ERP Detection Based on Smoothness Priors

Ali Mobaien, Reza Boostani, Mokhtar Mohammadi and Saeid Sanei, *Senior Member, IEEE*

***Abstract*—*Objective:* Detection of event-related potentials (ERPs) in electroencephalography (EEG) is of great interest in the study of brain responses to various stimuli. This is challenging due to the low signal-to-noise ratio of these deflections. To address this problem, a new scheme to detect the ERPs based on smoothness priors is proposed. *Methods:* The problem is considered as a binary hypothesis test and solved using a smooth version of the generalized likelihood ratio test (SGLRT). First, we estimate the parameters of probability density functions from the training data under the Gaussian assumption. Then, these parameters are treated as known values and the unknown ERPs are estimated under the smoothness constraint. The performance of the proposed SGLRT is assessed for ERP detection in post-stimuli EEG recordings of two oddball settings. We compared our method with several powerful methods regarding ERP detection. *Results:* The presented method performs better than the competing algorithms and improves the classification accuracy. *Conclusion:* SGLRT can be employed as a powerful means for different ERP detection schemes. *Significance:* The proposed scheme is opening a new direction in ERP identification which provides better classification results compared to several popular ERP detection methods.**

***Index Terms*—Event-related potentials, P300, smooth signal, detection, generalized likelihood ratio test.**

I. Introduction

EVENT-related potentials (ERPs) are small deflections in the brain activity signals which are related to the attention and perception levels of the brain processes. These small potentials can be evoked using sensory stimuli or even by motion and cognitive events [1]. Detecting and studying ERPs are of great interest in the field of biomedical signal processing, and can be used for medical diagnostics [2] or brain-computer interfaces (BCIs) [3]. Electroencephalography (EEG) is one of the most convenient tools to capture these potentials. EEG signals can be seen as a combination of desired signals (i.e. ERPs), the background EEG (the constant brain waves related to normal physiological activities), and noise (e.g. the measurement noise, eye blinks, and movement artifacts). If the non-ERP components are counted as noise, then the signal-to-noise ratio (SNR) of ERP is low, and therefore, the other sources often dominate the ERP waveform. Thus, detecting ERPs and tracking their parameters through the ongoing EEG are challenging [1]. Hence, new and effective signal processing techniques to more accurately detect and estimate the ERPs become necessary. ERP sequences are composed of several components and each component is related to a specific level of brain process [4]. EEG-ERP frequency domain analyses have indicated that these components have a smooth behavior compared to the wide-band background activities [5], [6]. This a priori on the desired signal structure can be used to improve the conventional ERP detection techniques.

Contribution: In this research, we propose a new scheme to detect ERPs from multichannel EEG signals based on smoothness priors. The main contribution of this work is to mathematically model and embed the smoothness constraint into an ERP detection scheme and provide a unified powerful method. To reach this goal, the problem is considered a binary hypothesis test, and it is solved by a two-step generalized likelihood ratio test (GLRT). In the first step, the GLRT solution, without considering smoothness priors, is obtained. It is shown that the achieved detector tends to be a linear discriminant analysis (LDA) method. In the second step, smoothness priors are added to the problem. Consequently, the solution leads to a smoothing operator that maps the data into a smooth subspace to better estimate the desired signal. Then, this operator is imposed on the GLRT solution and the proposed smooth GLRT (SGLRT) is achieved. Monte Carlo simulations over different ERP datasets show improved detection rates for the new method. The proposed SGLRT mainly relies on a single parameter that specifies the smoothness level of the desired signal, and it can easily be tuned for different ERPs.

The proposed method is evaluated over two ERP components namely P300 and mismatch negativity (MMN), focusing on P300 as one of the most studied ERPs. These components are more pronounced during the oddball paradigm. The subjects are exposed to a series of frequent (non-target) and infrequent (target) visual or auditory stimuli in such settings. Research findings indicate that when a subject is exposed to an infrequent stimulus, there is a small deflection in his/her EEG [5], [7]. The brain response to a rare stimulus can be used as a cue to control a BCI system. Generally, the studies on ERP signals can be divided into two categories. The first category focuses on ERP detection and tries to solve a binary classification problem to determine the target stimuli [3], suitable for BCI systems (e.g. a speller) [8]. The second category focuses on the estimation of ERP subcomponents and tracking their parameters (such as amplitude, latency, and width). In these studies, the target trials are usually known and the inter-trial variability of ERPs is of interest [9], suitable for diagnostic purposes [2]. The scope of this contribution is the detection of ERPs. According to the review article by Philip et al. [8], the most frequently employed classification methods for ERP (P300) elicitation are LDA, support vector machines (SVM), distance-based classifiers, artificial neural networks (ANN) and ensemble classifiers. Among recent studies regarding ERPs, the following works

Manuscript received August 20, 2022.

A. Mobaien and R. Boostani are with the Department of Electrical and Computer Engineering, Shiraz University, Shiraz, Iran (e-mail: a.mobaien@shirazu.ac.ir; boostani@shirazu.ac.ir).

M. Mohammadi is with the Department of Information Technology, College of Engineering and Computer Science, Lebanese French University, KR-Iraq (e-mail: mokhtar@lfu.edu.krd).

S. Sanei is with the School of Science and Technology, Nottingham Trent University, Nottingham NG11 8NS, U.K. (e-mail: saeid.sanei@ntu.ac.uk).

can be named. Yang et al. [10] proposed a simultaneous spatiotemporal equalization (STE) technique for whitening EEG data based on a multivariate autoregressive (MVAR) model, followed by a correlation detection (CD) method (STE-CD) for P300 detection. Campos et al. [11] developed an efficient algorithm to extract the underlying waveforms of ERPs using a three-step spatial filtering method. These waveforms are called principal ERPs (pERPs), and the method to represent ERP signals based on the pERPs is called pERP-reduction (pERP-RED). Oralhan [12] proposed a 3-dimensional convolutional neural network (3D-CNN) for P300 detection, and Zhang et al. [13] presented a deep neural network consisting of parallel spatial and temporal units (named STNN) for P300 detection. Most of these studies consider the ERP smoothness by roughly applying a low pass filter to the data. However, in this study, the smoothness is a built-in property of the presented method, which we believe, can be beneficial by enhancing the estimation of ERPs.

SGLRT is compared with LDA, SVM, STE-CD, pERP-RED, and STNN. The methods are evaluated over three publicly available datasets on P300 and MMN. The results demonstrate that the presented method can outperform the competing algorithms in different experimental setups and metrics. The rest of the article is as follows. In section II, the proposed method is rigorously investigated. In section III, the evaluation schemes, parameters selection, and the comparative results of the different methods are presented. Section IV includes the discussion on the competing methods. Finally, the paper is concluded in section V.

II. Method

Here, a two-step GLRT approach is developed to detect the activity of an ERP component in the background EEG. This method is usually used when the probability density function (PDF) parameters of observations are unknown. For a binary hypothesis test, GLRT is extracted as the likelihood ratio of observations, where the PDF unknown parameters are calculated using maximum likelihood (ML) estimation under the corresponding hypothesis. Then, the likelihood ratio is compared with a predefined threshold and the hypothesis is determined. The threshold is calculated based on a desirable false alarm rate [14].

A. Proposed Detection Scheme

The framework for the proposed SGLRT consists of two steps. First, the detection problem is solved by a GLRT approach without smoothness priors. Second, the smoothness priors are added to the problem and the solution is found.

1) GLRT Solution without Smoothness Priors: Here, we should decide whether the post-stimulus EEG recordings include ERP. This is to separate the target and non-target trials while the non-target trials are considered as the background activity. Furthermore, the background activity is modeled as a multivariate Gaussian noise with an unknown structure. The mixing model of signal and noise is assumed to be additive, widely used in the literature [15].

Let $X \in \mathbb{R}^{J\times N}$ be a post-stimulus data matrix, where J is the number of samples in time and N is the number of recorded channels. The data matrix is then vectorized by concatenating its columns, leading to $JN\times 1$ column vector $\underline{x}$. Each vectorized or matrix form of post-stimulus observations is known as a trial. Let Class c_0 be a collection of trials excluding ERP, and class c_1 be a collection of trials including ERP.

