Accounting for time dependence in large-scale multiple testing of event-related potential data

Ching-Fan Sheu, Emeline Perthame, Yuh-Shiow Lee, David Causeur

The Annals of Applied Statistics
2016, Vol. 10, No. 1, 219–245
DOI: 10.1214/15-AOAS888

ACCOUNTING FOR TIME DEPENDENCE IN LARGE-SCALE MULTIPLE TESTING OF EVENT-RELATED POTENTIAL DATA[1]

BY CHING-FAN SHEU[2,*], ÉMELINE PERTHAME[2,†], YUH-SHIOW LEE[‡] AND DAVID CAUSEUR[†]

National Cheng Kung University[*] and Agrocampus-Ouest[†] and National Chung Cheng University[‡]

Event-related potentials (ERPs) are recordings of electrical activity along the scalp time-locked to perceptual, motor and cognitive events. Because ERP signals are often rare and weak, relative to the large between-subject variability, establishing significant associations between ERPs and behavioral (or experimental) variables of interest poses major challenges for statistical analysis.

Noting that ERP time dependence exhibits a block pattern suggesting strong local and long-range autocorrelation components, we propose a flexible factor modeling of dependence. An adaptive factor adjustment procedure is derived from a joint estimation of the signal and noise processes, given a prior knowledge of the noise-alone intervals. A simulation study is presented using known signals embedded in a real dependence structure extracted from authentic ERP measurements. The proposed procedure performs well compared with existing multiple testing procedures and is more powerful at discovering interesting ERP features.

1. Introduction. High-throughput instrumental data such as event-related potentials [ERPs, see Handy (2004)] and functional magnetic resonance imaging (fMRI) [Poldrack, Mumford and Nichols (2011)] have become extensively used in both clinical and research settings. The former provides high temporal resolution to chart the time course of mental processes, whereas the latter implicates spatial areas in the brain that might be responsible for experimental effects. With the routine collection of massive amounts of data from ERP or fMRI studies, researchers must face the challenge of multiple comparison corrections: in sifting, simultaneously, through thousands of comparisons for significant effects, a balance must be struck between keeping a low false positive error rate while maintaining sufficient power for correct detection. How to achieve this objective for ERPs exhibiting a strong and complex dependence pattern over time is the focus of the present paper.

Two papers summarize the current status of mass univariate analysis of ERPs [Groppe, Urbach and Kutas (2011a, 2011b)]. These papers focused on comparing a

Received July 2014; revised October 2015.

[1]Supported in part by a Grant (MOST 103-2410-H-006-032-MY3) from the Ministry of Science and Technology of Taiwan to Ching-Fan Sheu.

[2]Ching-Fan Sheu and Émeline Perthame contributed equally to this work.

Key words and phrases. Dependence, ERP data, high-dimensional data, multiple testing.

variety of false discovery rate (FDR) control procedures [Benjamini and Hochberg (1995)] and permutation tests [e.g., Blair and Karniski (1993)], but they made no mention of the problem of dependent tests generated by the highly correlated ERPs over time. However, highly correlated data can severely affect the accuracy of FDR estimation and the stability of simultaneous testing (i.e., variances of discovery proportions) [Efron (2007)]. Consequently, ignoring dependence among test statistics also reduces the ability to detect true positives [Leek and Storey (2008)].

The pronounced pattern of temporal dependence observed in ERPs can induce a long-range regularity in the test statistics, resulting in spuriously low p-values outside of the support of the signal. Several different approaches can be taken to address the problem of dependent test statistics. Before the FDR controlling procedures became popular, Guthrie and Buchwald (1991) had proposed a test which considers significant only those runs of p-values lower than a preset threshold, for example, 0.05, whose lengths are unusually long with respect to a reference distribution for the lengths of such runs assuming an auto-regressive process under the null. The procedure, however, is not designed to control proportions of false positives. An alternative approach to dealing with correlation in multiple testing is to account for dependence by a hidden Markov model [Sun and Cai (2009)] assuming a latent class structure for the data. Another more general approach is to account for the multivariate dependence by some data reduction techniques involving latent variables [see SVA, for Surrogate Variable Analysis, by Leek and Storey (2008), LEAPP, for Latent Effect Adjustment After Primary Projection, by Sun, Zhang and Owen (2012) and, more recently, Allen, Grosenick and Taylor (2014)]. In genomic data analysis, a notable example of this approach is the factor analytic multiple testing procedure (FAMT) proposed by Friguet, Kloareg and Causeur (2009) under the assumption that the conditional covariance of the responses given the treatment variables can be well approximated by its factor components [Mardia, Kent and Bibby (1979)]. The FAMT procedure is especially applicable when accounting for unobserved processes whose effects can linearly affect responses.

These methods based on latent variables essentially differ in how the covariate's effect and the latent effects are disentangled in the estimation procedure. They all assume sparsity of the signal and that signal-free features can be identified to enable estimation of the factor structure of the noise dependence. In FAMT [Friguet, Kloareg and Causeur (2009)], a preliminary thresholding method on selection statistics is used to identify the set of null features. SVA [Leek and Storey (2008)] estimates the covariate's coefficients without first adjusting for correlation between the covariate and latent variables and then iteratively isolates the latent effects by downweighting the features for which the covariate's effect is nonzero. LEAPP [Sun, Zhang and Owen (2012)] splits the data in two and introduces a rotation matrix which transforms the data such that the covariate's effect is removed for all the rotated features except one. The latent effects are then estimated with

the former null rotated features using a mixed-effects regression estimation procedure. Finally, the factor structure is plugged into the estimation of the covariate's effect which is concentrated in the non-null rotated feature. The FAMT procedure [Causeur et al. (2011), Friguet, Kloareg and Causeur (2009)] has been modified for a dynamic factor-adjusted modeling of ERPs arising from the standard analysis of variance designs in Causeur et al. (2012). The method showed marked improvement over the standard procedures for ERP data analysis in detecting true signals in simulation studies.

However, none of the decorrelation procedures mentioned above make use of the highly regular time-dependence structure to disentangle the true signal from the noise as Guthrie and Buchwald (1991) have done. In the present paper, we illustrate how the regularity of the estimated signal can lead to a misidentification of support for the signal, which, in turn, produces an erroneous disentanglement of the covariate's effect and the latent effects. We therefore propose an estimation method that alternates between fitting the covariance factor structure and updating the estimated signal given the covariance between test statistics and a prior knowledge of signal-free time intervals.

This paper is organized as follows. Section 2 presents the linear model setting for significance analysis of ERP data. It also introduces two ERP studies: a comparison of mean ERP curves observed in two experimental conditions in a standard auditory oddball paradigm and a more complex directed forgetting experiment in which ERP time points are to be correlated with recognition memory performance which is the behavioral measure of interest. The strong time dependence among test statistics for the association between ERPs and the recognition performance is investigated. Section 3 proposes a factor regression model as a general framework to handle time dependence in large-scale significance analysis of ERP data. Section 4 proposes an adaptive factor-adjustment procedure, which iteratively captures the dependence of residual ERPs by a factor-analytic model and simultaneously corrects the estimation of the signal for the regularity induced by highly correlated responses. Section 5 presents the results of simulations comparing the proposed method against the classical procedures introduced in Section 2. Also included for comparison are two factor regression models: the surrogate variable analysis by Leek and Storey (2008) and the latent effect adjustment after primary projection method by Sun, Zhang and Owen (2012). In the final section, the ERP data arising from the auditory and the memory experiments are analyzed, respectively, using the proposed method. While the usual FDR-controlling procedures are unable to detect any meaningful difference between mean curves in the auditory experiment, the proposed method identifies significant intervals associated with the expected ERP component called mismatch negativity (MMN) [Näätänen (2003)] that has been well documented in the research literature. Similarly, in the memory experiment, no meaningful association between ERP time intervals and recognition performance is located by the standard procedures. Interestingly, the proposed method discovers a significant waveform correlation signal around 400 milliseconds (ms),

which could be explained by the FN400 component reported in the literature of ERPs and recognition memory [Rugg and Curran (2007)].

2. A general linear model framework for ERP data analysis. Event-related potentials (ERP) are voltage changes along the scalp time-locked to some physical or mental occurrence in the ongoing electrical brain activity recorded as an electroencephalogram (EEG). In the present section, a general framework is introduced for the significance analysis of ERP data, illustrated by two studies: the first one uses a standard paradigm, whereas the second one has a more sophisticated design involving a behavioral response as covariate.

Auditory oddball paradigm. In ERP studies, perhaps the most commonly used experimental task is the oddball paradigm. In this paradigm, typically two classes of stimuli (visual or auditory) are presented, one occurring frequently and the other occurring infrequently (odd). The subject is instructed to respond to the stimuli either actively (by pressing a button, say) or passively (by simply attending to them).

An auditory ERP study was performed at Kaohsiung Medical University in Taiwan to provide a test case data for the present investigation. College students with normal hearing are recruited for the study. The stimuli are two pure tones of 500 Hz and 1000 Hz. The former is presented 120 out of 150 trials, whereas the latter (odd) is presented only for 30 trials. The order of tone presentation is random. Participants were asked to fixate on a cross in the center of the screen and pay attention to the stimuli throughout the duration of the passive listening task. At each of four electrode locations (FZ, C3, C4 and O1), ERP waveform was obtained from each of the two tone conditions for each of the 13 participants.

Each curve comprises a total of 1000 ERP measurements, beginning at 100 ms before stimulus onset and terminating at 399 ms afterward with one record per 0.5 ms. For subsequent analysis, only the ERPs from the electrode FZ at the frontal medial scalp location will be used because maximal responses have typically been observed in this region [Näätänen (2003)]. Figure 1 displays the 26 ERP curves (two for each participant) obtained at electrode FZ.

The classical multivariate two-way analysis of variance model is used for the above significance analysis. For the ERP measurement Y_{jkt} for participant j, $j = 1, \ldots, n$, at time t in condition k ($k = 1$ for "500 Hz" and $k = 2$ for "1000 Hz"),

$$Y_{jkt} = \mu_t + \alpha_{jt} + \beta_{kt} + \varepsilon_{jkt}, \tag{2.1}$$

where $\alpha_t = (\alpha_{1t}, \ldots, \alpha_{nt})$ stands for the participant effect with $\sum_j \alpha_{jt} = 0$ and $\beta_t = (\beta_{1t}, \beta_{2t})$ for the condition effect with $\beta_{1t} + \beta_{2t} = 0$. At each time point t, the significance analysis of the difference curve is equivalent to testing the null hypothesis $H_0^{(t)} : \beta_{2t} = 0$, $t = 1, \ldots, T$.

