ERPs associated with human perceptual judgments have been previously used for visual target image search [2, 3, 4]. The major task of an ERP-based image search systems is to detect the ERPs associated with the target stimuli. The conventional approach for studying the ERPs is trial averaging. However, this approach is not sufficiently efficient for fast brain interfaces that exploit the ERPs, therefore recent research focuses on single-trial ERP detection [5].

The main challenges of single-trial ERP detection are high dimensionality and scarcity of training data. The ideal scenario is to have enough training samples under controlled circumstances. However, it is infeasible in real applications to have subjects perform a long training period with unchanging experimental conditions. When multiple EEG measurements are obtained from each individual at different times and possibly under changing experimental conditions, we normally can not fully control the circumstances under which the measurements are taken. There are considerable variations among individuals in the number and timing of observations. Therefore we need to seek a model suitable for aggregated data to capture the population characteristics and also the individual features as well.

A mixed effects model (MEM) [6] is a statistical hierarchical model. It was first proposed for the analysis of longitudinal time-series data [7]. There are two sources of variation in the MEM: between-individual variations and within-individual variation. By introducing multilevel random effects, the MEM easily handles data with multiple sources of variation, such as EEG data. Specifically for designs of aggregating data across multiple subjects/sessions, we can easily use the population-averaged parameters to specify the common EEG signal type (consistent pattern across subjects/sessions), and the subject-specific parameters to specify subject/session individuality (individual variety with the within- and between-subject/session variance). Therefore the MEM provides principled basis for combining historical and new data and is convenient for online adaptation. To this end, we apply this statistical approach to the classification of single-trial multichannel EEG sequences.

In this paper, we present a mixed-effects ERP detector that models single-trial ERP waveforms as varying individuals from a population; thus the classifier attempts to explain fluctuations in the baseline ERP waveform via a hierarchical Bayesian topology. To avoid numerical problems in high dimensional matrix computations, we determine low-dimensional calculations utilizing low-rank matrix properties.

This work was supported by DARPA and NGA under contract HM1582-05-C-0046 and by NSF under grants ECS-0524835, ECS-0622239, and IIS-0713690. It has been approved for public release, distribution unlimited.

This significantly reduces computational complexity. The paper illustrates the performance of the model in the analysis of cross session EEG dataset and compares the model to a broadly used approach – support vector machine(SVM) [8] for weighing the benefits of the proposed model. Results establish that the mixed-effects ERP detector can provide a reliable basis for visual target image search and EEG-based brain interfaces in general.

2. MIXED EFFECTS MODELS

2.1. Model description

A general MEM is written as, for individual i of N,

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\alpha} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i, \tag{1}$$

where $i = 1, \ldots, N$.

- $\mathbf{y}_i$ is an $n_i \times 1$ vector of observations for the ith individual and n_i is the number of observations for the ith individual.
- $\boldsymbol{\alpha}$ is defined as a $p \times 1$ population fixed effects vector.
- $\mathbf{X}_i$ is an $n_i \times p$ population design matrix of fixed effects.
- $\mathbf{b}_i$ is defined as a $k \times 1$ individual random effects vector. We assume that the random effects are independent and have a normal distribution $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D})$ where $\mathbf{D}$ is a $k \times k$ matrix of positive definite covariance.
- $\mathbf{Z}_i$ is an $n_i \times k$ individual design matrix of random effects.
- $\boldsymbol{\varepsilon}_i$ is an $n_i \times 1$ vector of independent and identically distributed (iid) errors with zero mean and positive definite within-individual variance. We assume that the error terms have a normal distribution $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \sigma^2\mathbf{I}_{n_i})$, where I_{n_i} denotes an $n_i \times n_i$ identity matrix.

Thus model (1) can be written as $\mathbf{y}_i \sim N(\mathbf{X_i}\boldsymbol{\alpha}, \sigma^2\mathbf{I}_{n_i} + \mathbf{Z}_i\mathbf{D}{\mathbf{Z}_i}^T)$. This means that $\mathbf{y}_i$ has a multivariate normal distribution with mean $\mathbf{X_i}\boldsymbol{\alpha}$ and covariance matrix $\sigma^2\mathbf{I}_{n_i} + \mathbf{Z}_i\mathbf{D}{\mathbf{Z}_i}^T$. In Equation (1), $\mathbf{y}_i$, $\mathbf{X}_i$ and $\mathbf{Z}_i$ are known and $\boldsymbol{\alpha}$, $\mathbf{b}_i$ and $\boldsymbol{\varepsilon}_i$ are unknown. The variance parameters, σ^2 and $\mathbf{D}$ are unknown and need to be estimated along with the population parameter $\boldsymbol{\alpha}$. These parameters are estimated using data from the whole population.

