Behav Res (2012) 44:635–643
DOI 10.3758/s13428-012-0230-0

A factor-adjusted multiple testing procedure for ERP data analysis

David Causeur · Mei-Chen Chu · Shulan Hsieh · Ching-Fan Sheu

Published online: 26 June 2012

Abstract Event-related potentials (ERPs) are now widely collected in psychological research to determine the time courses of mental events. When event-related potentials from treatment conditions are compared, often there is no a priori information on when or how long the differences should occur. Testing simultaneously for differences over the entire set of time points creates a serious multiple comparison problem in which the probability of false positive errors must be controlled, while maintaining reasonable power for correct detection. In this work, we extend the factor-adjusted multiple testing procedure developed by Friguet, Kloareg, and Causeur (*Journal of the American Statistical Association*, *104*, 1406–1415, 2009) to manage the multiplicity problem in ERP data analysis and compare its performance with that of the Benjamini and Hochberg (*Journal of the Royal Statistical Society B*, *57*, 289–300, 1995) false discovery rate procedure, using simulations. The proposed procedure outperformed the latter in detecting more truly significant time points, in addition to reducing the variability of the false discovery rate, suggesting that corrections for mass multiple testings of ERPs can be much improved by modeling the strong local temporal dependencies.

D. Causeur
Agrocampus Ouest,
Rennes, France

M.-C. Chu
Institute of Cognitive Science, National Cheng Kung University,
1 University Road,
Tainan City 701, Taiwan

S. Hsieh
Department of Psychology, National Cheng Kung University,
1 University Road,
Tainan City 701, Taiwan

C.-F. Sheu (✉)
Institute of Education, National Cheng Kung University,
1 University Road,
Tainan City 701, Taiwan
e-mail: csheu@mail.ncku.edu.tw

Keywords Event-related potentials · Multiple comparisons · False discovery rate · Factor-adjusted multiple testing

Introduction

Event-related brain potentials (ERPs) reflect voltage changes in an electroencephalogram that are time-locked to some physical or mental occurrence (Handy, 2004). As instrumentation technology in brain sciences continues to progress, ERPs have increasingly been used to study the time courses of mental processes underlying emotion, memory, attention, perception, language, and so on.

The quantification of the effects of experimental variables on ERPs, however, has been plagued by the problem of multiple comparisons (Tukey, 1953). For example, exploring whether there is any difference between ERP waveforms from two experimental conditions involves simultaneously testing many hypotheses, such that conventional significance testing can substantially raise the overall Type I error rate (see Groppe, Urbach, & Kutas, 2011a, for a review). In exploratory studies, researchers often have no a priori information on when or how long the differences should occur. Testing simultaneously for differences over the entire set of time points must control for the probability of false positive errors, while maintaining reasonable power for correct detection.

In this aspect, ERP analysis shares traits with other high-dimensional settings such as genomics (Storey & Tibshirani, 2003) and brain imaging (Genovese, Lazar & Nicols 2002) by needing to simultaneously analyze many related measured variables. Although the number of simultaneous tests in the analysis of single-channel ERPs is in the hundreds to thousands, as compared with millions of hypotheses in genomic studies, a sizable dependency exists between successive time samples of brain activity (Hunt, 1985). It is known that dependencies can alter the relative ordering of significance levels among hypotheses (Efron, 2007). However, most

statistical methods for performing multiple testing assume independence among the variables being tested. Ignoring the dependence among hypothesis tests can result in both unstable significance measures and bias caused by the confounding of dependent noise and the signal of interest. How to perform large-scale significance testing in the presence of arbitrarily strong dependence is among the most discussed topics in the high-dimensional data analysis literature (Efron, 2010).

Until recently, researchers in psychophysiology have used either the Guthrie–Buchwald (1991) test or the permutation test proposed by Blair and Karinski (1993) to conduct significance testing of difference ERPs. An attractive alternative, the false discovery rate (FDR) procedure of Benjamini and Hochberg (1995), which controls the average proportion of false rejections over the total number of rejections in multiple testing situations, adjusts *p*-values in a simple and efficient way. The method has gained popularity among neurophysiological researchers (e.g., Achim, 2001). However, in the presence of structural dependencies and correlated observations—features that are characteristic of ERPs—the simultaneous testing of many hypotheses can be difficult to adjust even with the Benjamin–Hochberg (BH) procedure and its extension (Benjamini & Yekutieli 2001), which modifies the BH method to account for a specific type of dependence across tests (called *positive dependence*). It ensures a better control of the FDR in situations of positive dependence but is also known to be very conservative. In ERP data analysis, the problem seems not to be the control of the FDR (the BH method performs well on highly autocorrelated data) but the power of the multiple-testing procedure. Curiously, Groppe, Urbach, and Kutas (2011a) paid little attention to the problem of dependent tests in their recent review of mass univariate analysis of ERPs, nor did they discuss the Guthrie-Buchwald method (1991), which explicitly takes dependence into account.

The factor-adjusted multiple testing (FAMT) procedure is a recently developed method for large-scale simultaneous hypothesis testing in high correlation situations (Friguet, Kloareg, & Causeur, 2009). Dependence is formalized as latent variables in the FAMT procedure,and multiple hypothesis testing is performed with adjustment for latent variables. The basic idea of the method is to derive modified test statistics under the assumption that the conditional covariance of the responses given the treatment variables has a certain factor-analytic structure (Mardia, Kent, & Bibby, 1979).

