Spatio-Temporal Modeling of EEG Data for Understanding Working Memory

Jinbo Bi JINBO@ENGR.UCONN.EDU
Tingyang Xu TINGYANG.XU@ENGR.UCONN.EDU
Chi-Ming Chen CHI-MING.CHEN@UCONN.EDU
University of Connecticut, Storrs, CT 06269 USA

Jason Johannesen JASON.JOHANNESEN@YALE.EDU
Yale University, New Haven, CT 06511 USA

Abstract

Electroencephalographic (EEG) recording provides a powerful measure of neural dynamics underlying human cognition, such as working memory. However, the analysis of multidimensional EEG data is challenging because it requires the modeling of temporal and spatial correlations in order to determine the EEG features most predictive of memory performance. Standard techniques, such as generalized estimating equations (GEE), can select features and account for sample correlation. However, they cannot explicitly model how a dependent variable relies on features measured at different memory stages and scalp locations. We propose an approach to automatically and simultaneously determine both the relevant spatial features and relevant temporal points that impact the response of a memory task. The proposed model can still correct for the non-*i.i.d* nature of the data, similar to GEE, by estimating the within-individual correlations. Our approach decomposes model parameters into a summation of two components and imposes separate block-wise LASSO penalties to each component when building a linear model in terms of multidimensional EEG features. An accelerated gradient descent algorithm is developed to efficiently solve the related optimization problem. We identified that the influential factors for working memory between healthy subjects and schizophrenia patients differ in frequency bands, scalp positions and information processing stages.

1. Introduction

Advances in science supporting the growth and adaptability, or ”neuroplasticity”, of human brain cells into late adulthood provide new promise for interventions designed to preserve and rehabilitate brain function (May, 2011). The merging of brain science and computer technology has created a consumer market for software designed to train brain functions, such as memory and attention, following the rationale that brain circuitry can be strengthened like muscles in response to repetitive exercise (Nahum et al., 2013). “Computer-based cognitive training” (CBCT) software can be purchased privately at low cost, can be used on mobile devices, is designed to be enjoyable and motivating, and can be self-administered without clinical oversight. However, there is an important and often overlooked shortcoming: CBCT cannot be assumed that compromised brain areas, or normally expected approaches to performing cognitive training exercises, will be utilized during this training. Instead, compensatory mechanisms may be used and reinforced during training. Underutilization of the damaged tissue may lead to further weakening, rather than strengthening, of its natural function.

Through the development of a brain-computer interface (BCI) enabled training program, we may address the critical limitation of current cognitive training software. The very first step, that is necessary if not essential, is to isolate the neural dysfunction associated with core cognitive impairments that CBCT should target. The identified target and important features for a cognitive function can then be incorporated into the design of a BCI prototype to improve CBCT. As working memory dysfunction is a core feature of many psychiatric disorders, it serves as a critical target for cognitive rehabilitation and hence is our study focus. Working memory requires network-level activation and coordination of neural activity between prefrontal cortical (PFC) and cortical association areas involved in sensory and attentional processes (Wang, 2010).

ICML Workshop on Statistics, Machine Learning and Neuroscience, Lille, France, 2015.

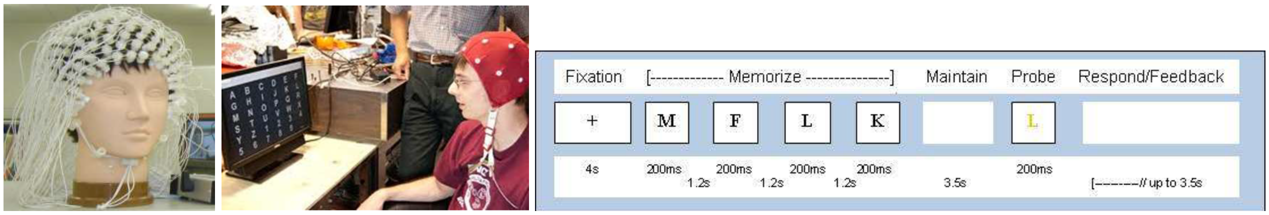

Figure 1. Illustration of EEG BCI apparatus and working memory test: (left) EEG recording montage; (middle) a BCI program called P300 speller; (right) a sample trial of Sternberg experiment depicting stages of information processing and time courses as extracted for EEG analysis based on memory span of 4 letters.

The cortical distribution of neural activity during working memory performance has been studied extensively using electroencephalographic (EEG) recording (Boonstra et al., 2013), enabling the evaluation of real-time changes in neural activity at distinct information processing stages (i.e., encoding, retention, retrieval) to behavioral performance (Klimesch, 1999). In this paper, we investigate the applications of an advanced machine learning approach in EEG analysis using data collected while participants performed a visual Sternberg working memory task.

EEG recording provides powerful methodology for studying neural dynamics of human cognition. EEG data is dimensional and complex, based on a time series of events sampled with high temporal resolution (i.e, millisecond level) and distributed spatially across multiple scalp locations (e.g., montages of 32 to 256 channels) (Figure 1(left and middle)). Given that research-grade bio amplification systems are capable of acquiring EEG at 1000Hz or higher, and that EEG is typically recorded for 10 minutes or more during psychophysiological experiments, the analysis of these data requires many decisions about the selection of time points and signal extraction methods used to best characterize the psychophysiological phenomena under investigation. We hence propose a new method, named by multiscale LASSO, which extends the widely used generalized estimating equations (GEE) by imposing the LASSO regularizer at multiple scales. This approach can jointly learn features and temporal dependency for outcome prediction with correction for the non-independent and identically distributed (*i.i.d.*) data using GEE's strategy. We use this algorithm to identify the most important EEG frequency bands, and information processing locations and stages for successful working memory.