2.2. Model parameter estimation

We use the expectation-maximization(EM) algorithm [9] to obtain maximum likelihood (ML) estimates for the model parameters, $\boldsymbol{\alpha}$ and $\mathbf{b}_i$, σ^2 and $\mathbf{D}$.

2.2.1. ML estimation of $\boldsymbol{\alpha}$ and $\mathbf{b}_i$

Writing $Var(\mathbf{y}_i)$ as $\mathbf{V}_i = \sigma^2\mathbf{I}_{n_i} + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T$, if all covariance parameters $\hat{\sigma}^2$ and $\hat{\mathbf{D}}$ were known, then $\mathbf{V}_i$ was known, we could estimate $\boldsymbol{\alpha}$ and $\mathbf{b}_i$. Assuming $\mathbf{y}_i$ is independent for each i, the joint density function of $\mathbf{y}_i$ is

$$f(\mathbf{y};\boldsymbol{\theta}) = \prod_{i=1}^{N} \frac{\exp[-\frac{1}{2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\alpha})^T\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\alpha})]}{(2\pi)^{\frac{n_i}{2}}|\mathbf{V}_i|^{\frac{1}{2}}} \tag{2}$$

where $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \mathbf{D}, \sigma)$. The log-likelihood function for the MEM is given by

$$\begin{aligned} l(\boldsymbol{\theta}) &= -\frac{1}{2}\{(\sum_{i=1}^{N} n_i)ln(2\pi) + \sum_{i=1}^{N}[ln|\mathbf{V}_i| \\ &+ (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\alpha})^T\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\alpha})]\}. \end{aligned} \tag{3}$$

If variances $\hat{\sigma}^2$ and $\hat{\mathbf{D}}$ were known, the log-likelihood function could be maximized by the generalized least squares estimator. Taking the derivative of $l(\boldsymbol{\theta})$ with respect to $\boldsymbol{\alpha}$ and equating to zero, we get

$$\hat{\boldsymbol{\alpha}} = (\sum_{i=1}^{N}\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i)^{-1}\sum_{i=1}^{N}\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{y}_i. \tag{4}$$

If $\boldsymbol{\alpha}$ was known, we can treat $\mathbf{b}_i$ as fixed effects and use least square estimation to obtain from Equation (1)

$$\hat{\mathbf{b}_i} = \mathbf{D}\mathbf{Z}_i^T\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\alpha}}). \tag{5}$$

2.2.2. EM algorithm for ML estimates of σ^2 and $\mathbf{D}$

M-step: If we were to observe $\mathbf{b}_i$ and $\boldsymbol{\varepsilon}_i$, we could easily obtain simple closed-form solution using ML estimation of variances,

$$\hat{\sigma}^2 = \sum_{i=1}^{N}\boldsymbol{\varepsilon}_i^T\boldsymbol{\varepsilon}_i / \sum_{i=1}^{N} n_i, \tag{6}$$

$$\hat{\mathbf{D}} = \sum_{i=1}^{N}\mathbf{b}_i\mathbf{b}_i^T/N. \tag{7}$$

E-step: If σ^2 and $\mathbf{D}$ were available, we could calculate the sufficient statistics as follows:

$$\begin{aligned} \textstyle\sum_{i=1}^{N}\boldsymbol{\varepsilon}_i^T\boldsymbol{\varepsilon}_i &= \textstyle\sum_{i=1}^{N}\hat{\boldsymbol{\varepsilon}}_i(\hat{\boldsymbol{\theta}})^T\hat{\boldsymbol{\varepsilon}}_i(\hat{\boldsymbol{\theta}}) \\ &+ \textstyle\sum_{i=1}^{N} tr\{Var[\boldsymbol{\varepsilon}_i|\mathbf{y}_i, \hat{\boldsymbol{\alpha}}(\hat{\boldsymbol{\theta}}), \hat{\boldsymbol{\theta}}]\}, \end{aligned} \tag{8}$$

$$\sum_{i=1}^{N}\mathbf{b}_i^T\mathbf{b}_i = \sum_{i=1}^{N}\{\hat{\mathbf{b}}_i(\hat{\boldsymbol{\theta}})^T\hat{\mathbf{b}}_i(\hat{\boldsymbol{\theta}}) + Var[\mathbf{b}_i|\mathbf{y}_i, \hat{\boldsymbol{\alpha}}(\hat{\boldsymbol{\theta}}), \hat{\boldsymbol{\theta}}]\}, \tag{9}$$