The utility of the method has been demonstrated on a benchmark microarray data set (Hedenfalk et al., 2001), which is widely used for the assessment of competing multiple-comparison procedures. The FAMT procedure improves on the performance of the BH procedure by stabilizing the ranking of the *p*-values for thousands of hypotheses under consideration. The similarity of correlations in gene sequences and temporal correlations suggests that the method should be applicable to the comparisons of ERP waveforms as well. The present study aims to adapt the FAMT procedure under dependence (Friguet et al., 2009) to the analysis of ERP waveforms. We propose a *dynamic* version of the FAMT procedure to account for the high correlations among ERP observations over time and perform simulation studies to illustrate the advantages of using this novel procedure for managing the multiplicity problem in ERP data analysis.

This article is organized as follows. We first introduce the BH procedure for multiple testing, followed by a discussion of the FAMT procedure under dependence method (Friguet et al., 2009). We perform two simulation studies: one to illustrate the impact of dependence on the performance of the BH procedure and another to assess and compare the performance of our dynamic FAMT procedure against that of the BH procedure. We describe the dynamic factor modeling of ERPs in the simulation studies. Conclusions are presented in the last section.

The Benjamini–Hochberg procedure

An important breakthrough in managing the multiple testing problem in high-dimensional settings is the method, due to Benjamini and Hochberg (1995), that bounds a particular measure of inaccuracy called the *false discovery rate* (FDR). This approach to multiple testing is to control for the expected proportion of incorrectly rejected null hypotheses (false discoveries) over the total number of rejections made:

$$FDR = E\left[\frac{FP}{NR}\middle| NR > 0\right] \Pr\{NR > 0\},$$

where *FP* is the number of false positives and *NR* is the number of rejections. The Benjamini-Hochberg procedure (1995) provides a useful definition of the scientifically relevant error quantity, leading to less conservative decision rules than those defined by traditional methods for multiple comparisons. The BH procedure for independent hypotheses can be described as follows:

1. Choose a value q such that $0 \leq q \leq 1$, where q is the acceptable level of false discovery rate. Typically, q is set to be the same as the classical α level of significance at .05.
2. Sort the observed *p*-values $p_1 \leq p_2 \leq \cdots p_m$, where m is the number of tests (hypotheses).
3. Let d be the largest index i for which $p_i \leq \left(\frac{i}{m}\right)q$, $i = 1, 2, \ldots, m$.
4. Reject all null hypotheses whose *p*-values$\leq p_d$.

Benjamini and Hochberg (1995) showed that the resulting FDR is smaller than or equal to $\left(\frac{m_0}{m}\right)q \leq q$, where m_0 is the (unknown) number of true null hypotheses.

The open-source statistical analysis software R (R Development Core Team, 2011) has a number of implementations of the BH procedure. For example, the function p.

adjust adjusts (raw) *p*-values for multiple comparisons on the basis of the BH procedure by setting the method option to "BH." The packages multtest (Pollard, Gilbert, Ge, Taylor & Dudoit 2004) and multcomp (Bretz, Hothorn, & Westfall, 2011) also contain functions that implement various versions of the BH procedure.

The application of FDR control in managing the multiplicity problem has been relatively rare in psychological research. Lage-Castellanos, Martínez-Montes, Hernández-Cabrera, and Galán (2010) reported a study evaluating the FDR method and permutation tests in ERP data analysis. It was found that the BH procedure for independent tests provided an effective method for dealing with the multiplicity problem. The BH procedure appeared to improve the probability of correct detection over the permutation tests, while controlling for the probability of spurious detections. A similar study focusing on comparing a variety of the BH procedures and permutation tests appeared more recently in psychophysiological research (Groppe, Urbach, & Kutas, 2011b).

Factor-adjusted multiple testing

Two observations motivate the study of the multiple testing problem under dependence: (1) 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). (2) Ignoring dependence among test statistics reduces the detection of true positives (Leek & Storey, 2008). In ERP studies, the identification of which time intervals of brain processes are related to experimental variables of interest clearly fits the problem context, since time samples of brain activity tend to be strongly dependent (Hunt, 1985). Recent statistical literature (see Friguet et al., 2009, for a review) has shown that dependence affects not the control of the expectation of the false discovery proportion (FDR) but its variance. In other words, ignoring dependence can lead to a control of the true false discovery proportion at a very low level, which results in a conservative procedure. Accounting for dependence by a factor model produces a less variable distribution of the false discovery proportion and, consequently, improves the true discovery proportion.

In the following, we describe the FAMT procedure in relation to ERP measurements. The m number of voltage differences over time is formalized as an m-dimension vector of normal random variables with respect to some treatment variables. The covariance matrix of differences is assumed to be positive definite, and constant with respect to the treatment variables.

Hsu (1992) was the first to propose the factor-analytic approach to simultaneous inference in the general linear model. He used a one-factor approximation of the correlation between test statistics for a set of linear contrasts in a univariate analysis of variance to derive a thresholding method for controlling the family-wise error rate. Friguet et al. (2009) considered the FDR and modeled the correlation structure of the responses by a multiple-factor structure, which was then used to obtain new test statistics for the significance of general linear contrasts. In essence, the factor-adjusted multiple comparison procedure is multiple comparisons on factor-adjusted data.