Problem Formulation: Consider $\underline{x}_0$ as a vector of observations to be tested, and $\underline{x}_k^i$ as k'th training trial of class $i \in \{0,1\}$. Here, the problem can be seen as a binary hypothesis test,

$$H_0 : \underline{x}_0 = \underline{v}_0 \quad \text{vs.} \quad H_1 : \underline{x}_0 = \underline{p} + \underline{v}_0$$

and for both hypotheses, we have the training trials as

$$\begin{cases} \underline{x}_k^0 = \underline{v}_k^0, \; k = 1,2,...,K_0 \\ \underline{x}_k^1 = \underline{p} + \underline{v}_k^1, \; k = 1,2,...,K_1 \end{cases}$$

where K_0 and K_1 are the numbers of available training data samples for the classes c_0 and c_1, respectively. In the above representation, $\underline{v}_0$, $\underline{v}_k^0$ ($k = 1,2,...,K_0$) and $\underline{v}_k^1$ ($k = 1,2,\ldots,K_1$) are $JN \times 1$ column vectors of noise which are assumed to be drawn from a multivariate normal distribution with unknown mean and covariance. These vectors are assumed to be independent and identically distributed (i.i.d.). However, the samples within each vector are correlated as represented by the covariance matrix. The noise vectors are assumed to be a combination of background EEG and other noise sources (e.g. the noise of the recording device). $\underline{p}$ is a column vector of length JN which indicates the presence of ERP, also, $\underline{p}$ is a smooth sequence.

Formation of the likelihood ratio: To derive GLRT, the likelihood ratio of PDFs under two hypotheses should be formed. To this aim, the test and train data samples are concatenated in a long vector and the PDF parameters of all data under each hypothesis are found. Since we have assumed that the trials are independent, there are only three parameters to be estimated under each hypothesis, mean of the data that does not contain ERP, mean of the data that contains ERP, and covariance matrix of normal distribution. Although these three parameters are the same for the two hypotheses, their estimations are different under each hypothesis. Here, the estimation of $\underline{p}$ is embedded in the estimation of the mean.

Let $\underline{x}_{all} = [\underline{x}_0^T, \underline{x}_1^{0^T}, \underline{x}_2^{0^T}, \ldots, \underline{x}_{K_0}^{0\;T}, \underline{x}_1^{1^T}, \underline{x}_2^{1^T}, \ldots, \underline{x}_{K_1}^{1\;T}]^T$ ($[.]^T$ denotes transpose operation). Then, PDF of observations under H_j, $j \in \{0,1\}$ is,

$$\begin{aligned} f\left(\underline{x}_{all}; \underline{\mu}_{0j}, \underline{\mu}_{1j}, R_j, H_j\right) = f\left(\underline{x}_0; \underline{\mu}_{jj}, R_j, H_j\right) \times \\ \prod_{k=1}^{K_0} f\left(\underline{x}_k^0; \underline{\mu}_{0j}, R_j, H_j\right) \times \prod_{k=1}^{K_1} f\left(\underline{x}_k^1; \underline{\mu}_{1j}, R_j, H_j\right) = \\ \alpha_j \exp\Big(-\tfrac{1}{2}\Big[(\underline{x}_0 - \underline{\mu}_{jj})^T R_j^{-1} (\underline{x}_0 - \underline{\mu}_{jj}) \\ + \textstyle\sum_{k=1}^{K_0} (\underline{x}_k^0 - \underline{\mu}_{0j})^T R_j^{-1} (\underline{x}_k^0 - \underline{\mu}_{0j}) \\ + \textstyle\sum_{k=1}^{K_1} (\underline{x}_k^1 - \underline{\mu}_{1j})^T R_j^{-1} (\underline{x}_k^1 - \underline{\mu}_{1j})\Big]\Big) \end{aligned} \quad (1)$$

where $\alpha_j = (1/(\sqrt{2\pi}^{JN} |R_j|^{1/2}))^{1+K_0+K_1}$ is the normalization coefficient of PDF function ($|.|$ denotes determinant operator), and $\underline{\mu}_{0j}$, $\underline{\mu}_{1j}$ and R_j are mean of the data that does not contain ERP, mean of the data that contains ERP, and the covariance matrix of normal distribution, respectively. For ML estimation of unknown parameters under H_j, we have:

$$\hat{\underline{\mu}}_{0j} = \frac{1}{K_0 + 1 - j}\left((1-j)\underline{x}_0 + \sum_{k=1}^{K_0} \underline{x}_k^0\right) \tag{2}$$

$$\hat{\underline{\mu}}_{1j} = \frac{1}{K_1 + j}\left(j\underline{x}_0 + \sum_{k=1}^{K_1} \underline{x}_k^1\right) \tag{3}$$

$$\begin{aligned}\hat{R}_j = \tfrac{1}{K_0+K_1-1}\Big((\underline{x}_0 - \underline{\mu}_{jj})(\underline{x}_0 - \underline{\mu}_{jj})^T \\ + \textstyle\sum_{k=1}^{K_0}(\underline{x}_k^0 - \underline{\mu}_{0j})(\underline{x}_k^0 - \underline{\mu}_{0j})^T \\ + \textstyle\sum_{k=1}^{K_1}(\underline{x}_k^1 - \underline{\mu}_{1j})(\underline{x}_k^1 - \underline{\mu}_{1j})^T\Big)\end{aligned} \tag{4}$$

Here, $\hat{R}_j$ is in fact the scaled and unbiased version of ML estimation, known as sample covariance matrix (SCM). Note that, the mean estimators are efficient and SCM is uniformly minimum variance unbiased (UMVU). Therefore, they give the best possible estimations from the available data samples. SCM is unbiased regardless of the data distribution. However, if the normality assumption is not met, it may not be UMVU [16]. By likelihood ratio and using the estimated parameters, GLRT can now be extracted as:

$$L_{GLRT}(\underline{x}_{all}) = \frac{f(\underline{x}_{all}; \hat{\underline{\mu}}_{01}, \hat{\underline{\mu}}_{11}, \hat{R}_1, H_1)}{f(\underline{x}_{all}; \hat{\underline{\mu}}_{00}, \hat{\underline{\mu}}_{10}, \hat{R}_0, H_0)} \tag{5}$$

The above equation is the exact GLRT solution for our hypothesis test. However, it can be simplified by excluding $\underline{x}_0$ for estimation of unknown parameters. It can be shown that under the Gaussian assumption, the estimation of mean and covariance are unbiased and consistent (i.e. by increasing the number of training samples, their estimations converge toward the true values) [16]. Hence, for large datasets, $\underline{x}_0$ can be ignored. In this case, $\hat{\underline{\mu}}_{00} = \hat{\underline{\mu}}_{01} = \hat{\underline{\mu}}_0$, $\hat{\underline{\mu}}_{10} = \hat{\underline{\mu}}_{11} = \hat{\underline{\mu}}_1$ and $\hat{R}_0 = \hat{R}_1 = \hat{R}$, and we have:

$$\hat{\underline{\mu}}_i = \frac{1}{K_i}\sum_{k=1}^{K_i} \underline{x}_k^i \tag{6}$$

$$\begin{aligned}\hat{R} = \tfrac{1}{K_0+K_1-2}\Big(\textstyle\sum_{k=1}^{K_0}(\underline{x}_k^0 - \hat{\underline{\mu}}_0)(\underline{x}_k^0 - \hat{\underline{\mu}}_0)^T \\ + \textstyle\sum_{k=1}^{K_1}(\underline{x}_k^1 - \hat{\underline{\mu}}_1)(\underline{x}_k^1 - \hat{\underline{\mu}}_1)^T\Big)\end{aligned} \tag{7}$$

where $i \in \{0, 1\}$. Now, using the simplified estimations to form the likelihood ratio and taking the natural logarithm, GLRT is achieved as,

$$\begin{aligned}L_{GLRT}(\underline{x}_0) = \Big[\tfrac{1}{2}(\underline{x}_0 - \hat{\underline{\mu}}_0)^T \hat{R}^{-1}(\underline{x}_0 - \hat{\underline{\mu}}_0) \\ -\tfrac{1}{2}(\underline{x}_0 - \hat{\underline{\mu}}_1)^T \hat{R}^{-1}(\underline{x}_0 - \hat{\underline{\mu}}_1)\Big] \underset{H_0}{\overset{H_1}{\gtrless}} \eta_{GLRT}\end{aligned} \tag{8}$$

In the above equation, the weighted distance between the trial under test (i.e. $\underline{x}_0$) and the mean of each class is calculated, then, their difference is compared with a threshold and $\underline{x}_0$ is assigned to the closer class. The representation of GLRT in (8) is equal to an LDA approach [17]. Note that LDA is suboptimal to the Bayes classifier when the distributions are Gaussian. In the Bayes classifier, a priori information (the probability of each class) and loss coefficients are used to form a fixed threshold. To designate each test data category, the likelihood ratio is compared with this threshold. Here, η_{GLRT} is considered as a floating threshold. However, it can be set by considering a priori information and loss coefficients. For example, if π_i is the occurrence probability of hypothesis $i \in \{0, 1\}$ and l_{ij}, $(i, j) \in \{0, 1\}$ is the loss of selecting H_i in case of H_j, then an appropriate value for the threshold is $\eta_{GLRT} = \ln((l_{10} - l_{00})\pi_0/(l_{01} - l_{11})\pi_1)$ [14].