where $\hat{\boldsymbol{\varepsilon}}_i(\hat{\boldsymbol{\theta}}) = \mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\alpha}}_i(\hat{\boldsymbol{\theta}}) - \mathbf{Z}_i\hat{\mathbf{b}}_i(\hat{\boldsymbol{\theta}})$ and $\hat{\mathbf{b}}_i(\hat{\boldsymbol{\theta}})$ were obtained from ML estimation. Based on $\boldsymbol{\varepsilon}_i|\boldsymbol{\theta} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_{n_i})$, $\mathbf{y}_i|\boldsymbol{\varepsilon}_i; \boldsymbol{\theta} \sim N(\mathbf{X}_i\boldsymbol{\alpha}, \mathbf{Z}_i\mathbf{D}{\mathbf{Z}_i}^T)$, and $\mathbf{y}_i|\boldsymbol{\theta} \sim N(\mathbf{X}_i\boldsymbol{\alpha}, \sigma^2\mathbf{I}_{n_i} + \mathbf{Z}_i\mathbf{D}{\mathbf{Z}_i}^T)$, we can derive

$$Var[\boldsymbol{\varepsilon}_i|\mathbf{y}_i, \hat{\boldsymbol{\alpha}}(\hat{\boldsymbol{\theta}}), \hat{\boldsymbol{\theta}}] = [(\mathbf{Z}_i\mathbf{D}{\mathbf{Z}_i}^T)^{-1} + (\sigma^2\mathbf{I}_{n_i})^{-1}]^{-1}. \tag{10}$$

Similarly, based on $\mathbf{b}_i|\boldsymbol{\theta} \sim N(\mathbf{0}, \mathbf{D})$, $\mathbf{y}_i|\mathbf{b}_i; \boldsymbol{\theta} \sim N(\mathbf{X}_i\boldsymbol{\alpha} + \mathbf{Z}_i\mathbf{b}_i, \sigma^2\mathbf{I}_{n_i})$, and $\mathbf{y}_i|\boldsymbol{\theta} \sim N(\mathbf{X}_i\boldsymbol{\alpha}, \sigma^2\mathbf{I}_{n_i} + \mathbf{Z}_i\mathbf{D}{\mathbf{Z}_i}^T)$, we can calculate

$$Var[\mathbf{b}_i|\mathbf{y}_i, \hat{\boldsymbol{\alpha}}(\hat{\boldsymbol{\theta}}), \hat{\boldsymbol{\theta}}] = (\mathbf{Z}_i{}^T\mathbf{Z}_i/\sigma^2 + \mathbf{D}^{-1})^{-1}. \tag{11}$$

Thus from (6) to (11), we obtain the variance parameter estimates as:

$$\hat{\sigma}^2 = \sum_{i=1}^{N} \hat{\boldsymbol{\varepsilon}}_i(\hat{\boldsymbol{\theta}})^T \hat{\boldsymbol{\varepsilon}}_i(\hat{\boldsymbol{\theta}}) \sum_{i=1}^{N} n_i + \frac{\sum_{i=1}^{N} tr\{[(\mathbf{Z}_i \mathbf{D} \mathbf{Z}_i{}^T)^{-1} + (\sigma^2 \mathbf{I}_{n_i})^{-1}]^{-1}\}}{\sum_{i=1}^{N} n_i}, \quad (12)$$

$$\hat{\mathbf{D}} = \frac{1}{N} \sum_{i=1}^{N} \{\hat{\mathbf{b}}_i(\hat{\boldsymbol{\theta}})^T \hat{\mathbf{b}}_i(\hat{\boldsymbol{\theta}}) + (\frac{\mathbf{Z}_i{}^T \mathbf{Z}_i}{\sigma^2} + \mathbf{D}^{-1})^{-1}\}. \quad (13)$$

Upon convergence of the EM iterations, we obtain $\hat{\sigma}^2$ and $\hat{\mathbf{D}}$.

3. DIMENSION REDUCTION FORMULAS

The model parameter estimation equations in the previous Section involves $n_i \times n_i$ matrix inverse and determinant calculations. One can reduce the dimensionality of these calculations to $k \times k$ $(k \ll n_i)$ using the following exact rank-reduction formulas.

3.1. Simplified formulas for log-likelihood function

Since $\mathbf{V}_i = \sigma^2 \mathbf{I}_{n_i} + \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^T$ involves an $n_i \times n_i$ low-rank matrix inversion. We can use the following dimension-reduction formulas to exploit the relevant rank-k subspace:

$$\mathbf{V}_i^{-1} = \sigma^{-2} \mathbf{I}_{n_i} - \sigma^{-2} \mathbf{I}_{n_i} \mathbf{Z}_i (\mathbf{D}^{-1} + \mathbf{Z}_i^T \sigma^{-2} \mathbf{I}_{n_i} \mathbf{Z}_i)^{-1} \mathbf{Z}_i^T \sigma^{-2} \mathbf{I}_{n_i} = \sigma^{-2} \mathbf{I}_k - \sigma^{-4} \mathbf{Z}_i^T \mathbf{Z}_i (\mathbf{D}^{-1} + \sigma^{-2} \mathbf{Z}_i^T \mathbf{Z}_i)^{-1}, \quad (14)$$

$$|\mathbf{V}_i| = \sigma^{2(n_i - k)} |\sigma^2 \mathbf{I}_k + \mathbf{D} \mathbf{Z}_i^T \mathbf{Z}_i|. \quad (15)$$