2. Problem description and data sets

The Sternberg data are considered ideal for testing feature selection algorithms as there are multiple information processing stages, frequency components, and sources of neural activity involved and no single dependent measure that, taken alone, would appropriately account for task performance. A feature selection procedure is needed to determine which EEG measures are most predictive of task performance based on the classification of correct vs. incorrect Sternberg trial responses. As our study data were collected in a clinical sample of patients with schizophrenia and healthy community members, we also seek to model differences in neural activation during working memory performance between groups. EEG features identified in this way represent the vectors that optimally distinguish correct from incorrect trial performance for healthy controls and schizophrenia patients, respectively, and produce coefficients with valences indicating whether lesser or greater activity on the selected feature is associated with the specified classification outcome.

Our study sample consisted of 37 individuals meeting the diagnostic criteria for schizophrenia (SZ) and 6 healthy normal (HN) adults enrolled in clinical trial NCT00923078 (https://clinicaltrials.gov). The study was conducted under oversight of VA Connecticut Healthcare System (VACHS) and Yale University institutional review boards. All participants provided written informed consent. Our analysis uses a subset of the sample collected from the parent study.

Inclusion was limited to individuals aged from 18 to 70, native English speaking, with stable housing for minimum of 30 days. SZ sample members had minimum of 30 days since discharge from last hospitalization, 30 days since last change in psychiatric medications, and were receiving mental health services through VACHS or Yale affiliated outpatient facilities. Subjects were excluded based on current diagnosis of alcohol or substance abuse disorders, history of brain trauma or neurological disease, mental retardation or premorbid intelligence < 70, and auditory or visual impairment that would interfere with study procedures. Any current or past Axis I diagnosis of psychiatric disorders was exclusionary for HN sample enrollment.

EEG was recorded using a 64-channel BioSemi ActiveTwo (BioSemi B.V., Amsterdam, Netherlands) bio-amplifier and electrode system with sensors located according to the 10-20 system. Additional off-cap electrodes were placed bilaterally at mastoids (reference), the outer canthi of both eyes (horizontal electrooculogram; HEOG), and above and below the right orbit (vertical electrooculogram; VEOG). Continuous EEG was monitored online in ActiView V6.05 and acquired at a 1024 Hz sampling rate with a bandpass filter setting of 0.16-100 Hz. The Sternberg task was administered using Neurobehavioral Systems' Presentation software (Neurobehavioral Systems, Inc., Albany, CA), with behavioral responses captured using two buttons of a Cedrus RB-834 response pad (Cedrus Corporation, San Pedro, CA). Total EEG set up time was approximately 30 minutes.

As shown in Figure 1(right), the Sternberg working memory task (Raghavachari, et al., 2001) used in this study consisted of sequentially presented letters, with span widths of 4-8 randomly generated letters each (1.2s inter-stimulus interval). A total of 90 trials were administered for each individual in three blocks of 30, each lasting approximately 8 minutes. In each trial, a 3.5s retention period followed each stimulus set (i.e., the encoding stage), terminating with a response probe to which participants indicated using one of two response pad buttons whether the probe letter was or was not presented in the set (i.e., the retrieval stage). Auditory feedback was given to indicate correct, incorrect, and time-out trials. The early 4s period before the stimulus set started was considered as the baseline stage which was used for participants to prepare for the memory task.

Amplitudes of EEG signals were extracted for each trial and each individual by Morlet wavelet decomposition on 98 scales from 0.5 Hz to 100 Hz. In this study, an EEG record consisted of totally 60 features extracted from five frequency bands (δ: 0.5 - 4 Hz, θ: 4 - 8 Hz, α: 8 - 12 Hz, β: 14 - 28 Hz, and γ: 30 - 58 Hz), three brain regions (Fz, Cz and Oz), and the four memory stages (baseline, encode, retain and retrieve). Within each brain region and each memory stage, outliers of the EEG feature amplitudes were excluded (outside of 2 standard deviation from the mean). A binary label associated with each record indicated whether the individual answered correctly (0) or incorrectly (+1) in the trial. Because the training data contained multiple trials of a single individual, the records are not expected to be *i.i.d.*. Generalized estimating equations (GEE) are commonly used to estimate the sample correlation simultaneously while constructing predictive models. However, they build a generalized linear model (GLM) using the 60 features as a vector, and hence ignore the spatio-temporal structure of the data.

3. The proposed algorithm and analysis

3.1. Proposed formulations

In our approach, data in the t-th record of subject i are aligned into a $d \times p$ matrix, denoted by

$$\mathbf{X}_{(i;t)} = [\mathbf{x}_{s_1}^{(i)}, \mathbf{x}_{s_2}^{(i)}, \cdots, \mathbf{x}_{s_p}^{(i)}]$$

where p refers to the number of information processing stages (4 stages in our cases) and $\mathbf{x}_{s_1}^{(i)}, \cdots, \mathbf{x}_{s_p}^{(i)}$ refer to the 15 features (5 frequency bands extracted from the 3 regions) collected at each stage. Then a linear model can be represented as

$$y_t^{(i)} = tr\left(\mathbf{X}_{(i;t)}^{\top}\mathbf{W}\right)$$

where $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_p]$ represents the weights for each feature across the rows and the weights for each memory stage across the columns.