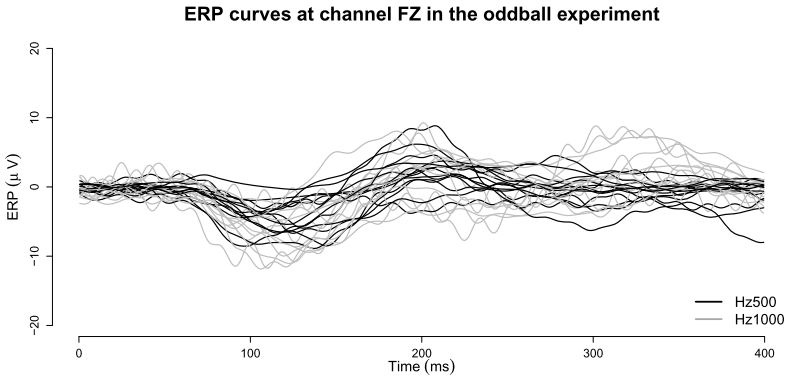

FIG. 1. *ERP amplitude over time recorded at each half-ms after the onset at the frontal medial electrode location FZ in the auditory oddball experiment.*

Directed forgetting paradigm. People may be motivated to forget unpleasant events that have happened to them and their ability to do so successfully could be linked to emotional well-being [Weiner (1968)]. The directed forgetting paradigm refers to experimental procedures by which participants can be instructed to intentionally forget previously studied information [Johnson (1994)].

The study used the item method similar to that described by Lee, Lee and Fawcett (2013) to investigate the time course of directed forgetting. The experiment consists of two phases. In the study phase, twenty participants were instructed (with a "+" or "X" cue, respectively) to either remember or forget a stimulus word that had been displayed briefly on a computer screen. ERPs were recorded throughout each of 90 trials—half for to-be-remembered (TBR) and half for to-be-forgotten (TBF)—each lasting for one thousand milliseconds. Subsequently, participant's ability to recognize whether or not the word had been presented before was tested (old or new). In the test phase, 90 new words were mixed with 90 old words. The proportion of hits (a correct "old" response to a word that had indeed been presented) minus that of false alarms (an "old" response to a "new" word that had not been presented before) was used as a measure of recognition performance. ERP amplitudes recorded once per ms from nine electrode positions—3 each from frontal, central and posterior regions—during the study phase were analyzed. The ERPs were first averaged over trials by condition for each participant.

It is conjectured that brain activations would be different depending on whether people were cued to remember or to forget and that the difference could be inferred from over which time intervals ERPs for a condition are found to be significantly correlated with the recognition performance. At each electrode position (channel) on the scalp, the research question can be cast as a large-scale significance analysis of the statistical relationships between ERPs and the recognition performance x, for each instruction condition: for the ERP measurement Y_{jkt} for participant j,

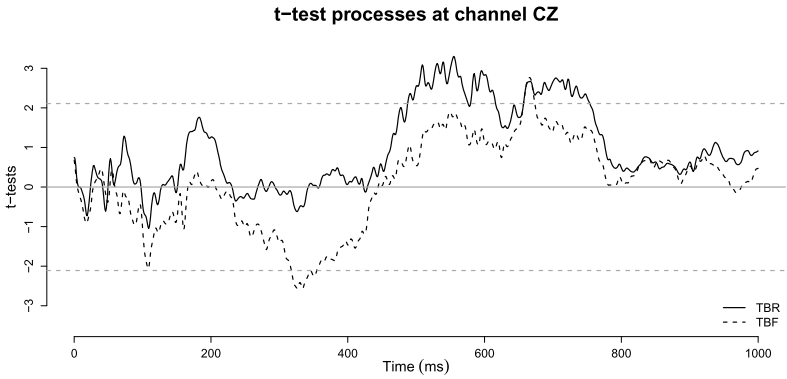

FIG. 2. *t-tests for the significance of β_t at channel CZ for the TBR (solid curve) and TBF (dashed) condition, respectively. The top and bottom horizontal lines give, respectively, the 2.5th and 97.5th quantiles of the null distribution.*

$j = 1, \ldots, n$, at time t in condition k ($k = 1$ for TBR and $k = 2$ for TBF),

$$Y_{jkt} = \mu_t + \alpha_{jt} + \gamma_{kt} + \beta_{kt} x_{jk} + \varepsilon_{jkt}, \tag{2.2}$$

where $\alpha_t = (\alpha_{1t}, \ldots, \alpha_{nt})$ stands for the participant effect with $\sum_j \alpha_{jt} = 0$ and $\gamma_t = (\gamma_{1t}, \gamma_{2t})$ for the instruction condition effect with $\gamma_{1t} + \gamma_{2t} = 0$. At each time point t in a condition k, the analysis of the relationship between ERP measurement Y_{kt} and the recognition performance x is equivalent to testing the null hypothesis $H_0^{(kt)} : \beta_{kt} = 0$, $k = 1, 2$, $t = 1, \ldots, T$. The observed values of the corresponding t-statistics at channel CZ are displayed in Figure 2.

For each condition and especially for TBF, significant time points are rare, as indicated by the t-statistics in Figure 2. More importantly, the curves show a strong regularity inconsistent with the expected profile of a sequence of independently distributed Student's t-variables. The strong dependence among tests is known to affect the joint null distribution of test statistics. This strong temporal regularity is also confirmed by Figure 3: The histogram, on the top of the left panel, shows that a large portion of the residual correlations of model (2.2) at channel CZ are strongly positive; the image plot of the residual correlation matrix, on the top right, shows an apparent autocorrelation component generating a larger number of correlations near one along the diagonal as well as blocks of strong positive correlations—with intervals of highly inter-correlated time points and an increasing lag-1 autocorrelation over time. Similar patterns of dependence are observed at other channels. As a reference, 40 trajectories are generated according to a first-order autoregressive process. The autocorrelation parameter is set to 0.99, as estimated on ERP data at channel CZ. The bottom panels of Figure 3 show a histogram and an image plot of estimated correlations among variables for this dataset. The pattern of dependence observed on ERP data appears more complex than what could be produced by a first-order autoregressive structure.

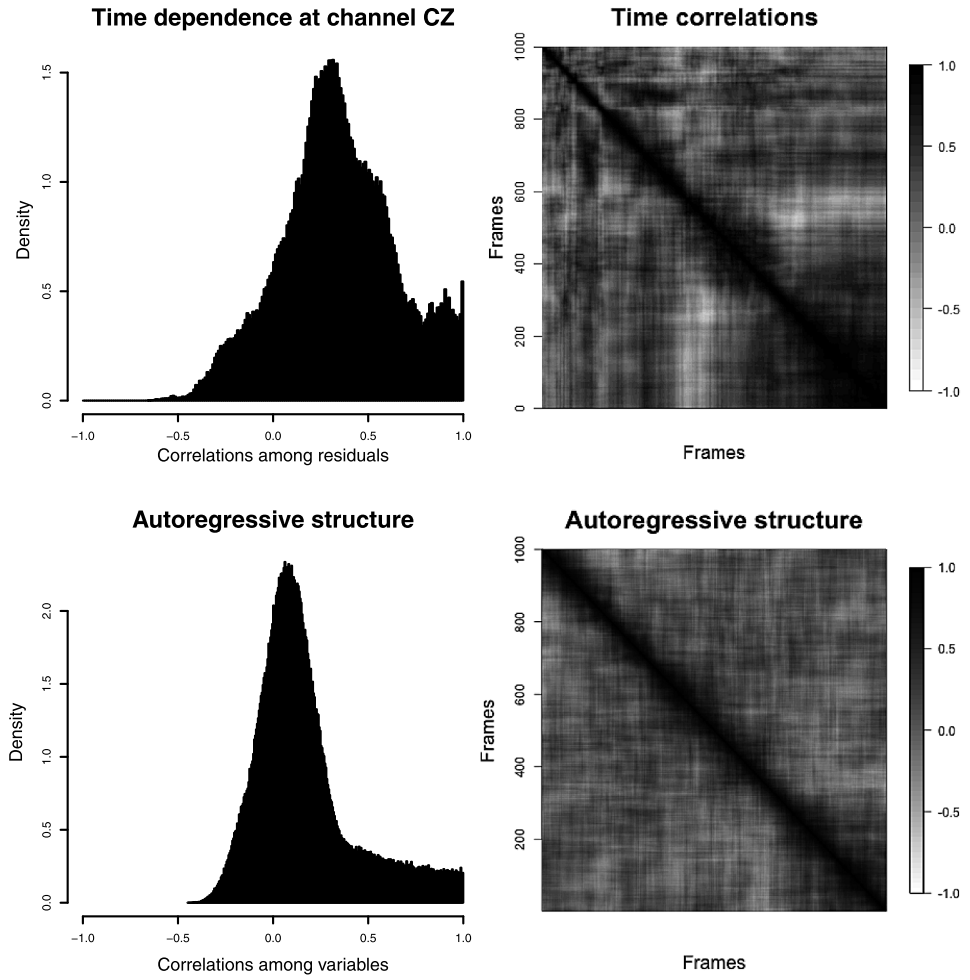

FIG. 3. *Top of the left panel*: *histogram of correlations among residuals of model* (2.2) *at channel* CZ *over time. Top of the right panel*: *image plot of the correlations among residuals over time frames. Bottom of the left panel*: *histogram of correlations among variables generated according to a first-order autoregressive process. Bottom of the right panel*: *image plot of the correlations among variables generated according to a first-order autoregressive process.*

Multivariate Analysis of Variance modeling of ERPs. The following general framework for the significance analysis of ERPs explicitly accounts for the time dependence. Let Y_{it} be the measured ERP for subject $i = 1, \ldots, n$, at time t, with $t = 1, \ldots, T$, where T is the number of frames. For example, a trial lasting for 1000 ms with an ERP recording per 10 ms yields 100 frames. A multivariate linear model is assumed for the relationship between the ERPs and covariates $x_i = (x_{i1}, \ldots, x_{ip})'$, adjusted for the effect of other covariates $z_i = (z_{i1}, \ldots, z_{ir})$

when necessary:

$$Y_{it} = \mu_t + \beta_t' x_i + b_t' z_i + \varepsilon_{it}, \tag{2.3}$$

where μ_t is the intercept at time t, β_t and b_t are the p- and r-vectors of regression coefficients associating the ERP at time t with x and z, respectively, and ε_{it} is the random error term, normally distributed with mean 0 and standard deviation σ_t. Typically, independence is assumed among the errors ε_{it}: for each participant i, the random vector $\varepsilon_i = (\varepsilon_{i,1}, \varepsilon_{i,2}, \ldots, \varepsilon_{i,T})'$ is assumed to be normally distributed with mean 0 and variance $D_\sigma = \operatorname{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_T^2)$, where $\operatorname{diag}(\cdot)$ stands for the matrix operator transforming a vector into a diagonal matrix whose diagonal entries are given by elements of the vector. To account for time dependence in ERP data, the independence assumption for ε in model (2.3) is relaxed by assuming $\operatorname{Var}(\varepsilon) = \Sigma = D_\sigma^{1/2} R D_\sigma^{1/2}$, where R is a $T \times T$ residual correlation matrix.