If matrix $\mathbf{D}$ is nonsingular, we can have the log of the determinant as a function of $\mathbf{D}^{-1}$.

$$\ln |\mathbf{V}_i| = \ln |\sigma^2 \mathbf{D}^{-1} + \mathbf{Z}_i^T \mathbf{Z}_i| - \ln |\mathbf{D}^{-1}| + (n_i - k) \ln \sigma^2. \quad (16)$$

3.2. Simplified formulas for σ^2 and $\mathbf{D}$

To avoid inverse matrices in Equation (10) and (11), by using matrix inversion lemma, we have the following simplification,

$$[(\mathbf{Z}_i \mathbf{D} \mathbf{Z}_i{}^T)^{-1} + (\sigma^2 \mathbf{I}_{n_i})^{-1}]^{-1} = \sigma^2 \mathbf{I}_{n_i} - \sigma^4 \mathbf{I}_{n_i} \mathbf{V}_i^{-1} \quad (17)$$

$$(\mathbf{Z}_i{}^T \mathbf{Z}_i / \sigma^2 + \mathbf{D}^{-1})^{-1} = \mathbf{D} - \mathbf{D} \mathbf{Z}_i^T \mathbf{V}_i^{-1} \mathbf{Z}_i \mathbf{D}. \quad (18)$$

Therefore Equation (12) and (13) can be simplified as follows

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} \hat{\boldsymbol{\varepsilon}}_i(\hat{\boldsymbol{\theta}})^T \hat{\boldsymbol{\varepsilon}}_i(\hat{\boldsymbol{\theta}})}{\sum_{i=1}^{N} n_i} + \sigma^2 - \frac{\sigma^4 \sum_{i=1}^{N} tr(\mathbf{V}_i^{-1})}{\sum_{i=1}^{N} n_i} \quad (19)$$

$$\hat{\mathbf{D}} = \frac{\sum_{i=1}^{N} [\hat{\mathbf{b}}_i(\hat{\boldsymbol{\theta}})^T \hat{\mathbf{b}}_i(\hat{\boldsymbol{\theta}})]}{N} + \mathbf{D} - \frac{\mathbf{D}(\sum_{i=1}^{N} \mathbf{Z}_i^T \mathbf{V}_i^{-1} \mathbf{Z}_i)\mathbf{D}}{N}. \quad (20)$$

Using (14), we also obtain

$$\mathbf{Z}_i^T \mathbf{V}_i^{-1} \mathbf{Z}_i = \mathbf{Z}_i^T \mathbf{Z}_i (\sigma^2 \mathbf{I}_k + \mathbf{D} \mathbf{Z}_i^T \mathbf{Z}_i)^{-1}. \quad (21)$$

If $(\mathbf{Z}_i^T \mathbf{Z}_i)^{-1}$ exists, we can have

$$\mathbf{Z}_i^T \mathbf{V}_i^{-1} \mathbf{Z}_i = [\sigma^2 (\mathbf{Z}_i^T \mathbf{Z}_i)^{-1} + \mathbf{D}]^{-1}. \quad (22)$$

Furthermore from (20),

$$\sum_{i=1}^{N} \mathbf{Z}_i^T \mathbf{V}_i^{-1} \mathbf{Z}_i = \sigma^{-2} \sum_{i=1}^{N} \mathbf{Z}_i^T \mathbf{Z}_i - \sigma^{-4} \sum_{i=1}^{N} [(\mathbf{Z}_i^T \mathbf{Z}_i) (\mathbf{D}^{-1} + \sigma^{-2} \mathbf{Z}_i^T \mathbf{Z}_i)^{-1} (\mathbf{Z}_i^T \mathbf{Z}_i)^T], \quad (23)$$

3.3. Simplified formulas for α and $\mathbf{b}_i$

From (4),

$$\sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i = \sigma^{-2} \sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{X}_i - \sigma^{-4} \sum_{i=1}^{N} [(\mathbf{X}_i^T \mathbf{Z}_i) (\mathbf{D}^{-1} + \sigma^{-2} \mathbf{Z}_i^T \mathbf{Z}_i)^{-1} (\mathbf{X}_i^T \mathbf{Z}_i)^T], \quad (24)$$

$$\sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{y}_i = \sigma^{-2} \sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{y}_i - \sigma^{-4} \sum_{i=1}^{N} [(\mathbf{X}_i^T \mathbf{Z}_i) (\mathbf{D}^{-1} + \sigma^{-2} \mathbf{Z}_i^T \mathbf{Z}_i)^{-1} (\mathbf{Z}_i^T \mathbf{y}_i)], \quad (25)$$