GEE estimates the parameters of a GLM taking into account the correlations in the training examples. Similar to GLM, there exists a link function that can be used to translate the nonlinear model into a linear model. The expectation of the outcome $y_t^{(i)}$ for subject i at time t is computed as:

$$E(y_t^{(i)}) = \mu_t^{(i)} = g^{-1}(\eta_t^{(i)}), \tag{1}$$

where $\mu_t^{(i)}$ represents the mean model, g^{-1} is the inverse of a link function g in a GLM (McCullagh & Nelder, 1989), and $\eta_t^{(i)} = \left(\mathbf{x}_t^{(i)}\right)^{\top}\mathbf{w}$. The variance of $y_t^{(i)}$ is computed as $\text{var}(y_t^{(i)}) = \text{var}(\mu_t^{(i)})/\phi$ where ϕ is a scaling parameter that may be known or estimated.

GEE presumes a so-called working correlation structure, typically denoted by $\mathbf{R}(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ is a parameter to be determined from data. The common choices of $\mathbf{R}(\boldsymbol{\alpha})$ include exchangeable, tri-diagonal and AR-1 formula (Liang & Zeger, 1986). The exchangeable correlation structure, also called *equi-correlation*, assumes that $corr(y_{it}, y_{it'}) = \alpha$ for all $t \neq t'$. The tri-diagonal structure uses a tridiagonal matrix as $\mathbf{R}(\boldsymbol{\alpha})$ where $corr(y_{it}, y_{it'}) = \alpha$ if $t' = t \pm 1$ or 0 otherwise. The AR formula assumes a correlation structure along continuous time analogous to the first-order autoregressive process, and uses $corr(y_{it}, y_{it'}) = \alpha^{|t-t'|}$.

To estimate the regression coefficients $\mathbf{w}$, the quasi-likelihood methods of GEE minimize a loss function that is defined via the model deviance (Olsson, 2002). The model deviance measures the difference between the log-likelihood of the estimated mean model and that of the observed values. For an arbitrary distribution, it may not correspond to an explicit function. For the exponential family, it takes a special form, which is still complicated. Hence, GEE solves the *estimating equations* that are defined by setting the derivatives of the loss function to 0. Although the deviance cannot be written out explicitly, the estimating equations can always be computed by the following equation assuming there are m subjects and without loss of generality each subject has n repeated trial records in the data:

$$L(\mathbf{w}, \boldsymbol{\alpha}) = \sum_{i=1}^{m} \left(\mathbf{D}^{(i)}\right)^{\top} \left(\boldsymbol{\Sigma}^{(i)}\right)^{-1} \mathbf{s}^{(i)} = 0. \tag{2}$$

where $\mathbf{D}^{(i)} = \partial\boldsymbol{\mu}^{(i)}/\partial\mathbf{w}$ is a $d \times n$ matrix where $\boldsymbol{\mu}^{(i)}$ combines all $\mu_t^{(i)}, \forall t = 1, \cdots, n$ into a vector, $\mathbf{s}^{(i)} =$

$\mathbf{y}^{(i)} - \boldsymbol{\mu}^{(i)}(\mathbf{w})$. The $n \times n$ matrix $\boldsymbol{\Sigma}^{(i)}$ is the estimated covariance between the different trials of an individual as:

$$\boldsymbol{\Sigma}^{(i)}(\boldsymbol{\alpha}) = \left(\mathbf{A}^{(i)}\right)^{1/2} \mathbf{R}(\boldsymbol{\alpha}) \left(\mathbf{A}^{(i)}\right)^{1/2} / \phi \qquad (3)$$

where $\mathbf{A}^{(i)}$ is defined as the $n \times n$ diagonal matrix with var$(\mu_t^{(i)})$ as the t-th diagonal element. Algorithms are given in (Liang & Zeger, 1986) to compute $\mathbf{w}$ for the different choices of $\mathbf{R}(\boldsymbol{\alpha})$.

To apply GEE to our model, we replace $\eta_t^{(i)}$ used in GEE by the following formula

$$\eta_t^{(i)} = tr\left(\mathbf{X}_{(i;t)}^{\top} \mathbf{W}\right). \qquad (4)$$

Substituting Eq.(4) into Eq.(1) yields a formulation similar to GEE Eq.(2). Although the model deviance cannot be written out, for notational convenience, we denote by $Dev^{(i)}(\mathbf{W}, \boldsymbol{\alpha})$ the deviance occurred on subject i. GEE minimizes $\sum_{i=1}^{m} Dev^{(i)}(\mathbf{W}, \boldsymbol{\alpha})$ for the optimal $\mathbf{W}$ and $\boldsymbol{\alpha}$.

Now, to discover the most influential features and the memory stages in predicting the response y of a trial, (and also to control the model capacity,) we apply regularizers to the model parameters. We first decompose $\mathbf{W}$ into a summation of two components as $\mathbf{W} = \mathbf{U} + \mathbf{V}$ and apply different regularizers to $\mathbf{U}$ and $\mathbf{V}$. The block-wise LASSO, such as the $\ell_{1,2}$ matrix norm, is widely used in multi-task learning or feature selection with group structures, but has not been explored within the GEE setting. To the best of our knowledge, it has not been studied in spatio-temporal analytics how to produce shrinkage effects simultaneously on both features and contingent temporal records through proper regularization. The general $\ell_{1,p}$ matrix norm (Zhang et al., 2010) calculates the sum of the ℓ_p norms of the rows in a matrix. Regularizers based on the $\ell_{1,p}$ norms encourage row sparsity by shrinking the entire rows to have zero entries.