The method can be described in the following steps:

1. Consider a model for the ERP measurements, Y_t, of a subject over time $t=1,\ldots, m$,

$$Y = f(t, x; \theta_t) + \varepsilon_t, \quad Var(\varepsilon_t) = \sum = BB' + \Psi_t$$

where $f(t, x; \theta_t)$ represents the fixed-effects component of the model dependent on time, treatment variables x, and parameters θ_t. It is assumed that the conditional variance of the responses can be decomposed into an $m \times m$ diagonal matrix of unique variances Ψ and an $m \times q$ matrix of factor loadings B (Bartholomew, 1987). In other words, there exist a small number ($q<m$) of latent variables (factors) that account for common information in the responses.

2. Given the normally distributed factors $Z_1, \cdots, Z_q$ with expectation $\mathbf{0}$ and their corresponding loadings $b_1, \cdots, b_q$, the factor-adjusted ERP observations are of the form:

$$Y - \left(b_1 Z_1 + \cdots + b_q Z_q\right) = f(t, x; \theta_t) + e_t,$$
$$Var(e_t) = \psi_t^2 \mathrm{I}$$

where ψ_t^2 are the unique variances and I is an identity matrix. Friguet et al. (2009) fit the model by the maximum likelihood method, using the EM algorithm proposed by Rubin and Thayer (1982). They also derived an explicit formula linking the inflation of the variance of the number of false discoveries to the amount of correlation among the multiple tests and devised an iterative method to select the number of factors on the basis of minimizing the variance inflation.

3. Assuming the factor analysis model (Ψ, B, Z) for Y_t (Mardia et al., 1979) and given a sample of subjects n, one can define, for $t=1,\ldots, m$, factor-adjusted test statistics for $H_{0,t} : \lambda'\theta_t=0$, where λ is a vector of coefficients defining an arbitrary linear contrast of interest:

$$T_{z,t} = \frac{\sigma_t}{\psi_t}\left[T_t - \frac{b_t^{'}}{\sigma_t}\tau(Z)\right] \quad ,$$

where $\tau(Z) = \sqrt{n} S_{zx} S_{xx}^{-1'} \lambda / \sqrt{\lambda' S_{xx}{}^{-1} \lambda}$, σ_t is the conditional standard deviation (SD) of Y_t given x, S_{xx} is the sample variance–covariance matrix of treatment variables, S_{zx} is the sample covariance matrix of the factors

and the treatment variables, b_t is the tth row of B, and T_t is the normalized estimate of the linear contrasts (i.e., unadjusted test statistics).

Furthermore, the distribution of T_z and its expected value and variance are known.

Friguet et al. (2009) showed that the reduced dependence among the modified test statistics T_z leads to a large gain in power and stable estimates of FDR measures such as the variance of the number of false discoveries. The expression of the variance of the number of false discoveries is given in Friguet et al. The expression shows that this variance is minimal when the tests are independent. We can determine the number of factors in the model by monitoring the variance of the number of false discoveries. This variance decreases when factors are added in the model until the proper number of factors is included, at which point the residual errors are uncorrelated. The FAMT procedure has been implemented in R for genomic studies (Causeur, Friguet, Houée, & Kloareg, 2011). The main goal of this article is to propose a dynamic extension of the FAMT procedure to improve upon how mass univariate analysis of ERP data is currently practiced. The extension is *dynamic* because it captures a time dependence and a different adjustment is made for each time point's comparison.

The novel procedure performs as well as the BH procedure in controlling the global FDR, while permitting more positive discoveries.

Simulation studies

Ultimately, an assessment of competing statistical procedures is best performed with real, benchmark data sets in which different ERP components have been known to differ significantly, and when it is known when and for how long they differ between different experimental conditions. Lacking access to any such benchmark ERP data, we resort to conducting simulation studies in which true differences between ERPs are known a priori.

Simulation of ERP curves

We consider the standard paired comparison design with the same group of subjects performing a experimental task in two conditions: conditions 1 (control, say) and 2 (treatment). The length of an ERP waveform is 800 ms, with one observation per milliseconds. The number of subjects is 16. These values are chosen to match a typical experimental paradigm in ERP data collection. The basic data unit is a vector of consecutive ERPs (in microvolts, μV) over time that are averaged over trials for a single channel recorded from an individual subject for one condition. The vector is made up of three components: (1) a negative peak (half-cycled sine wave) centered at 400 ms, spanning from 350 to 450 ms in latency, (2) a slow-going sine wave, (3) an autocorrelated noise component. The first two components are constant across subjects. They represent the structural dependence of the ERP waveforms. The first-order autoregressive process is used to capture the time dependence of the ERP waveforms (Hunt, 1985). The difference in the heights (amplitudes) of the peaks between the ERPs in the interval (350 ms, 450 ms) of the two conditions defines the critical region for correct detection. Except for the noise process, scaling factors, and some minor details, our simulated ERP data are similar to those described in Yeung, Bogacz, Holroyd, and Cohen (2004). In fact, their computer programs are used to generate the true waveform signals and the true difference curve shown in Fig. 1. (These MATLAB programs are available at http://www.cs.bris.ac.uk/home/rafal/phasereset/.) The thick part of the curve in the right panel indicates the time points for which the statistical power of a paired t-test (based on 16 subjects and at the .05 level of significance) is greater than 0.75 (with 0.96 at the peak).