Model (2.3) explicitly introduces two kinds of covariates: x, whose effects on ERPs are of primary interest, and z, which can be viewed as auxiliary covariates. In the directed forgetting experiment, z contains the subject effect and the main effect of the instruction condition (TBR or TBF). The recognition performance is the only covariate of interest x. In fact, this special case $p = 1$ covers a wide range of situations in which x is a numeric covariate (such as a behavior score) or a categorical variable for representing two-group comparisons, which are the most frequently used experimental designs for ERP studies [Handy (2004)] (the situation in which $p > 1$ occurs, for example, when the covariate of interest is a k-group variable, with $k > 2$). For ease of discussion, we will refer to the $T \times p$ matrix β, whose rows are the p-vectors β_t, as the signal.

For ERP data, the signal is usually both rare and weak: rare because for most time points t, the null hypothesis $H_{0,t} : \beta_t = 0$ is true (i.e., signal is absent for most of the observation duration), and weak because, with respect to the moderate number of subjects in a typical ERP experiment and the amount of residual variability in ERP curves, the odds are not in favor of successful detection of time points for which $H_{0,t} = 0$ does not hold. According to the general linear model theory, the selection of significant time points is based on the $T \times p$ observed signal, $\hat{\beta}$, whose rows $\hat{\beta}_t$ are obtained by the ordinary least squares estimation of model (2.3):

$$\hat{\beta}_t = \left(x' P_z x\right)^{-1} x' P_z Y_t, \tag{2.4}$$

where $P_z = I_n - Z(Z'Z)^{-1}Z'$, Z is the $n \times (r+1)$ matrix whose ith row is $(1, z_i')$, $Y_t = (Y_{1t}, \ldots, Y_{nt})'$ and x is the $n \times p$ matrix whose ith row is x_i. The vector $\mathcal{T} = (\mathcal{T}_t)_{t=1,\ldots,T}$ of test statistics for the set of null hypotheses $H_{0,t}$ is given by the following expression of F-statistics:

$$\mathcal{T}_t = \frac{1}{p} \frac{\hat{\beta}_t' x' P_z x \hat{\beta}_t}{\hat{\sigma}_t^2}, \tag{2.5}$$

where $\hat{\sigma}_t^2$ is the standard degree-of-freedom corrected estimate of the residual variance in model (2.3).

Under the null hypothesis $H_{0,t}$, each component $\mathcal{T}_t$ of $\mathcal{T}$ is distributed according to an F-distribution with p and $d = n - p - r - 1$ degrees of freedom. For the directed forgetting experiment, $p = 1$. It explains the use of Student's t-tests there, as they are obtained as the signed square root of the test statistics $\mathcal{T}_t$. In the following, p_t stands for the p-value associated with $\mathcal{T}_t$.

It is important to note that, in the present multivariate linear model framework, the dependence structure of the test statistics is directly inherited from that in the residual correlation R of model (2.3): under the family-wise null hypothesis $H_0 = \bigcap_t H_{0,t}$, the components of $\mathcal{T}$ are indeed F-statistics with the following correlation structure:

$$\operatorname{Cor}(\mathcal{T}_t, \mathcal{T}_{t'}) = r_{tt'}^2 \left(\frac{1}{p} + \frac{1}{d} \right) \frac{p(d-4)}{p+d-2} \approx_{d \to +\infty} r_{tt'}^2,$$

where $r_{tt'}$ is the generic term of the matrix R.

Multiple testing. The collection of p-values $(p_t)_{t=1,\dots,T}$ is generally the only input for multiple testing procedures. Among them, the method proposed by Guthrie and Buchwald (1991) is the first to address the issues of the time dependence in ERPs by assuming a first-order autoregressive correlation structure for t-tests. The method is designed to prevent erroneous detections of short significant intervals rather than to control for any Type I error rate.

In contrast, most multiple testing methods consist in rejecting the null $H_{0,t}$ if $p_t \leq p^*$, where the threshold p^* is chosen to guarantee that the corresponding number V of erroneous rejections of the null is controlled. The most common methods, which are designed for a moderate number of simultaneous tests, such as for post-hoc comparisons in analysis of variance, aim at controlling the family-wise error rate defined as $\text{FWER} = \mathbb{P}(V \geq 1)$ to guarantee that $\text{FWER} \leq \alpha$ for a preset level α. However, FWER-controlling procedures are usually far too conservative when the number of tests, T, becomes large. In the last two decades, the questions raised by large-scale significance analysis have generated a plethora of simultaneous testing procedures and thresholding methods for high-dimensional data [see Efron (2010), van der Laan and Dudoit (2007) for a review of the popular procedures and Groppe, Urbach and Kutas (2011a, 2011b), Lage-Castellanos et al. (2010), specifically, for ERP data analysis]. A new family of methods aims to control, instead of FWER, the false discovery rate (FDR), defined as the expected proportion of erroneous rejections of the null among the positive tests: $\text{FDR} = \mathbb{E}(\text{FDP})$, where the false discovery proportion FDP is 0 if the number R of rejections is itself 0 and $\text{FDP} = V/R$ if $R > 0$ [Benjamini and Hochberg (1995)]. More relevant for the current work are methods that control the FDR by the Benjamini–Hochberg (BH) procedure for correlated tests. The best known among these is the Benjamini and Yekutieli (2001) (BY) procedure, which modifies the BH procedure to control the FDR under some specific assumptions of

positive dependence among tests. We note that testing ERPs from the directed forgetting experiment by controlling the FDR at the 0.05 level using the original BH procedure, no significant time points are found at channel CZ, where t-statistics are displayed in Figure 2.

The negative impact of dependence on the accuracy of multiple testing procedures, especially due to the instability of ranking, has generated a great deal of research interest. A direct approach to handle the dependence among test statistics is through modeling dependence structures in the data. In genomic data analysis, many researchers [Friguet, Kloareg and Causeur (2009), Leek and Storey (2008), Sun, Zhang and Owen (2012)] proposed modeling the dependence among tests using a latent factor model to decorrelate the test statistics so as to restore the consistent ranking in p-values.

3. Time dependence among test statistics. First, we propose a flexible factor modeling of the residual correlations in model (2.3) to account for the complex dependence pattern of the ERPs over time. Then, we proceed to model jointly signals and dependence so as to obtain sharper test statistics after eliminating, as much as possible, the impact of dependence.

We assume there exist q latent factors, $f = (f_1, \ldots, f_q)'$, normally distributed with mean 0 and variance, I_q, such that, conditional on z_i, x_i and f_i, the ERP measurement Y_{it} for subject i at time t:

$$Y_{it} = \mu_t + b_t' z_i + \beta_t' x_i + \lambda_t' f_i + e_{it}, \tag{3.1}$$

where λ_t is the q-vector of factor loadings for Y_t and e_{it} is the specific random error term, normally distributed with mean 0 and variance ψ_t^2. Moreover, it is assumed that the specific errors e_{it} are mutually independent, which induces the following decomposition of the residual covariance matrix Σ:

$$\Sigma = \Lambda \Lambda' + \Psi, \tag{3.2}$$

where Λ is the $T \times q$ matrix whose tth row is λ_t' and Ψ is the diagonal matrix whose tth diagonal element is ψ_t^2. In other words, latent factors are introduced to capture linearly the time dependence among residuals of model (2.3) [Causeur et al. (2012)].

To illustrate the ability of model (3.1) to fit the complex dependence pattern observed in Figure 3, models with 1, 5 and 10 factors, respectively, are estimated for the residual correlations of model (2.2) at channel CZ using the EM algorithm described in Friguet, Kloareg and Causeur (2009). The results are compared in Figure 4, showing that the general dependence structure can be well approximated with a moderate number of factors (with respect to the number of time points in the data). A variety of methods can be used to choose the number of factors for latent variable models. We discuss this issue further in Section 5.

In the supplementary material [Sheu et al. (2016)], a simulation study based on model (3.1) is conducted to demonstrate the impact of time dependence on the