2) SGLRT - Improving GLRT by Exploiting ERPs Smoothness: Here, it is proposed to use a smooth estimation of the desired signal from observation vector $\underline{x}_0$ in GLRT instead of $\underline{x}_0$ itself. This idea comes from the fact that ERPs have slow fluctuations compared to the background EEG. Furthermore, to test each incoming data vector $\underline{x}_0$, it is compared with the mean signals of each class which are achieved by temporal averaging. Hence, we suggest mapping $\underline{x}_0$ into a smooth subspace and increasing the SNR by eliminating high frequency and noisy contents. Therefore, the EEG under test is pulled closer to its corresponding class mean.

Smooth signal PDF: It is assumed that $\underline{x}_0 = \underline{s} + \underline{n}$, where $\underline{s}$ is an unknown smooth template and $\underline{n}$ is a noise component covering the corresponding frequency band and is sampled from a zero-mean Gaussian distribution with known covariance matrix $\hat{R}$. The smooth template includes the ERP component under H_1 and excludes it under H_0. In either case, it can be estimated from $\underline{x}_0$ under smoothness constraint. Here also an ML estimation approach is employed. Note that in this representation, $\underline{s}$ is the unknown mean of the multivariate normal distribution that $\underline{x}_0$ comes from, and we are going to estimate this mean by using only one trial. It is straightforward to show that without smoothness constraint $\hat{\underline{s}} = \underline{x}_0$.

Consider $\underline{s} = \Phi\underline{s}'$, where Φ is a structural matrix chosen to form $\underline{s}$ based on $\underline{s}'$. Here, $\underline{s}'$ is a template formed by the concatenation of the temporal vectors from the main sources contributing to the ERP components in the source space. Φ gives the freedom to form a convenient structure for $\underline{s}$ based on the underlying sources of ERP. For example, assume that the P300 is best recorded in two positions on the scalp (one from the frontal region and the other from the central). In this situation, the concatenation of P300 sources can be modeled as $\underline{s}'$, and these sources can be projected to the channels of interest by a linear combination of their time samples through Φ. This example can be extended to more independent sources over the scalp (see Appendix B for more details). Now, the PDF of each observation vector $\underline{x}_0$ can be modeled as:

$$f(\underline{x}_0; \Phi\underline{s}', \hat{R}) = \beta \exp\left(\frac{-1}{2}(\underline{x}_0 - \Phi\underline{s}')^T \hat{R}^{-1}(\underline{x}_0 - \Phi\underline{s}')\right) \tag{9}$$

where $\beta = 1/(\sqrt{2\pi}^{JN}|\hat{R}|^{1/2})$ is the normalization factor of PDF function. In the above representation, $\underline{s}'$ is the unknown smooth template to be estimated.

Desired signal estimation under smoothness constraint: In modeling the smoothness constraint, we used the fact that the jth derivative of a smooth sequence is bounded and its norm is smaller than a threshold [18], [19]. Here, a matrix form of the discrete approximation of the differentiation operator is adopted. Let d_j be the finite impulse response approximation of length M for the jth order derivative operator. For $j > 2$, this operator can be achieved iteratively by $d_j = d_{j-1} * d_1$, where $d_1 = [1, -1]$ and "$*$" indicates the convolution operator. For a signal length $L > M$, $D_j \in \mathbb{R}^{(L+M-1)\times L}$ is defined as a Toeplitz matrix form of d_j (the derivative operator $d_j = [d_{j1}, d_{j2}, \ldots, d_{jM}]$ is shifted along rows of D_j).

$$D_j^T = \begin{bmatrix} d_{j1} & d_{j2} & \cdots & d_{jM} & 0 & \cdots & 0 \\ 0 & d_{j1} & d_{j2} & \cdots & d_{jM} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & d_{j1} & d_{j2} & \cdots & d_{jM} \end{bmatrix}$$

The left multiplication of D_j by any column-wise signal of length L is equivalent to the jth order derivative approximation of the signal [19].

The ML estimation of $\underline{s}'$ is equal to the following optimization problem which is regularized by smoothness priors.

$$\hat{\underline{s}'} = \arg\min_{\underline{s}'} (\underline{x}_0 - \Phi\underline{s}')^T \hat{R}^{-1} (\underline{x}_0 - \Phi\underline{s}'), \text{ s.t. } \|D_j \Phi \underline{s}'\|^2 \leq \delta^2 \tag{10}$$

Here δ^2 indicates the smoothness bound. The problem is a convex optimization problem with a quadratic constraint and the Lagrangian form of the problem is:

$$\hat{\underline{s}'} = \arg\min_{\underline{s}'} \left\{ (\underline{x}_0 - \Phi\underline{s}')^T \hat{R}^{-1} (\underline{x}_0 - \Phi\underline{s}') + \lambda \|D_j \Phi \underline{s}'\|^2 \right\} \tag{11}$$

where $\lambda \geq 0$ is the Lagrange coefficient. This problem is in the form of a constrained weighted least square error (CWLSE) problem [20] and its solution is obtained by:

$$\hat{\underline{s}'} = (\Phi^T \hat{R}^{-1} \Phi + \lambda \Phi^T D_j^T D_j \Phi)^{-1} \Phi^T \hat{R}^{-1} \underline{x}_0 \tag{12}$$

and $\hat{\underline{s}} = \Phi\hat{\underline{s}'}$. The Lagrange coefficient can be obtained in two ways. When the smoothness bound is presumed or known, an optimal λ can be calculated using singular value decomposition (SVD) [19], [20], and when δ^2 is unknown, methods such as the L-curve are used [19], [21]. Let $\Psi = \Phi(\Phi^T \hat{R}^{-1} \Phi + \lambda \Phi^T D_j^T D_j \Phi)^{-1} \Phi^T \hat{R}^{-1}$, then, Ψ can be seen as a projection matrix which maps $\underline{x}_0$ into a smooth subspace and pulls it in a direction to minimize the CWLSE. By increasing λ from 0, the smoothness of the estimated signal increases. Fig. 1b shows the estimated signal $\hat{\underline{s}}$ from $\underline{x}_0$ for different values of λ, $\Phi = I_{JN}$ (an identity matrix of JN dimension) and derivative order $j = 2$. In this work, the differentiation operator is adopted to model the smoothness constraint. In practice, this operator may be replaced by other operators that can quantify signal fluctuations (e.g. weighted differentiators) since eventually, it is the value of λ that controls the smoothness level of the estimated ERP.

Imposing smoothness priors on GLRT to achieve SGLRT: Note that it is necessary to map the mean vectors into the same space as data is mapped. This guarantees that the weighted distance is measured in a common space. Now, the proposed SGLRT method can be obtained by replacing $(\underline{x}_0 - \hat{\underline{\mu}}_i)$ by $\Psi(\underline{x}_0 - \hat{\underline{\mu}}_i)$, $i \in \{0, 1\}$ in (8), and performing some simplifications,

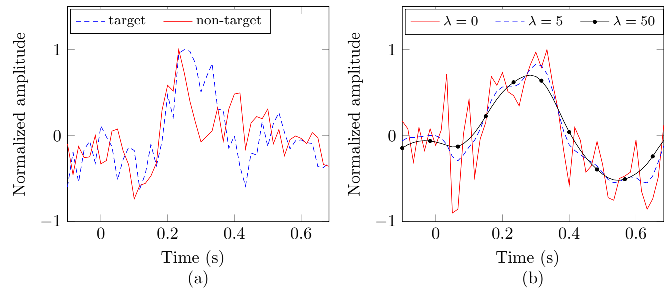

Fig. 1. Selected ERP signals from dataset II of BCI competition III, subject B and channel C_z. (a) A target versus non-target trial averaged over 15 repetitions. From the target trial, the P300 component can be identified as a smooth positive signal centered around 300-400ms post-stimuli. From the non-target trial, the ERP components contributing to attention are observable. (b) Estimated ERP from a target trial for different values of λ.

$$L_{SGLRT}(\underline{x}_0) = \underline{x}_0^T \Psi^T \hat{R}^{-1} \Psi \hat{\underline{p}} + c \underset{H_0}{\overset{H_1}{\gtrless}} \eta_{SGLRT} \tag{13}$$

where $\hat{\underline{p}} = (\hat{\underline{\mu}}_1 - \hat{\underline{\mu}}_0)$ is an estimation of ERP component from training data, and $c = (1/2)(\hat{\underline{\mu}}_0^T \Psi^T \hat{R}^{-1} \Psi \hat{\underline{\mu}}_0 - \hat{\underline{\mu}}_1^T \Psi^T \hat{R}^{-1} \Psi \hat{\underline{\mu}}_1)$ is a constant value. Equation (13) calculates a weighted correlation of observation vector $\underline{x}_0$ and $\hat{\underline{p}}$ in a smooth subspace to detect the presence of ERP. One may consider that $(\underline{x}_0 - \hat{\underline{\mu}}_i)$ is in fact an estimation of centralized (zero mean) noise under H_i. Therefore, $\Psi(\underline{x}_0 - \hat{\underline{\mu}}_i)$ can be seen as the projection of noise into a smooth subspace that increases SNR. Here, the SGLRT algorithm to detect ERPs is summarized in Algorithm 1.