In our parameter matrix $\mathbf{W}$, rows correspond to features and columns correspond to the stages. If we apply the $\ell_{1,2}$ norm to $\mathbf{U}$ (row-wisely), the optimal solution of $\mathbf{U}$ will contain rows with all zero entries. Thus, a selected subset of features in the p stages will be used in the predictive model to predict the trial response. The $\ell_{1,2}$ norm of $\mathbf{V}^{\top}$ (column-wisely) encourages to select among columns of $\mathbf{V}$. If the k-th column of $\mathbf{V}$ contains the largest values in the selected columns, the trial response is most contingent on the features of that k-th stage. Overall, we solve the following optimization problem for the best model parameters $\mathbf{W}$ which is computed as $\mathbf{U} + \mathbf{V}$:

$$\min_{\mathbf{U},\mathbf{V}} \quad \sum_i Dev^{(i)}(\mathbf{U} + \mathbf{V}, \boldsymbol{\alpha}) + \lambda_1 \|\mathbf{U}\|_{1,2} + \lambda_2 \|\mathbf{V}^{\top}\|_{1,2} \qquad (5)$$

where $\mathbf{W}$ in the deviance is simply replaced by $\mathbf{U} + \mathbf{V}$.

The optimization of Eq.(5) is challenging. In general, even solving the GEE formulation is not easy as it involves not only the mean model but also the variance term $\boldsymbol{\Sigma}^{(i)}$. The algorithm that solves the GEE (i.e., the estimating equations) applies the Newton-Raphson method in the iterative reweighted least squares (IRLS) procedure (Fu, 2003) to estimate $\mathbf{w}$ and $\boldsymbol{\Sigma}^{(i)}$. However, this method does not solve any formula that uses regularizers. By modifying the Newton-Raphson method or shooting algorithm (Fu, 2003), it can be extended only to the regularizers that are decomposable into individual parameters w_j. For instance, the ℓ_1 vector norm of $\mathbf{w}$ can be decomposed into the summation of individual $|w_j|$, $j = 1, \cdots, d$. The $\ell_{1,2}$ matrix norm, unfortunately, cannot be decomposed in such a way. Therefore, we have developed an accelerated gradient descent method based on the fast iterative shrinkage-thresholding algorithm (FISTA) (Beck & Teboulle, 2009). As GEE can be proved to be a convex problem in terms of $\mathbf{w}$, our formulation is convex as well in terms of $\mathbf{U}$ and $\mathbf{V}$ with a fixed $\boldsymbol{\alpha}$. Our FISTA algorithm can be proved (omitted) to find the global optimal solution $\mathbf{W}$ of Eq.(5) when a consistent estimator of $\boldsymbol{\alpha}$ is provided by GEE.

3.2. Optimization algorithm

To solve Eq.(5), we design an alternating optimization algorithm that alternates between optimizing two working sets of variables: one set consisting of $\mathbf{U}$ and $\mathbf{V}$ and the other consisting of $\boldsymbol{\alpha}$.

(a) Find $\mathbf{U}$ and $\mathbf{V}$ when $\boldsymbol{\alpha}$ is fixed

When $\boldsymbol{\alpha}$ is fixed, the objective function of Eq.(5), denoted by $f(\mathbf{U}, \mathbf{V})$, is convex with a continuously differentiable part $\ell(\mathbf{U}, \mathbf{V})$ that is the deviance and a non-differential part $R(\mathbf{U}, \mathbf{V})$ that constitutes the two regularizers. We hence have

$$f(\mathbf{U}, \mathbf{V}) = \ell(\mathbf{U}, \mathbf{V}) + R(\mathbf{U}, \mathbf{V}).$$

We develop a FISTA algorithm in the following iterative procedure to find optimal $\mathbf{U}$ and $\mathbf{V}$.

Denote the iterates at the k-th iteration by $\mathbf{U}_k$ and $\mathbf{V}_k$. Let $\nabla_{\mathbf{U}} \ell(\mathbf{U}, \mathbf{V})$, $\nabla_{\mathbf{V}} \ell(\mathbf{U}, \mathbf{V})$ be the partial derivative of $\ell(\mathbf{U}, \mathbf{V})$ with respect to $\mathbf{U}$ and $\mathbf{V}$, respectively, For any given point $(\tilde{\mathbf{U}}, \tilde{\mathbf{V}})$, the following $Q_{L,\tilde{\mathbf{U}},\tilde{\mathbf{V}}}(\mathbf{U}, \mathbf{V})$ is a *well-defined* proximal map for the nonsmooth R

$$\begin{aligned} Q_{L,\tilde{\mathbf{U}},\tilde{\mathbf{V}}}(\mathbf{U}, \mathbf{V}) = {} & \ell(\tilde{\mathbf{U}}, \tilde{\mathbf{V}}) + R(\mathbf{U}, \mathbf{V}) \\ & + < \nabla_{\mathbf{U}} \ell(\tilde{\mathbf{U}}, \tilde{\mathbf{V}}), \mathbf{U} - \tilde{\mathbf{U}} > + \frac{L}{2} \|\mathbf{U} - \tilde{\mathbf{U}}\|_F^2 \\ & + < \nabla_{\mathbf{V}} \ell(\tilde{\mathbf{U}}, \tilde{\mathbf{V}}), \mathbf{V} - \tilde{\mathbf{V}} > + \frac{L}{2} \|\mathbf{V} - \tilde{\mathbf{V}}\|_F^2. \end{aligned}$$

If $\ell(\mathbf{U}, \mathbf{V})$ has Lipschitz continuous gradient with Lipschitz modulus L. Then, according to the Lemma 2.1 in

(Beck & Teboulle, 2009), the inequality

$$f(\mathbf{U},\mathbf{V}) \le Q_{L,\tilde{\mathbf{U}},\tilde{\mathbf{V}}}(\mathbf{U},\mathbf{V}). \tag{6}$$

holds indicating that $Q_{L,\tilde{\mathbf{U}},\tilde{\mathbf{V}}}(\mathbf{U},\mathbf{V})$ is the upper bound of $f(\mathbf{U},\mathbf{V})$.