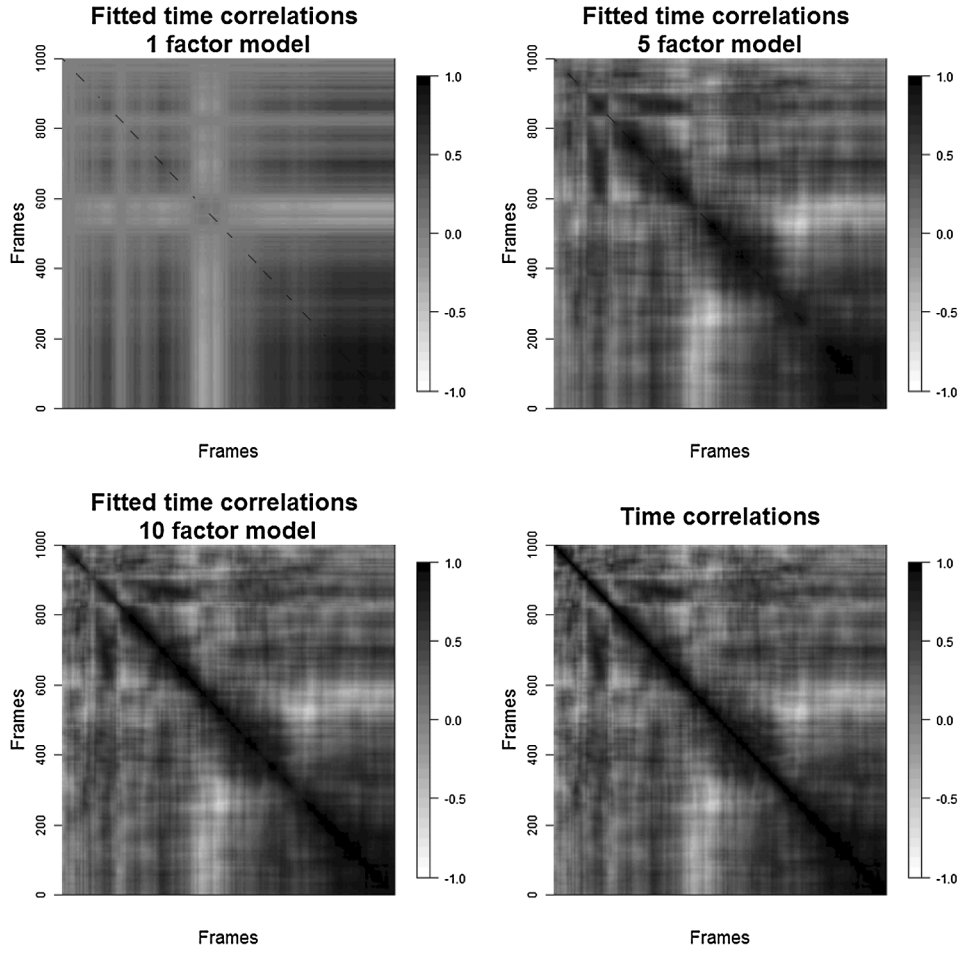

FIG. 4. *Image plots of the fitted correlation matrix of the residuals of model* (2.2) *at channel CZ using factor models with* 1, 5 *and* 10 *factors*, *top and bottom left panels*, *respectively. The bottom right panel reproduces the right panel of Figure* 3.

ability of multiple testing procedures to identify a predetermined true signal. The instability of significant findings discovered by procedures ignoring dependence is highlighted. Indeed, for highly dependent test statistics, the chance of declaring at least one time point as significant is much lower than that for independent statistics; moreover, when positives are found, the expected false discovery proportion is much larger.

4. Joint modeling of signal and dependence. The method proposed here employs an iterative scheme to update estimates of signals and estimates of model parameters for dependence structure in turn. At each step, based on the known signal-free time points $\mathcal{T}_0$, the process of estimation errors outside of $\mathcal{T}_0$ is up-

graded by making use of its correlation with the counterpart in $\mathcal{T}_0$. This method improves over the previous factor modeling approach in detecting ERP signals [Causeur et al. (2012)].

First, let $\Delta = \hat{\beta} - \beta$ denote the $T \times p$ matrix of estimation errors whose tth row is $\delta_t = (x' P_z x)^{-1} x' P_z \varepsilon_t$, and $\varepsilon_t = (\varepsilon_{1t}, \ldots, \varepsilon_{nT})'$ is the n-vector of residual errors in model (2.3). Let vec$(\cdot)$ be the matrix operator transforming a matrix into a vector by concatenating its rows. The pT-vector, vec(Δ), is distributed according to a normal distribution with mean 0 and covariance $V_\delta = \Sigma \otimes (x' P_z x)^{-1}$, where $\otimes$ is the Kronecker matrix product.

Correction of the signal estimation based on a prior knowledge. From cumulative empirical experience with ERP studies, researchers are likely to have gained some notion for when a signal should begin and how long it should last for an experimental condition. Lacking such a prior knowledge, one can use the preliminary results of a multiple testing procedure to screen for time points at which the signal is unlikely to be present, that is, $\beta_t = 0$ for t belonging to the collection of measurement occasions $\mathcal{T}_0$. Thus, the estimation error δ_t, for $t \in \mathcal{T}_0$, is not confounded with the true signal β_t: $\Delta_0 = \hat{\beta}_0$, where Δ_0 (resp. $\hat{\beta}_0$) is the submatrix of Δ (resp. $\hat{\beta}$) restricted to $t \in \mathcal{T}_0$. This allows us to partition Δ into two submatrices:

$$\tilde{\Delta} = \begin{pmatrix} \Delta_0 \\ \Delta_{-0} \end{pmatrix}, \tag{4.1}$$

where Δ_{-0} is the submatrix of Δ with rows δ_t, $t \neq \mathcal{T}_0$, and rearrange V_δ correspondingly:

$$\tilde{V}_\delta = \begin{pmatrix} \Sigma_{0,0} & \Sigma'_{-0,0} \\ \Sigma_{-0,0} & \Sigma_{-0,-0} \end{pmatrix} \otimes \left(x' P_z x\right)^{-1},$$

where $\Sigma_{0,0}$ (resp. $\Sigma_{-0,-0}$) is the submatrix of Σ restricted to rows and columns corresponding to time points t in $\mathcal{T}_0$ (resp. $t \notin \mathcal{T}_0$) and $\Sigma_{-0,0}$ is the submatrix of Σ restricted to rows corresponding to $t \notin \mathcal{T}_0$ and columns corresponding to $t \in \mathcal{T}_0$.

For each $t \notin \mathcal{T}_0$, we predict δ_t from Δ_0 by its best linear predictor:

$$\begin{aligned} \operatorname{vec}(\hat{\Delta}_{-0}) &= \left[\hat{\Sigma}_{-0,0} \otimes \left(x' P_z x\right)^{-1}\right]\left[\hat{\Sigma}_{0,0} \otimes \left(x' P_z x\right)^{-1}\right]^{-1} \operatorname{vec}(\Delta_0) \\ &= \left[\hat{\Sigma}_{-0,0} \otimes \left(x' P_z x\right)^{-1}\right]\left[\hat{\Sigma}_{0,0}^{-1} \otimes \left(x' P_z x\right)\right] \operatorname{vec}(\Delta_0) \\ &= \left[\hat{\Sigma}_{-0,0} \hat{\Sigma}_{0,0}^{-1}\right] \otimes I_p \operatorname{vec}(\Delta_0), \end{aligned}$$

where I_p is the $p \times p$ identity matrix and $\hat{\Sigma}_{-0,0}$ and $\hat{\Sigma}_{0,0}$ are estimators of $\Sigma_{-0,0}$ and $\Sigma_{0,0}$, respectively.

Equivalently, in matrix form,

$$\hat{\Delta}_{-0} = \hat{\Sigma}_{-0,0} \hat{\Sigma}_{0,0}^{-1} \Delta_0. \tag{4.2}$$

Because a matrix inversion is involved, the choice of an estimator of $\Sigma_{0,0}$ is critical for numerical computation. Here, we can use the factor model (3.1) to estimate the

whole matrix Σ by $\hat{\Sigma} = \hat{\Lambda}\hat{\Lambda}' + \hat{\Psi}$, where $\hat{\Psi}$ is the diagonal matrix of estimated specific variances $\hat{\Psi}_t^2$ and $\hat{\Lambda}$ is the $T \times q$ matrix of estimated loadings ($q \ll T$). The partition (4.1) of Δ results in corresponding partitions of Ψ and Λ:

$$\tilde{\Lambda} = \begin{pmatrix} \Lambda_0 \\ \Lambda_{-0,} \end{pmatrix}, \qquad \tilde{\Psi} = \begin{pmatrix} \Psi_0 & 0 \\ 0 & \Psi_{-0} \end{pmatrix}.$$

Estimators of $\Sigma_{0,0}$ and $\Sigma_{-0,-0}$ are derived as

$$\hat{\Sigma}_{-0,0} = \hat{\Lambda}_{-0}\hat{\Lambda}_0', \qquad \hat{\Sigma}_{0,0} = \hat{\Psi}_0 + \hat{\Lambda}_0\hat{\Lambda}_0'.$$

Note that computing $\hat{\Sigma}_{0,0}^{-1}$ in expression (4.2) involves only the inversion of a $q \times q$ matrix according to Woodbury's identity [Press et al. (2007)]:

$$\hat{\Sigma}_{0,0}^{-1} = \hat{\Psi}_0^{-1} - \hat{\Psi}_0^{-1}\hat{\Lambda}_0(I_q + \hat{\Lambda}_0'\hat{\Psi}_0^{-1}\hat{\Lambda}_0)^{-1}\hat{\Lambda}_0'\hat{\Psi}_0^{-1}.$$

An estimate $\hat{\Delta}^{(1)}$ for Δ can be obtained by substituting $\hat{\Delta}_{-0}$, given by expression (4.2), for Δ_{-0} in (4.1). A new estimate for β is then obtained by correcting the current estimate $\hat{\beta}$ for the predicted estimation error:

$$\hat{\beta}^{(1)} = \hat{\beta} - \hat{\Delta}^{(1)}.$$

The new estimate is used to update the calculation of the residual errors $\hat{\varepsilon}$:

$$\hat{\varepsilon}^{(1)} = P_z\big(Y - x\hat{\beta}^{(1)'}\big).$$

A new factor decomposition of the covariance of the updated residual errors is again derived, producing a new estimate for Δ and, in turn, a new estimate $\hat{\beta}^{(k)}$ of the signal, where the superscript k indicates the step in the iteration. The calculation continues until a predetermined convergence criterion is reached for the estimation of β.

Decorrelation of test statistics by adaptive factor adjustment (*AFA*). The literature on the estimation of factor models, particularly for psychometric applications, is extensive [see Mardia, Kent and Bibby (1979) for a review]. The maximum likelihood estimation introduced by Jöreskog (1967) is especially suitable for the linear model framework of the present approach. Unfortunately, the direct maximization of the multivariate normal likelihood is intractable. A fast and efficient Expectation–Maximization (EM) algorithm [Rubin and Thayer (1982)], presented in detail in Friguet, Kloareg and Causeur (2009), is adapted for the present setting. Once estimates of the factor model parameters are obtained, estimates of the factors F are given by Thompson's scores [Thomson (1951)].

A critical issue for factor modeling of ERPs is choosing the optimal number of factors to retain. Extracting too many factors could render the estimates of the residual specific variances $\hat{\Psi}_t^2$ artificially smaller than expected, inflating false positives as a result. Observing that the variance of the number of false positives is an increasing function of the amount of dependence among the test statistics, Friguet,

Kloareg and Causeur (2009) derive a closed-form expression for the variance inflation $\mathcal{V}_k$ of the k-factor model for this dependence. These authors assess the number of factors by estimating the variance $\mathcal{V}_k$ of the number of false positives when the tests are calculated with the k-factor-adjusted residuals: $\hat{\varepsilon} - \hat{F}_k \hat{\Lambda}_k$ for each k-factor model (Λ_k, Ψ_k). Finally, the retained number of factors is $\hat{k} = \arg\min_k \mathcal{V}_k$. In contrast, the number of factors is determined via parallel analysis [Buja and Eyuboglu (1992)] in surrogate variable analysis [Leek and Storey (2008)] and latent effect adjustment after primary projection [Sun, Zhang and Owen (2012)].

