




## Bayesian Inferences on Neural Activity in EEG-Based Brain-Computer Interface

Tianwen Ma, Yang Li, Jane E. Huggins, Ji Zhu & Jian Kang


**To cite this article:** Tianwen Ma, Yang Li, Jane E. Huggins, Ji Zhu & Jian Kang (2022) Bayesian Inferences on Neural Activity in EEG-Based Brain-Computer Interface, Journal of the American Statistical Association, 117:539, 1122-1133, DOI: [10.1080/01621459.2022.2041422](https://doi.org/10.1080/01621459.2022.2041422)

**To link to this article:** <https://doi.org/10.1080/01621459.2022.2041422>

 [View supplementary material](#) 



 [Published online: 18 Mar 2022.](#)

 [Submit your article to this journal](#) 

 [Article views: 1103](#)

 [View related articles](#) 

 [View Crossmark data](#) 

 [Citing articles: 3](#) [View citing articles](#) 



# Bayesian Inferences on Neural Activity in EEG-Based Brain-Computer Interface

Tianwen Ma<sup>a</sup>, Yang Li<sup>b</sup>, Jane E. Huggins<sup>c</sup>, Ji Zhu<sup>b</sup>, and Jian Kang<sup>a</sup> 

<sup>a</sup>Department of Biostatistics, University of Michigan, Ann Arbor, MI; <sup>b</sup>Department of Statistics, University of Michigan, Ann Arbor, MI; <sup>c</sup>Department of Physical Medicine and Rehabilitation and Department of Biomedical Engineering, University of Michigan, Ann Arbor, MI

## ABSTRACT

A brain-computer interface (BCI) is a system that translates brain activity into commands to operate technology. A common design for an electroencephalogram (EEG) BCI relies on the classification of the P300 event-related potential (ERP), which is a response elicited by the rare occurrence of target stimuli among common nontarget stimuli. Few existing ERP classifiers directly explore the underlying mechanism of the neural activity. To this end, we perform a novel Bayesian analysis of the probability distribution of multi-channel real EEG signals under the P300 ERP-BCI design. We aim to identify relevant spatial temporal differences of the neural activity, which provides statistical evidence of P300 ERP responses and helps design individually efficient and accurate BCIs. As one key finding of our single participant analysis, there is a 90% posterior probability that the target ERPs of the channels around visual cortex reach their negative peaks around 200 milliseconds poststimulus. Our analysis identifies five important channels (PO7, PO8, Oz, P4, Cz) for the BCI speller leading to a 100% prediction accuracy. From the analyses of nine other participants, we consistently select the identified five channels, and the selection frequencies are robust to small variations of bandpass filters and kernel hyper parameters. Supplementary materials for this article are available online.

## ARTICLE HISTORY

Received May 2021  
Accepted February 2022

## KEYWORDS

Bayesian analysis;  
Brain-computer interface;  
Gaussian process; Neural  
activity

## 1. Introduction

### 1.1. Background

A brain-computer interface (BCI) is a device that interprets brain activity to operate technology. An electroencephalogram (EEG)-based BCI speller system is a particular BCI device that enables a person to “type” words without using a physical keyboard by recording EEG brain activity. It has been used for assisting people with disabilities, such as amyotrophic lateral sclerosis (ALS), with regular communication (Wolpaw et al. 2018). The brain activity is measured with EEG signals, which have the features of noninvasiveness, low cost, and high temporal resolution.

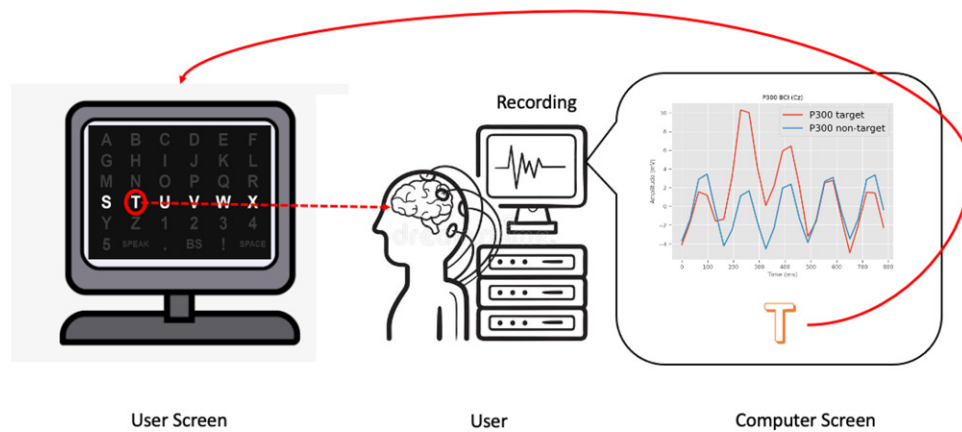
The conventional BCI framework is based on the P300 event-related potential (ERP) BCI design, known as the P300 ERP-BCI design (Farwell and Donchin 1988). However, we also include other types of ERPs that help interpret and classify the brain activity. An ERP is a signal pattern in the brain activity in response to an external event. The P300 ERP is a particular ERP that occurs in response to a rare, but relevant event (i.e., highlighting a group of characters on the screen). The relevant (target) P300 ERP has a *positive* deflection in voltage with the latency (the delay from the onset of the event to the first response peak) around 300 ms (Rodden and Stemmer 2008). The right-most plot in Figure 1 shows the typical target and nontarget P300 ERPs from a real participant.

There are three challenges in making valid inferences on brain activity in the P300 ERP-BCI system. First, the signal-to-noise ratio of the EEG signals is quite low. A typical P300

ERP-BCI system requires collecting data from multi-dimensional input and repeated sequences of events. Second, to reduce the time to complete the sequence of events necessary to present all the keys on the virtual keyboard, we minimize the time between adjacent events within each sequence and between adjacent sequences. Thus, the time between events is shorter than the time required to produce a P300 ERP response. Therefore, the observed EEG signal is a mixture of overlapping ERP responses, which may or may not contain a P300 ERP. No formal statistical methods can resolve this mixture and make valid inferences on the overlapping responses. Finally, during the calibration time in the current P300 ERP-BCI system, participants may experience variations in attention from fatigue to boredom, leading to missed or delayed responses that may obscure statistical inferences.

### 1.2. Conventional Framework with Motivating Dataset

The conventional P300 ERP-BCI design presents a sequence of events on a virtual keyboard and analyzes the EEG signals in a fixed time response window after each event to make a *binary* decision whether a P300 ERP response is produced by that event, which forms the fundamental basis of the P300 ERP-BCI operation. For multi-channel EEG signals, channel-specific EEG signal segments are concatenated for binary classification. Here, an EEG channel is defined as an electrode capturing brain activity. Multiple electrodes are placed on the scalp to achieve stable prediction accuracy. The binary classification results are then converted into character-level probabilities. We denote



**Figure 1.** An illustration of the conventional procedure of the P300 ERP-BCI operation. The P300 ERP-BCI design presents a sequence of events on a virtual screen to the user. The user focuses on a specific character and responds to different events eliciting different brain signals (P300 or no P300). These brain signals are recorded by the EEG machine. Classifiers are then constructed to analyze EEG signals in a fixed time response window after each event to make a binary decision whether a P300 ERP response is produced. The binary classification results are converted into character-level probabilities, and the character with the highest probability is shown on the screen.

“key” and “target key” as a generic character to be typed and the specific character that the user wants to type, respectively. Usually, events within each sequence cover all the possible keys, but multiple keys can exist in each event. Thus, the P300 ERP-BCI is designed to identify the unique key from the intersection of all events that produce P300 ERP responses within each sequence. Finally, the conventional P300 ERP-BCI design presents a fixed number of events (stimuli) with a fixed number of sequences before the final decision is made. Figure 1 describes the procedure of the conventional P300 ERP-BCI operation.

To better illustrate the framework, we briefly introduce the motivating dataset following the experimental protocol by (Thompson, Gruis, and Huggins 2014). It is part of the database of noninvasive experimental data in the P300 ERP-BCI experiments conducted at the University of Michigan Direct Brain Interface Laboratory (UM-DBI). Under the protocol mentioned above, each participant copied a multi-character phrase during the experimental session. The dataset of each participant consisted of the training (calibration) data and the testing (free-typing) data. We created a participant-specific classifier with the training data and tested on the free-typing data. The study adopted the row-and-column paradigm (RCP) design developed by Farwell and Donchin in 1988. The BCI display screen was a  $6 \times 6$  grid of characters. Each event was either a row stimulus or a column stimulus. The order of the row and column stimuli was random, and it looped through all rows and columns every consecutive 12 stimuli, called a sequence. For each character of interest, participants were asked to mentally count when they saw a row or column stimulus containing the character of interest and to ignore stimuli that did not include the current character of interest. Thus, each sequence always had two events (stimuli) that were supposed to elicit P300 ERPs (one row and one column) out of every 12 events. In particular, the left side of Figure 1 shows 36 characters in a  $6 \times 6$  grid with the fourth row being highlighted.

Many state-of-the-art machine learning (ML) methods such as stepwise linear discriminant analysis (swLDA) (Donchin, Spencer, and Wijesinghe 2000; Krusienski et al. 2008), logistic regression (LR) (Viana, Batista, and Melges 2014), random forest (RF) (Okumuş and Aydemir 2017), support vector machine

(SVM) (Kaper et al. 2004), convolutional neural network (CNN) (Cecotti and Graser 2010), independent component analysis (ICA) (Xu et al. 2004), and recent XGBoost (Leoni et al. 2021) have successfully constructed binary classifiers for P300-ERPs. These discriminant approaches treat target or nontarget stimuli as the response variable and the truncated-and-concatenated EEG signal segments as feature vectors. Although these approaches are straightforward to implement, it is difficult for them to make statistical inferences about brain activity with overlapping P300 ERP responses. The functional graphical model (Qiao, Guo, and James 2019, FGM) is a powerful tool to model the conditional dependency over functional variables and it has been used to model multiple-subject EEG data in an alcoholism study for functional connectivity analysis. However, FGM cannot be directly adopted in our study due to the differences in the goal of analysis and the data structure.

As a flexible tool for Bayesian nonparametrics and machine learning, the Gaussian Process (GP), a stochastic process where every finite collection of its realizations follows a multivariate normal distribution, has been widely used for modeling functional and dependent data over time and space (Rasmussen 2003). Different extensions of GPs have been proposed for different neuroscience applications. In particular, for feature selection in scalar-on-image regression, the soft-thresholded GP prior (Kang, Reich, and Staicu 2018) models sparse, continuous and piece-wise smooth functions. This prior has also been extended to model the sparsity and dependence in the effects of nodes over a graph in the framework of Bayesian network marker selection (Cai, Kang, and Yu 2020). However, none of these existing GPs can be directly applied to detection of our P300 ERPs in EEG signals.