Algorithm 1 SGLRT algorithm to detect ERPs

Training Phase

1: Compute $\hat{\underline{\mu}}_0$, $\hat{\underline{\mu}}_1$, and $\hat{R}$ based on (6) and (7).

2: Compute $\Psi = \Phi(\Phi^T \hat{R}^{-1} \Phi + \lambda \Phi^T D_j^T D_j \Phi)^{-1} \Phi^T \hat{R}^{-1}$.

3: Compute $\hat{\underline{p}} = (\hat{\underline{\mu}}_1 - \hat{\underline{\mu}}_0)$ and $c = (1/2)(\hat{\underline{\mu}}_0^T \Psi^T \hat{R}^{-1} \Psi \hat{\underline{\mu}}_0 - \hat{\underline{\mu}}_1^T \Psi^T \hat{R}^{-1} \Psi \hat{\underline{\mu}}_1)$.

Test Phase

1: Calculate L_{SGLRT} for input data $\underline{x}_0$ based on (13).

2: If $L_{SGLRT}(\underline{x}_0) > \eta_{SGLRT}$, $\underline{x}_0$ contains the desired ERP, otherwise, it does not.

The proposed framework could be simply extended to multiple hypothesis testing if the signal part of each hypothesis is smooth compared to noise. In a multiple decision task, the likelihood ratio of different pairs of hypotheses is formed, and by comparing all the pairs, the most probable one is selected (this is also the solution for optimal Bayes classifier if the loss functions are assumed uniform) [14]. Consequently, the decision-making scheme is very similar to the binary test. Hence, the smoothness priors can be simply applied to each pair of hypotheses to increase the SNR level.

B. Validity of the Assumptions

An underlying assumption of this study is the smoothness of ERPs compared to background activities. This is supported by several studies. For instance, consider P300 which consists of two subcomponents namely p3a and p3b. p3a arrives first and is mostly related to attention while p3b is mostly related to memory update. In an extensive review article by Polich [5], the strong relationship between the theta (4-8Hz) and alpha (8-12Hz) band activities of EEG signals and the P300 subcomponents are discussed. In another study by Spencer and Polich [22], spectral analysis of the p3b from an auditory oddball task showed a strong increase in the theta band activity. Furthermore, other studies reported alpha suppression observed during sensory stimuli and also cognitive tasks related to attention and memory [6]. This phenomenon is called alpha event-related desynchronization (ERD). Alpha ERD along with delta (1-4Hz) and theta-range event-related synchronization (ERS) reveals an energy shift from higher toward lower frequencies when an ERP is elicited [5]. This indicates the slow variations of induced ERPs compared to wide-band (0-60Hz) background activities.

In the method derivation, non-target trials are considered background activity. This may be questionable as we know there are also ERP components related to attention in response to both target and non-target stimuli (e.g. p3a, see Fig. 1a). From another perspective, ERPs can be categorized into two general classes, endogenous and exogenous responses. Exogenous responses relate to the processing of external stimuli by the sensory system and can not be prevented by intention. However, endogenous responses are considered as a result of information processing related to the paradigm [23]. From this, one may conclude that in any designed experiment for emerging ERPs through sensory stimuli, there are exogenous responses in non-target trials (for instance, visual evoked potential in visual P300 speller paradigms). Note that in the proposed SGLRT, the exogenous responses are hidden in the class means. This means the proposed SGLRT can be used either for separating the target from non-target trials or for separating EEG data containing ERPs from EEG data not containing ERPs at all. It is also worth mentioning that (13) is a weighted correlation between the observation vector $\underline{x}_0$ and the difference wave in $\underline{\hat{p}}$, and in the difference wave any exogenous responses are canceled out. Furthermore, the background activity is modeled as Gaussian noise. This choice is based on the idea that the background activity results from the superposition of the signals from numerous neurons which are quasi-randomly activated. Based on the central limit theorem, its distribution tends toward a Gaussian function. This assumption is also tested using Henze-Zirkler's and Mardia's multivariate normality tests [24] on the employed datasets. For P300 data, the first test approves Gaussian assumption at significance level $p = 0.05$, while the other approves the skewness test. However, for MMN data, only Mardia's skewness test is fulfilled.

In the derivation of the GLRT solution, it is mentioned that the test trial can be excluded from the estimation of PDF parameters if the number of training trials is large enough. For a specific dataset, the sufficient number of data can be investigated using a t-test on the estimated parameters, including and excluding the test trial. When there are no meaningful changes in the estimated parameters by increasing the data size from a certain point, the sufficient number of data is found. For the studied datasets in this work, using more than twenty trials is enough to approve the assumption at $p = 0.05$.

III. Method Evaluation

A. Data

1) P300: The presented method is investigated using two publicly available P300 datasets from [25] and [26] (dataset II of BCI competition III). The first dataset consists of ten subjects while trying to spell 10 (5 for training and 5 for testing) characters. Subjects were exposed to a 6×6 matrix of characters while focusing on the desired one. The third to tenth subjects (S3, S4, . . . , S10) used the row-column speller scheme where all rows and columns of this matrix were successively and randomly intensified for 100ms, followed by the blank matrix for 60ms. Two out of 12 intensifications (one row and one column) contained the desired character. This led to 20 target and 100 non-target trials. Each stimulus was repeated 15 times. 8 channels captured EEG signals at 256Hz. In the second dataset, data was recorded from two subjects (SA and SB) while trying to spell 185 (85 for training and 100 for testing) characters. Again, subjects used the row-column speller with 100ms intensification of rows/columns followed by the blank matrix for 75ms. This led to 370 target and 1850 non-target trials. Each stimulus was repeated 15 times. 64 channels captured EEG signals which then passed through a 0.1 to 60Hz bandpass filter, and were finally digitized at 240Hz.

2) MMN: This dataset is obtained from a publicly released project called SPM [27], [28]. It is a 128-channel single-subject EEG dataset acquired from an auditory oddball paradigm. There are 480 non-target trials and 120 target trials in the dataset. SPM12 software is used to preprocess and segment the data. The data are passed through a 0.1 to 60Hz bandpass filter and down-sampled to 240Hz.

B. Statistical Methods and Metrics

The proposed SGLRT is evaluated and compared with ~~the~~ other methods based on three metrics: accuracy, F1 score, and Kappa coefficient. Four experiments are performed for method evaluation. In the first experiment, SGLRT is compared with LDA, SVM, STE-CD, and pERP-RED (the represented data by the pERP-RED are fed to an LDA classifier) in terms of classification accuracy. In this regard, a 10-fold cross-validation approach is applied to two equal-size groups of randomly chosen trials where one contains the ERP component and the other ~~one~~ does not. Training sets of different sizes are used to perform this task, and for each case, at least four random and non-overlapping groups of data are generated. The final results are the averages over the results of these groups. Furthermore, to see if the improvements made by SGLRT are statistically meaningful, the significance of its results is also tested by paired t-tests. To this aim, SGLRT is considered as the reference method (noted by "ref."), and the p-values that indicate the significance of its results compared with the competing methods are reported (see Tables I and II where the p-values are reported in parentheses). The competing methods are also evaluated in different SNRs using the P300 datasets. For this, the trials are averaged over the first k_r repetitions which increases SNR (or decreases noise level) approximately k_r times. Moreover, finding a proper λ value and its effect on the classification performance is investigated.

In the second experiment, F1 scores and Kappa coefficients of the competing methods are obtained for the three datasets in different repetitions. The results include all subjects within each dataset. For P300, training and testing sets are chosen based on the description of each dataset. For MMN, the first half of the samples are used for training and the rest for testing. Moreover, to see the performance of SGLRT and its competitors in situations with a large number of different data, an inter-subject experiment is carried out on the first P300 dataset. In this experiment, the methods are tested using the data of each subject while they are trained by the data of the other seven subjects. The results for all repetitions (i.e. $k_r = 1$ to 15) are accumulated and the corresponding F1 scores and Kappa coefficients are reported.

There are two objectives for performing the third experiment. Firstly, to evaluate the presented method when the number of EEG channels is not limited, and to compare it with the methods that do not apply this limitation. Secondly, to compare the method with a recent deep learning structure. To these goals, SGLRT is compared with SVM, and STNN over the second P300 dataset with the same criterion as in [13] (i.e. classification accuracy over the data testing set when the method is trained by the training set samples, using all 64 channels).