From (5),

$$\mathbf{Z}_i^T \mathbf{V}_i^{-1} \mathbf{y}_i = \sigma^{-2} \mathbf{Z}_i^T \mathbf{y}_i - \sigma^{-4} (\mathbf{Z}_i^T \mathbf{Z}_i)(\mathbf{D}^{-1} + \sigma^{-2} \mathbf{Z}_i^T \mathbf{Z}_i)^{-1} (\mathbf{Z}_i^T \mathbf{y}_i), \quad (26)$$

$$\mathbf{Z}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i = \sigma^{-2} \mathbf{X}_i^T \mathbf{Z}_i{}^T - \sigma^{-4} (\mathbf{Z}_i^T \mathbf{Z}_i)(\mathbf{D}^{-1} + \sigma^{-2} \mathbf{Z}_i^T \mathbf{Z}_i)^{-1} (\mathbf{X}_i^T \mathbf{Z}_i)^T, \quad (27)$$

4. MIXED-EFFECTS ERP DETECTOR CONSTRUCTION

We want to construct a single-trial ERP detector using MEMs. We adopt the receiver operating characteristic (ROC) curve [10] to assess performance quantitatively with a scalar measure.

4.1. Data preparation – cross session data

Two subjects were recruited for the study. The images were presented at very high rate of 100 ms/image in the rapid serial visual presentation paradigm. The subjects performed target detection by clicking on a button (which typically occur after 500ms and do not create overlapping brain activity with portion used for our detectors) as soon as they saw a target. At the same time, we monitored and recorded their EEG signals by a 32-channel Biosemi system. The sampling rate was

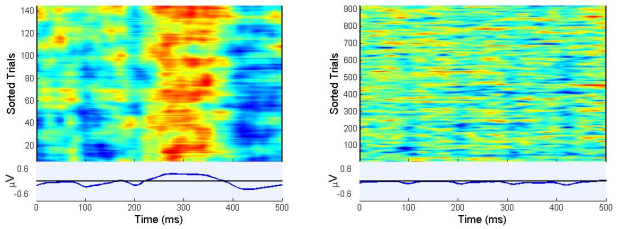

Fig. 1. Images of ERP and non-ERP signals associated with targets (left) and distractors (right). Time-zero corresponds to stimulus onset in each trial. The bottom traces are the EEG signals averaged over trials.

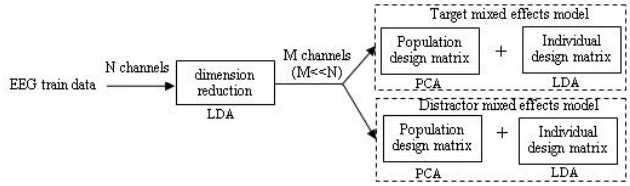

Fig. 2. Training structure of the mixed effects ERP detector

256Hz. Each subject had one morning session and one afternoon session each day and five days in a row. Each session contained 200 trials and each trial was 5 seconds. Images were randomly displayed and 75% of the trials contained a single target instance. Figure 1 shows the image plots of the ERP and non-ERP signals corresponding to target and distractor stimuli for subject 1 at channel 1. One can observe a clear ERP pattern corresponding to targets while no pattern corresponding to distractors. To evaluate cross session performance, we aggregated the data across sessions. We simulated the condition as a realistic scenario that each session used as a test set was appended to the previous training session and the new training session (aggregated training session) would be used for future sessions. For instance, we trained on $session$ 1 and tested on $session$ 2; then we trained on $sessions$ $1+2$ and tested on $session$ 3 and so on until trained on $sessions$ $1+2+...+9$ and tested on $session$ 10.

The procedures of data pre-processing [11] were applied to obtain statistically independent (as much as possible) training data. We segmented the EEG data into the task-relevant epochs (500 ms after each image trigger), filtered the data through a 1-45 Hz bandpass filter and conducted data normalization. We adopted a disjoint windowing scheme in the training and a sequential windowing scheme in the test. For both training and test, we removed the distractors in the interval of one second before and after the targets and selected the targets with following button responses within 1.5 seconds.

4.2. ERP detector – MEM

Our goal in classification is to build a mixed effects ERP detector to accurately discriminate the EEG signals containing the ERP versus the non-ERP. Figure 2 shows the training procedures of the models, which are used under the likelihood ratio test framework for detection.