### 1.3. Our Contributions

To the best of our knowledge, we are among the first to study the probability distribution of multi-trial EEG signals from real participants in BCI experiments using a Bayesian generative model. Our Bayesian analysis explores the mechanism of neural activity in response to external stimuli. Our model *explicitly* addresses the challenge of overlapping ERPs between adjacent stimuli,

and the model can be applied to multi-channel EEG signals without signal concatenation nor segmentation. We develop a new GP-based prior to the spatial-temporal varying trajectories of P300 ERP responses. The proposed prior facilitates selecting important time windows in which the average brain activity in response to the target stimuli and nontarget stimuli is different (split) or the same (merge); thus, it is termed the split-and-merge GP (SMGP). We make fully posterior inferences on participant-and-channel-specific P300 ERPs in a fixed EEG response window.

Based on our Bayesian analysis, we first aim to identify significant split time windows for frontal, central, parietal, parietal-occipital, and occipital channels. We do not expect to identify significant split time windows for channels close to ears. We study the neural activity patterns among both healthy controls and participants with the Amyotrophic Lateral Sclerosis (ALS) disease under the ERP-BCI design. Finally, we perform the brain region ranking by the participant-specific information criterion. We hypothesize brain regions associated with the cognitive function as well as the visual function will be selected with high reproducibility across participants (Brunner et al. 2010). In addition, we expect that the signal to detect target P300 ERPs for the participant with ALS is weaker than healthy controls, but it should still be significant for classification. Finally, we expect that it may take longer for senior participants than for the young participants to reach the peak of target P300 ERP responses (Polich, Howard, and Starr 1985).

The article is organized as follows: Section 2 presents the model for the probability distribution of EEG signals under the P300 ERP-BCI design along with the prior specifications. Section 3 develops the method for posterior inference. Sections 4 and 5 present the analyses of the multi-channel EEG data from real BCI users and simulations, respectively. Section 6 concludes the paper with a brief discussion.

## 2. Bayesian Modeling of EEG-BCI Data

### 2.1. Notation and Problem Setup

We begin with the notation. Denote by  $\mathbb{R}$  the real line. For any interval  $\mathcal{A} \subset \mathbb{R}$ , let  $\mathbb{I}_{\mathcal{A}}(t) = 1$  if  $t \in \mathcal{A}$  and 0 otherwise. Denote by  $\mathcal{N}(\mu, \Sigma)$  a normal distribution with mean  $\mu$  and variance (covariance)  $\Sigma$ . Denote by  $\mathcal{GP}(\mu, \kappa)$  the GP with the mean function  $\mu$  and the covariance kernel  $\kappa$ . All the time variables in this manuscript are multiples of a prespecified unit time.

Our model focuses on the multi-channel EEG data for one participant. Suppose a total of  $L$  target characters are typed for BCI calibration in the training data. For each character  $l$  ( $l = 1, \dots, L$ ), the BCI generates  $I$  sequences of  $J$  ( $J = 12$ ) stimuli consisting of six row stimuli, denoted as  $1, \dots, 6$  and six column stimuli, denoted as  $7, \dots, 12$  on the  $6 \times 6$  keyboard in a random order (Figure 2(a)). Let  $i$  ( $i = 1, \dots, I$ ) index the sequence. For the  $i$ th sequence of the  $l$ th target character, let  $\mathbf{W}_{l,i} = (W_{l,i,1}, \dots, W_{l,i,12})^\top$  represent the starting time points of the  $J$  stimuli (stimulus-occurring indicators) and take values from permutations of  $\{1, \dots, 12\}$ . For example,  $\mathbf{W}_{l,i} = (8, \dots, 3, 2, \dots, 11)^\top$  indicates that the first row, ..., the last row, the first column ... and the last column appear in the 8th

stimulus, ... 3rd stimulus, 2nd stimulus, ..., and 11th stimulus, respectively. Let  $\mathbf{Y}_l = (Y_{l,1}, \dots, Y_{l,12})^\top$  represent the stimulus-type indicators, where  $Y_{l,j} \in \{0, 1\}$  with the constraint  $\sum_{j=1}^6 Y_{l,j} = \sum_{j=7}^{12} Y_{l,j} = 1$ . The event  $Y_{l,j} = 1$  indicates the  $l$ th target letter is located in the  $j$ th row stimulus for  $j = 1, \dots, 6$  and the  $(j - 6)$ th column stimulus for  $j = 7, \dots, 12$ . Thus, each possible value of  $\mathbf{Y}_l$  uniquely determines one target character on the  $6 \times 6$  keyboard. For example,  $\mathbf{Y}_l = (\underbrace{0, 0, 0, 1, 0, 0}_{\text{row}}, \underbrace{0, 1, 0, 0, 0, 0}_{\text{column}})^\top$

indicates that the target letter is ‘‘T’’ located at the fourth row and the second column. We drop the sequence index  $i$  for  $\mathbf{Y}_l$  because the stimulus-type indicators are always the same given the same character  $l$ . For all the sequences, the time domain of the EEG signals are registered to  $[0, T]$ . Finally, suppose we consider  $E$  channels of EEG signals and let  $e$  ( $e = 1, \dots, E$ ) index the channel, and we denote  $X_{l,i,e}(t)$  as the observed EEG signal intensity of the  $i$ th sequence and  $l$ th target character from channel  $e$  at time  $t \in [0, T]$ .

### 2.2. A Bayesian Generative Model

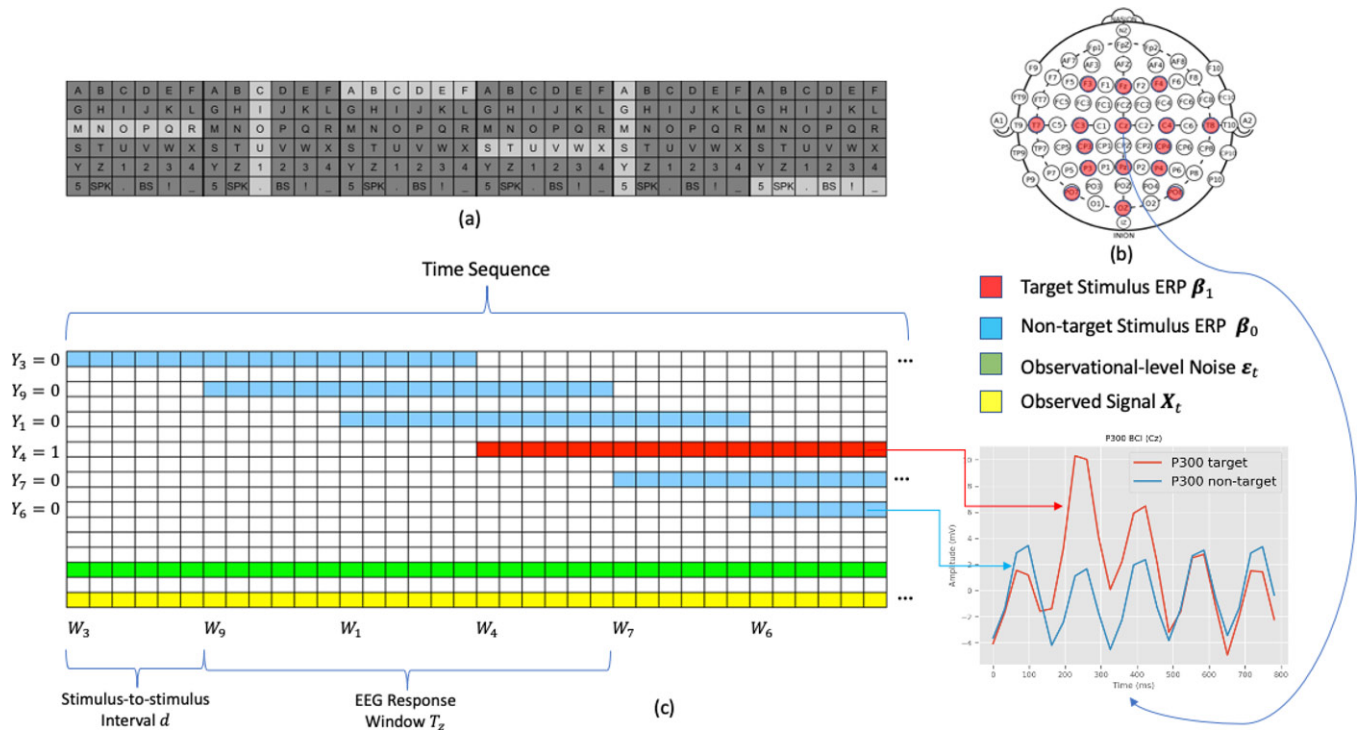
Suppose we are interested in making inferences on the P300-ERP in a window of length  $T_z$  right after the onset of the stimulus. We refer to  $T_z$  as the response window length and assume  $T_z$  is a multiple of  $d$  for simplicity, where  $d$  is the stimulus-to-stimulus interval. The total length of time  $T$  per sequence is then defined as  $T = T_z + (J - 1)d$ . We consider the observed EEG signals  $X_{l,i,e}(t)$  as a mixture of the  $J$  stimulus-induced potentials given stimulus-type indicators  $\mathbf{Y}_l$  and stimulus-occurring times  $\mathbf{W}_{l,i}$  as follows: For any  $t \in [0, T]$ ,

$$\begin{aligned} X_{l,i,e}(t) &= M_{l,i,e}(t) + \epsilon_{l,i,e}(t), \quad \tau_{l,i,j} = t - (W_{l,i,j} - 1)d, \\ M_{l,i,e}(t) &= \sum_{j=1}^J [\beta_{1,e}(\tau_{l,i,j}) Y_{l,j} + \beta_{0,e}(\tau_{l,i,j})(1 - Y_{l,j})] \mathbb{I}_{[0, T_z]}(\tau_{l,i,j}), \end{aligned} \quad (1)$$

where  $M_{l,i,e}(t)$  is the expected EEG signals at time  $t$  from channel  $e$  induced by  $J$  stimuli that occur at different time points. The two unknown functions  $\beta_{1,e}(\tau)$  and  $\beta_{0,e}(\tau)$  ( $\tau \in [0, T_z]$ ) represent the average brain activity responses to the target and the nontarget stimulus, respectively. To simplify the problem, we assume that the shape and magnitude of ERP functions only depend on the stimulus-type indicators, regardless of the stimulus location or the stimulus order. The random noise  $\epsilon_{l,i,e}(t)$  characterizes the intrinsic brain activity of channel  $e$  that is, unrelated to the stimulus responses. Assuming that  $\epsilon_{l,i,e}(t)$  is spatially-correlated across channels and temporally dependent, we consider the following additive model:

$$\begin{aligned} \epsilon_{l,i,e}(t) &= \zeta_{l,i,e} + \varepsilon_{l,i}(t), \\ \zeta_{l,i} &= (\zeta_{l,i,1}, \dots, \zeta_{l,i,E})^\top \sim \mathcal{N}(0, C_s), \\ \varepsilon_{l,i}(t) &= \rho_{l,0} + \sum_{m=1}^q \rho_{l,m} \varepsilon_{l,i}(t - md) + \varepsilon_{l,i,0}(t), \\ \varepsilon_{l,i,0}(t) &\sim \mathcal{N}(0, \sigma_x^2), \end{aligned}$$

where  $\zeta_{l,i,e}$  is the channel-specific random effect and  $\zeta_{l,i,1}, \dots, \zeta_{l,i,E}$  jointly follows a multivariate normal distribution with the mean zero and the covariance matrix  $C_s$ . The temporal