The last experiment is dedicated to the effect of structural matrix Φ on the performance of SGLRT. In this experiment, the classification accuracy of letters (CAL) is obtained for SA and SB of the second P300 dataset using all 64 channels. To this aim, after model training by the training set, L_{SGLRT} is calculated for rows and columns data of each letter in the test set, and this letter is predicted by the intersection of the row and the column with the highest score. Furthermore, the data transmission capacity of SGLRT, often called as information transfer rate (ITR) [8], [12], is reported for this experiment.

C. Data Preparation and Parameter Selection

For the P300 dataset, trials are extracted using 700ms windows of post-stimuli signals, and for the MMN dataset, suggested by [28], windows from -100ms to 400ms regarding stimuli are used. In general, the analyzing window should be long enough to include the complete ERP component, although the start and finish points can slightly be altered. For the first two experiments the P300 data from eight channels (F_z, C_z, P_3, P_z, P_4, PO_7, PO_8, and O_z) are used since they are the most P300-related informative channels [25]. For MMN data, ten channels from the scalp EEG (every 14 channel from the first one) including C_{21} (corresponding to F_z [28]) are considered. For LDA, SVM, and STE-CD, the data are

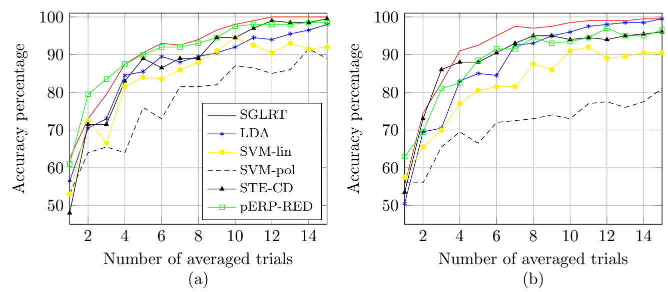

Fig. 2. The performance of the competing methods in different SNRs, and sample size 20. (a) P300-S4 data and (b) P300-S8 data.

bandpass filtered, using a 0.1 to 15Hz Butterworth filter of order 4. However, the raw data are fed to SGLRT and pERP-RED as they benefit from built-in noise reduction procedures. Finally, the P300 and MMN data are down-sampled at rates $1/4$ and $1/3$ (resulting in 45, 42, and 40 temporal samples per channel for the three described datasets), and the trials are normalized to have zero mean and unit variance in each channel.

Here, a list of selected parameters in this study is presented. For SGLRT, η_{SGLRT} is set to zero (the same as the LDA, STE-CD, and pERP-RED threshold). $\lambda = 5$ and 2 are used for P300 and MMN datasets, and the derivative order is set to 2. The structural matrix $\Phi = I_{JN}$ (except for the fourth experiment) implies that a separate ERP source for each recording channel has been considered (see Appendix B). The pERP-RED-based classifier is implemented in R using the pERPred package [11]. Here, the set of selected parameters for pERP-RED is shown as $\{N_p, V_p\}$ where N_p is the number of pERPs and V_p is the percent variation used in the electrode PCA step. The set of chosen parameters is $\{4, 85\}$ for P300 data, and $\{3, 80\}$ for MMN data. The MATLAB platform is used for the implementation of other methods. To realize the SVM classifier, the "svmtrain" function is employed while using the least square method as its optimization technique, and a linear (SVM-lin), also an order three polynomial (SVM-pol) function as its kernels. For the realization of STE-CD, the Arfit package [29] is used. The minimum and maximum orders of the MVAR model are set to 4 and 16, respectively (suggested by [10]). For each case, the optimizer of Schwarz's Bayesian criterion is chosen as the best order. Note that, for each method, different settings have been examined through 10-fold cross-validations for several sample sizes, and the best set of parameters has been chosen based on paired t-tests of the results.

D. Results

1) First Experiment: Fig. 2 shows the performance of competing methods in different SNRs (i.e. different k_r values). The results have been presented for P300-S4 and S8 data, and a sample size of 20. The proposed method provides higher accuracy almost in all SNR levels (except for S4 and $k_r < 5$, where pERP-RED performs slightly better). By increasing the SNR, the performance of all methods increases, as expected. Table I indicates the averaged performance (over different

TABLE I

MEAN ACCURACY (P-VALUE ≤) OF THE COMPETING METHODS FOR DIFFERENT DATASETS IN FIVE SAMPLE SIZES

Data	Method	Size of Training Set				
		30	60	120	240	360
P300, SA	SGLRT	77.1 (ref.)	78.9 (ref.)	80.9 (ref.)	82.3 (ref.)	82.5 (ref.)
	LDA	74.1 (0.02)	76.3 (0.01)	78.6 (0.01)	81.1 (0.01)	82.3 (0.23)
	SVM-lin	70.9 (0.01)	69.1 (0.01)	70.0 (0.01)	75.3 (0.01)	78.7 (0.01)
	SVM-pol	66.0 (0.01)	71.1 (0.01)	74.2 (0.01)	76.1 (0.01)	76.5 (0.01)
	STE-CD	73.5 (0.01)	76.0 (0.01)	78.2 (0.01)	80.2 (0.01)	81.1 (0.01)
	pERP-R.	77.9 (0.78)	78.4 (0.28)	79.4 (0.01)	79.7 (0.01)	80.2 (0.01)
P300, SB	SGLRT	83.3 (ref.)	85.1 (ref.)	87.8 (ref.)	87.9 (ref.)	88.0 (ref.)
	LDA	76.6 (0.01)	80.5 (0.01)	84.7 (0.01)	85.7 (0.01)	86.1 (0.01)
	SVM-lin	74.4 (0.01)	72.4 (0.01)	75.8 (0.01)	81.4 (0.01)	83.1 (0.01)
	SVM-pol	65.5 (0.01)	72.6 (0.01)	77.3 (0.01)	79.7 (0.01)	80.0 (0.01)
	STE-CD	79.9 (0.01)	82.7 (0.01)	85.9 (0.01)	86.3 (0.01)	84.4 (0.01)
	pERP-R.	80.7 (0.01)	82.9 (0.01)	85.8 (0.01)	85.9 (0.01)	86.1 (0.01)
MMN	SGLRT	74.3 (ref.)	75.9 (ref.)	78.8 (ref.)	NA	NA
	LDA	72.7 (0.01)	69.0 (0.01)	70.4 (0.01)	NA	NA
	SVM-lin	72.1 (0.05)	68.9 (0.01)	65.7 (0.01)	NA	NA
	SVM-pol	60.0 (0.01)	66.1 (0.01)	73.1 (0.01)	NA	NA
	STE-CD	69.1 (0.01)	71.1 (0.01)	75.8 (0.02)	NA	NA
	pERP-R.	61.7 (0.01)	68.5 (0.01)	71.7 (0.01)	NA	NA

TABLE II

MEAN ACCURACY (P-VALUE ≤) OF THE COMPETING METHODS FOR P300 DATA (S3 TO S10) AND SAMPLE SIZE 20

Subj.	Method					
	SGLRT	LDA	SVM-lin	SVM-pol	STE-CD	pERP-RED
S3	78.7 (ref.)	77.7 (0.21)	69.9 (0.01)	74.0 (0.02)	76.6 (0.09)	83.9 (1.00)
S4	91.0 (ref.)	86.5 (0.01)	84.0 (0.01)	77.7 (0.01)	87.2 (0.01)	90.7 (0.38)
S5	90.9 (ref.)	86.1 (0.01)	86.0 (0.01)	73.7 (0.01)	90.0 (0.17)	88.7 (0.02)
S6	68.9 (ref.)	69.8 (0.74)	61.0 (0.01)	53.8 (0.01)	71.8 (0.97)	65.4 (0.04)
S7	78.6 (ref.)	71.9 (0.01)	77.2 (0.26)	66.0 (0.01)	76.4 (0.08)	82.5 (1.00)
S8	91.8 (ref.)	87.4 (0.01)	81.9 (0.01)	71.1 (0.01)	88.7 (0.01)	88.4 (0.01)
S9	86.0 (ref.)	87.8 (1.00)	88.5 (0.01)	65.3 (0.01)	83.6 (0.04)	85.5 (0.29)
S10	88.0 (ref.)	83.8 (0.01)	85.9 (0.04)	73.1 (0.01)	89.5 (0.90)	89.4 (0.88)
Avg.	84.2 (ref.)	81.4 (0.02)	78.3 (0.01)	69.3 (0.01)	83.0 (0.09)	84.3 (0.53)

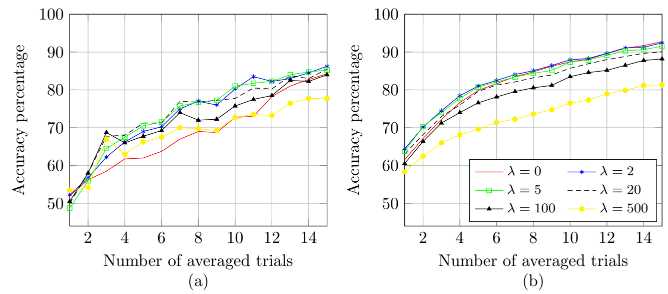

Fig. 3. The effect of λ on SGLRT performance for P300-SA data. (a) Training-set size is 20. (b) Training-set size is 360.