Starting from an initial point $(\mathbf{U}_0,\mathbf{V}_0)$, we iteratively search for the optimal solution. At each iteration k, we first use the iterates $(\mathbf{U}_{k-1},\mathbf{V}_{k-1})$ and $(\mathbf{U}_{k-2},\mathbf{V}_{k-2})$ to compute (at the first iteration, $(\tilde{\mathbf{U}}_1,\tilde{\mathbf{V}}_1) = (\mathbf{U}_0,\mathbf{V}_0)$)

$$\begin{aligned}\tilde{\mathbf{U}}_k &= \mathbf{U}_{k-1} + \left(\frac{t_{k-1}-1}{t_k}\right)(\mathbf{U}_{k-1}-\mathbf{U}_{k-2})\\ \tilde{\mathbf{V}}_k &= \mathbf{V}_{k-1} + \left(\frac{t_{k-1}-1}{t_k}\right)(\mathbf{V}_{k-1}-\mathbf{V}_{k-2}),\end{aligned} \tag{7}$$

where t_k is a scalar and updated at each iteration as:

$$t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}. \tag{8}$$

Then, we solve the following problem

$$\begin{aligned}\min_{\mathbf{U},\mathbf{V}} \quad & < \nabla_{\mathbf{U}}\ell_k, \mathbf{U}-\tilde{\mathbf{U}}_k > + \frac{L}{2}\|\mathbf{U}-\tilde{\mathbf{U}}_k\|_F^2\\ & + < \nabla_{\mathbf{V}}\ell_k, \mathbf{V}-\tilde{\mathbf{V}}_k > + \frac{L}{2}\|\mathbf{V}-\tilde{\mathbf{V}}_k\|_F^2\\ & + R(\mathbf{U},\mathbf{V})\end{aligned} \tag{9}$$

for a solution $(\mathbf{U}_k,\mathbf{V}_k)$, where $\nabla_{\mathbf{U}}\ell_k$ and $\nabla_{\mathbf{V}}\ell_k$ are respectively the partial derivatives of ℓ computed at $(\tilde{\mathbf{U}}_k,\tilde{\mathbf{V}}_k)$, and L acts as a learning step size.

Since there is no interacting term between $\mathbf{U}$ and $\mathbf{V}$ in Eq.(9), the problem can be decomposed into two separate subproblems as follows:

$$\min_{\mathbf{U}} < \nabla_{\mathbf{U}}\ell_k, \mathbf{U}-\tilde{\mathbf{U}}_k > + \frac{L}{2}\|\mathbf{U}-\tilde{\mathbf{U}}_k\|_F^2 + \lambda_1\|\mathbf{U}\|_{1,2}, \tag{10}$$

$$\min_{\mathbf{V}} < \nabla_{\mathbf{V}}\ell_k, \mathbf{V}-\tilde{\mathbf{V}}_k > + \frac{L}{2}\|\mathbf{V}-\tilde{\mathbf{V}}_k\|_F^2 + \lambda_2\|\mathbf{V}^\top\|_{1,2}. \tag{11}$$

The two subproblems share the same structure and thus can be solved following the same procedure. Hence, we only show how to solve (10) for the best $\mathbf{U}$.

Eq.(10) is equivalent to the following problem

$$\min_{\mathbf{U}} \frac{1}{2}\left\|\mathbf{U} - \left(\tilde{\mathbf{U}}_k - \frac{1}{L}\nabla_{\mathbf{U}}\ell_k\right)\right\|_F^2 + \frac{\lambda_1}{L}\|\mathbf{U}\|_{1,2}$$

after omitting constants, and this problem has a closed-form solution where each row of $\mathbf{U}_k$, $\mathbf{U}^k_{(i,)}$ is:

$$\mathbf{U}^k_{(i,)} = \max\left(0, 1 - \frac{\lambda_1}{L\|\mathbf{P}^{(k)}_{(i,)}\|_2}\right)\mathbf{P}^{(k)}_{(i,)},$$

and $\mathbf{P}^{(k)} = \tilde{\mathbf{U}}_k - \frac{1}{L}\nabla_{\mathbf{U}}\ell_k$. The gradient vector $\nabla_{\mathbf{U}}\ell_k$ (i.e., the gradient of the deviance) can be computed by Eq.(2) with the fixed $\boldsymbol{\alpha}$, i.e.

$$\nabla_{\mathbf{U}}\ell_k = \text{reshape}\left(\sum_{i=1}^{m}\left(\mathbf{D}^{(i)}\right)^\top\left(\mathbf{\Sigma}^{(i)}\right)^{-1}\mathbf{s}_k^{(i)}\right) \tag{12}$$

where $\mathbf{s}_k^{(i)} = \mathbf{y}^{(i)} - \boldsymbol{\mu}^{(i)}$, and $\mu_t^{(i)} = g^{-1}(tr(\mathbf{X}_{(i;t)}^\top(\tilde{\mathbf{U}}_k + \tilde{\mathbf{V}}_k)))$. Given our response variable is binary (classification), the logistic link function will be used and all the parameters can be computed as: $\mathbf{D}^{(i)} = \frac{\partial\boldsymbol{\mu}^{(i)}}{\partial\boldsymbol{\eta}^{(i)}} \times \frac{\partial\boldsymbol{\eta}^{(i)}}{\partial\text{vect}(\tilde{\mathbf{U}}_k)} = \mathbf{A}^{(i)}\left[\text{vect}\left(\mathbf{X}_{(i;1)}\right),\ldots,\text{vect}\left(\mathbf{X}_{(i;n)}\right)\right]^\top$, and $\mathbf{\Sigma}^{(i)} = \left(\mathbf{A}^{(i)}\right)^{1/2}\mathbf{R}(\boldsymbol{\alpha})\left(\mathbf{A}^{(i)}\right)^{1/2}/\phi$ where $\mathbf{A}^{(i)} = \text{diag}\left(\frac{\exp(\eta_t^{(i)})}{\left(1+\exp(\eta_t^{(i)})\right)^2}\right)$.