**Figure 2.** (a). A figure showing a  $6 \times 6$  grid screen of the ERP-BCI speller system, where only one row or one column was being flashed gray for each stimulus. (b). A figure from *Wikimedia Commons* (URL) by Brylie Christopher Oxley / CC0, 2017, demonstrating a 64-channel EEG locations using the International 10–20 standard developed by Jasper (1958). Channels marked with red were used in our ERP-BCI design. (c). An illustration of the data generative mechanism of a single-channel EEG sequence under the ERP-BCI design. Red, blue, green, and yellow blocks represented target responses, nontarget responses, background noise irrelevant to stimuli, and observed signals ( $X_t$ ).  $W, Y$  were stimulus-occurring indicators and stimulus-type indicators. We assumed each stimulus-related potential could be characterized by  $\beta_1$  or  $\beta_0$  with a long and fixed response window; the observed signal was generated when we aligned different signal components and summed up at each time point. For example, given the target character was “T,” the fourth stimulus was the target one. The graph in the bottom right of the figure illustrates the empirical ERP estimates from channel C2 based on a real participant, where target and nontarget ERP estimates were averaged over 570 and 2850 EEG signal segments, respectively. A significant magnitude difference between target and nontarget ERPs was observed around 300 ms poststimulus.

random effect  $\varepsilon_{l,i}(t)$  is assumed to follow an autoregressive model of order  $q$  and noise variance  $\sigma_x^2$ . For a given channel  $e$  and a letter  $l$ , Figure 2(c) illustrates the proposed Bayesian generative model for half the length of a sequence. Among the consecutive six stimuli, there exists one target stimulus at the 4th stimulus.

### 2.3. The Split-and-Merge GP

To identify the time window that contains major differences in brain activity responses between target and nontarget stimuli, we develop a new GP-based model to model the joint prior distribution of  $\beta_{0,e}(\tau)$  and  $\beta_{1,e}(\tau)$ , for  $\tau \in [0, T_z]$ , named as the split-and-merge GP (SMGP). For  $k = 0, 1$ , we assume that  $\{\beta_{k,1}(\tau), \dots, \beta_{k,E}(\tau)\}$  are independent and marginally follow the same prior distribution specified by the SMGP. For simplicity, we drop the channel-specific subscript  $e$  to specify the SMGP as follows:

$$\beta_k(\tau) = \alpha_k(\tau)\zeta(\tau) + \alpha_0(\tau)\{1 - \zeta(\tau)\}, \quad (2)$$

where  $\alpha_k(\tau) \sim \mathcal{GP}(0, \kappa_\alpha)$  and  $\zeta(\tau) \in [0, 1]$ . Note that  $\beta_0(\tau) = \alpha_0(\tau)$  and  $\beta_1(\tau)$  is the weighted average between  $\alpha_1(\tau)$  and  $\alpha_0(\tau)$  by  $\zeta(\tau)$ . When  $\zeta(\tau) = 0$ ,  $\beta_0(\tau) = \beta_1(\tau)$  with probability one, that is, the two processes are merged; when  $\zeta(\tau) = 1$ ,  $\beta_0(\tau) \neq \beta_1(\tau)$  with probability one. Thus, we refer to  $\zeta(\tau)$  as the split-and-merge indicator process. Let  $\mathcal{W}_s = \{\tau : \zeta(\tau) > \zeta_0\}$  and  $\mathcal{W}_m = \{\tau : \zeta(\tau) \leq \zeta_0\}$  represent the split time window

and the merge time interval, respectively, where  $\zeta_0$  is a hyperparameter. For efficient posterior inference on  $\mathcal{W}_s$  and  $\mathcal{W}_m$ , we define the truncated GP (TGP) similar to the ordinary GP as follows. A time-continuous stochastic process  $\{\zeta(\tau), \tau \in \mathcal{T}\}$  is a truncated GP if and only if for every finite set of indices  $\tau_1, \dots, \tau_p$  in the index set  $\mathcal{T}$ ,  $\zeta_{\tau_1}, \dots, \zeta_{\tau_p}$  follows a multivariate truncated Gaussian distribution, where the truncated domain has the block rectangular shape. In this case, we assign a TGP prior with mean 0.5 and covariance kernel  $\kappa_\zeta$  truncated on  $[0, 1]$  to  $\zeta(\tau)$ , that is,  $\zeta(\tau) \sim \mathcal{TGP}_{[0,1]}(0.5, \kappa_\zeta)$ .

## 3. Posterior Inference

### 3.1. Model Representation and Prior Specification

Let  $\mathcal{MN}(M, U, V)$  denote a matrix normal distribution with location matrix  $M$  and two scale matrices  $U$  and  $V$  (Dawid 1981). We rewrite Equation (1) in the form of matrix normal distribution such that

$$X_{l,i} \sim \mathcal{MN}(M_{l,i}, C_t, C_s), \quad (3)$$

where  $X_{l,i} = (X_{l,i,e})_{e=1}^E$  and  $M_{l,i} = (M_{l,i,e})_{e=1}^E$  are matrix-wise observed EEG signals and predicted EEG signals using convolution for the  $i$ th sequence,  $l$ th target character, respectively.  $C_s$  and  $C_t$  are the spatial and temporal covariance matrices jointly characterizing the random error  $\varepsilon_{l,i} = (\varepsilon_{l,i,e})_{e=1}^E$ , respectively.

Equation (3) can be expressed as

$$\text{vec}(\mathbf{X}_{l,i}) \sim \mathcal{N}(\text{vec}(\mathbf{M}_{l,i}), C_s \otimes C_t), \quad (4)$$

where  $\otimes$  is the Kronecker product and  $\text{vec}(\cdot)$  is the vectorization operator that converts the matrix to the column vector. The log-likelihood of the matrix normal model is

$$\sum_{l,i} -\frac{T}{2} \log \det(C_s) - \frac{E}{2} \log \det(C_t) - \frac{1}{2} \text{tr} \left[ C_s^{-1} (\mathbf{X}_{l,i} - \mathbf{M}_{l,i})^T C_t^{-1} (\mathbf{X}_{l,i} - \mathbf{M}_{l,i}) \right]. \quad (5)$$

Therefore, we rewrite the mean structure of  $\mathbf{M}_{l,i}$  with convolution as follows:

$$\begin{aligned} \text{vec}(\mathbf{X}_{l,i}) &\sim \mathcal{N}(\text{diag}(G_{l,i})\text{vec}(\boldsymbol{\beta}), C_s \otimes C_t), \quad i = 1, \dots, I, \\ & \quad l = 1, \dots, L, \\ \boldsymbol{\beta} &= (\boldsymbol{\beta}_e)_{e=1}^E, \quad \boldsymbol{\beta}_e = (\boldsymbol{\beta}_{1,e}^T, \boldsymbol{\beta}_{0,e}^T)^T = S(\boldsymbol{\zeta}_e)\boldsymbol{\alpha}_e = A(\boldsymbol{\alpha}_e)\boldsymbol{\zeta}_e, \\ \boldsymbol{\alpha} &= (\boldsymbol{\alpha}_e)_{e=1}^E, \quad \boldsymbol{\alpha}_e = (\boldsymbol{\alpha}_{1,e}^T, \boldsymbol{\alpha}_{0,e}^T)^T, \end{aligned} \quad (6)$$

where  $\boldsymbol{\beta}_{1,e}, \boldsymbol{\beta}_{0,e}$  are channel-specific response functions to target and nontarget stimuli after we have applied the SMGP prior.  $\boldsymbol{\alpha}_{1,e}, \boldsymbol{\alpha}_{0,e}$  are channel-specific response functions to target and nontarget stimuli before selection. They follow the  $\mathcal{GP}(\mathbf{0}, \boldsymbol{\kappa}_\alpha)$  with the scale parameters  $\sigma_{0,1,e}^2, \sigma_{0,0,e}^2$ . We use a  $\gamma$ -exponential function shown in Equation (7) to specify the kernel covariance function.