An autoregressive model for ERPs

We now state more formally the model from which ERP curves are drawn for our simulation studies. Let Y_{ijt} be the value (amplitude) of the ERP waveform at time t for the ith subject of condition j, for j=1, 2. The following model is assumed for Y_{ijt}:

$$Y_{ijt} = s_j(t) + \varepsilon_{ij}(t),$$

where t=1,…, T identifies the time points (digitizing a continuous curve), $s_j(t)$ represents the true signal of condition j, and $\varepsilon_{ij}(t)$ is the random error term assumed to be normally distributed with a mean of 0 and an SD of σ. An autocorrelation of lag 1 structure is assumed for the dependence of the residual process $\varepsilon_{ij}(t)$ along time: Cor $\left[\varepsilon_{ij}(t-1), \varepsilon_{ij}(t)\right] = \rho$, for t=1,…, T. Large-scale significance testing of the mean ERP difference between the two conditions consists in the simultaneous tests of the null hypothesis $H_0^{(t)} : s_1(t) = s_2(t)$.

To illustrate the impact of autoregressive dependence on significance testing of difference potentials, we simulated 1,000 data sets, each of which consisted of 16 ERP difference curves following a normal distribution whose mean was the true difference curve (at each time point) with an SD of one, and a first-order autoregressive covariance structure with autocorrelation ρ. For each comparison, the p-value of a paired t-test was obtained from evaluating whether or not the difference was significantly different from zero at that

particular time point. The resulting *p*-value cutoffs were adjusted by the BH procedure to control the false discovery rate at the .05 level. For each simulated data set, the significant time points were compared with the true difference curve, and the cases of false and true discoveries counted. True discoveries correspond to significant time points in the interval of peak values identified by the thick curve on the right panel of Fig. 1. A total of 800 tests were performed for each of 1,000 simulation runs. We computed the proportions of true and false discoveries for each simulated data set. The means, *SD*s, and medians of these proportions are reported in Table 1. Since the number of no discoveries (either true or false) is also an important indicator of the instability (variability) of error quantity one wishes to control, we also calculated the proportions of zero discoveries in 1,000 simulation runs.

Table 1 summarizes the results for ρ=0 and ρ=0.99. The value 0.99 is chosen not only because it highlights the dependence problem, but also because it matches what we have observed in empirical ERP data. What is observed is the correlation of the raw data, whereas what is modeled is the correlation of error component. The two are usually close. It is noted that a large autocorrelation produces high instability in the distribution of error rates. The most striking features of the results are a large proportion of cases with no true detections for high dependency and a marked increase of the *SD* of the true discovery proportion when the autocorrelation goes from 0 to 0.99. The *SD* for FDR increases by about a factor of 1.5, that for the true discovery rate by about a factor of 3. The proportion of cases where no true discoveries were made increases by a factor of 26 when going from no dependence to ρ=0.99 in this simulation study. These observations are consistent with many recent findings in the theory of multiple testing for highly correlated data in high-dimensional settings (see Efron, 2007; Friguet et al., 2009; Leek & Storey, 2008).

The simulation results above clearly illustrate the failure of the BH procedure for highly correlated simultaneous testing situations. The procedure proposed by Benjamini and Yekutieli (2001) is known to offer better control of the FDR in situations of positive dependence across tests, but at the cost of being too conservative. For comparing ERPs using the BH procedure, the issue is not the control of the FDR but the loss of statistical power. Therefore, we do not consider the Benjamini and Yekutieli (2001) procedure in this study. In the following, we describe a dynamic extension of the FAMT procedure (Friguet et al., 2009) to account for time dependence in ERP curves so that the variability of error rates can be reduced and the power of positive detections increased. The same 1,000 ERP data sets with ρ=0.99 will be used to assess the performance of this novel procedure.

Dynamic factor modeling of ERPs

The main difference between the static scenario addressed by the FAMT procedure in genomic data and the dynamic ERP data is the pattern of dependence over time. In the procedure to be described below, we first model this time dependence by a first-order autoregressive correlation and then attempt to decompose this dependence by a factor-analytic model. Although it is not explicitly specified in the dynamic factor model, it turns out that each factor captures time dependence at a given location on the whole interval of observations. In other words, instead of extending the original FAMT procedure globally, it is necessary to adapt the procedure to adjust ERP observations locally in time (point-wise, in fact).