Once a factor regression model (3.1) [Causeur et al. (2012), Friguet, Kloareg and Causeur (2009), Leek and Storey (2008), Sun, Zhang and Owen (2012)] is fitted to a set of dependent data for multiple testing, the new test statistics $\tilde{\mathcal{T}}$ (presumably independent) for testing the collection of nulls $H_{0,t}, t = 1, \ldots, T$, will be referred to as factor-adjusted test statistics.

In summary, the adaptive factor-adjusted multiple testing procedure we propose alternates between the estimation of the signal corrected for the predicted estimation error (by factor modeling the dependence structure), and the calculation of factor-adjusted test statistics, which are then used to update the current knowledge of $\mathcal{T}_0$. Starting from a given $\mathcal{T}_0^k$ at the kth step of the procedure with the current estimate $(\hat{\Psi}_k, \hat{\Lambda}_k; \hat{F}_k)$ of the factor parameters, the $(k+1)$th step consists of two parts:

- Calculate the predicted estimation error $\hat{\Delta}^{(k+1)}$ and update the estimate of the signal by $\hat{\beta}^{(k+1)} = \hat{\beta} - \hat{\Delta}_{(k+1)}$. Consequently, the residual error is also updated: $\hat{\varepsilon}^{(k+1)} = P_{\tilde{z}}(Y - x\hat{\beta}^{(k+1)'})$;
- Estimate $(\hat{\Psi}_{k+1}, \hat{\Lambda}_{k+1}; \hat{F}_{k+1})$ of the factor model based on $\hat{\varepsilon}^{(k+1)}$. Factor-adjusted tests statistics are derived and $\mathcal{T}_0$ is, in turn, updated. The update of $\mathcal{T}_0^k$ should favor the selection of time points for which no-signal is expected with a high confidence, yielding potentially a large number of false positives, rather than a more stringent rule, which would cover more accurately the true $\mathcal{T}_0$, but also with a higher chance of including the support of the signal. We suggest adopting the following rule: $\mathcal{T}_0^{k+1} = \{t = 1, \ldots, T, \tilde{p}_t^{(k+1)} \geq 0.2\}$, where $\tilde{p}_t^{(k+1)}$ is the current factor-adjusted p-value at time t.

The iteration terminates at step k such that $\mathcal{T}_0^{k+1} = \mathcal{T}_0^k$.

Our experience with the method suggests that different choices of the threshold (here 0.2 for the rule above) on the p-values to update $\mathcal{T}_0$ do not alter the final result, provided that the choice is not too extreme: a very small value tends to erase the signal and over-control the FDR, whereas a value near 1 tends to produce the same results as the estimation method chosen to initialize the method, that is, ordinary least-squares.

In a multiple testing setting for ERP data analysis, estimating jointly the signal and the residual covariance model to decorrelate the test statistics can be associated with any thresholding procedure depending on whether the overall Type I

error, FDR or FWER, is to be controlled. Because the BH procedure [Benjamini and Hochberg (1995)] is widely considered as the gold standard under independence, we choose it to correct the p-values produced by the AFA method. This combination of adaptive factor adjustment estimation with the BH procedure is hereafter referred to as the AFA multiple testing procedure.

An illustration. A single run of the simulation study presented in supplementary material [Sheu et al. (2016)], with true nonzero signal on the time interval $[450, 550]$ and peak amplitude $\max_t \beta_t = 5$ at $t = 500$, is selected to demonstrate how the regular pattern of the estimated signal induced by the strong time dependence in V_δ can generate confusion between the true signal β and the estimation error Δ. The solid curve in the upper panel of Figure 5 represents the values of t-statistics based on the Ordinary Least Squares estimation (OLS) of the signal. The

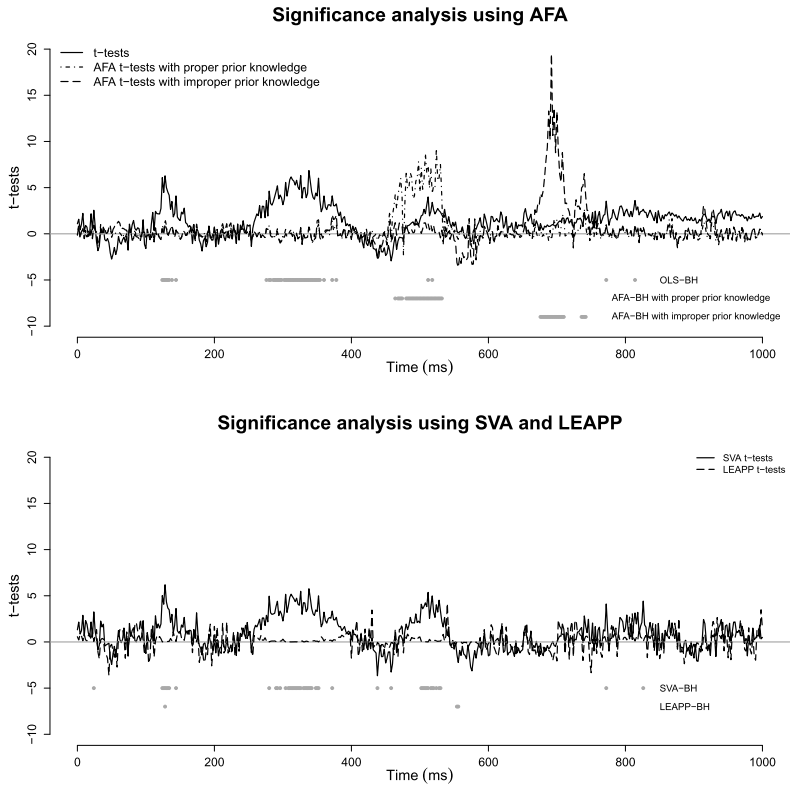

FIG. 5. *t-statistics for a single simulation run. Top panel: the OLS estimation of the signal (solid curve) and t-statistics after Adaptive Factor Adjustment, based on $\mathcal{T}_0 = [1, 100] \cup [901, 1000]$ (dotted) and $\mathcal{T}_0 = [450, 550]$ (dashed). Bottom panel: t-statistics after SVA (solid) and LEAPP (dashed) adjustment. The gray dots above the x-axis indicate significant time points identified by the BH method controlling the FDR at 0.05 level.*

BH procedure based on the raw p-values not only fails to locate the true signal but also incorrectly detects significant intervals outside of the support of β. This is an example of the inconsistent ranking of unadjusted p-values resulting from applying multiple testing procedures while ignoring dependence.

The bottom panel of Figure 5 shows that neither the surrogate variable analysis (SVA) [Leek and Storey (2008)] nor the latent effect after primary projection (LEAPP) [Sun, Zhang and Owen (2012)] succeeds in properly disentangling the signal from the time dependent noise, resulting in erroneous identifications of significant intervals.

To illustrate the impact of the prior knowledge of $\mathcal{T}_0$ on the proposed estimation procedure, assume first that no signal is expected in $\mathcal{T}_0 = [1, 100] \cup [901, 1000]$, which is a proper prior knowledge. The dot-dashed curve of the top panel in Figure 5 represents the values of t-statistics obtained by the AFA method based on the former prior knowledge. The significant interval detected after BH correction for the final factor-adjusted statistics with a control of the FDR at level 0.05 indicates that the support of the signal is here consistently estimated. The dashed curve in the same plot displays the values of t-statistics obtained by the AFA method based on the incorrect prior knowledge of signal-free interval $\mathcal{T}_0 = [450, 550]$, that is, exactly where the true signal lies. With such a misguided prior for input, AFA clearly fails by locating significant intervals where the true signal is absent.

To investigate the sensitivity of the significance analysis to the choice of $\mathcal{T}_0$, we have implemented the AFA method on the 1000 datasets in the simulation study of supplementary material [Sheu et al. (2016)], with true nonzero signal on the time interval $[450, 550]$ and peak amplitude $\max_t \beta_t = 5$ at $t = 500$, with a fixed length of 200 ms for $\mathcal{T}_0 = [t_0 - 100; t_0 + 100]$ and a center t_0 moving from 100 to 900 ms. For each choice of $\mathcal{T}_0$, the AFA significance analysis is assessed by the Positive Predictive Value (PPV), also called precision, defined as the expected proportion of correct rejections of the null among the positives. Figure 6 displays the PPV curve along with t_0. It confirms that, as soon as the prior knowledge of $\mathcal{T}_0$ does not intersect too much of the interval of nonzero signal, the precision of the method is very good. In the present situation, when more than 50% of $\mathcal{T}_0$ is in the support of the true signal, then the method fails to detect the peak.

5. A comparative study. The performance of the AFA procedure is compared against that of existing multiple testing procedures chosen either because they are widely used or because they are specifically designed to account for dependence. The BY procedure [Benjamini and Yekutieli (2001)] guarantees the control of the FDR under specific dependence assumptions, although it does not correct for the impact of correlation by adjusting the raw p-values. The SVA [Leek and Storey (2008)] and LEAPP [Sun, Zhang and Owen (2012)] procedures are representatives of recently developed approaches based on a factor regression model similar to model (3.1) to decorrelate test statistics. The method by Causeur et al. (2012) has been superseded by the present AFA and is, therefore, not included for comparison.

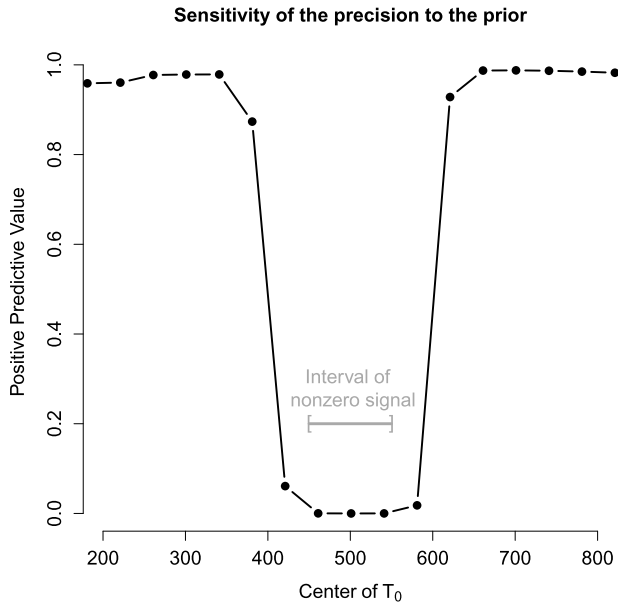

FIG. 6. *Positive Predictive Value (PPV) of the AFA procedure for different choices of the prior knowledge of* $\mathcal{T}_0 = [t_0 - 100; t_0 + 100]$ *along* t_0. *The PPV are calculated using* 1000 *simulated ERP datasets with true nonzero signal on the time interval* [450, 550] *and true peak amplitude* $\max_t \beta_t = 5$ [*see supplementary material Sheu et al.* (2016)].