$$k(x_i, x_j) = \sigma_0^2 \exp \left\{ - \left( \frac{\|x_i - x_j\|_2^2}{s_0} \right)^{\gamma_0} \right\}, \quad (7)$$

where  $0 \leq \gamma_0 < 2, s_0 > 0$ . In practice, we treat them as the hyper parameters and select the optimal pair by the Bayes factor (Kass and Raftery 1995).  $\boldsymbol{\zeta}_e$  follows the truncated normal distribution  $\mathcal{TN}_{\mathcal{D}}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  with the prior mean  $\mathbf{0.5}$  and the prior covariance matrix  $\boldsymbol{\Sigma}_{\boldsymbol{\zeta}}$  on the truncated domain  $[0, 1]^{T_z}$ . We use the method by Li and Ghosh (2015) for efficient sampling.  $S, A$  are linear transformations that map  $\boldsymbol{\alpha}_e, \boldsymbol{\zeta}_e$  to  $\boldsymbol{\beta}_e$ .  $G_{l,i}$  is the linear transformation that maps  $\boldsymbol{\beta}_e$  to the predicted EEG signals via convolution. For  $C_s$ , we decompose  $C_s$  as  $\sigma_x^2 \tilde{C}_s$ , where  $\sigma_x^2$  follows the inverse gamma distribution  $\Gamma^{-1}(a_s, b_s)$  with the shape parameter  $a_s$  and the rate parameter  $b_s$ , and  $\tilde{C}_s$  is a positive definite matrix characterized by the distance measure among selected channels. To simplify, we assume all selected channels share the same distance such that  $\tilde{C}_s$  has a compound symmetry structure dependent on the scalar parameter  $\rho_s$ . We use an adaptive rejection sampling method (Gilks and Wild 1992) to sample  $\rho_s$ , where it is originally generated from the uniform distribution  $U(0, 1)$ . For  $C_t(\rho_t)$ , we assume  $\rho_t$  follows a discrete uniform distribution  $\mathcal{U}_d(\mathcal{V}_{\rho_t})$ , where  $\rho_t$  is a two-dimension vector and takes values from a discrete set  $\mathcal{V}_{\rho_t}$  for which the correlation matrix is invertible, that is,  $\|\rho_t\|_1 < 1$ . Finally, the prior specification is as follows:

$$\begin{aligned} \boldsymbol{\alpha}_{1,e} &\sim \mathcal{GP}(\mathbf{0}, \sigma_{1,e}^2 \boldsymbol{\kappa}_\alpha), \quad \boldsymbol{\alpha}_{0,e} \sim \mathcal{GP}(\mathbf{0}, \sigma_{0,e}^2 \boldsymbol{\kappa}_\alpha), \\ \boldsymbol{\zeta}_e &\sim \mathcal{TN}_{[0,1]}(\mathbf{0.5}, \boldsymbol{\Sigma}_{\boldsymbol{\zeta}}), \\ \sigma_x^2 &\sim \Gamma^{-1}(a_s, b_s), \quad \rho_s \sim U(0, 1), \quad \rho_t \sim \mathcal{U}_d(\mathcal{V}_{\rho_t}). \end{aligned} \quad (8)$$

### 3.2. Markov Chain Monte Carlo

We perform the standard Markov chain Monte Carlo (MCMC) method to sample parameters from their posterior conditional distribution given the training set. We adopt the Gibbs sampler to simulate the posterior distribution of  $\boldsymbol{\alpha}, \boldsymbol{\zeta}, \sigma_x^2, \rho_s$ , and  $\rho_t$ . Since  $\boldsymbol{\zeta}$  takes continuous values between 0 and 1, we average the posterior samples of  $\boldsymbol{\beta}_1, \boldsymbol{\beta}_0$  whenever  $\boldsymbol{\zeta}$  samples are smaller than the threshold  $\zeta_0$  for the explicit split-and-merge effect, where  $\zeta_0$  is a hyper parameter, and it takes discrete values in  $\{0.1, 0.2, \dots, 0.8, 0.9\}$  and the optimal one is selected by the Bayes factor. For the convergence check, we run multiple chains with different seed values, and evaluate the conditional log-likelihood and Gelman-Rubin statistic of each parameter (Gelman and Rubin 1992). Details of the Gibbs sampling scheme can be found in the supplementary materials.

### 3.3. Posterior Predictive Probability for Character Classification

Under the RCP design, the selection of the target character requires the selection of the target row among six candidate rows and the target column among six candidate columns. Let  $\mathbf{W}^*, \mathbf{Y}^*$ , and  $\mathbf{X}^*$  be  $I^*$  sequences of stimulus-occurring indicators, stimulus-type indicators, and  $I^*$  sequences of matrix-wise EEG signals from new observations given the same target character  $\omega$ , respectively. Let  $\Theta$  be the parameter set defined in Equation (1). Let  $\mathbf{y}^\omega \in \{0, 1\}, r^\omega, c^\omega$  be the stimulus-type indicator, row index, and column index associated with the target character  $\omega$ , respectively. The probability of  $\omega$  as the target character is

$$\begin{aligned} \Pr(\mathbf{Y}^* = \mathbf{y}^\omega | \mathbf{X}^*, \mathbf{W}^*, \mathbf{X}, \mathbf{W}, \mathbf{Y}) &= \int \Pr(\mathbf{Y}^* = \mathbf{y}^\omega | \Theta; \mathbf{X}^*, \mathbf{W}^*) \pi(\Theta | \mathbf{X}, \mathbf{W}, \mathbf{Y}) d\Theta \\ &= \int \Pr\left(\mathbf{Y}^* = \mathbf{y}^\omega, y_{r^\omega}^\omega = y_{c^\omega}^\omega = 1, y_j^* = 0, \right. \\ & \quad \left. j \notin \{r^\omega, c^\omega\} | \Theta; \mathbf{X}^*, \mathbf{W}^*\right) \pi(\Theta | \mathbf{X}, \mathbf{W}, \mathbf{Y}) d\Theta \end{aligned}$$

where  $\Pr(\mathbf{Y}^* = \mathbf{y}^\omega, y_{r^\omega}^\omega = y_{c^\omega}^\omega = 1, y_j^* = 0, j \notin \{r^\omega, c^\omega\} | \Theta; \mathbf{X}^*, \mathbf{W}^*)$  is proportional to

$$\begin{aligned} \Pr(\mathbf{Y}^* = \mathbf{y}^\omega) &\prod_{i=1}^{I^*} \pi(\mathbf{X}_i^* | \Theta; y_{r^\omega}^\omega = y_{c^\omega}^\omega = 1, y_j^* = 0, \\ & \quad j \notin \{r^\omega, c^\omega\}, \mathbf{W}_i^*). \end{aligned}$$

Here,  $\Pr(\mathbf{Y}^* = \mathbf{y}^\omega) = 1/36$  is the predictive prior on each candidate character if we do not have prior knowledge about the inferred target character. In practice, when we need multiple sequences to select the target character, we compute the cumulative character-based posterior conditional probability vector by multiplying sequence-specific posterior conditional likelihood estimates together.

### 4. Analysis of EEG-BCI Data

We performed the analysis of EEG-BCI data and demonstrated the detailed results from one real BCI participant, referred to as Participant A. Since the primary goal of our analysis was

to identify the spatial-temporal pattern of P300 ERP response signals, participants with clear signal patterns (larger signal-to-noise ratios) were preferred. We selected ten participants among the total population under the RCP design such that the number of sequence to achieve 100% accuracy on the training data with the logistic model was smaller than five. The steps of real-data analysis were as follows: First, we fitted the model to all 16 channels using the spatial dependency correlation of the compound symmetry structure. We identified the spatial-temporally activated locations. Next, we performed the channel selection based on our method and fitted the model to the data for the selected channels using the same spatial dependency assumption. Then, we fitted six existing ML methods to the dataset and compared the prediction accuracy of our method to the other ML methods to evaluate the goodness of model fit. Finally, we provided the cross-participant, sensitivity and reproducibility analyses.

#### 4.1. Dataset and Preprocessing

For the training session, each participant was asked to wear an EEG cap with 16 channels corresponding to different regions on the brain surface and sit approximately 0.8 m from a 17-inch monitor with the BCI display. Figure 2(b) shows the spatial distribution of channels. Channels marked with red were used for recording and analysis purposes. The abbreviated names were F3, Fz, F4, T7, C3, Cz, C4, T8, CP3, CP4, P3, Pz, P4, PO7, PO8, and Oz (Thompson, Gruis, and Huggins 2014). For the calibration dataset, each participant copied a 19-character phrase “THE\_QUICK\_BROWN\_FOX” including three spaces. The stimulus presentation and recording were controlled using the BCI2000 software platform (Schalk et al. 2004). An event was defined as a row stimulus or column stimulus, which highlighted for 31.25 ms and paused for 125 ms afterwards, and the total of 156.25 ms was referred to as the stimulus-to-stimulus interval  $d$ . We defined the 12 stimuli flashing all rows and columns as a sequence and defined multiple sequences as a super-sequence. In our P300 ERP-BCI design, a super-sequence corresponded to the EEG signals associated with the given target character. During the training session, each super-sequence included 15 sequences, and a total of 19 super-sequences were collected. Extra time was recorded after the last stimulus in the super-sequence. The length of each super-sequence was about 29,000 ms with the sampling rate of 256 Hz.

The data preprocessing steps are summarized as follows: First, we applied a notch filter at 60 Hz to remove the power line noise and a band-pass filter between 0.5 Hz and 6 Hz to all 16 channels and then down-sampled raw signals with a decimation factor of eight. Second, we truncated each character-specific super-sequence into 15 sequence segments, where each sequence segment contained 12 consecutive stimuli and subsequent signals of 20 time points to record the entire ERP response to the last stimulus within the single sequence. Each sequence segment contained 2500 ms, 80 sampling points.

#### 4.2. Model Settings

To evaluate the model performance, we chose the odd sequences in the calibration dataset as the training set and used the even

sequences as the testing set. This splitting scheme reduced the overlap between adjacent sequences and attenuated the effect of any shift in attention compared to a random training-testing-split scheme. Since it took time for participants to be familiar with the study design or identify the target characters, we excluded the first sequence of each super-sequence from the training set. Therefore, for the SMGP method, the training set and testing set both ended up with 133 (7 sequences for 19 characters) 80-dimension sequence segments for each channel. We used the cumulative character-level accuracy at seven sequences for prediction evaluation.

For other existing ML methods, we truncated the original character-specific super-sequence into 180 stimulus signal segments in addition to the same band-pass filter, down-sampling procedure, and splitting scheme, where each stimulus signal segment started from the onset of a single stimulus and lasted for 780 ms, that is, 25 sampling points. Therefore, the training set and testing set both contained 1596 (19 characters, each contained 7 sequences of 12 stimuli) 400-dimension concatenated truncated signal segments for all 16 channels.

For the SMGP method,  $\kappa_\alpha$  was generated from a  $\gamma$ -exponential kernel with hyper parameters  $s_0 = 0.5$ ,  $\gamma_0 = 1.8$ ,  $\sigma_{0,1}^2 = 1$ , and  $\sigma_{0,0}^2 = 1$ . For feature selection and prediction of the swLDA method, the inclusion and exclusion probabilities were 0.1 and 0.15, and at most 30% of the feature vector was selected. We ran the MCMC algorithm for 2000 iterations with 1000 burn-ins for three chains with different seed values. We concluded that the algorithm converged, as the Gelman–Rubin statistics for the parameters of interest were all smaller than 1.1.