The basic idea of a dynamic factor modeling of ERP signals is to capture the pattern of the error process ε by a linear combination of latent factors $f_k(t)$:

$$\varepsilon_{ij}(t) = b_{ij}^{(1)} f_1(t) + b_{ij}^{(2)} f_2(t) + \cdots + b_{ij}^{(q)} f_q(t) + e_{ij}(t),$$

where q is the number of factors, $b_{ij} = \left(b_{ij}^{(1)}, b_{ij}^{(2)}, \cdots, b_{ij}^{(q)}\right)'$ is the q-vector of factor loadings for $\varepsilon_{ij}(t)$, and $e_{ij}(t)$ is the specific error term (to account for that part of the variable that is unique), which is assumed to be normally distributed

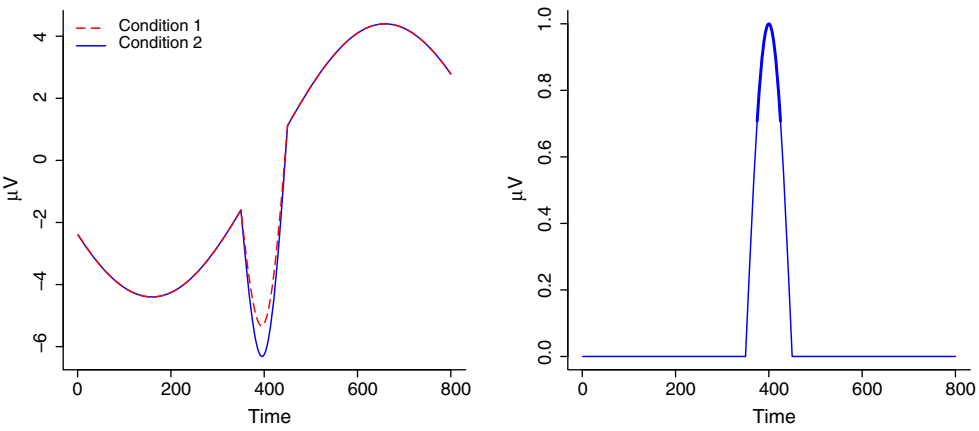

Fig. 1 True ERP waveforms and true difference curve in the simulation study. The left panel shows the true waveforms for the two conditions. The right panel shows that the true difference in the two conditions is zero everywhere, except for the time interval (350 ms, 450 ms)

Table 1 Results of a simulation showing the impact of dependence over time on discovery rates, based on the Benjamini–Hochberg procedure

	Mean	*SD*	Median	Zero discovery
	ρ=0.00			
False discovery	3.30	8.51	0.00	74.40
True discovery	37.90	14.09	37.25	2.00
	ρ=0.99			
False discovery	2.62	12.56	0.00	92.90
True discovery	38.48	42.75	0.00	52.30

Means, standard deviations (*SD*s), and medians for proportions (%) of false and true discoveries are based on 1,000 simulation runs for no time dependence (ρ=0) and for ρ=0.99. The proportion (%) of runs in which no discoveries were made is also tabulated.

with a mean of 0 and an *SD* of ψ_{ij}. Moreover, it is assumed that the specific errors are independent along time.

The first step to separate the ERP observations generated by the error process, $\varepsilon_{ij}(t)$, from the true ERP signals is to consider observations only for those time points where the usual (two-sided) paired *t*-tests yield no significant difference between conditions.

Different factor-analytic models of this observed error process can then be selected with different numbers of factors. Here, the same iterative method of choosing the number of factors on the basis of the criterion of minimizing variance inflation is used. We set the maximum number of factors to 12 because the rank of the empirical covariance matrix is 15 (=16−1) and because, by trial and error, we have found that this upper limit not only keeps the computational burden down, but also captures enough dependence structure in the data.

Suppose the number of factors chosen is q; then, the latent components of the model are clearly identified, and the model can be written as:

$$Y_{ij}(t) - \left(b_{ij}^{(1)} f_1(t) + b_{ij}^{(2)} f_2(t) + \ldots + b_{ij}^{(q)} f_q(t)\right) = s_j(t) + e_{ij}(t).$$

At each time point, the factor-adjusted data on the left-hand side of the equation can now be used for testing the null hypotheses $H_0^{(t)} : s_1(t) = s_2(t)$ with an independent error structure $e_{ij}(t)$. This turns dependent tests with unadjusted data into independent tests with adjusted data.

However, to guard against overfitting the model with too many factors, especially in the present small sample situation, we select only 1 factor among as many as up to 12 factors in the model determined earlier. In other words, to adjust the data at each time point, we select the best-fitting 1-factor model (among the q number of 1-factor models) by minimizing the Bayesian information criterion (BIC; Schwarz, 1978).

It can be checked whether the factor chosen for a given time point is consistent with the association between the factors and locations in the entire intervals. For example, the (absolute) correlation curves between each factor of a five-factor model and ERP values are displayed on Fig. 2 for one simulated data set. For each time point in the horizontal axis of Fig. 2, we compute five separate (absolute) correlations. For each factor of the five-factor model, we compute the correlation between the 16 observed ERPs and the adjusted ERPs predicted by the respective one-factor model. It clearly shows that, corresponding to a respective time interval, each factor is highly correlated with ERP values. Note that this association between one factor and one time interval results from the fit and is of course not prespecified in the model. This association between each factor and a time interval explains why, at each time point, the *t*-test is adjusted only from the effect of the most appropriate factor. This avoids overfitting of the model used for testing at each time point, which would occur if nonrelevant factors were used. Overfitting actually worsens the properties of multiple testing procedures since it results in an underestimation of the residual variance and, consequently, an uncontrolled Type I error rate (minimizing BIC is known to select models more parsimoniously).