To summarize, we compare the performance of the following six different methods for multiple testing of ERP data (the level of FDR control is set at 0.05 for all comparisons):

1. BH: the Benjamini–Hochberg procedure [Benjamini and Hochberg (1995)] applied to the raw p-values.
2. BY: the Benjamini–Yekutieli procedure [Benjamini and Yekutieli (2001)] applied to the raw p-values.
3. GB: the Guthrie–Buchwald procedure [Guthrie and Buchwald (1991)] applied to the raw p-values with a graphical threshold equal to 0.05.
4. LEAPP: the latent effect adjustment after primary projection [Sun, Zhang and Owen (2012)] with control of the FDR using BH. The default options of the `R` package `leapp` [Sun, Zhang and Owen (2014)] are used for the model and the number of factors.
5. SVA: the surrogate variable analysis procedure [Leek and Storey (2008)] with control of the FDR using BH. The model and the number of factors are set to the default options of the `R` package `sva` [Leek et al. (2014)].
6. AFA: the proposed adaptive factor adjustment method with a control of the FDR. The prior input for $\mathcal{T}_0$ is the set of time points for which the p-value of the usual t-test is greater than or equal to 0.2. The number of factors is determined,

individually for each simulated dataset, by minimizing the variance inflation criterion [Friguet, Kloareg and Causeur (2009)] as implemented in the R package ERP [Causeur and Sheu (2014)].

In the following simulation study, $n \times T$ ERP data are generated according to model (3.1), with $n = 20$ and $T = 1000$ matching the number of participants and time frames of the directed forgetting experiment. For each simulation run, the only covariate x is the centered recognition performance as observed in the TBR condition of the experiment. We set $\mu_t = 0$, $b_t = 0$, for all $t = 1, \ldots, T$, and use the sample estimates from the observed ERP curves at electrode CZ for the residual standard deviations $\sigma_t, t = 1, \ldots, T$. The residual correlation R is derived from the 5-factor model displayed in the top right panel of Figure 4. The true signals $t \mapsto \beta_t$ have the same bell shape on the same support from 450 to 550 ms with varying peak heights starting at zero, and then from 1.5 to 12.5 in equal step sizes of 0.1. Figure 1 in supplementary material [Sheu et al. (2016)] of this paper shows the corresponding powers of the individual t-tests of $H_{0,t} : \beta_t = 0$ for a Type-I error rate $\alpha = 0.05$. For each signal amplitude, 1000 ERP datasets are generated.

For each simulation run, the procedures are assessed by the FDR, the Positive Predictive Value (PPV) defined as the expected proportion of correct rejections of the null among the positives, and the probability of no rejection $\text{PNR} = \mathbb{P}(R = 0)$, defined as the expected proportion of datasets for which no null is rejected, where R is the number of rejections of the null. Figure 7 summarizes the results. The top left panel of Figure 7 shows that AFA method inherits from the BH procedure good properties in terms of FDR control, which is far from true for either GB, LEAPP or SVA, especially when the true signal is weak or moderate. In addition, this control of the FDR is not affected by the instability caused by dependence as reported in Section 2. The bottom panel of Figure 7 shows that the probability of no rejection of AFA is among the lowest, provided that the signal is moderate to high, which guarantees the positive FDR, namely, the expected False Discovery Proportion given at least one positive time point has been found, is close to the FDR. In the top right panel of the same figure, the PPV curve confirms the superiority of AFA for detecting moderate to high peaks.

Moreover, a multiple testing procedure may be considered as a tool to detect time points at which a signal is above the threshold from the biomedical signal processing perspective. In this context, sensitivity refers to the lowest signal amplitude at which detection becomes possible for the method under consideration. We define the sensitivity of a multiple testing procedure as the minimum peak height for which $1 - \text{PNR}$ exceeds 0.1:

$$\text{Sensitivity} = \min\Big\{\max_t \beta_t, 1 - \text{PNR} \geq 0.10\Big\}. \tag{5.1}$$

Similarly, the resolution of a multiple testing procedure is the extent to which it can detect the time points for nonzero signal. We define the resolution of a method

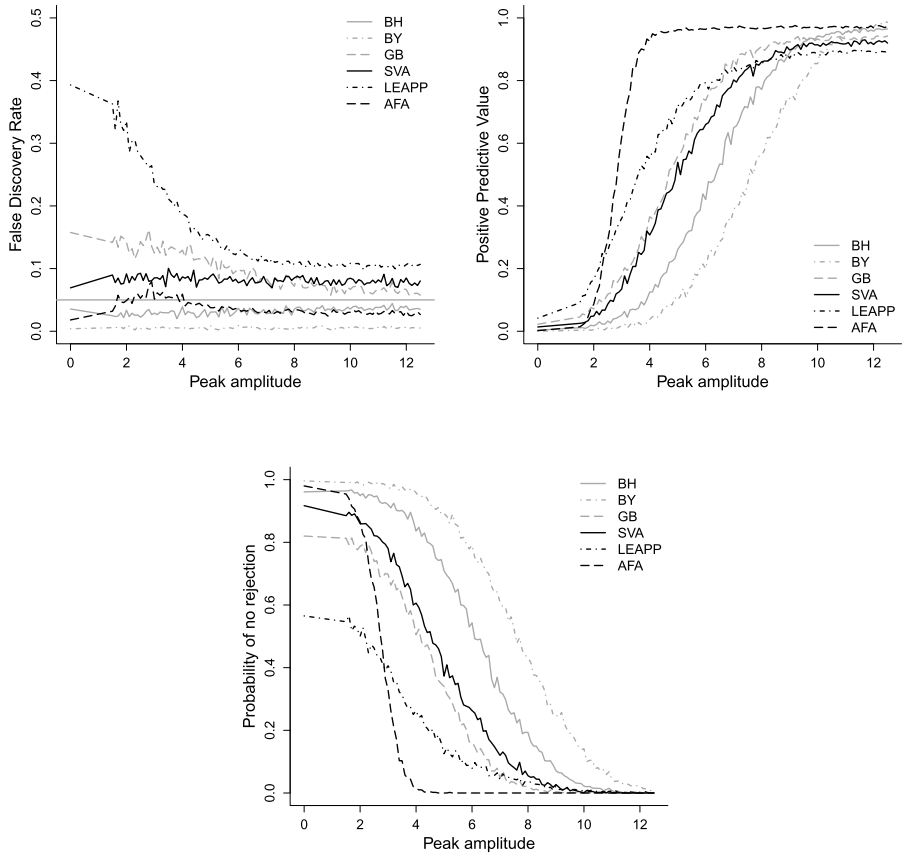

FIG. 7. *False Discovery Rate* (*top left*), *Positive Predictive Value* (*top right*) *and Probability of No Rejection* (*bottom*) *for comparing* 6 *multiple testing procedures using* 1000 *simulated ERP data set at each peak amplitude. FDR is expected to be near and below* 0.05, *PPV large and PNR small.*

as the minimum peak height for which PPV exceeds 0.9:

$$\text{Resolution} = \min\Big\{\max_t \beta_t, \text{PPV} \geq 0.90\Big\}. \tag{5.2}$$

The sensitivity and resolution measures of the 6 methods are calculated from the simulated data sets and presented in Table 1. Overall, the AFA procedure outperforms the other methods according to the assessment measures considered in this comparison. Table 1 shows that all three decorrelation methods, SVA, LEAPP and AFA, are quite sensitive, especially LEAPP, which is consistent with the display in the top left panel of Figure 7 that, for low to moderate signal amplitudes, it fails control the FDR at the required level. In terms of the resolution, the AFA method has the lowest level and GB, the second lowest level, confirming that an explicit modeling of the residual time dependence is useful to disentangle the signal from the noise. The same conclusion is also evident from Figure 8, in which

TABLE 1
Sensitivity and resolution of 6 *multiple testing procedures from the simulation study*

	Multiple testing methods					
	BH	**BY**	**GB**	**SVA**	**LEAPP**	**AFA**
Sensitivity	3.6	4.9	0.0	1.5	0.0	1.9
Resolution	9.1	10.4	7.9	8.9	>12.5	3.7

the Root Mean Squared Error (RMSE) for β estimation by Ordinary Least Squares (OLS) fitting, SVA, LEAPP and AFA, are computed from the simulation datasets and presented as box plots. As expected, OLS, which ignores dependence, performs the worst on average. All three decorrelation methods have markedly lower RMSE than that of OLS, with the AFA method achieving the smallest error, on average, at any level of the signal amplitude compared. The increasing difficulty of the LEAPP method to extract signal from noise as the signal amplitude increases is again confirmed by the box plots elevating from left to right in the bottom left panel in Figure 8.

6. Analysis of ERP data.

Auditory oddball experiment. The auditory oddball paradigm introduced in Section 2 is commonly used to calibrate newly acquired ERP instruments against that of standard ones which have been in use in the laboratory. From the auditory ERP curves collected in a passive listening task, experimenters can expect

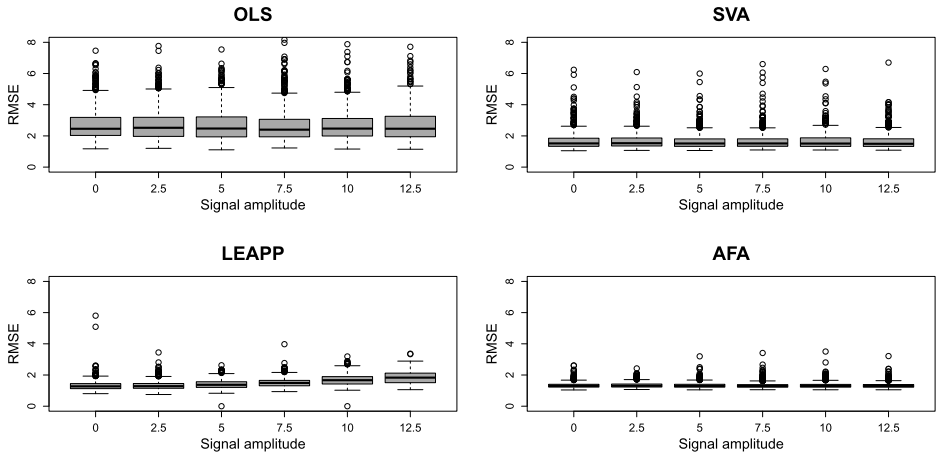

FIG. 8. *Root Mean Squared Error for the estimation of* β *is computed from* 1000 *datasets at each signal amplitude*, *respectively*, *using each of four methods*: *OLS*, *SVA*, *LEAPP and AFA*.