To rank the importance of the channels, we propose the following statistics based on the SMGP model fitting of multi-channels EEG data:

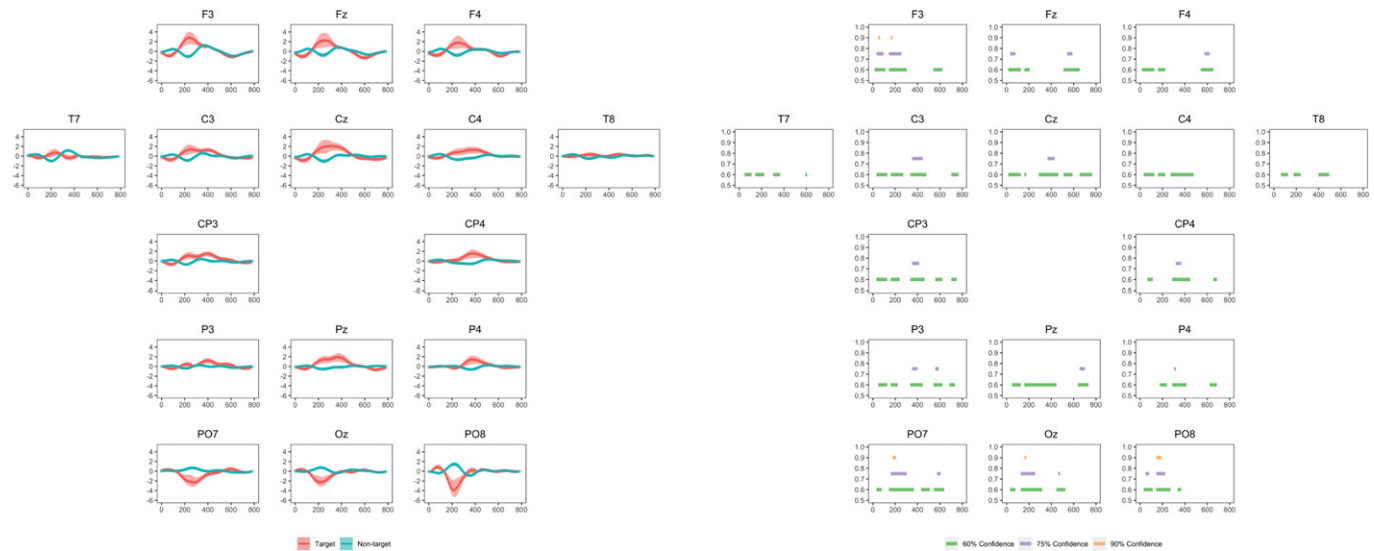
$$R_e^2 = \frac{\text{var}\{\mathbb{E}\{X_e(t)|M_e(t)\}}{\text{var}\{X_e(t)\}}, \quad (9)$$

where the numerator and the denominator explained the variability of the convolution components in Equation (1) across sequences and the variability of the observed signals across sequences, respectively. Under our model assumption,  $R_e^2$  took values between 0 and 1. To examine the proposed information criterion, we included from the optimal two and up to five channels for sub-channel analyses. For each combination of channels, we refitted the model and reported the prediction accuracy.

#### 4.3. Single-Participant Results

We focus on the results of Participant A in this section.

*ERP Estimates:* The left panel of Figure 3 showed the mean estimated ERP functions of target and nontarget stimuli and their 95% credible bands based on the 16-channel model fitting result. Channel-specific plots are arranged by their relative spatial locations. In general, we saw a clear separation of target against nontarget ERP functions for all channels except channel T8. Between 400 ms and 500 ms poststimulus, the target ERP functions gradually declined to zero and collapsed with nontarget ERP functions, which shows that our SMGP prior worked well in this case.



**Figure 3.** *Left Panel:* Channel-specific ERP function estimates of target and nontarget stimuli with the 95% credible bands of Participant A. *Right Panel:* Channel-specific significant temporal intervals by varying thresholds of median split probabilities of Participant A. The result was produced by the 16-channel model fitting results. The varying thresholds included 0.6, 0.75, and 0.9. We arranged the channel-specific plots by their spatial locations. The upper and lower rows represented the front and back of the head. A “z” (zero) referred to a channel placed on the mid-line sagittal plane of the skull. Channels with even numbers (2, 4, 6, 8) referred to the electrode placement on the right side of the head, whereas channels with odd numbers (1, 3, 5, 7) referred to those on the left.

**Split Windows:** The right panel of Figure 3 showed channel-specific significant split time windows with varying thresholds of median split probabilities of 0.6, 0.75, and 0.9. We rearranged channel-specific brain activity plots by their spatial locations. With 90% posterior probability, the split time windows appeared at 50–65 ms and 160–175 ms for channel F3, at 170–205 ms for channel PO7, at 160–170 ms for channel Oz, and at 150–190 ms for channel PO8 poststimulus. These significant split time windows corresponded to the first negative peaks of their target ERP curve estimates. For channel Cz, the split time windows appear at 370–430 ms poststimulus with 75% posterior probability, which approximately corresponded to the first positive peaks of the target P300 ERP response curve estimates. For channel Pz, the split time windows appeared at 650–700 ms poststimulus with 75% posterior probability. For channels T7, C4, and T8 close to ears, moderate differences in brain activity between target and nontarget stimuli were observed, but no split time window was identified with more than 60% posterior probability. A common gap of split time windows around 150 ms was observed, which corresponded to the time points where target and nontarget ERP functions first crossed. For time points when target and nontarget ERP functions were merged, fewer points were generally selected by the SMGP prior.

**Interpretation:** Two common patterns were observed among the results of the ERP estimates. First, the target ERPs of the frontal and central channels (channel names starting with “F” and “C”) shared the negative drop around 100 ms and reached their first peak with the latency around 250 ms, which corresponded to the N100 and P300 pattern described by Rodden and Stemmer in 2008. Second, the target ERPs of parietal-occipital and occipital channels (channel names starting with “PO” and “O”) reached their negative peaks around 200 ms poststimulus, and they gradually collapsed with nontarget ERP functions *without* reaching a positive peak. Since channels PO7, PO8, and Oz represented the locations of the visual cortex, observing

only the negative peaks might be indicative of the pattern of the N2 signal (Folstein and Van Petten 2008). Several discrepancies were also observed. First, the lengths of the split time windows differed among channels. For example, the central channels and frontal channels had the split time window between the onset of the stimulus and 500 ms poststimulus and between the onset of the stimulus and 400 ms poststimulus, respectively. Second, the shapes of ERP functions differed among channels. For example, channels C3, CP3, and P3 had secondary peaks around 400 ms poststimulus, while target ERP functions of other channels collapsed with the nontarget ones without clear secondary peaks. Those secondary peaks might be indicative of the pattern of the P3b signals (van Dinteren et al. 2014).

**Channel Ranking and Prediction:** According to the 16-channel joint fitting result of the SMGP method and the proposed information criterion  $R_c^2$  in Equation (9), the top five selected channels for Participant A were PO7, PO8, Oz, P4, and Cz. We compared the prediction accuracy of our SMGP method to other ML methods for Participant A to evaluate the goodness of our model fit. Table 1 summarizes the cumulative testing prediction accuracy, comparing the SMGP method to other ML methods at seven sequences for the top five selected channels and all 16 channels, where the best prediction accuracy values are in bold in each row in Table 1. The SMGP method achieved 100% accuracy with channels PO8 and PO7, and maintained 100% with more channels included. It performed better than other ML methods. The SMGP method, swLDA and XGBoost performed perfectly when all channels were used.

**Sensitivity and Reproducibility:** We performed the sensitivity analysis for the dataset of Participant A by changing the hyper parameters of the  $\gamma$ -exponential kernel. We assigned 0.4, 0.5, and 0.6 to the scale parameter  $s_0$  and 1.7, 1.8, and 1.9 to the gamma parameter  $\gamma_0$ . We selected channels PO7, PO8, Oz, P4, and Cz for the sensitivity analysis. Figures S3 and S4, supplementary materials showed the P300 ERP function estimates



**Table 1.** Cumulative prediction accuracy of Participant A for 19 characters comparing the SMGP method with  $\zeta_0 = 0.4$  to other ML methods at seven sequences for the top five selected channels and all 16 channels.

Channels	SMGP	CNN	SVM	Logistic	RF	swLDA	XGBoost
PO8, PO7	<b>1.00</b>	0.89	0.95	0.95	0.95	0.95	0.95
PO8, PO7, Oz	<b>1.00</b>	0.89	1.00	1.00	0.95	1.00	0.95
PO8, PO7, Oz, P4	<b>1.00</b>	0.89	1.00	0.95	1.00	1.00	0.95
PO8, PO7, Oz, P4, Cz	<b>1.00</b>	0.89	1.00	0.95	1.00	1.00	0.95
All Channels	<b>1.00</b>	0.89	0.95	0.95	0.95	1.00	1.00

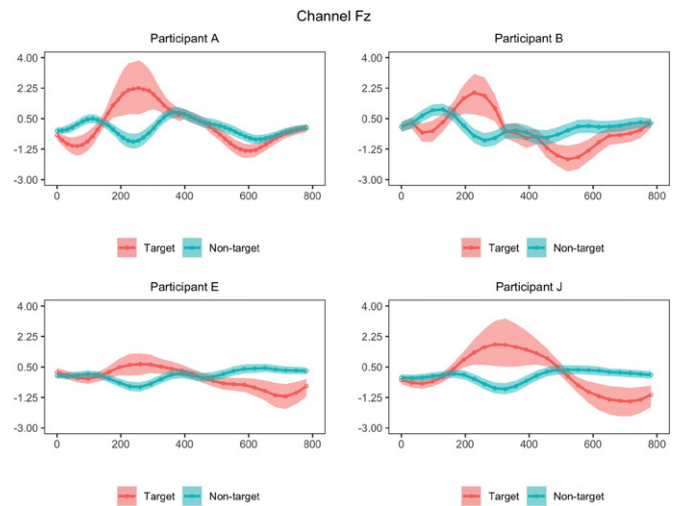
with 95% credible bands and channel-specific significant temporal intervals by different thresholds of median split probabilities for channels Cz and PO8 under nine variations of kernel hyper parameters. Overall, the combination of  $s_0$  and  $\gamma_0$  did not affect either ERP function estimates very much. For channel Cz, we observed the split window with the threshold of 0.90 when  $s_0$  and  $\gamma_0$  were in the middle of the hyper parameter space. Table S4, supplementary materials shows the prediction accuracy with channels PO8, PO7, Oz, P4, and Cz at seven sequences under nine combinations of kernel hyper parameters. The analysis suggested that a combination of moderate  $s_0$  and  $\gamma_0$  produced the best prediction performance for Participant A.