SNRs) of different methods for P300-SA, P300-SB, and MMN data, and for five different sample sizes. ~~As can be seen,~~ Evidently, SGLRT provides meaningfully higher detection rates compared to all the other methods for all ~~sets of data~~ the datasets (except for P300-SA and sample size 30, where pERP-RED performs slightly better). Among all the competing methods, pERP-RED has the closest performance to SGLRT. SVM-pol has the weakest average results for small ~~amounts of~~ training ~~data.~~ datasets However, by increasing the size of ~~the~~ training set, its performance improves. Note that, for P300 data the performance gap between SGLRT and LDA is higher for smaller ~~amounts of~~ training data sets. However, by increasing the size of the training set, LDA approaches SGLRT. This is due to better estimation of LDA parameters (i.e., mean and covariance) for larger ~~amounts of~~ training data sizes. It can also be seen that SGLRT has good performance in detecting a weak ERP component such as MMN, even if the Gaussian assumption is not fully met (see subsection II-B). Table II presents the averaged results of different methods for P300-S3 to S10, and sample size 20. In most cases, the proposed SGLRT outperforms the competing methods except for pERP-RED. The average results of all eight subjects indicate the close performance of SGLRT compared to pERP-RED and its superior performance compared to the other methods. In Fig. 3, the effect of λ values on SGLRT performance is assessed with regard to the size of the training set and SNR. Note that for λ values near zero and for $\Phi = I_{JN}$, SGLRT tends to LDA (see (12) and (13)). In general, the results suggest $2 \leq \lambda \leq 5$ as proper values. The results also suggest larger λ values for smaller training sets and smaller values for larger ones. This may be due to the reason that SGLRT calculates a weighted distance between the input data and the mean of each class in the smooth subspace to predict the data label. For larger training sets the estimated means are smoother. Therefore, smaller λ values are required. By increasing λ from proper values, the performance of SGLRT degrades. The drop is more noticeable for higher SNRs as the smoothness of both input data and class means increases. Results interpretation suggests an interesting solution in terms of adaptive smoothness for improving the SGLRT performance. This means an adaptive λ value can be used based on the data SNR level. As a rule of thumb, for high SNR situations, λ should be set near zero, and by decreasing the SNR, λ should be increased.

2) Second Experiment: Table III represents F1 scores and Kappa coefficients of the intera-subject test in the second experiment. For the first P300 dataset, SGLRT outperforms all the other methods for $k_r = 1$ and 5 except for pERP-RED. For $k_r = 10$ and 15, SGLRT outperforms all the other methods (except for $k_r = 10$ and for pERP-RED, where their performances are the same). For the second P300 dataset, SGLRT outperforms all other methods in $k_r = 1$. For the other numbers of repetitions, LDA shows the best performance while SGLRT is the second-best method. For the MMN dataset SGLRT outperforms all the other methods. Table IV depicts the results of inter-subject experiment. These results show that SGLRT outperforms the other approaches in five cases. The second best method is pERP-RED which performs the best in three cases.

3) Third Experiment: For this experiment the results are presented in Table V. For the results of the deep structure refer to Table 8 in [13] (STNN-6,2 represents the best results in [13]). Evidently, SGLRT performs better than SVM and STNN for both $k_r = 5$ and $k_r = 15$.

4) Fourth Experiment: Table VI represents the results of this experiment for two structural matrices $\Phi = I_{JN}$ and $\Phi = \Phi_x$, in which Φ_x is generated based on xDAWN algorithm (see Appendix B for more details). Evidently, SGLRT can benefit from a well-defined structural matrix to improve

TABLE III
F1 Scores, Kappa Coefficients of the Competing Methods over Different Datasets in Different Number of Repetitions (k_r)

Data	k_r	SGLRT	LDA	SVM-lin	SVM-pol	STE-CD	pERP-R.
P300,	1	0.16, 0.04	0.09, -0.01	0.22, 0.04	0.02, 0.01	0.05, -0.08	0.26, 0.08
1st	5	0.62, 0.57	0.34, 0.29	0.47, 0.36	0.11, 0.09	0.43, 0.37	0.65, 0.58
Data-	10	0.83, 0.80	0.76, 0.72	0.70, 0.64	0.28, 0.24	0.66, 0.62	0.83, 0.80
set	15	0.91, 0.89	0.85, 0.83	0.77, 0.72	0.28, 0.24	0.75, 0.71	0.88, 0.86
P300,	1	0.39, 0.23	0.37, 0.21	0.37, 0.20	0.17, 0.11	0.36, 0.19	0.36, 0.18
2nd	5	0.62, 0.53	0.63, 0.55	0.57, 0.47	0.33, 0.26	0.54, 0.43	0.59, 0.48
Data-	10	0.74, 0.68	0.77, 0.73	0.69, 0.63	0.51, 0.45	0.69, 0.62	0.71, 0.65
set	15	0.81, 0.77	0.84, 0.81	0.76, 0.70	0.61, 0.55	0.77, 0.72	0.78, 0.73
MMN	1	0.68, 0.59	0.50, 0.37	0.57, 0.43	0.36, 0.31	0.50, 0.38	0.61, 0.49

TABLE IV
F1-Scores, Kappa Coefficients of the Competing Methods for Inter-subject Experiment on P300 Data (S3 to S10)

Subj.	SGLRT	LDA	SVM-lin	SVM-pol	STE-CD	pERP-RED
S3	0.36, 0.21	0.29, 0.16	0.46, 0.31	0.23, 0.12	0.20, 0.00	0.46, 0.31
S4	0.75, 0.69	0.67, 0.60	0.60, 0.50	0.46, 0.39	0.30, 0.12	0.64, 0.55
S5	0.40, 0.25	0.48, 0.36	0.48, 0.36	0.37, 0.30	0.18, -0.07	0.41, 0.28
S6	0.47, 0.36	0.38, 0.29	0.34, 0.21	0.24, 0.17	0.24, 0.05	0.47, 0.36
S7	0.68, 0.62	0.58, 0.50	0.47, 0.37	0.30, 0.22	0.32, 0.15	0.70, 0.64
S8	0.79, 0.74	0.76, 0.71	0.70, 0.62	0.57, 0.50	0.47, 0.35	0.73, 0.67
S9	0.72, 0.66	0.72, 0.66	0.59, 0.48	0.35, 0.26	0.24, 0.04	0.69, 0.61
S10	0.76, 0.70	0.71, 0.63	0.61, 0.51	0.35, 0.28	0.35, 0.17	0.70, 0.62

TABLE V
Accuracy of the Competing Methods over Dataset II of BCI Competition III for Subject A and B in 5 and 15 Repetitions

Method	$k_r = 5$			$k_r = 15$		
	SA	SB	Avg.	SA	SB	Avg.
SGLRT	87.4	92.5	90.0	96.8	96.7	96.7
SVM-lin	72.8	82.9	77.9	86.5	87.2	86.8
SVM-pol	84.4	84.8	84.6	89.1	88.4	88.7
STNN-6, 2	88.4	89.8	89.2	NA	NA	NA

its performance (e.g. using Φ_x improves CAL by 4% for SB and $k_r = 15$). Note that in this experiment, the probability of choosing the correct letter by chance is 1/36 while in the former experiments the chance level is 1/2.