The calculation of the Lipschitz modulis L can be computational expensive. We therefore follow the similar argument in (Gong et al., 2012) to find a proper approximation L_k at each iteration k starting from $L_0 > 0$. Recall that the Lipschits constant L is defined:

$$L = \max_{\mathbf{W}} \lambda_{\max}\left(\nabla\nabla\ell_{\mathbf{W}}\right)$$

where $\lambda_{\max}(\cdot)$ indicates the maximum singular value of the formula ℓ. Decompose the Hessian matrix $\nabla\nabla\ell_{\mathbf{W}}|_{\mathbf{W}\to 0}$ into $\mathbf{M}^\top\mathbf{M}$ where $\mathbf{M} \in \mathbb{R}^{dp\times q}$ and q is the rank of the Hessian matrix. We can find the upper bound of L by:

$$L \le ||\mathbf{M}||_{\infty,1}||\mathbf{M}^\top||_{\infty,1} \tag{13}$$

where $||\mathbf{M}||_{\infty,1}$ computes the maximum value of the ℓ_1-norm across all rows of $\mathbf{M}$. Intead of using L, we compute the upper bound $\tilde{L}$ in Eq.(13) in our iterations.

Algorithm 1 summarizes the steps for finding optimal $\mathbf{U}$ and $\mathbf{V}$ with fixed $\boldsymbol{\alpha}$.

Algorithm 1 Search for optimal $\mathbf{U}$ and $\mathbf{V}$ with fixed $\boldsymbol{\alpha}$

```
Input: $\mathbf{X}$, $\mathbf{y}$, $\mathbf{\Sigma}$, $\lambda_1$, $\lambda_2$
Output: $\mathbf{U}$, $\mathbf{V}$
1. Set $k$ = 0, compute $\tilde{L}$, and $t_1 = 1$;
2. Solve Eq.(9) to obtain $\mathbf{U}_k$ and $\mathbf{V}_k$.
3. Compute $t_{k+1}$ by Eq.(8).
4. Compute $\tilde{\mathbf{U}}_{k+1}$ and $\tilde{\mathbf{V}}_{k+1}$ by Eq.(7).
5. $k = k + 1$.
Repeat 2 $\sim$ 5 until convergence.
```

(b) Find α when U *and* V are fixed

When $\mathbf{U}$ and $\mathbf{V}$ are fixed, the regularizers no longer appear in the objective of Eq.(5). Eq.(5) is degenerated into just the GEE formula with $\boldsymbol{\alpha}$ as the variables. Hence, $\boldsymbol{\alpha}$ can be estimated via the standard GEE procedure, i.e., from the current Pearson residuals defined by:

$$r_t^{(i)} = \frac{y_t^{(i)} - tr\left(\left(\mathbf{X}_{(i;t)}\right)^\top (\mathbf{U}+\mathbf{V})\right)}{(\sigma_{t,t}^{(i)})^{(1/2)}}.$$

where $\sigma_{t,t}^{(i)}$ is the t-th diagonal entry in the matrix $\mathbf{\Sigma}^{(i)}$ (Liang & Zeger, 1986). The specific estimator of $\boldsymbol{\alpha}$ depends on the choices of $\mathbf{R}(\boldsymbol{\alpha})$. This GEE-based procedure has been shown to find a *consistent* estimate of $\boldsymbol{\alpha}$ (Liang & Zeger, 1986).

Let $N = mn$ be the total number of data records, and $q = dp$ be the practical number of parameters in $\mathbf{W}$. A general approach to estimating $\mathbf{R}$ is given by:

$$R_{j,k} = \sum_i \frac{r_j^{(i)} r_k^{(i)}}{N-q}, \tag{14}$$

for $j = 1, \cdots, n$, and $k = 1, \cdots, n$. In addition, the scaler parameter ϕ in Eq.(3) can be estimated as follows:

$$\phi = (N-q) / \sum_i \sum_t (r_t^{(i)})^2. \tag{15}$$

Algorithm 2 shows the overall procedure for solving (5).

Algorithm 2 Main algorithm - Jointly select features and temporal points

```
Input: $\mathbf{X}$, $\mathbf{y}$, $\lambda_1$, $\lambda_2$
Output: $\mathbf{U}$, $\mathbf{V}$
1. Set $\mathbf{R}(\alpha) = \mathbf{I}$;
2. Solve for $\mathbf{U}$ and $\mathbf{V}$ using Algorithm 1;
3. Estimate $\alpha$ using a proper estimator in (Liang & Zeger, 1986) and compute $\mathbf{R}(\alpha)$ by Eq.(14) and $\phi$ by Eq.(15)).
Repeat 2 $\sim$ 3 until convergence.
```

4. Experimental results

In this section, we discuss the preliminary results we obtained on our EEG trial data. In our study data, schizophrenia (SZ) patients went through three sessions of the Sternberg trials, and healthy normal (HN) members were only included in the first session. There were 90 trials in each session for each individual. However, very few patients participated all sessions and many trial records had missing values or significant level of noise or outliers, for which we had to clean the data carefully. After data cleaning, there were 1131 trials for 14 SZ in session 1, 761 trials for 9 SZ in session 2, and 1191 trials for 14 SZ in session 3. Each patient had 74 to 94 trials, and 83 on average. The rate of incorrect responses for the SZ patients was 27.2%. There were 519 trials for 6 HN participants. Each participant had 82 to 90 trials, and 87 on average. The rate of incorrect responses for HN participants was 14.7%. Note that the current study data contained a limited sample of subjects from the parent study. Additional efforts will be needed to clean and process the full dataset and repeat the analyses reported here.