#### 4.4. Cross-Participant Comparison

First, we applied our information criterion to each of the selected ten participants to identify the top five channels by the information criterion in Equation (9), and selected the ultimate top five channels based on the frequency. Then, we identified spatial-temporal patterns of the neural activity based on selected ten participants. Among 10 participants, we selected four typical participants to compare the neural activity patterns between participants with ALS and controls as well as between younger and older participants.

We performed two sensitivity analyses on channel ranking with respect to bandpass filters and kernel hyper parameters. Overall, the channel selection results were robust. For bandpass filters, we always identified channels PO7, PO8, and Oz, followed by channels P4 and Cz. For kernel hyper parameters, we always identified channels PO7, PO8, and Oz, followed by channels P4 and P3. For common neural patterns, target ERPs of frontal and central channels shared the negative drops between 100 ms and 150 ms and reached their first positive peaks around 300 ms poststimulus. Target ERP functions gradually declined to zero and collapsed with nontarget ERP functions between 600 ms and 800 ms poststimulus. Target ERP functions of parietal-occipital, and occipital channels only reached their negative peaks between 200 ms and 250 ms poststimulus without reaching further positive peaks.

In comparing the results of Participant E with ALS to the three healthy controls (A, B, and J), Figure 4 showed the ERP function estimates of channels Fz of the four participants. We identified a common positive peak for target ERP functions around 300 ms poststimulus although Participant E had the smallest peak magnitude of  $0.6 \mu V$  compared to the remaining three above  $2.0 \mu V$ . Finally, we compared the neural activity patterns of two young participants (A and B, around 25 years old) with two senior participants (E and J, around 60 years



**Figure 4.** ERP function estimates of target and nontarget stimuli with 95% credible bands of Participants A, B, E, and J at channel Fz. Participants A and B were young female healthy controls, while Participants E and J were elderly men, of whom only E was diagnosed with ALS.

old). The split-and-merge time windows (SMTW) of frontal channels appeared significantly different between the young and senior participants. On channel Fz, target ERP functions of all participants showed another negative peak after the first major positive peak. For young participants (A and B), target ERP functions merged with nontarget ERP functions after the second negative peak within the 800 ms poststimulus window; however, for senior participants (E and J), target ERP functions were significantly below nontarget ERP functions. One reason is that generally, it takes longer for senior participants to achieve the target P300 response peak (Pavarini et al. 2018). Therefore, for a senior participant, if the ERP response window is set to be longer, target ERP functions may merge with nontarget ERP functions after 800 ms.

#### 5. Simulations

We performed several simulation studies to make statistical inferences and compare the prediction accuracy of our method to other ML methods. To make the simulated data resemble the real data, we assumed the simulated data with an additive signal-and-noise effect. For the signal component, we applied the convolution rule, and designed the ERP functions based on Hoffmann et al. (2008). For the noise component, we considered both Gaussian and student-t distributions to mimic different tail distributions with variances close to the real data. We also considered the autoregressive correlation structure to model the temporal association of the background noise. Finally, we considered a scenario where, given true stimulus-type indicators, a subset of target stimuli was randomly selected as nontarget ones. This pattern mimicked a situation when participants missed target stimuli due to an attention shift in practical BCI use.

Section 5.1 presents a multi-channel simulation study to examine channel ranking and selection by our information criterion, and to evaluate the SMTW with our inference-based criterion. Section 5.2 presents the single-channel simulation study

with different misspecification scenarios to test the robustness of our analysis.

### 5.1. Channel Selection and Ranking

*Setup:* We randomly generated stimulus-occurring indicators and stimulus-type indicators with 19 characters of interest, “THE\_QUICK\_BROWN\_FOX,” including three spaces. To evaluate the performance and the channel ranking, we designed two groups of prespecified mean response functions (MRFs 1 and 2). MRF 1 had different temporal separation effects, while MRF 2 had channel-specific SNR values (Figure S1, supplementary materials). We considered a true generative scenario with two levels of noise variance, that is,  $\sigma_x^2 \in \{20, 40\}$ . We simulated the noise assuming a temporal relationship of AR(2) with the parameter  $\rho_t = (0.5, 0)$  and a spatial dependency relationship of compound symmetry structure with the parameter  $\rho_s = 0.5$ . The EEG signals were generated with a response window of length 935 ms, that is, 30 time points. We performed 100 dataset replications for this scenario. For each dataset, we generated five sequences per character for training and testing.

*Model Settings and Diagnostics:* All simulated datasets were fitted with Equation (3). A feature vector was defined as a three-dimensional super-sequence matrix with five replications and the channel-specific response window was of length 935 ms, that is, 30 time points. The covariance kernel  $\kappa_\alpha$  was assumed with a  $\gamma$ -exponential kernel. The length-scale, gamma, and scaling of nontarget stimuli were  $s_0 = 0.5$ ,  $\gamma_0 = 1.8$ , and  $\sigma_{0,0}^2 = 0.5$ , respectively. For simulation studies with MRF 1, the peak ratios of target to nontarget stimuli were all 5; for simulation studies with MRF 2, the peak ratios of target to nontarget stimuli were 5, 2, and 1, respectively. We ran the MCMC for 2000 iterations with 1000 burn-ins. The MCMC convergence was assessed by running three chains with different seeds and initial values. The Gelman-Rubin statistics for the parameters of interest were smaller than 1.1, indicating an approximate convergence for each model fit.

*Results:* To evaluate the SMTW, we defined two quantities, the inference-based split window ratio (ISWR) and the inference-based merge window ratio (IMWR) as follows:

$$\text{ISWR}(\xi) = \frac{|\{t : \hat{\zeta}(t) > \zeta_0 \ \& \ \zeta(t) = 1\}|}{|\{t : \zeta(t) = 1\}|},$$

$$\text{IMWR}(\xi) = \frac{|\{t : \hat{\zeta}(t) \leq \zeta_0 \ \& \ \zeta(t) = 0\}|}{|\{t : \zeta(t) = 0\}|}.$$

Since the swLDA method explicitly performed feature selection, we defined the estimation-based selection window ratio (ESWR) and the estimation-based exclusion window ratio (EEWR) as follows:

$$\text{ESWR}(\xi) = \frac{|\{t : \hat{\zeta}(t) = 1 \ \& \ \zeta(t) = 1\}|}{|\{t : \zeta(t) = 1\}|},$$

$$\text{EEWR}(\xi) = \frac{|\{t : \hat{\zeta}(t) = 0 \ \& \ \zeta(t) = 0\}|}{|\{t : \zeta(t) = 0\}|}.$$

Table 2 summarized the channel-specific ISWR, IMWR of the SMGP method and the ESWR, EEWR of the swLDA method,

**Table 2.** Channel-specific ISWR, IMWR of the SMGP method and the ESWR, EEWR of the swLDA method.

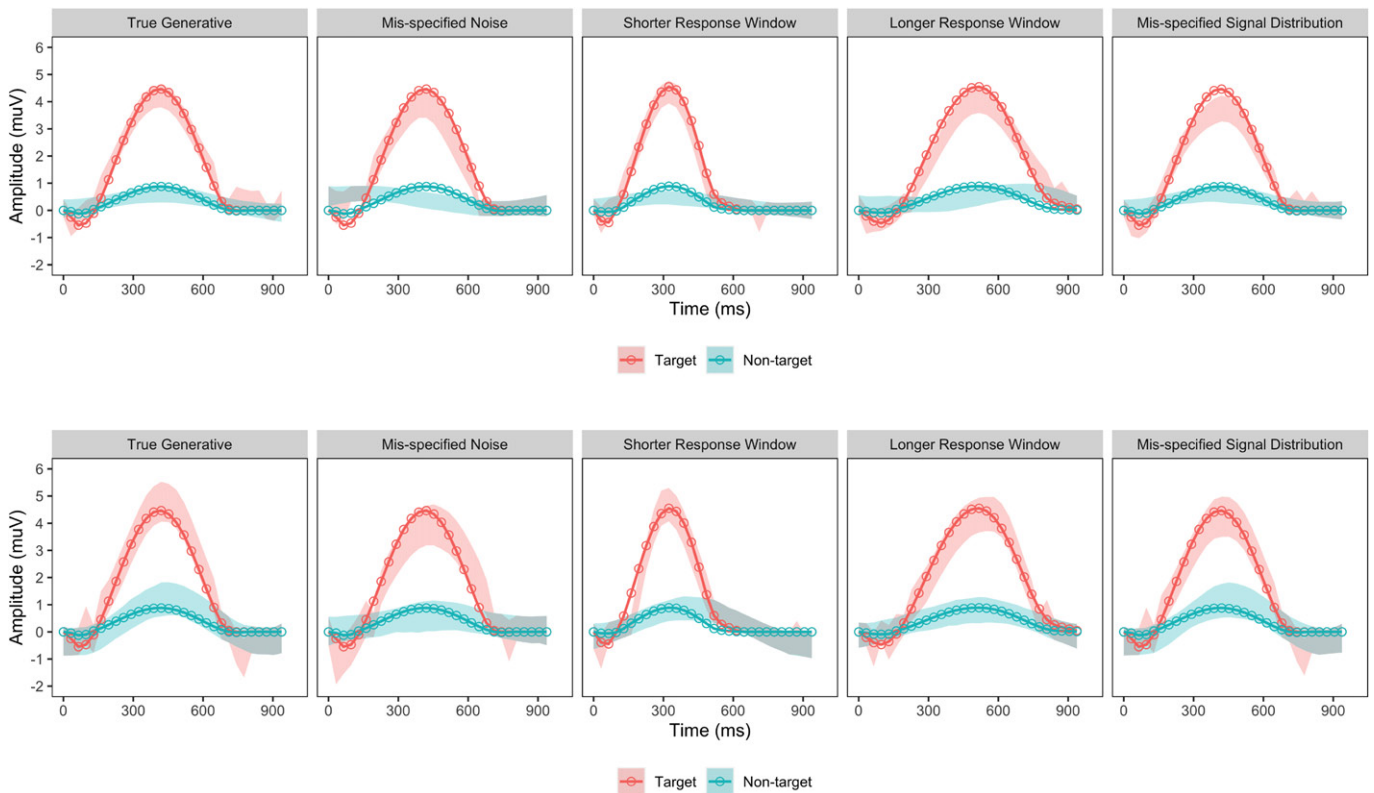
Methods	Testing sequences		
	3	4	5
<b>SMGP</b>	<b>0.91 (0.07)</b>	<b>0.96 (0.04)</b>	<b>0.99 (0.03)</b>
Neural network	0.76(0.10)	0.87(0.08)	0.92(0.07)
SVM	0.81(0.09)	0.89(0.07)	0.94(0.06)
Logistic regression	0.76(0.08)	0.87(0.07)	0.91(0.06)
Random forest	0.76(0.10)	0.86(0.08)	0.92(0.06)
swLDA	0.85(0.08)	0.93(0.06)	0.97(0.04)
XGBoost	0.67(0.11)	0.77(0.09)	0.85(0.08)
Channels	SMGP	swLDA	
	ISWR	IMWR	ESWR EEWR
1	0.98 (0.03)	0.56 (0.11)	0.32 (0.07) 0.69 (0.08)
2	0.99 (0.03)	0.56 (0.12)	0.32 (0.07) 0.75 (0.09)
3	0.99 (0.02)	0.59 (0.11)	0.26 (0.07) 0.8 (0.09)