In Table VI, ITR (bits/min) is also reported for each case, which is calculated based on the following equation [12],

$$ITR = \frac{60}{T}\left[\log_2 L + P\log_2 P + (1-P)\log_2(\frac{1-P}{L-1})\right] \tag{14}$$

where $0 < P < 1$ is the probability of selecting a target command (P is calculated based on the classification accuracy), L is the number of possible commands that a BCI system can generate, and T (sec) is the required time to generate a command. In our case $L = 36$, and $T = ((12 \times k_r - 1) \times 0.175 + 0.700) + t_{gs}$ for k_r repetitions, where $tgs = 2.5$sec is the required time for gaze shifting. As can be seen, for a lower number of repetitions, ITR is higher. However, the applicability of BCI systems in low repetitions may not be taken for granted as the low accuracy makes the users frustrated. Since for a specific BCI scheme and a certain number of repetitions, L and T are constants, then, the method with higher classification accuracy has a higher ITR as well. By this logic, it can be concluded that our proposed method

TABLE VI
Classification Accuracy of Letters (CAL) and ITR (bits/min) of SGLRT for $\Phi = I_{JN}$ and $\Phi = \Phi_x$

k_r	SA				SB			
	$\Phi = I_{JN}$		$\Phi = \Phi_x$		$\Phi = I_{JN}$		$\Phi = \Phi_x$	
	CAL	ITR	CAL	ITR	CAL	ITR	CAL	ITR
1	17	3.0	18	3.3	33	9.6	33	9.6
5	55	8.3	63	10.3	75	13.6	73	13.1
10	82	8.9	82	8.9	91	10.7	93	11.1
15	94	7.9	94	7.9	92	7.6	96	8.2

provides higher ITR compared to the other methods mentioned in the former experiments. Note that for this specific BCI scheme and for $k_r = 15$, the best achievable ITR is 9.0.

IV. Discussion

This work presents a new scheme to detect ERPs based on smoothness priors. The main goal of this study is to improve conventional methods by adding this extra information to the desired signal morphology. The proposed SGLRT benefits from an internal smoothing operator that can estimate the desired signal based on an adjustable smoothness level. This allows for a more specific and accurate trial-by-trial ERP estimation and detection by increasing the SNR. SGLRT mainly relies on the data covariance and mean to separate the target from non-target trials. Hence, enhancing their estimations also leads to improved detection rates. This can be achieved by increasing the size of the training set since estimations of mean and covariance matrix are unbiased and consistent (see II-A1, Formation of the likelihood ratio, also see Table I).

In the derivation of the proposed method, the EEG data matrix is vectorized by concatenating its columns. This may lead to discontinuities in the transition between channels. To mitigate these abrupt changes, the selected windows should be long enough to let the ERP components tend toward zero at the start and finish points of the selected sequence in each channel. In addition, some other points help mitigate these discontinuities. First, we have used some sort of baseline correction to zero-mean the data before segmentation. Second, by averaging over more repetitions, the discontinuity level between channel transitions decreases. Finally, the smoothing operator is designed to allow the ERP components to pass while the abrupt changes are blocked. Therefore, the method itself mitigates these discontinuities.

Conventional methods usually consider smoothness priors by roughly passing the data through a bandpass filter. In contrast, the proposed method employs a smoothing filter which is specifically designed to maximize the separability of target and non-target trials. Some of the new methods use dynamic models for ERPs and incorporate temporal smoothness and spatial correlations. For example, the authors in [2] developed a cooperative particle filtering approach for ERP tracking over trials. These methods proved to be very effective. However, those models might not exactly follow the structure of ERPs. In our method, there is no constraint on ERP morphology and different components with different shapes (polynomial of any order, Gaussian, etc.) and latencies can be considered. The only constraint is the smoothness of the ERP component, and its level is adjustable for different ERPs through the λ value. It should be noted that the proposed method filters the data in the time domain, and tries to find a smooth subspace to enhance ERPs. Several approaches in the literature use spatial filtering to find a subspace for better emerging of ERPs (for instance, the well-known xDAWN algorithm [30]), or to estimate underlying sources of ERPs (for instance, the pERP-RED method). Although these methods are very effective, smoothness priors are not considered for signal structure in time. For instance, in [11], the authors say *"A second limitation is that the derived pERPs are not penalized in time to attain a desired degree of smoothness."*. One idea that comes to mind is to consider smoothness in a joint time-space domain, or add smoothness priors to the previously developed spatial filtering methods and achieve even more powerful schemes regarding ERP estimation or detection.

The proposed SGLRT can be directly applied to many ERP-based BCI systems (e.g. P300 mind speller). Furthermore, accurate detection of ERPs can help better the separation of ERP subcomponents (e.g. p3a and p3b in P300) in a later stage, which can then be used by various BCI applications and mental activity evaluations. It can also be used as a template for ERP localization in a forward source localization solution [9], [31]. In addition, the proposed method can be considered as a general framework for signal activity detection since by increasing the sampling frequency by a certain amount, many bandlimited signals can be considered smooth compared to wide-band noises. The proposed method can also be implemented online. But before that, a session of recordings is required for model training. Moreover, in the online testing phase, the system needs to collect data after applying the stimuli several times (i.e. for several repetitions), which is time-consuming. This is the current issue in all ERP-based systems. However, by training better models like SGLRT, we can improve the speed and performance of these systems.

For the SGLRT training, there are three computationally intensive steps, 1) covariance matrix estimation, 2) covariance matrix inversion, and 3) computation of the smoothing matrix Ψ. The first two steps can be performed efficiently with complexity $O((K_0 + K_1)(J^2N + JN^2))$ and $O(J^3 + N^3 + J^2N^2)$ (see Appendix A), and Ψ can be computed with complexity $O(3J^3N^3)$ if $\Phi = I_{JN}$, and with complexity $O(J^3(3N^2M_s + 3NM_s^2 + M_s^3))$ otherwise. LDA and pERP-RED share the first two steps with SGLRT while pERP-RED also requires three steps of spatial filtering to extract the underlying ERPs and another step to represent the data based on those ERPs. STE-CD does not require covariance matrix estimation and inversion, instead, it fits a multivariate autoregressive (MVAR) model to the data and pre-whitens the data before classification. Generally, model training of the competing methods in our study is completed in a time scale of seconds. Here, the training run-times (average of 100 runs) of the competing algorithms are reported for the most computationally intensive case of our study using a 2.13GHz Core i3 CPU with a 3GB RAM, which are 7.18, 1.51, 0.24, 0.25, 9.08, and 4.20 seconds for SGLRT, LDA, SVM-lin, SVM-pol, STE-CD, and pERP-RED, respectively. The test phase of SGLRT requires a fairly low computational complexity ($O(JN)$) as it calculates a linear function of the input data to predict their labels.

Finally, SGLRT has some limitations. First, the estimated smooth ERPs are a combination of all ERP subcomponents. Hence, this method is useful for an ERP detection scheme and cannot estimate separate ERP subcomponents (such as p3a and p3b in P300). Second, the proposed method is covariance-based. Hence, increasing the number of EEG channels and/or sampling rate increase the computational burden (see Appendix A for some solutions to mitigate this shortcoming). It should also be noted that since SGLRT imposes no constraint on the number of EEG channels, the optimal channels can be selected to reduce the computational burden. Third, the proposed scheme is derived based on the assumption that the data follows a normal distribution. If this assumption is not met, the proposed method performance may degrade.

V. Conclusion

Contributions: In this study, a novel and powerful method for ERP detection from multichannel EEG signals was presented. The method was derived through a two-step GLR test and based on the smoothness priors of ERPs. **Findings:** We applied SGLRT to three real ERP datasets and compared it with LDA, SVM, STE-CD, pERP-RED, and a deep structure (STNN). For these comparisons, three experiments were carried out and the methods were evaluated by three metrics (accuracy, F1 score, and Kappa coefficient). Based on the results, we found out that in many cases SGLRT can improve the detection of ERPs for small and large datasets over different numbers of trial repetitions. This is very promising as the ERP-based BCI systems mainly suffer from lack of classification accuracy. Hence, it is an important step toward real-life applicability of these systems. **Future lines:** In future studies, an approach based on adaptive smoothness corresponding to different noise levels will be investigated. Moreover, a joint spatiotemporal smoothing filter, which may improve the results, is of interest. For this, a set of spatial filters can be used before the proposed method to reduce space dimension. However, how these filters affect smoothness priors is not known. At last, the role of structural matrix Φ and underlying sources of ERPs should be investigated more.

Appendix A
Covariance Estimation and Inversion in SGLRT

SGLRT, LDA, and pERP-RED are covariance-based methods. Hence, a good estimation of the covariance matrix plays an important role in their performances. Covariance estimation is challenging for high dimension data and usually requires a large dataset. For example, in the case of the JN data dimension, $K > JN$ independent trials are required to gain a full rank and positive definite estimation. One approach to address this issue is the matrix normal distribution assumption for EEG data [32]. By this assumption, the estimation of the covariance matrix is obtained by $\hat{R} = \hat{V} \otimes \hat{U}$, where $\hat{U}$ is the ML estimation of the $J \times J$ temporal covariance matrix, $\hat{V}$ is the ML estimation of the $N \times N$ spatial covariance matrix, and $\otimes$ is the Kronecker product. This assumption reduces the required number of trials for a positive definite covariance estimation to $K > \max\{J/N, N/J\}$. In this study, this model is used for P300 data. However, for MMN data, it is used whenever it could provide better results. Furthermore, the first k_r trials are treated independently for covariance estimation (rather than using their average), which provides more data. From a practical point of view, an infinitesimal value can also be added to the diagonal elements of the covariance matrix to stabilize its inversion. SGLRT requires the inversion of the covariance matrix of the JN dimension, which can be computationally expensive. As mentioned, the matrix normal distribution is a well-suited assumption for EEG data [32]. Therefore, the inverse of the estimated covariance matrix can be obtained by $\hat{R}^{-1} = \hat{V}^{-1} \otimes \hat{U}^{-1}$, which dramatically decreases the computational burden. Moreover, the calculation of Ψ in (13) involves inversion of $(\Phi^T \hat{R}^{-1}\Phi + \lambda\Phi^T D_j^T D_j \Phi)$. This summation could be inverted in a recursive manner if the summation itself and at least one of its terms (i.e. $\Phi^T \hat{R}^{-1}\Phi$ or $\lambda\Phi^T D_j^T D_j \Phi$) are not singular [33]. In a special case when $\Phi = I_{JN}$, the inverse of the first term is $\hat{R}$. Therefore, there is no need to invert any matrix [33].