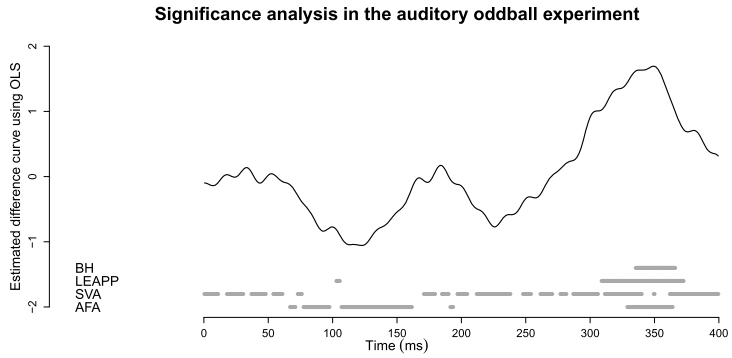

FIG. 9. *Significance analysis of the difference curve* (1000 *Hz*–500 *Hz*) *in the auditory oddball experiment. Significant time points identified by four multiple testing procedures are indicated by gray dots above the x-axis*: *BH*, *LEAPP*, *SVA*, *AFA using the same notation as in Section* 5.

to find two signature components: (1) the auditory evoked potential (AEP) peaking between 80 and 120 ms after stimulus onset and maximal over fronto-central scalp locations in either tone conditions and (2) a mismatch negativity (MNN) for the difference curves (ERPs for odd tone trials minus that for frequent tone trials) peaking between 100 and 200 ms from onset. The AEP [Rosberg, Butrous and Ford (2008)] is primarily an exogenous component which can be elicited by any discernible auditory stimulus without any task demand. The MNN [Näätänen (2003)] is elicited by any discriminable change ("odd") in some repetitive aspect of auditory stimulation ("frequent").

In biological psychiatry, a reduced AEP and a decrease in the amplitude of MMN have both been reported in patients with schizophrenia. Therefore, whether statistical tests can achieve reliable verification of the above two signature ERP components in the test case data need to be carefully assessed and compared. This verification is of fundamental importance if these components are to be considered as electrophysiological markers for further assessment of psychiatric and neurological disorders [Williams et al. (2005)]. Figure 9 shows the results of significance analyses of the difference curve, using the data introduced in Section 2, by four methods presented in Section 5: BH, LEAPP, SVA and AFA. Note that the determination of the number of factors by minimization of the variance inflation criterion in AFA gives 2 factors, whereas the parallel analysis used by LEAPP and SVA gives 3 factors. The comparison of the four methods shows that MMN is only identified as significant by AFA. The three other methods point out a late peak around 300 ms and SVA is noticeably too liberal here.

Directed forgetting experiment. Our ability to recognize words that we have been told to forget evidently relies more on familiarity than does recognition of words we were told to remember [see, e.g., Gardiner, Gawlik and Richardson-Klavehn (1994)]. Empirical studies of recognition memory using ERPs have in-

dicated that the early phase of recognition involving familiarity is associated with modulations of the ERP component FN400, an enhanced positivity for old items relative to new items observed from approximately 400 to 600 ms after stimulus onset. The finding that the FN400 component increases gradually with recognition confidence [Rugg and Curran (2007)] also suggests that this component is an index of familiarity. Although the directed forgetting experiment introduced in Section 2 is exploratory in nature, previous research indicates that one would expect significant time intervals around 400 to 600 ms. Qualitatively, one would also expect late significant time intervals for the TBF condition for electrodes in the posterior locations. Confirmation of these predictions is an important step forward in understanding the neurophysiological mechanism regarding intentional control of remember and forgetting.

However, a naive application of the BH method to the ERP and behavioral data from the directed forgetting experiment failed to identify any significant time points at any of the 9 electrode locations. To apply the AFA method, we selected, for frontal and central electrodes, a prior knowledge of $\mathcal{T}_0 = [1, 200] \cup [901, 1000]$ ms, and, for the posterior locations, $\mathcal{T}_0 = [1, 200]$ ms [e.g., Paz-Caballero and Menor (1999)]. After examining the variance inflation criterion for each channel, the number of factors was set to 2. The top and bottom panels of Figure 10 display the correlation curves at channel CZ of the two instruction conditions based on the OLS and the AFA estimations of the signal, respectively. Note that the AFA method reveals a positive significant waveform, with a large peak in both conditions in the interval [400, 700] ms, which is preceded by a negative significant peak only in the TBF condition.

Figure 11 displays a spatial representation on the scalp of the correlation curves based on the AFA estimation of the signal. Significant positive peaks mainly occurred from 400 to 700 ms for both conditions at each of 9 locations. In addition, the analysis by the AFA method confirms significant negative peaks appearing in most locations but only in the `TBF` condition. This inflexion of the correlation curves around 400 ms, clearer in the TBF condition, implicates the relationship between instruction and the modulation of the FN400 component.

7. Conclusions. Mass univariate analysis of event-related brain potentials [Groppe, Urbach and Kutas (2011a)] has long been recognized as a challenging problem because ERP signals are often rare, occurring only in brief moments during trials, and weak, relative to the large between-subject variability [see Donoho and Jin (2008), Jin (2009) for the rare and weak terminology]. When testing simultaneously for significance over a large number of measurements over time [Woolrich et al. (2009)], the need to control for the probability of false positive errors is pitted against that for maintaining reasonable power for correct detection. Controversies have erupted when researchers appeared to favor one need over another in their statistical methodology [Vul et al. (2009)].

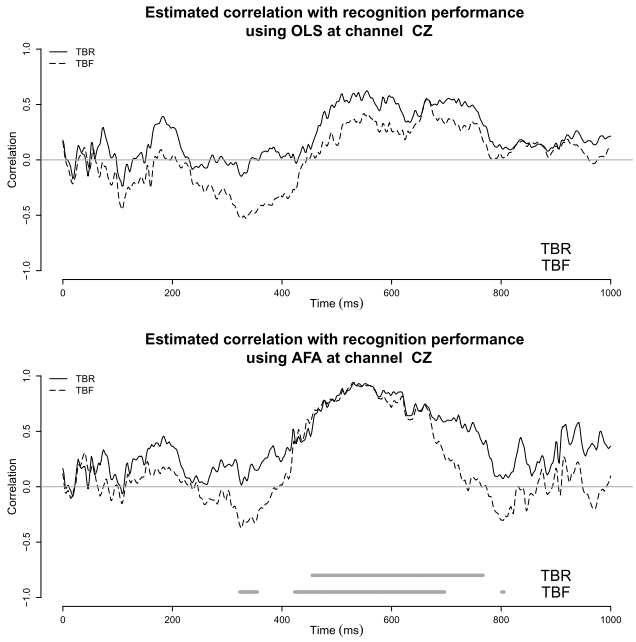

FIG. 10. *Correlations between the ERPs and the recognition performance for the two conditions* (*solid curve for TBR*, *dashed for TBF*) *based on the OLS* (*top panel*) *and the AFA* (*bottom panel*) *estimations of the signal. Significant time points are indicated by gray dots above the x-axis.*

Compounding the challenge, ERPs are highly dependent over time, not only causing the performance of multiple testing procedures under the independence assumption to be unstable but also masking the location as well as the size of the true signal even after tests are corrected for dependence. The adaptive factor adjusted method meets the challenge posed by mass univariate ERP analysis within a multivariate linear model framework by a factor modeling of the time dependence and a joint modeling of signal and noise processes, given a prior input on the intervals in which the signal is absent. An iterative scheme is devised to estimate model parameters and the methodology is implemented in an R package available from Comprehensive R Archive Network (CRAN) at cran.r-project.org/web/packages/ERP.

Although permutation tests [Blair and Karniski (1993), Westfall and Young (1993)] have also been widely used in ERP data analysis, we have not reviewed them here because a recent study [Lage-Castellanos et al. (2010)] reported that the BH method [Benjamini and Hochberg (1995)] and the local FDR method [Efron (2007)] provided the best balance (compared against the permutation test) between Type I and Type II error in situations when there is no a priori information about when and where ERP differences occur. In light of their conclusion, the results of our comparative study reported in Section 6 are particularly encouraging: the proposed adaptive factor-adjusted method surpassed all other five methods in keeping

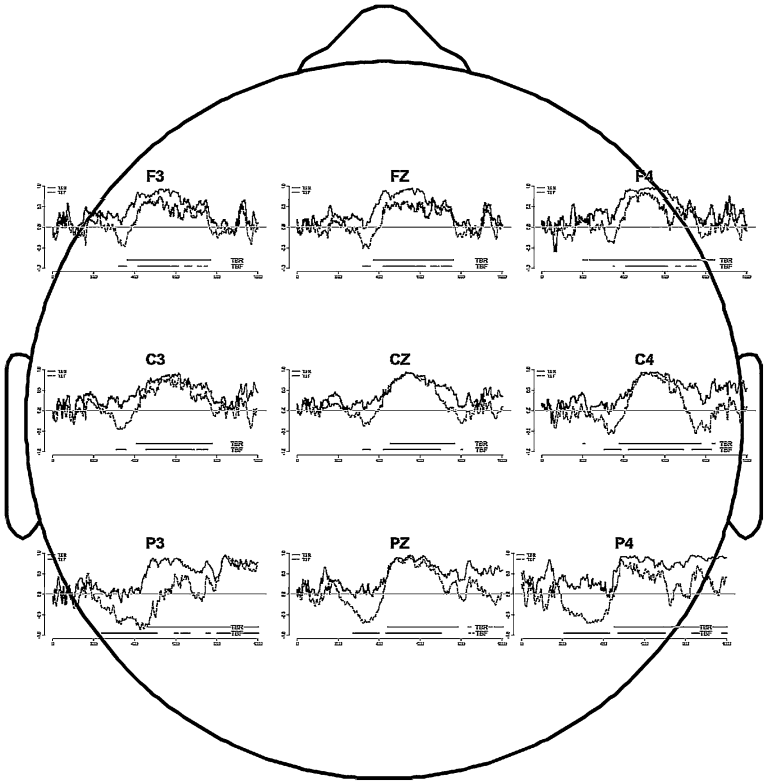

FIG. 11. *Correlation curves* (*solid for TBR*, *dashed for TBF*) *based on the AFA estimation of the signal with corresponding significant intervals for* 9 *channels*.

the FDR under control and maintaining power of correct detection. Furthermore, the exploratory data analysis of the directed forgetting experiment demonstrated that the AFA method is ideally suited for detecting weak ERP signals embedded in a complex and highly dependent noise process.