To be consistent with researchers' belief that isolated peaks or troughs with only a single time point in ERPs are very rarely truly significant, we prefer kernel smoothing (Wand & Jones, 1995) the factor-adjusted *t*-statistics with a bandwidth of 20 time points (equal to 2.5 % of the total observations in the current setup) before computing the corresponding *p*-values.

Finally, the BH procedure is used to correct the resulting *p*-value cutoffs by controlling the FDR at the .05 level. The left panel of Fig. 3 shows the raw adjusted *t*-statistic before smoothing. The right panel shows the smoothed *t*-statistics

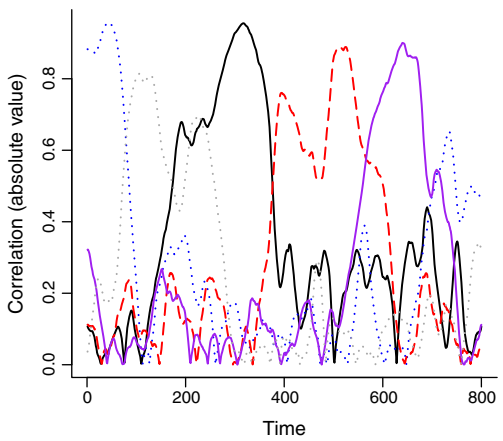

Fig. 2 Correlation curves between factors in a five-factor model and ERP values for one simulated data set. Corresponding to a respective time interval, each correlation curve displays a high (absolute) correlation between the ERP values and one of the factors

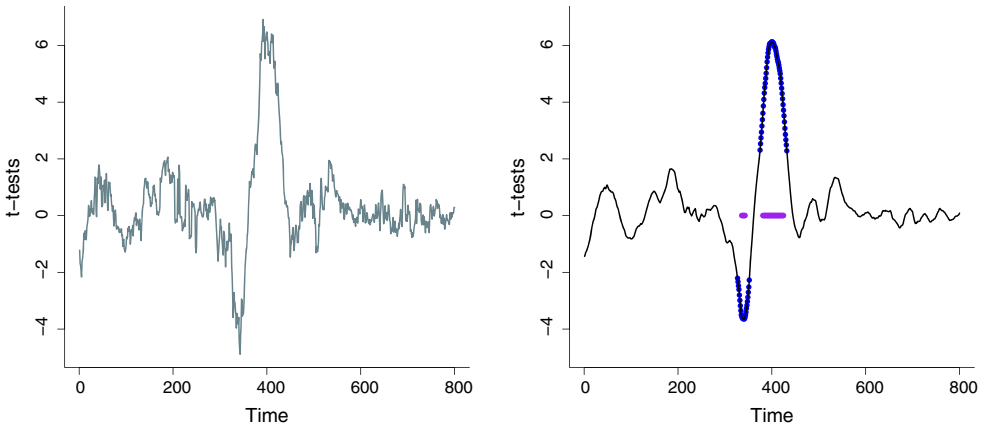

Fig. 3 Left panel: factor-adjusted *t*-statistics. Right panel: kernel-smoothed factor-adjusted *t*-statistics. The significant time points are marked with blue points. The segments of purple points are the significant time points after Benjamini–Hochberg correction to control the false discovery rate at the .05 level

and the significant time intervals identified after the correction with the BH procedure at the end. The motivation for introducing a final smoothing step in our method comes from a careful examination of the *t*-test curves after removal of the time dependency, using the factor model. Restoring independence moves the *t*-test curves closer to the true signals but also makes them wigglier. This variability can sometimes generate isolated *t*-test values far from their neighboring values. These isolated values do not make sense and tend to inflate the number of false discoveries. It is true that a large amount of smoothness generally introduces dependency. However, in our method, the smoothing corrections are very local and do not affect the large-scale shape of the *t*-test curves.

We summarize, in six concrete steps, our proposed procedure for significance testing of ERP potentials in a paired design as discussed above and comment on how each step can be performed with either an existing function in R or by a suite of R functions created by the first author. These programs will be made available to users upon request.

1. At each time point, perform a paired *t*-test. (This step can be performed with t.test in R.)
2. For those ERP observations at time points where paired tests showed no significant difference between conditions, center and scale them together. These observations contain information about residual ERPs. For those ERP observations at time points where paired tests showed significant difference between conditions, center and scale them within each condition, respectively.
3. Fit a series of factor-analytic models to the entire scaled and centered ERPs in step 2, starting from a 1-factor model and ending with a 12-factor model. Determine the number of factors for the best-fitting model. Retain the corresponding factor scores for later use.
4. Calculate the factor-adjusted *t*-statistics for testing the difference potentials (the condition effect in the model). This is achieved by first selecting the best-fitting one-factor model at each time point, considering only the factors selected in the previous step. (This and the previous steps can be performed with a suit of R functions in the FAMT package; Causeur et al., 2011.) Note that in factor adjustment, we rescale factor loadings b_k and specific variances Ψ_k: Let s_k be the standard deviation used to scale the data in factor extraction; then, Ψ_k is multiplied by s_k^2 and b_k by s_k.
5. Choose an interval (window or bandwidth), compute the weighted average factor-adjusted *t*-statistics, and calculate the corresponding *p*-values. (This step can be performed using an R function such as ksmooth for kernel smoothing.)
6. Adjust the *p*-value cutoffs by the BH procedure to control the FDR at a desired level and identify the significant time interval for difference potentials. (This step can be performed with the p.adjust function in R.)