NOTE: *Upper Panel:* Cumulative prediction accuracy for the multi-channel simulation study under the true generative mechanism with  $\sigma_x^2 = 20$ ,  $\rho_t = (0.5, 0)$ ,  $\rho_s = 0.5$  comparing the SMGP method to other ML methods. The split threshold of SMGP method was  $\zeta_0 = 0.5$ . Point estimates and standard errors averaged over 100 datasets were reported. Results of the SMGP method were marked in bold. Overall, the SMGP method had the highest and most precise prediction accuracy. *Lower Panel:* The ISWR, IMWR of the SMGP method and the ESWR, EEWR of the swLDA method for the multi-channel simulation study under the true generative mechanism with  $\sigma_x^2 = 20$ ,  $\rho_t = (0.5, 0)$ . Channel-specific point estimates and standard errors averaged over 100 datasets were reported.

and the cumulative prediction accuracy over the number of testing sequences with  $\sigma_x^2 = 20$  comparing the SMGP method to other ML methods. The ISWR of the SMGP method was close to 100%, which indicated that our method identified relevant temporal features better than the swLDA method. Our method also had the highest and most precise prediction accuracy among all methods. Similar results were obtained when we used  $\sigma_x^2 = 40$ . Plots of ERP function estimates for both  $\sigma_x^2 = 20, 40$ , prediction accuracy, and the SMGP prior evaluation for  $\sigma_x^2 = 40$  were shown in the Supplementary Material. For simulation studies with varying SNR values, the means and standard errors of  $R_c^2$  estimates were 20.52(1.55), 9.94(1.07), 4.81(0.82) for  $\sigma_x^2 = 20$ , and 10.66(1.05), 4.90(0.68), 2.48(0.53) for  $\sigma_x^2 = 40$  (values multiplied by 100). The information criterion ranked three channels successfully for all the datasets, indicating that the information criterion worked well.

### 5.2. Misspecification Scenarios

*Setup:* The stimulus-occurring indicators and stimulus-type indicators were generated randomly following the same rule as in Section 5.1. We illustrated the design of the prespecified mean response functions in Figure 5. For the data generative mechanism, we considered the following five scenarios with the AR(2) temporal correlation parameter  $\rho_t = (0.5, 0)$  and two levels of the noise variance  $\sigma_x^2 = 10, 20$ . (i) The true generative mechanism scenario simulated the data completely from Equation (1). (ii) The misspecified noise scenario simulated the data from Equation (1) with the noise following a Student-t distribution with 5 degrees of freedom. (iii) The scenario of the shorter response window length simulated the data with prespecified mean response functions of length 780 ms, that is, 25 time points. (iv) The scenario of the longer response window length simulated the data with prespecified mean response



**Figure 5.** The upper and lower panels showed the 95% credible bands of ERP functions to target and nontarget stimuli under five simulation scenarios with true parameter  $\sigma_x^2 = 10, \rho = (0.5, 0)$  and  $\sigma_x^2 = 20, \rho = (0.5, 0)$ , respectively. The split threshold was  $\zeta_0 = 0.5$ . The dots and curves were the true curve values. For the true generative scenario, the credible bands covered the entire true curve. For the misspecified scenarios, the credible bands almost covered the true curves.

functions of length 1090 ms, that is, 35 time points. (v) The misspecified signal scenario simulated the data with a disproportionate distribution of target and nontarget stimuli. Given true stimulus-type indicators, a subset (10%) of target stimuli was randomly treated as nontarget ones by mistake so that it produced the incorrect target P300 ERPs. The replication size, training sequences, and testing sequences were the same as in Section 5.1.

*Model Settings and Diagnostics:* All simulated datasets were fitted with the proposed model with the estimated response window of length 935 ms, that is, 30 time points. The covariance kernel  $\kappa_\alpha$  was set to an exponential squared kernel. The length-scale, the scaling of target stimuli, and the scaling of nontarget stimuli were  $s_0 = 0.5, \sigma_{0,1}^2 = 10$ , and  $\sigma_{0,0}^2 = 0.5$ , respectively. We ran the MCMC for 2000 iterations with 1000 burn-ins. The MCMC convergence was assessed by running three chains with different seeds and initial values. The Gelman-Rubin statistics for the parameters of interest were smaller than 1.1, indicating an approximate convergence.

*Results:* Figure 5 showed the estimated ERP functions for target and nontarget stimuli under five scenarios with true parameters  $\sigma_x^2 = 10$  (the upper panel) and  $\sigma_x^2 = 20$  (the lower panel). For the true generative scenario, the credible bands covered the entire true curves. For the misspecified scenarios, credible bands almost covered the true curves. The posterior distributions of  $\sigma_x$  and  $\rho$  concentrated around the true values. Table 3 summarizes the ISWR, IMWR of the SMGP method and the ESWR, EEWR of the swLDA method under five scenarios

with  $\sigma_x^2 = 10$  (the upper panel) and  $\sigma_x^2 = 20$  (the lower panel). Both point estimates and standard errors over 100 datasets were computed. In the single-channel setting, both the ISWR and IMWR of our method were higher than the ESWR and EEWR of the swLDA method. This result implied that our method identified time windows better than the swLDA method. We also summarized the cumulative prediction accuracy under five scenarios comparing the SMGP method to other ML methods. The prediction accuracy of the SMGP method among the misspecified scenarios was consistently higher than the other ML methods, suggesting that our analysis was relatively robust to moderate model mis-specifications.

### 6. Discussion

We have applied a new Bayesian generative framework to model the conditional distribution of multi-sequence EEG signals from real participants under the P300 ERP design. Our Bayesian analysis explored the mechanism of brain activity in response to external stimuli by directly considering the overlapping ERPs between adjacent stimuli without signal concatenation and segmentation. We developed a new GP-based prior to identify the spatial-temporally activated intervals with the split-and-merge GP (SMGP) prior. We proposed an information criterion for channel ranking and confirmed it with existing literature.

We made fully posterior inferences on participant-and-channel specific P300 ERPs with the SMGP prior given a fixed EEG response window. Although past studies by (D’Avanzo et al. 2011; Mowla et al. 2018) have developed Bayesian and

**Table 3.** The detection accuracy of the SMTW of the SMGP and swLDA methods for the single-channel simulation study under five scenarios with  $\sigma_x^2 = 10$ ,  $\rho_t = (0.5, 0)$  in the upper panel and  $\sigma_x^2 = 20$ ,  $\rho_t = (0.5, 0)$  in the lower panel.

$\sigma_x^2 = 10$	SMGP		swLDA	
	ISWR	IMWR	ESWR	EEWR
True generative	0.89 (0.07)	0.96 (0.05)	0.53 (0.08)	0.75 (0.07)
Misspecified noise	0.86 (0.07)	0.94 (0.06)	0.48 (0.07)	0.78 (0.06)
Shorter window	0.91 (0.07)	0.96 (0.04)	0.64 (0.07)	0.79 (0.07)
Longer window	0.86 (0.06)	0.96 (0.06)	0.46 (0.07)	0.72 (0.09)
Misspecified signal	0.86 (0.07)	0.96 (0.04)	0.49 (0.07)	0.76 (0.07)
$\sigma_x^2 = 20$	SMGP		swLDA	
	ISWR	IMWR	ESWR	EEWR
True generative	0.86 (0.07)	0.94 (0.07)	0.47 (0.07)	0.79 (0.08)
Misspecified noise	0.82 (0.08)	0.91 (0.08)	0.41 (0.07)	0.81 (0.07)
Shorter window	0.88 (0.08)	0.95 (0.05)	0.55 (0.08)	0.84 (0.06)
Longer window	0.82 (0.07)	0.93 (0.08)	0.41 (0.08)	0.77 (0.08)
Misspecified signal	0.83 (0.08)	0.94 (0.07)	0.43 (0.07)	0.81 (0.07)

NOTE: The split threshold of the SMGP method was  $\zeta_0 = 0.5$ . Point estimates and standard errors averaged over 100 datasets were reported.

frequentist filtering methods to estimate amplitude and latency of P300 ERP responses, their results were based on single-trial (sequence) EEG signals, and both methods discarded the spatial dependence among channels. Our SMGP method handles multi-channel, multi-sequence, overlapping EEG signals, produces mean P300 ERP estimates with 95% credible bands, and achieves comparable prediction accuracy. When we compare the ERP function estimates of channel Pz for the three methods, they share a small negative drop in amplitude around 100 ms poststimulus, followed by a major positive peak between 200 ms and 450 ms poststimulus. Then, the ERP function estimates gradually decline to zero. The identification of channel-specific SMTW provides statistical evidence for the scientific findings of P300 ERP responses.

In terms of channel ranking and selection, the study by McCann et al. (2015) pointed out that the difference in P300 ERP-BCI communication efficiency was subtle with five or more channels. Both studies performed channel ranking and selection from the same cohort of participants. They identified Cz, Pz, PO7, PO8, and Oz as the top selected channels, which overlapped with our identification of PO7, PO8, Oz, and Cz. These shared selection results provide statistical evidence for spatial distributions of P300 ERP responses. In particular, the finding that channels PO8, PO7, and Oz appear the most frequently supports the finding that the performance of a P300 speller is associated with eye gaze (Brunner et al. 2010). Finally, the participant-specific channel selection helps establish user-specific profiles for efficient brain-computer communications. Thus, we can incorporate user-specific channel selection to design the EEG cap, which increases the implementation speed.