It is expected that the same estimation procedure can be applied to many multiple testing situations with strong dependence: either along wavelength in Near InfraRed Spectroscopy (NIRS) or spatially distributed in function Magnetic Resonance Imaging (fMRI).

Acknowledgments. We thank Dr. Ching-Hung Lin and Dr. Yi-Hsien Wang of the Psychology Department at Kaohsiung Medical University, Taiwan, for sharing with us their auditory oddball ERP data. We are very grateful to Dr. David Allbritton of the Psychology Department at DePaul University, for his comments on an earlier draft of the manuscript.

SUPPLEMENTARY MATERIAL

Accounting for time dependence in large-scale multiple testing of event-related potential data: Online supplement. The impact of ERP time dependence on multiple testing results (DOI: 10.1214/15-AOAS888SUPP; .pdf). To demonstrate the impact of time dependence on the ability of multiple testing procedures to identify a predetermined true signal, a simulation study is conducted in which ERP data are generated according to model (3.1). This simulation study compares the GB procedure [Guthrie and Buchwald (1991)] and two FDR-controlling procedures: BH [Benjamini and Hochberg (1995)] and BY [Benjamini and Yekutieli (2001)]. The results highlight the instability of multiple testing results when using methods ignoring dependence among tests.

REFERENCES

ALLEN, G. I., GROSENICK, L. and TAYLOR, J. (2014). A generalized least-square matrix decomposition. *J. Amer. Statist. Assoc.* **109** 145–159. MR3180553

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **57** 289–300. MR1325392

BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** 1165–1188. MR1869245

BLAIR, R. C. and KARNISKI, W. (1993). An alternative method for significance testing of waveform difference potentials. *Psychophysiology* **30** 518–524.

BUJA, A. and EYUBOGLU, N. (1992). Remarks on parallel analysis. *Multivariate Behavioral Research* **27** 509–540.

CAUSEUR, D. and SHEU, C. F. (2014). ERP: Significance analysis of Event-Related Potentials data. R package version 1.0.1.

CAUSEUR, D., FRIGUET, C., HOUÉE-BIGOT, M. and KLOAREG, M. (2011). Factor analysis for multiple testing (FAMT): An R package for large-scale significance testing under dependence. *Journal of Statistical Software* **40** 1–19.

CAUSEUR, D., CHU, M. C., HSIEH, S. and SHEU, C. F. (2012). A factor-adjusted multiple testing procedure for ERP data analysis. *Behavior Research Methods* **44** 635–643.

DONOHO, D. and JIN, J. (2008). Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proc. Natl. Acad. Sci. USA* **105** 14790–14795.

EFRON, B. (2007). Correlation and large-scale simultaneous significance testing. *J. Amer. Statist. Assoc.* **102** 93–103. MR2293302

EFRON, B. (2010). *Large-Scale Inference*: *Empirical Bayes Methods for Estimation*, *Testing*, *and Prediction*. *Institute of Mathematical Statistics* (*IMS*) *Monographs* **1**. Cambridge Univ. Press, Cambridge. MR2724758

FRIGUET, C., KLOAREG, M. and CAUSEUR, D. (2009). A factor model approach to multiple testing under dependence. *J. Amer. Statist. Assoc.* **104** 1406–1415. MR2750571

GARDINER, J. M., GAWLIK, B. and RICHARDSON-KLAVEHN, A. (1994). Maintenance rehearsal affects knowing, not remembering: Elaborative rehearsal affects remembering, not knowing. *Psychonomic Bulletin & Review* **1** 107–110.

GROPPE, D. M., URBACH, T. P. and KUTAS, M. (2011a). Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review. *Psychophysiology* **48** 1711–1725.

GROPPE, D. M., URBACH, T. P. and KUTAS, M. (2011b). Mass univariate analysis of event-related brain potentials/fields II: Simulation studies. *Psychophysiology* **48** 1726–1737.

GUTHRIE, D. and BUCHWALD, J. S. (1991). Significance testing of difference potentials. *Psychophysiology* **28** 240–244.

HANDY, T. (2004). *Event-Related Potentials*. MIT Press, Cambridge, MA.

JIN, J. (2009). Impossibility of successful classification when useful features are rare and weak. *Proc. Natl. Acad. Sci. USA* **106** 8859–8864. MR2520682

JOHNSON, H. M. (1994). Processes of successful intentional forgetting. *Psychological Bulletin* **116** 274–292.

JÖRESKOG, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika* **32** 443–482. MR0221659

LAGE-CASTELLANOS, A., MARTÍNEZ-MONTES, E., HERNÁNDEZ-CABRERA, J. A. and GALÁN, L. (2010). False discovery rate and permutation test: An evaluation in ERP data analysis. *Stat. Med.* **29** 63–74. MR2751379

LEE, Y. S., LEE, H. M. and FAWCETT, J. M. (2013). Intentional forgetting reduces color-naming interference: Evidence from item-method directed forgetting. *Journal of Experimental Psychology. Learning, Memory and Cognition* **39** 220–236.

LEEK, J. T. and STOREY, J. D. (2008). A general framework for multiple testing dependence. *Proc. Natl. Acad. Sci. USA* **105** 18718–18723.

LEEK, J. T., JOHNSON, W. E., PARKER, H. S., JAFFE, A. E. and STOREY, J. D. (2014). SVA: Surrogate Variable Analysis. R package version 3.6.0.

MARDIA, K. V., KENT, J. T. and BIBBY, J. M. (1979). *Multivariate Analysis*. Academic Press, London.

NÄÄTÄNEN, R. (2003). Mismatch negativity: Clinical research and possible applications. *Int. J. Psychophysiol.* **48** 179–188.

PAZ-CABALLERO, M. D. and MENOR, J. (1999). ERP correlates of directed forgetting effects in direct and indirect memory tests. *European Journal of Cognitive Psychology* **11** 239–260.

POLDRACK, R. A., MUMFORD, J. A. and NICHOLS, T. E. (2011). *Handbook of Functional MRI Data Analysis*. Cambridge Univ. Press, Cambridge. MR2839490

PRESS, W. H., TEUKOLSKY, S. A., VETTERLING, W. T. and FLANNERY, B. P. (2007). *Numerical Recipes*: *The Art of Scientific Computing*, 3rd ed. Cambridge Univ. Press, Cambridge. MR2371990

ROSBERG, T., BUTROUS, N. N. and FORD, J. M. (2008). Reduced auditory evoked potential component N100 in schizophrenia—A critical review. *Psychiatric Research* **161** 259–274.

RUBIN, D. B. and THAYER, D. T. (1982). EM algorithms for ML factor analysis. *Psychometrika* **47** 69–76. MR0668505

RUGG, M. D. and CURRAN, T. (2007). Event-related potentials and recognition memory. *Trends Cogn. Sci.* (*Regul. Ed.*) **11** 251–257.

SHEU, C. F., PERTHAME, E., LEE, Y. S. and CAUSEUR, D. (2016). Supplement to "Accounting for time dependence in large-scale multiple testing of event-related potential data." DOI:10.1214/15-AOAS888SUPP.

SUN, W. and CAI, T. T. (2009). Large-scale multiple testing under dependence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 393–424. MR2649603

SUN, Y., ZHANG, N. R. and OWEN, A. B. (2012). Multiple hypothesis testing adjusted for latent variables, with an application to the AGEMAP gene expression data. *Ann. Appl. Stat.* **6** 1664–1688. MR3058679

SUN, Y., ZHANG, N. R. and OWEN, A. B. (2014). LEAPP: Latent effect adjustment after primary projection. R package version 1.1.

THOMSON, G. H. (1951). *The Factorial Analysis of Human Ability*. London Univ. Press, London.

VAN DER LAAN, M. J. and DUDOIT, S. (2007). *Multiple Testing Procedures with Applications to Genomics*. Springer, New York.

VUL, E., HARRIS, C., WINKIELMAN, P. and PASHLER, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspect. Psychol. Sci.* **4** 274–290.

WEINER, B. (1968). Motivated forgetting and the study of repression. *J. Pers*. **36** 213–234.
WESTFALL, P. H. and YOUNG, S. S. (1993). *Resampling-Based Multiple Testing*: *Examples and Methods for P-Value Adjustment*. Wiley, New York.
WILLIAMS, L. M., SIMMS, E., CLARK, C. R., PAUL, R. H., ROWE, D. and GORDON, E. (2005). The test-retest reliability of a standardized neurocognitive and neurophysiological test battery: "neuromarker". *Int. J. Neurosci*. **115** 1605–1630.
WOOLRICH, M. W., BECKMANN, C. F., NICHOLS, T. E. and SMITH, S. M. (2009). Statistical Analysis of FMRI Data. In *fMRI techniques and protocols* (M. Filippi, ed.). Humana Press, New York.

C.-F. SHEU
INTITUTE OF EDUCATION
NATIONAL CHENG KUNG UNIVERSITY
1 UNIVERSITY ROAD
TAINAN 701
TAIWAN
E-MAIL: csheu@mail.ncku.edu.tw

É. PERTHAME
D. CAUSEUR
AGROCAMPUS OUEST, IRMAR UMR 6625 CNRS
65 RUE DE ST-BRIEUC, CS 84215
35042 RENNES CEDEX
FRANCE
E-MAIL: emeline.perthame@agrocampus-ouest.fr
david.causeur@agrocampus-ouest.fr

Y.-S. LEE
DEPARTMENT OF PSYCHOLOGY
NATIONAL CHUNG CHENG UNIVERSITY
168 UNIVERSITY ROAD
CHIAYI 621
TAIWAN
E-MAIL: psyysl@ccu.edu.tw