Appendix B
Guidelines for Designing the Structural Matrix Φ

Consider $\underline{s}_m \in \mathbb{R}^{J\times 1}$, $m = 1, 2, \ldots, M_s$ as the underlying sources contributing to ERPs, where $\underline{s}' \in \mathbb{R}^{JM_s\times 1}$ is formed by their concatenation. It is assumed that the recorded ERP in each of the N channels is explained by a linear combination of these sources. Let a_{mn} $(n = 1, 2, \ldots, N)$ be the weight to map the mth underlying source to the nth channel, and let $\underline{z}$ be a row vector of $J-1$ zeros, then, $\phi_n \in \mathbb{R}^{J\times JM_s}$ is formed as a Toeplitz matrix form of $[a_{1n}, \underline{z}, a_{2n}, \underline{z}, \ldots, a_{mn}, \underline{z}] \in \mathbb{R}^{1\times JM_s}$, which is a matrix of coefficients that maps $\underline{s}'$ to the nth channel. Now, $\Phi \in \mathbb{R}^{JN\times JM_s}$ can be obtained by concatenation of ϕ_n matrices. Various methods can be adopted to obtain the coefficients of matrix Φ. For instance, one may use pERP-RED to extract the underlying sources of ERPs and then estimate the coefficients based on these sources [11]. Another approach can be by using the spatial filters extracted by the xDAWN algorithm [30]. If $U_x \in \mathbb{R}^{N\times N}$ is the set of filters extracted by xDAWN, then, the mth row and the nth column of U_x^{-1} contains a_{mn}. From [30], the first filters are mostly contributed to ERPs while the last ones mostly extract noise. Therefore, it is suggested to use only the first four filters (i.e. $M_s = 4$). Note that, if it is desired to select a subset of recording channels contributing the most to the underlying sources of ERPs and it is desired to consider a separate source for each selected channel, then, Φ should be equal to an identity matrix.

References

[1] S. J. Luck, *An Introduction to the Event-Related Potential Technique*. MIT press, 2014.

[2] S. Monajemi *et al.*, “Cooperative particle filtering for tracking ERP subcomponents from multichannel EEG,” *Entropy*, vol. 19, no. 5, p. 199, 2017.

[3] H. Cecotti and A. Graser, “Convolutional neural networks for P300 detection with application to brain-computer interfaces,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 433–445, 2010.

[4] G. F. Woodman, “A brief introduction to the use of event-related potentials in studies of perception and attention,” *Atten. Percept. Psycho*, vol. 72, no. 8, pp. 2031–2046, 2010.

[5] J. Polich, “Updating P300: An integrative theory of p3a and p3b,” *Clin. Neurophysiol.*, vol. 118, no. 10, pp. 2128–2148, 2007.

[6] F. Boiten *et al.*, “Event-related desynchronization: The effects of energetic and computational demands,” *Electroen. Clin. Neuro.*, vol. 82, no. 4, pp. 302–309, 1992.

[7] S. H. Patel and P. N. Azzam, “Characterization of N200 and P300: Selected studies of the event-related potential,” *Int. J. Medical Sci.*, vol. 2, no. 4, p. 147, 2005.

[8] J. T. Philip and S. T. George, “Visual P300 mind-speller brain-computer interfaces: A walk through the recent developments with special focus on classification algorithms,” *Clin. EEG Neurosci.*, 2019.

[9] D. Jarchi *et al.*, “A new spatiotemporal filtering method for single-trial estimation of correlated ERP subcomponents,” *IEEE Trans. Biomed. Eng.*, vol. 58, no. 1, pp. 132–143, 2010.

[10] C. Yang *et al.*, “The spatio-temporal equalization for evoked or event-related potential detection in multichannel EEG data,” *IEEE Trans. Biomed. Eng.*, 2019.

[11] E. Campos *et al.*, “Principle ERP reduction and analysis: Estimating and using principle ERP waveforms underlying ERPs across tasks, subjects and electrodes,” *NeuroImage*, vol. 212, p. 116630, 2020.

[12] Z. Oralhan, “3D input convolutional neural networks for P300 signal detection,” *IEEE Access*, vol. 8, pp. 19 521–19 529, 2020.

[13] Z. Zhang *et al.*, “Spatial-temporal neural network for P300 detection,” *IEEE Access*, vol. 9, pp. 163 441–163 455, 2021.

[14] S. M. Kay, *Fundamentals of Statistical Signal Processing: Detection Theory*. Upper Saddle River, NJ: Prentice-Hall, 1998.

[15] J. Candy, *Model-Based Signal Processing*. Wiley-IEEE Press, 2005.

[16] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice Hall PTR, 1993.

[17] G. J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons, 2004, vol. 544.

[18] W. Gersch, “Smoothness priors,” in *New Directions in Time Series Analysis*. New York: Springer, 1993, pp. 113–146.

[19] R. Sameni, “Online filtering using piecewise smoothness priors: Application to normal and abnormal electrocardiogram denoising,” *Signal Process.*, vol. 133, pp. 52–63, 2017.

[20] K. W. Cheung *et al.*, “A constrained least squares approach to mobile positioning: algorithms and optimality,” *EURASIP J. Adv. Sig. Pr.*, vol. 2006, no. 1, p. 020858, 2006.

[21] P. C. Hansen, “The L-curve and its use in the numerical treatment of inverse problems,” *IMM, Department of Mathematical Modelling, Technical University of Denmark*, 1999.

[22] K. M. Spencer and J. Polich, “Poststimulus EEG spectral analysis and P300: attention, task, and probability,” *Psychophysiology*, vol. 36, no. 2, pp. 220–232, 1999.

[23] G. P. Jacobson, “Exogenous and endogenous auditory brain events occurring between 50-200 ms: Past, present and future applications,” in *Seminars in Hearing*, vol. 20, no. 01, 1999, pp. 63–75.

[24] S. Korkmaz *et al.*, “MVN: an R package for assessing multivariate normality.” *R J.*, vol. 6, no. 2, p. 151, 2014.

[25] C. Guger *et al.*, “How many people are able to control a P300-based brain–computer interface (BCI)?” *Neurosci. Lett.*, vol. 462, no. 1, pp. 94–98, 2009.

[26] J. R. Wolpaw *et al.* (2005) BCI competition III webpage. [Online]. Available: http://www.bbci.de/competition/iii/

[27] J. Ashburner *et al.*, “SPM12 manual,” 2014. [Online]. Available: https://www.fil.ion.ucl.ac.uk/spm/data/eeg_mmn/

[28] M. I. Garrido *et al.*, “Dynamic causal modelling of evoked potentials: A reproducibility study,” *NeuroImage*, vol. 36, no. 3, pp. 571–580, 2007.

[29] T. Schneider and A. Neumaier, “Algorithm 808: Arfit—A Matlab package for the estimation of parameters and eigenmodes of multivariate autoregressive models,” *ACM Trans. Math. Software. (TOMS)*, vol. 27, no. 1, pp. 58–65, 2001.

[30] B. Rivet *et al.*, “xDAWN algorithm to enhance evoked potentials: Application to brain–computer interface,” *IEEE Trans. Biomed. Eng.*, vol. 56, no. 8, pp. 2035–2043, 2009.

[31] L. Spyrou and S. Sanei, “Source localization of event-related potentials incorporating spatial notch filters,” *IEEE Trans. Biomed. Eng.*, vol. 55, no. 9, pp. 2232–2239, 2008.

[32] P. Gonzalez-Navarro *et al.*, “Spatio-temporal EEG models for brain interfaces,” *Signal Process.*, vol. 131, pp. 333–343, 2017.

[33] K. S. Miller, “On the inverse of the sum of matrices,” *Math. Mag.*, vol. 54, no. 2, pp. 67–72, 1981.