Potential future directions would improve our work. First, we could modify the stimulus presentation paradigm from the current RCP design to the checkerboard design (Townsend et al. 2010). The checkerboard design avoids the refractory effect (Martens et al. 2009) in the RCP design, where participants might miss or fail to produce the second regular P300 ERP response when two target stimuli are too close. Second, we could measure the participant-specific brain connectivity under the no-control (NoC) condition to specify the prior spatial

covariance matrix. For Participant A, we could assume a multi-block compound symmetry structure to estimate within-block, intra-block correlation parameters, and the scalar parameter  $\sigma^2$ . Third, it is also of interest to adjust the potential confounders in the model for single participant analysis, which may include preferences over certain characters to type and the duration of BCI use. This analysis requires a new study design to collect data on those information. In addition, we could develop the framework of a multi-subject analysis to incorporate the age effect by modifying the priors.

Overall, the proposed generative modeling approach performs innovative statistical inferences on brain activity and provides a promising platform to develop the simulation study framework to test other online P300 ERP-BCI study designs. The Bayesian framework also incorporates prior information such as character-to-character relationships to increase the spelling speed.

## Supplementary Materials

The online supplementary materials include details of the MCMC algorithm in Section S1, additional results of the simulation study in Section S2, additional results of sensitivity analyses in Section S3, and additional results of Participants B, E, and J in Section S4.

## Acknowledgements

The authors would like to thank the Editor Professor Heping Zhang, the Associate Editor and reviewers for their helpful comments and constructive suggestions, which led to a much-improved manuscript. The authors would like to acknowledge the participants in our BCI experiments. The content is solely the responsibility of the authors and does not necessarily represent the official views of NICHD, NIH, NIDRR, or the Department of Education.

## Funding

This work was partially supported by grants NSF IIS2123777 (Kang, Huggins, and Zhu), NIH R01DA048993 (Kang), NIH R01MH105561 (Kang), and NIH R01GM124061 (Kang). Collection of the recorded data was supported in part by the National Institute of Child Health and Human Development (NICHD), the National Institutes of Health (NIH) under grant R21HD054697 and by the National Institute on Disability and Rehabilitation Research (NIDRR) in the Department of Education under grant H133G090005.

## ORCID

Jian Kang  <http://orcid.org/0000-0002-5643-2668>

## References

- Brunner, P., Joshi, S., Briskin, S., Wolpaw, J. R., Bischof, H., and Schalk, G. (2010), "Does the 'P300' Speller Depend on Eye Gaze?," *Journal of Neural Engineering*, 7, 056013. [1124,1132]
- Cai, Q., Kang, J., and Yu, T. (2020), "Bayesian Network Marker Selection via the Thresholded Graph Laplacian Gaussian Prior," *Bayesian Analysis*, 15, 79–102. [1123]
- Cecotti, H., and Graser, A. (2010), "Convolutional Neural Networks for P300 Detection with Application to Brain-computer Interfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33, 433–445. [1123]
- Dawid, A. P. (1981), "Some Matrix-variate Distribution Theory: Notational Considerations and a Bayesian Application," *Biometrika*, 68, 265–274. [1125]

- Donchin, E., Spencer, K. M., and Wijesinghe, R. (2000), "The Mental Prosthesis: Assessing the Speed of a P300-based Brain-computer Interface," *IEEE Transactions on Rehabilitation Engineering*, 8, 174–179. [1123]
- D'Avanzo, C., Schiff, S., Amodio, P., and Sparacino, G. (2011), "A Bayesian Method to Estimate Single-trial Event-related Potentials with Application to the Study of the P300 Variability," *Journal of Neuroscience Methods*, 198, 114–124. [1131]
- Farwell, L. A., and Donchin, E. (1988), "Talking off the Top of Your Head: Toward a Mental Prosthesis Utilizing Event-related Brain Potentials," *Electroencephalography and Clinical Neurophysiology*, 70, 510–523. [1122]
- Folstein, J. R., and Van Petten, C. (2008), "Influence of Cognitive Control and Mismatch on the N2 Component of the ERP: A Review," *Psychophysiology*, 45, 152–170. [1128]
- Gelman, A., and Rubin, D. B. (1992), "Inference from Iterative Simulation using Multiple Sequences," *Statistical Science*, 7, 457–472. [1126]
- Gilks, W. R., and Wild, P. (1992), "Adaptive Rejection Sampling for Gibbs Sampling," *Journal of the Royal Statistical Society, Series C*, 41, 337–348. [1126]
- Hoffmann, U., Vesin, J.-M., Ebrahimi, T., and Diserens, K. (2008), "An Efficient P300-based Brain-computer Interface for Disabled Subjects," *Journal of Neuroscience Methods*, 167, 115–125. [1129]
- Jasper, H. H. (1958), "The Ten-twenty Electrode System of the International Federation," *Electroencephalography and Clinical Neurophysiology*, 10, 370–375. [1125]
- Kang, J., Reich, B. J., and Staicu, A.-M. (2018), "Scalar-on-Image Regression via the Soft-Thresholded Gaussian Process," *Biometrika*, 105, 165–184. [1123]
- Kaper, M., Meinicke, P., Grossekaethefer, U., Lingner, T., and Ritter, H. (2004), "BCI Competition 2003-data Set IIb: Support Vector Machines for the P300 Speller Paradigm," *IEEE Transactions on Biomedical Engineering*, 51, 1073–1076. [1123]
- Kass, R. E., and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795. [1126]
- Krusienski, D. J., Sellers, E. W., McFarland, D. J., Vaughan, T. M., and Wolpaw, J. R. (2008), "Toward Enhanced P300 Speller Performance," *Journal of Neuroscience Methods*, 167, 15–21. [1123]
- Leoni, J., Strada, S. C., Tanelli, M., Jiang, K., Brusa, A., and Proverbio, A. M. (2021), "Automatic Stimuli Classification from ERP Data for Augmented Communication via Brain-Computer Interfaces," *Expert Systems with Applications*, 184, 115572. [1123]
- Li, Y., and Ghosh, S. K. (2015), "Efficient Sampling Methods for Truncated Multivariate Normal and Student-t Distributions Subject to Linear Inequality Constraints," *Journal of Statistical Theory and Practice*, 9, 712–732. [1126]
- Martens, S., Hill, N., Farquhar, J., and Schölkopf, B. (2009), "Overlap and Refractory Effects in a Brain-computer Interface Speller Based on the Visual P300 Event-related Potential," *Journal of Neural Engineering*, 6, 026003. [1132]
- McCann, M. T., Thompson, D. E., Syed, Z. H., and Huggins, J. E. (2015), "Electrode Subset Selection Methods for an EEG-based P300 Brain-computer Interface," *Disability and Rehabilitation: Assistive Technology*, 10, 216–220. [1132]
- Mowla, M. R., Huggins, J. E., Natarajan, B., and Thompson, D. E. (2018), "P300 Latency Estimation Using Least Mean Squares Filter," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1976–1979. IEEE. [1131]
- Okumuş, H., and Aydemir, Ö. (2017), "Random Forest Classification for Brain Computer Interface Applications," in *2017 25th Signal Processing and Communications Applications Conference (SIU)*, pp. 1–4. IEEE. [1123]
- Pavarini, S. C. I., Brigola, A. G., Luchesi, B. M., Souza, É. N., Rossetti, E. S., Fraga, F. J., Guarisco, L. P. C., Terassi, M., Oliveira, N. A., Hortense, P., Pedroso, R. V., and Ottaviani, A. C. (2018), "On the Use of the P300 as a Tool for Cognitive Processing Assessment in Healthy Aging: A Review," *Dementia & Neuropsychologia*, 12, 1–11. [1129]
- Polich, J., Howard, L., and Starr, A. (1985), "Effects of Age on the p300 Component of the Event-Related Potential from Auditory Stimuli: Peak Definition, Variation, and Measurement," *Journal of Gerontology*, 40, 721–726. [1124]
- Qiao, X., Guo, S., and James, G. M. (2019), "Functional Graphical Models," *Journal of the American Statistical Association*, 114, 211–222. [1123]
- Rasmussen, C. E. (2003), "Gaussian Processes in Machine Learning," in *Summer School on Machine Learning*, eds. O. Bousquet, U. von Luxburg, G. Rätsch, pp. 63–71, Berlin: Springer. [1123]
- Rodden, F. A., and Stemmer, B. (2008), "A Brief Introduction to Common Neuroimaging Techniques," in *Handbook of the Neuroscience of Language*, eds. B. Stemmer, and H. A. Whitaker, pp. 57–67, Amsterdam: Elsevier. [1122]
- Schalk, G., McFarland, D. J., Hinterberger, T., Birbaumer, N., and Wolpaw, J. R. (2004), "BCI2000: A General-purpose Brain-computer Interface (BCI) System," *IEEE Transactions on Biomedical Engineering*, 51, 1034–1043. [1127]
- Thompson, D. E., Gruis, K. L., and Huggins, J. E. (2014), "A Plug-and-play Brain-computer Interface to Operate Commercial Assistive Technology," *Disability and Rehabilitation: Assistive Technology*, 9, 144–150. [1123,1127]
- Townsend, G., LaPallo, B. K., Boulay, C. B., Krusienski, D. J., Frye, G., Hauser, C., Schwartz, N. E., Vaughan, T. M., Wolpaw, J. R., and Sellers, E. W. (2010), "A Novel P300-based Brain-computer Interface Stimulus Presentation Paradigm: Moving beyond Rows and Columns," *Clinical Neurophysiology*, 121, 1109–1120. [1132]
- van Dinteren, R., Arns, M., Jongasma, M. L., and Kessels, R. P. (2014), "P300 Development across the Lifespan: A Systematic Review and Meta-analysis," *PloS one*, 9, e87347. [1128]
- Viana, S. S., Batista, D. M., and Melges, D. B. (2014), "Logistic Regression Models: Feature Selection for P300 Detection Improvement," in *XXIV Brazilian Congress on Biomedical Engineering—CBEB (Vol. 2014)*, pp. 979–982. [1123]
- Wolpaw, J. R., Bedlack, R. S., Reda, D. J., Ringer, R. J., Banks, P. G., Vaughan, T. M., Heckman, S. M., McCane, L. M., Carmack, C. S., Winden, S., McFarland, D. J., Sellers, E. W., Shi, H., Paine, T., Higgins, D. S., Lo, A. C., Patwa, H. S., Hill, K. J., Huang, G. D., and Ruff, R. L. (2018), "Independent Home Use of a Brain-computer Interface by People with Amyotrophic Lateral Sclerosis," *Neurology*, 91, e258–e267. [1122]
- Xu, N., Gao, X., Hong, B., Miao, X., Gao, S., and Yang, F. (2004), "BCI Competition 2003-Data Set IIb: Enhancing P300 Wave Detection Using ICA-based Subspace Projections for BCI Applications," *IEEE Transactions on Biomedical Engineering*, 51, 1067–1072. [1123]