We validated the proposed approach by comparing it to the most relevant method, which was the GEE (Liang & Zeger, 1986). We experimented with the different correlation structures including exchangeable, tri-diagonal, AR-1 and independent formula. The receiver operating characteristic (ROC) curves were used to evaluate the performance of each resultant classifier and the area under the ROC curve (AUC) was reported in Table 1 (Fawcett, 2006). We separated our analysis for SZ and HN with the hypothesis that SZ patients may use different mechanisms or brain functions to perform memory tasks from those of HN participants. We hence built classifiers to separate trials with correct responses from those with incorrect responses, respectively, for SZ and HN. We then compared the features selected for use in the SZ classifiers and HN classifiers.

For each of the SZ and HN datasets, 1/3 of the records were randomly chosen from every subject to form the test data and the rest of the records were used in training. The hyper-parameters λ_1 and λ_2 in our approach and GEE (one parameter) were tuned in a two-fold cross validation within the training data. In other words, the training records were further split in half: one used to build a classifier with a chosen parameter value from a range of 1 to 10 with a step size 0.1; and the other used to test the resultant classifier. We chose the parameter values that gave the best two-fold cross validation performance, which were $\lambda_1 = 5.9$ and $\lambda_2 = 10$ for SZ and $\lambda_1 = 2$ and $\lambda_2 = 3.1$ for HN.

Table 1 provides the AUC comparison results (shown in percentages) between the two methods and for different datasets and sample correlation assumptions. The results in Table 1 show that our approach outperformed the traditional GEE in almost all comparison scenarios in terms of classification accuracy. Most importantly, our approach was able to select along two dimensions: among the features and among the memory information processing stages. Traditional GEE did not have any shrinkage effect to select features. The advanced version of GEE used in our experiments implemented a ℓ_1 regularizer, so it could select among all 60 features. Because it did not use the spatio-temporal structure of the 60 features, it was unable to model along the different dimensions (locations versus temporal stages).

We noticed that both GEE and our approach performed the best when using independent sample-correlation as-

Table 1. Comparison of AUC values (in percentage) between our approach and the GEE method on both healthy normal and schizophrenia data and for all different assumptions of correlation structures. (ind - independent sample-correlation structure.)

	GEE				Our Approach			
Population	AR(1)	Exchangeable	Tri-diagonal	ind	AR(1)	exchangeable	Tri-diagonal	ind
Healthy Normal (HN)	54.1	52.2	55.5	57.3	55.1	54.9	55.0	68.0
Schizophrenia (SZ)	60.3	55.5	43.6	65.0	62.6	60.0	48.2	66.3

sumption, which was naturally against our intuition because there were multiple trials from a single individual and these trials were expected to correlate. The equi-correlated (exchangeable) assumption assumed that the correlation among all trials was equal and indicated by a constant. Together with AR-1 and Tri-diagonal correlation structures, these assumptions were slightly worse than the independent correlation assumption. However, we also noticed that the trials were not labeled in sequence in our data so the algorithms would not be able to model and distinguish the correlations between consecutive trials from those of far-apart trials. (The trials that an individual performed in a short continuous timeframe may correlate more strongly than trials far apart.)

We include two figures to demonstrate the selected features and stages in the classifiers constructed by our approach. The selected features for SZ patients are shown in Figure 2. The selected features for HN participants are shown in Figure 3. An obvious observation is that the two populations selected quite different features but the most important information processing stages were the same. Some of the selected EEG features replicate those early reports, including upward modulation of γ in SZ patients and engagement of α during encoding and retention periods (Chen et al., 2014; Herrmann et al., 2004).

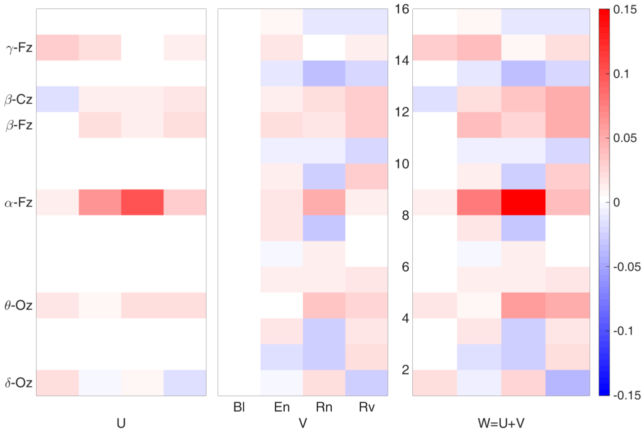

Figure 2. Columns and rows selected by the classifier for separating correct versus incorrect Sternberg trials of SZ patients. Red (blue) color indicates that the corresponding features were positive (negative) predictors of the incorrect response. Features with white color were not used in the classifier.