The procedure is applied to the same 1,000 sets of highly dependent simulated ERP data with ρ=0.99, and the results are summarized in Table 2.

It is noted that the true discovery rate increases by almost a factor of two with the new procedure, as compared with the BH procedure. In fact, the average true discovery rate is

Table 2 Results of a simulation showing the improvement of dynamic factor modeling over no factor adjustment on discovery rates

	Mean	*SD*	Median	Zero discovery
	ρ=0.99- Without factor adjustment			
False discovery	2.62	12.56	0.00	92.90
True discovery	38.48	42.75	0.00	52.30
	ρ=0.99- Dynamic factor adjustment			
False discovery	4.31	13.97	0.00	85.80
True discovery	68.92	39.53	49.20	22.00

Means, standard deviations (*SD*s), and medians for proportions (%) of false and true discoveries based on 1,000 simulation runs using the dynamic factor adjusted procedure and no adjustment. The proportions (%) of runs in which no false (or positive) discoveries were made are also tabulated.

about 69 %, which is reasonably close to the expected power of 75 % by design. Although the new procedure does not seem to reduce the variability of discovery rates as indicated by the sizes of *SD*s, the proposed procedure does provide more stable and reliable findings from one data set to another by reducing the proportion of runs in which no true discovery is made by about a factor of two and a half.

Conclusion

One of the most prominent features of ERPs is that they are highly correlated measurements over time. The significance test of different potentials developed by Guthrie and Buchwald (1991) explicitly recognizes this feature and accounts for it. Similarly, in adapting the FAMT procedure to manage the multiplicity problem in ERP data analysis, we have found it important to concentrate on factor-adjusting the test statistics locally (point-wise, in fact, and choosing one of the best-fitting factors in the global model), in combination with employing a smoothing technique to reduce the possibility of falsely declaring isolated peaks and valleys in ERPs significant. The dynamic ranges of the two adjustments are quite different, with the latter being much broader than the former. How to choose an appropriate amount of smoothing will require further investigation. At present, we also lack a formal proof that our stratagem of choosing only one factor for adjustment at one time point neither overfits nor underfits the data. Our experience with the simulation does show that this strategy is simple and fast, works reasonably well most of the time, and yields consistent results.

The Guthrie–Buchwald test also considers the dependence among tests, but it is, strictly speaking, not an FDR-controlling multiple-testing procedure. Guthrie and Buchwald (1991) suggested that autoregressive time processes are very regular, which can induce long-range intervals of false discoveries. For several different values of the autoregressive coefficient, they provided a table that can be used to determine over which length a run of successive time points with p-values lower than .05 is deemed significant. Because this Type I error rate is not the same as in the usual multiple testing procedures, we have decided to exclude this method from the present study. Moreover, the factor model is here fitted to an AR correlation structure, but it is not designed only for this kind of time dependence, whereas the Guthrie–Buchwald tables for significant lengths of runs are established for AR(1) error process only. We have programmed the Guthrie–Buchwald method in R and applied it to the simulated data sets with ρ=.99. In 74.05 % of the simulations, the test finds no significant intervals (with what Guthrie and Buchwald called a graphical threshold of .05). The mean false discovery proportion is .027 (the median is 0. and the *SD* is .14). The mean true discovery proportion is .24 (the median is 0, and the *SD* is .43).

In this work, we have not considered autoregressive moving average (ARMA) models for ERP data, and we also readily admit that the correlation structure of ERP data is likely to be much more complex than autocorrelation per se. However, the moving average (MA) part of an ARMA dependence structure should be much easier to capture by a factor model than the autoregressive part (which is nonlinear). This explains why we have focused on the autoregressive structure in our simulation studies. Moreover, it is quite straightforward to define simple autoregressive models, whereas the scope of ARMA processes is so large that it could not be covered without doing a complex simulation study. The main purpose of the study is to demonstrate how our novel method can improve on ERP data analysis in a commonly accepted, albeit simplistic, setup.

The simulation results suggested that the proposed procedure performed very well—in power gains and in stabilizing true discoveries—as compared with using the BH procedure (1995), which, as currently practiced, does not dynamically adjust latent variables for dependence.

In our simulation studies, we have chosen to work with a relatively small sample size of 16 subjects, although doubling the sample size would have certainly made the modeling, statistically speaking, less challenging. Given that most practicing researchers would be loathe to double their time and effort in data collection to increase power, our simulation results shown in Table 2 rather impressively demonstrate the improvements our proposed dynamic FAMT procedure can make over the BH procedure in mass univariate analysis of ERPs in a paired-comparison paradigm. To extend the procedure for two independent groups or other experimental designs is conceptually straightforward, although some programming efforts will be required. On the other hand, to extend the current procedure to multichannel ERPs will demand further effort in statistical modeling.

Author Note This work was supported, in part, by a Hubert Curien research grant to David Causeur and Grant NSC 100-2410-H-006-027-MY3 to Ching-Fan Sheu from the National Science Council of Taiwan. We thank Andrew Heathcote, David Allbritton, and an anonymous reviewer for a number of helpful comments and suggestions. Correspondence concerning this article should be addressed to Ching-Fan Sheu, Institute of Education, National Cheng Kung University, 1 University Road, Tainan City 701, Taiwan (e-mail: csheu@mail.ncku.edu.tw).