4.2.1. Feature dimension reduction

We apply linear discriminant analysis (LDA) [10] on multichannel EEG data for dimension reduction and congregate the selected channel projections to form a feature vector as the basis for the MEM based generative models for classification. For each target sample $\mathbf{x}_i^t$ or distractor sample $\mathbf{x}_i^d$, we can measure the mean $(\mathbf{M}_t, \mathbf{M}_d)$ and variance $(\mathbf{C}_t, \mathbf{C}_d)$ for the target cluster and distractor cluster,

$$\mathbf{M}_t = 1/N_t \sum_{i=1}^{N_t} \mathbf{x}_i^t \tag{28}$$

$$\mathbf{M}_d = 1/N_d \sum_{i=1}^{N_d} \mathbf{x}_i^d \tag{29}$$

$$\mathbf{C}_t = 1/N_t \sum_{i=1}^{N_t} (\mathbf{x}_i^t - \mathbf{M}_t)(\mathbf{x}_i^t - \mathbf{M}_t)^T \tag{30}$$

$$\mathbf{C}_d = 1/N_d \sum_{i=1}^{N_d} (\mathbf{x}_i^d - \mathbf{M}_d)(\mathbf{x}_i^d - \mathbf{M}_d)^T. \tag{31}$$

For linear projections, $f(\mathbf{x}_i^t) = \mathbf{w}^T\mathbf{x}_i^t$ and $f(\mathbf{x}_i^d) = \mathbf{w}^T\mathbf{x}_i^d$, we consider maximizing the objective function

$$J(\mathbf{w}) = \mathbf{w}^T\mathbf{S}_b\mathbf{w}/\mathbf{w}^T\mathbf{S}_w\mathbf{w}. \tag{32}$$

where between cluster scatter matrix is $\mathbf{S}_b = (\mathbf{M}_t - \mathbf{M}_d)(\mathbf{M}_t - \mathbf{M}_d)^T$ and within cluster scatter matrix is $\mathbf{S}_w = (\mathbf{C}_t + \mathbf{C}_d)$. The solution is given by a generalized eigendecomposition expression:

$$\mathbf{w} = eig(\mathbf{S}_w^{-1}\mathbf{S}_b). \tag{33}$$

We select the subset of the eigenvectors associated with the largest eigenvalues as the optimal projections. The number of the eigenvectors to be retained, (N_eigs1_LDA), is selected via crossvalidation.

4.2.2. Population design matrix

We develop the population design matrices for the target cluster and the distractor cluster respectively using principal component analysis (PCA) [10]. To create the target basis matrix, we first calculate the covariance matrix of target training data and calculate the eigenvectors and eigenvalues. Then we sort the columns of the eigenvector matrix according to their associated eigenvalues and select a subset of eigenvectors as the basis vectors by their cumulative energy contents. Similarly we can develop the population design matrix for the distractor cluster. The percentage of the cumulative energy $(Perc\_eigs\_PCA)$ can be chosen by the users. Here we use the same percentage of the eigen-energy for both target and distractor population design matrices.

4.2.3. *Individual design matrix*

We develop the individual design matrix using LDA as described in Section 4.2.1. We just simply select a subset of the eigenvectors associated with the large eigenvalues as the individual design matrix. Here we set the number of the eigenvector equals one for simplicity (We investigated larger number of the eigenvectors with crossvalidation, but the performances were close).

4.2.4. *MEM training*

After we have the target population design matrix $\mathbf{X}_i^T$, the distractor population design matrix $\mathbf{X}_i^D$ and the individual design matrix $\mathbf{Z}_i$, we can estimate the model parameters,$\boldsymbol{\alpha}$,$\mathbf{b}_i$ and variance parameters, σ^2,$\mathbf{D}$ using EM algorithm for ML estimation as discussed in Section 2.2. We derive a target MEM and a distractor MEM after the training as shown in Figure 2.

4.2.5. *MEM test*

After we have the MEM models for the target cluster and the distractor cluster, we can subject the test patterns to the models. For each test pattern, we have

$$\mathbf{y}_i^{Test} = \mathbf{X}_i\boldsymbol{\alpha} + \mathbf{Z}_i\mathbf{b}_i^{Test} + \boldsymbol{\varepsilon}_i \tag{34}$$

where $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D})$ and $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \sigma^2\mathbf{I}_{n_i})$. Since we have $p(\mathbf{y}_i^{Test}|\boldsymbol{\alpha}, \mathbf{b}_i^{Test}) \sim N(\mathbf{X}_i\boldsymbol{\alpha}+\mathbf{Z}_i\mathbf{b}_i^{Test}, \sigma^2\mathbf{I}_{n_i})$, we can maximize the posterior and obtain the optimal individual random effect parameter for the test pattern,

$$\mathbf{b}_i^{Test^*} = \mathbf{D}\mathbf{Z}_i^T\mathbf{V}_i^{-1}(\mathbf{y}_i^{Test} - \mathbf{X}_i\boldsymbol{\alpha}). \tag{35}$$