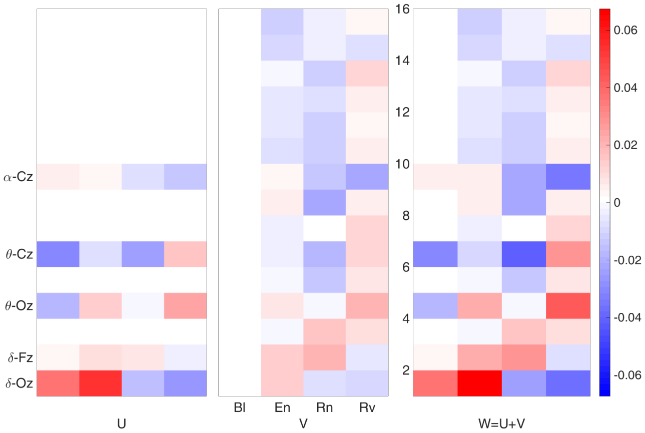

Figure 3. Columns and rows selected by the classifier for separating correct versus incorrect Sternberg trials of HN participants. Red (blue) color indicates that the corresponding features were positive (negative) predictors of the incorrect response. Features with white color were not used in the classifier.

Based on our models, the two groups showed remarkably different patterns, with EEG activity in higher frequency bands during the encoding stage associated with incorrect trial responses in SZ (Figure 2). However, these features were positive predictors of trial accuracy in healthy participants (Figure 3), for whom engagement of low frequency activity was associated with incorrect responses. It appears that the SZ patients used more brain areas in the memory tasks than the HN participants. Frontal γ was previously identified as important for both SZ and HN subjects, but was not selected for HN participants in our new model, which may warrant further investigation. On the other hand, among the selected three stages of both groups, the features during the retention stage tended to receive the largest weights in magnitude on average. All these results will require careful examination in new studies to confirm the validity and replicate on independent samples.

5. Conclusion

We have proposed a new learning formulation for spatio-temporal analytics. Unlike existing methods, the proposed approach can simultaneously determine the temporal contingency and the influential features to predict an outcome of related samples. The model parameter matrix used by the predictive model is computed by the summation of two component matrices: one matrix reflects the selection among features; and the other characterizes the dependency along the temporal line. Moreover, our approach simultaneously models the sample correlations in the longitudinal data while constructing a predictive model. A new accelerated gradient descent algorithm can efficiently solve the related optimization problem.

There are multiple limitations of the current work. The sample size of our study was small, especially the number of healthy normal participants was only 6, for which the validity of selected features may require additional examination when the sample is augmented. It may require a significantly more effort from neuroscientists to interpret and investigate the results stemming from new machine learning approaches. Moreover, our approach can be easily extended to handle more dimensions. For instance, we may organize the data into a three dimensional matrix with one dimension for stages, one dimension for scalp locations and one dimension for frequency bands. Then we split the model parameter matrix into the sum of three components and apply different regularizers to them to emphasize and separate the selection among the different dimensions, which we will leave for future work.

Acknowledgments

This work was supported by a VA grant I21 RX001731 and a NSF grant IIS-1320586 of United States. Jinbo Bi was also supported by NIH grant R01DA037349 and NSF grants DBI-1356655, IIS-1407205 and IIS-1447711.

References

Beck, A. and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *Siam Journal on Imaging Sciences*, 2(1):183–202, 2009. ISSN 1936-4954.

Boonstra, Tjeerd W., Powell, Tamara Y., Mehrkanoon, Saeid, and Breakspear, Michael. Effects of mnemonic load on cortical activity during visual working memory: linking ongoing brain activity with evoked responses. *International journal of psychophysiology : official journal of the International Organization of Psychophysiology*, 89(3):409–418, 2013.

Chen, Chi-Ming A., Stanford, Arielle D., Mao, Xiangling, Abi-Dargham, Anissa, Shungu, Dikoma C., Lisanby, Sarah H., Schroeder, Charles E., and Kegeles, Lawrence S. Gaba level, gamma oscillation, and working memory performance in schizophrenia. *NeuroImage.Clinical*, 4:531–539, 2014.

Fawcett, T. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

Fu, Wenjiang J. Penalized estimating equations. *Biometrics*, 59(1):pp. 126–132, 2003. ISSN 0006341X.

Gong, Pinghua, Ye, Jieping, and Zhang, Changshui. Robust multi-task feature learning. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '12, pp. 895–903, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1462-6. doi: 10.1145/2339530.2339672.

Herrmann, Christoph S., Senkowski, Daniel, and Rottger, Stefan. Phase-locking and amplitude modulations of eeg alpha: Two measures reflect different cognitive processes in a working memory task. *Experimental Psychology*, 51(4):311, 2004.

Klimesch, Wolfgang. EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Research Reviews*, 29(2):169–195, 1999.

Liang, K. Y. and Zeger, S. L. Longitudinal data-analysis using generalized linear-models. *Biometrika*, 73(1):13–22, 1986. ISSN 0006-3444.

May, Arne. Experience-dependent structural plasticity in the adult human brain. *Trends in cognitive sciences*, 15 (10):475–482, 2011.

McCullagh, P. and Nelder, J. A. *Generalized linear models (Second edition)*. London: Chapman & Hall, 1989.

Nahum, Mor, Lee, Hyunkyu, and Merzenich, Michael M. Principles of neuroplasticity-based rehabilitation. *Progress in brain research*, 207:141, 2013.

Olsson, Ulf. *Generalized linear models*, volume 18. 2002.

Wang, Xiao-Jing. Neurophysiological and computational principles of cortical rhythms in cognition. *Physiological Reviews*, 90(3):1195–1268, 2010.

Zhang, Yu, Yeung, Dit-Yan, and Xu, Qian. Probabilistic multi-task feature selection. In *Proceedings of NIPS'10*, pp. 2559–2567, 2010.