References

Achim, A. (2001). Statistical detection of between-group differences in event-related potentials. *Clinical Neurophysiology, 112,* 1023–1034. doi:10.1016/S1388-2457(01)00519-3

Bartholomew, D. J. (1987). *Latent variable models and factor analysis*. London: Griffin.

Benjamini, Y., & Yekutieli, D. (2001). The control of false discovery rate in multiple testing under dependency. *The Annals of Statistics, 29,* 1165–1188. doi:10.1214/aos/1013699998

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, *57*, 289–300. Retrieved from http://www.jstor.org

Blair, R., & Karinski, W. (1993). An alternative method for significance testing of waveform difference potentials. *Psychophysiology, 30,* 518–524. doi:10.1111/j.1469-8986.1993.tb02075.x

Bretz, F., Hothorn, T., & Westfall, P. (2011). *Multiple comparisons using R*. Boca Raton, FL: Chapman & Hall/CRC. doi:10.1201/9781420010909

Causeur, D., Friguet, C., Houée, M., & Kloareg, M. (2011). Factor analysis for multiple testing (FAMT): An R package for large-scale significance testing under dependence. *Journal of Statistical Software*, *40*(14), 1–19. Retrieved from http://cran.r-project.org/

Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association, 102,* 93–103. doi:10.1198/016214506000001211

Efron, B. (2010). *Large-scale inference: Empirical Bayes methods for estimation, testing and prediction*. Cambridge: Cambridge University Press.

Friguet, C., Kloareg, M., & Causeur, D. (2009). A factor model approach to multiple testing under dependence. *Journal of the American Statistical Association, 104,* 1406–1415. doi:10.1198/jasa.2009.tm08332

Genovese, C. R., Lazar, N. A., & Nicols, T. (2002). Thresholding of statistical maps in functional neuroimaging using false discovery rate. *NeuroImage, 15,* 870–878. doi:10.1006/nimg.2001.1037

Groppe, D. M., Urbach, T. P., & Kutas, M. (2011a). Mass univariate analysis of event-related brain potentials/fields: I. A critical tutorial review. *Psychophysiology, 48,* 1711–1725. doi:10.1111/j.1469-8986.2011.01273.x

Groppe, D. M., Urbach, T. P., & Kutas, M. (2011b). Mass univariate analysis of event-related brain potentials/fields: II. Simulation studies. *Psychophysiology, 48,* 1726–1737. doi:10.1111/j.1469-8986.2011.01272.x

Guthrie, D., & Buchwald, J. S. (1991). Significance testing of difference potentials. *Psychophysiology, 28,* 240–244. doi:10.1111/j.1469-8986.1991.tb00417.x

Handy, T. (2004). *Event-related potentials: A methods handbook*. Cambridge, MA: MIT Press.

Hedenfalk, I., Duggan, D., Chen, Y. D., Radmacher, M., Bittner, M., Simon, R., & Trent, J. (2001). Gene expression profiles in hereditary breast cancer. *The New England Journal of Medicine, 344,* 539–548. doi:10.1056/NEJM200102223440801

Hsu, J. C. (1992). The factor analytic approach to simultaneous inference in the general linear model. *Journal of Computational and Graphical Statistics, 1,* 151–168. doi:10.2307/1390839

Hunt, E. (1985). Mathematical models of the event-related potential. *Psychophysiology, 22,* 395–402. doi:10.1111/j.1469-8986.1985.tb01621.x

Lage-Castellanos, A., Martínez-Montes, E., Hernández-Cabrera, J. A., & Galán, L. (2010). False discovery rate and permutation test: An evaluation in ERP data analysis. *Statistics in Medicine, 29,* 63–74. doi:10.1002/sim.3784

Leek, J. T., & Storey, J. D. (2008). A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences, 105,* 18718–18723. doi:10.1073/pnas.0808709105

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate analysis*. London: Academic Press. doi:10.1002/zamm.19810610315

Pollard, K. S., Gilbert, H. N., Ge, Y., Taylor, S., & Dudoit, S. (2004). *multtest: Resampling-based multiple hypothesis testing. R package version 2.8.0*. Retrieved from http://cran.r-project.org/

R Development Core Team (2011). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved from http://cran.r-project.org/

Rubin, D. B., & Thayer, D. T. (1982). EM algorithms for ML factor analysis. *Psychometrika, 47,* 69–76. doi:10.1007/BF02293851

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6,* 461–464. doi:10.1214/aos/1176344136

Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genome-wide studies. *Proceedings of the National Academy of Sciences, 100,* 9440–9445. doi:10.1073/pnas.1530509100

Tukey, J. W. (1953). *The problem of multiple comparisons [Mimeographed notes]*. Princeton, NJ: Princeton University.

Wand, M. P., & Jones, M. C. (1995). *Kernel smoothing*. London: Chapman & Hall.

Yeung, N., Bogacz, R., Holroyd, C. B., & Cohen, J. D. (2004). Detection of synchronized oscillations in the electroencephalogram: An evaluation of methods. *Psychophysiology, 41,* 822–832. doi:10.1111/j.1469-8986.2004.00239.x