After we obtain $\mathbf{b}_{i,target}^{Test^*}$ for the target model and $\mathbf{b}_{i,distractor}^{Test^*}$ for the distractor model using appropriate design eigenvectors in (35), we can employ the likelihood ratio test using the respective model log-likelihood estimates:

$$\begin{aligned} l(\mathbf{b}_i^{Test^*}) &= \ln[N(\mathbf{X}_i\boldsymbol{\alpha} + \mathbf{Z}_i\mathbf{b}_i^{Test^*}, \sigma^2\mathbf{I}_{n_i}] + \ln[N(\mathbf{0}, \mathbf{D})] \\ &= ln(C_{\sigma^2}) + ln(C_{\mathbf{D}}) - [\mathbf{y}_i^{Test} - (\mathbf{X}_i\boldsymbol{\alpha} + \mathbf{Z}_i\mathbf{b}_i^{Test^*})]^T \\ &\times \frac{[\mathbf{y}_i^{Test} - (\mathbf{X}_i\boldsymbol{\alpha} + \mathbf{Z}_i\mathbf{b}_i^{Test^*})]}{2\sigma^2} - \frac{\mathbf{b}_i^{Test^{*T}}\mathbf{D}^{-1}\mathbf{b}_i^{Test^*}}{2} \end{aligned}$$

where C_{σ^2} and $C_{\mathbf{D}}$ are constants. The discriminant value of the MEM (the estimates of target likelihood) is the difference between the log likelihood values of the target and distractor models.

4.3. Parameter regularization

We employ 10-fold cross validation [10] for parameter regularization for each subject. We train the MEM classifier, choosing the optimal N_eigs1_LDA in dimension reduction and Perc_eigs_PCA in population design matrix from discrete sets that give the best validation performance. Validation performance is the average of the the area under the ROC curve

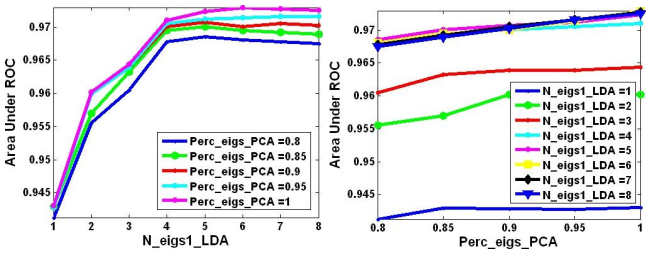

Fig. 3. 10-fold cross validation results for MEM model parameter selections on subject 1. Left panel is varying N_eig1_LDA for fixed Perc_eigs_PCA and right panel is varying Perc_eigs_PCA for fixed N_eig1_LDA.

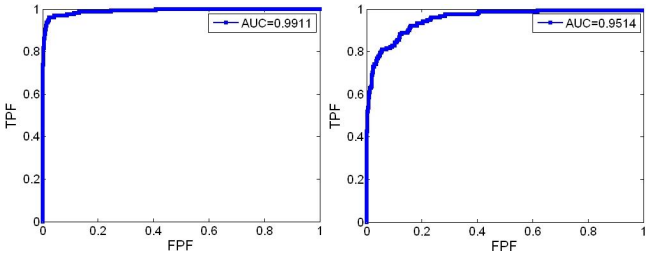

Fig. 4. ROC cures for subject 1 (left) and 2 (right) on Test Session 9 using MEM. FPF is false positive fraction and TPF is true positive fraction

(AUC) of nine classifiers, each of which is trained on a different nine-fold training set, and evaluated on a one-fold validation set. We also apply the same technique for parameter (kernel size σ^2 and cost parameter C) regularization for SVM [11].

5. RESULTS

5.1. MEM on cross section data

Exhaustive 10-fold crossvalidation to search for the optimal parameters on a discrete set of $N\_eig1\_LDA = [1, 2, 3, 4, 5, 6, 7, 8]$ and $Perc\_eigs\_PCA = [80\%, 85\%, 90\%, 95\%, 100\%]$ was performed for the MEM.The results are shown in Figure 3 for subject 1 session 1. In this case, the optimal parameters are $N\_eig1\_LDA = 6$ and $Perc\_eigs\_PCA = 100\%$ for subject 1. The same procedure was applied to each training session for each subject.

In the experiments, both subjects achieved high detection performance. Figure 4 shows the ROC curves of the test results for two subjects on their respective test session 9. We can see that both subjects achieve high detection rates with $AUC = 0.99$ for subject 1 and $AUC = 0.95$ for subject 2.

5.2. Comparison of MEM and SVM

We also conducted exhaustive 10-fold crossvalidation to search for the optimal parameters on a discrete set of $\sigma^2 = [0.05, 0.1$

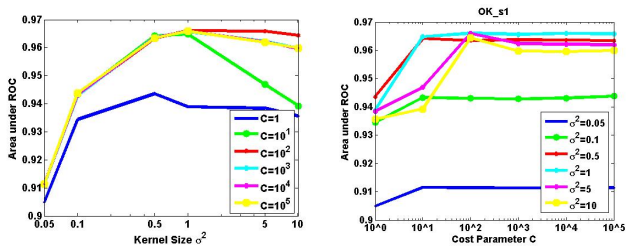

Fig. 5. 10-fold cross validation results for SVM model parameter selections on subject 1. Left panel is varying σ^2 for fixed C and right panel is varying C for fixed σ^2.

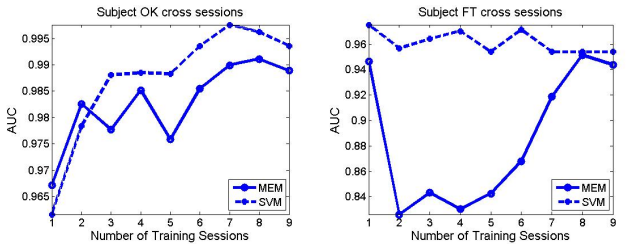

Fig. 6. Area under ROC(AUC) for different number of training sessions for subject 1 (left) and 2 (right) using MEM and SVM. (By the time of submission, the last three SVM test results for subject 2 are not ready. Those are corresponding to the last three dots in dash line of right panel. They will be ready by the time of paper acceptance.)

$, 0.5, 1, 5, 10]$ and $C = [10^0, 10^1, 10^2, 10^3, 10^4, 10^5]$ for the SVM. We obtained optimal parameters on each training session for each subject. Figure 5 shows the validation results for subject 1 session 1. In this case, the optimal parameters are $\sigma^2 = 1$ and $C = 100$. We applied the same procedure to each training session for each subject.

Both the MEM and the SVM achieve high detection performance and the AUC exhibits a generally increasing trend with the inclusion of additional training data from subsequent sessions except the SVM on subject 2. Figure 6 demonstrates the increasing AUC. The SVM shows slightly better performance than the MEM as expected based on that discriminative methods often result in better performance than generative models in classifications due to optimal decision boundaries [12]. However, the MEM is more adaptive for the purpose of online training.

6. DISCUSSION

Our results on the cross session dataset of the real application demonstrate the viability of the MEM on cross session single-trial ERP detection. Given a reasonable amount of cross session training data, the mixed ERP model achieve excellent generalization performance in terms of high AUC. The results prove that the MEM approaches the problem of characterizing between session variation in signal statistics. Our dimension(rank)-reduced version of the mixed effects model significantly lowers the computational complexity from $O(n_i^2)$ to $O(k^2)$ $(k << n_i)$. Comparing to the SVM, a discriminative learning method, the MEM is more adaptive because it offers to encode prior knowledge about the data structure in a very direct way. Future work will investigate the application of the mixed effect ERP models for online training in the ERP-based brain interfaces by imposing hyper priors across sessions to allows the fixed effect parameters and random effect variances to be different for each session.

7. REFERENCES

[1] Barry S. Oken, *Evoked Potentials in Clinical Medicine*, chapter 15, LippincottRaven, Philadelphia, PA, 1997.

[2] J.S. Johnson and B.A.Olshausen, "Timecouse of neural signatures of object recognition," *Journal of Vision*, vol. 3, pp. 499–512, sep 2003.

[3] S. Makeig, A. Delorme, M. Westerfield, J.Townsend, E. Courchense, and T. Sejnowski, "Electroencephalographic brain dynamics following visual targets requiring manual responses," *Public Library of Science Biology*, vol. 2, no. 6, pp. 747–762, 2004.

[4] R.V.Rullen and S.J. Thorpe, "The time course of visual processing:from early perception to decition-making," *Journal of Cognitive Neuroscience*, vol. 13, no. 4, pp. 454–461, 2001.

[5] P.Sajda, A.D. Gerson, M.G. Philiastides, and L.C. Parra, *Towards Brain-Computer Interfacing*, chapter Single-trial analysis of EEG during rapid visual discrimination: Enabling cortically-coupled computer vision, MIT Press, 2007.

[6] E. Demidenko, *Mixed Models Theory and Application*, John Wiley&Sons, Inc., New Jersey, NY, 2004.

[7] N.M. Laird and J.H. Ware, "Random-effect models for longitudinal data," *Biometrics*, vol. 38, pp. 963–974, 1982.

[8] B.Scholkopf and A.J. Smola, *Learning with kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, New Jersey, NY, 2001.

[9] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood with incomplete data via the e-m algorithm," *Journal of the Royal Statistical Society*, vol. B39, pp. 1–38, 1977.

[10] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, Wiley, Inc., New Jersey, NY, 2001.

[11] Y. Huang, D. Erdogmus, S. Mathan, and M. Pavel, "Large-scale image database triage via eeg evoked responses," in *Proc. of IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Las Vegas, NV, 2008, pp. 429–432.

[12] Tony Jebara, *Machine Learning Distriminative and Generative*, Kluwer Academic Publishers Group, Dordrecht, Netherlands, 2004.