計畫名稱：Massive multiple testing of event-related potentials and brain imaging data: A factor analytic approach to dealing with temporal and spatial dependence

計畫類別：個別型計畫
計畫編號：NSC－103－2410－H－032－MY3
執行期間：103年8月1日至106年7月31日

執行機構及系所：國立成功大學教育研究所

計畫主持人：許清芳
共同主持人：李玉琇，龔俊嘉

除繳交成果報告外，另繳交出席國際學術會議心得報告及發表之論文

處理方式：得立即公開查詢

中華民國106年10月30日
Final Report

Massive multiple testing of event-related potentials and brain imaging data: A factor analytic approach to dealing with temporal and spatial dependence
Grant number: 103-2410-H-032-MY3
Project duration: 08/01/2014 - 07/31/2017

Principal investigator: Ching-Fan Sheu
Affiliation: Institute of Education, National Cheng Kung University
E-Mail: csheu@mail.ncku.edu.tw

Introduction

The goal of the project was to extend the adaptive factor analytic (AFA) modeling approach proposed by Causeur et al. [2012] to massive multiple testing of event-related potentials (ERPs) data and functional magnetic resonance imaging (fMRI) data in the multivariate general linear model framework. We augmented the AFA procedure with a novel scheme of boosting detection of rare and weak signal involving some prior knowledge of noise and signal. The new approach was applied to massive multiple testing of correlations between difference ERP curves and a targeted behavior performance using unpublished data collected from directed forgetting experiment similar to [Lee et al., 2013]. Comparing against several leading decorrelational methods [Leek & Storey, 2008; Sun et al., 2012; Allen et al., 2014], the factor analytic procedure possesses markedly improved power for detecting true positives, while controlling the false discovery rate at conventional levels. The advantage of the proposed procedure is a consequence of its ability to take into account the strong time dependence observed in the ERP data. An attempt to adapt the AFA procedure to fMRI multiple testing problem by accounting for spatial and/or spatio-temporal dependence was redirected toward multi-channel ERP context after realizing that
fMRI data show weak temporal resolution with complex spatial dependence. In addition, it appears that investigating the AFA approach to time-frequency analysis of ERPs might bring better insight into how to model efficiently spatio-temporal patterns without being bogged down by the size of fMRI data.

**Objectives**

The project extended the AFA modeling approach developed by Causeur et al. [2012] based on the procedure for genomic data analysis [Friguet et al., 2009] to improve upon univariate mass analysis of ERPs. A new estimation scheme which iterates between the estimation of signals and that for the residual covariance structure is proposed. The former uses the linear model theory to predict the signals, while the latter depends, *a priori*, that noise-alone observations are available; and that the unknown residual covariance structure can be decomposed into a factor-analytic common variance structure plus an independent, residual (unique) covariance structure. This decorrelation approach together with the de-coupling of signal plus noise and noise alone processes gains power over several leading decorrelational approaches for detecting true positives at the same level of global false discovery rate. Sheu et al. [2016] summarized some of the research results.

**Literature review**

Two papers summarize the current status of mass univariate analysis of ERPs [Groppe et al., 2011a,b]. These papers focused on comparing a variety of false discovery rate (FDR) control procedures [Benjamini & Hochberg, 1995] and permutation tests (e.g., [Blair & 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 & 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 & Buchwald, 1991 had proposed a test which considers significant only those runs of p-values lower than a preset threshold, e.g., 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.

The FAMT procedure Friguet et al., 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. For example, the ERP data arising from the auditory oddball and the memory experiments are analyzed using the proposed method. While the usual FDR-controlling procedures are unable to detect any meaningful difference between mean curves in the oddball study, the proposed method identifies significant intervals which can be associated to well-known ERP components. Similarly, in the memory experiment, no meaningful association between ERPs and recognition performance is pointed out by the standard procedures. Interestingly, the proposed method discovers a significant waveform correlation signal around 400ms, which could be explained by the FN400 component reported in the ERP literature on recognition memory Rugg & Curran, 2007.

Methods

A general linear model framework for ERP data analysis

Multivariate Analysis of Variance modeling of ERPs

The following general framework for the significance analysis of ERPs ex-
plicitly 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 1,000 milliseconds (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:

$$Y_{it} = \mu_t + \beta_t' x_i + b_t' z_i + \epsilon_{it}, \quad (1)$$

where $\mu_t$ is the intercept at time $t$, $\beta_t$ and $b_t$ are the $p$- and $r$-vectors of regression coefficients associating the ERP at time $t$ with $x$ and $z$, respectively, and $\epsilon_{it}$ is the random error term, normally distributed with mean 0 and standard deviation $\sigma_t$. Typically, independence is assumed among the errors $\epsilon_{it}$: for each participant $i$, the random vector $\epsilon_i = (\epsilon_{i1}, \epsilon_{i2}, \ldots, \epsilon_{iT})'$ is assumed to be normally distributed with mean 0 and variance $D_\sigma = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_T^2)$, where diag(.) 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 $\epsilon$ in model (1) is relaxed by assuming $\text{Var}(\epsilon) = \Sigma = D_\sigma^{1/2} R D_\sigma^{1/2}$, where $R$ is a $T \times T$ residual correlation matrix.

Model (1) 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 $\beta$, whose rows are the $p$-vectors $\beta_t$, 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 (1):

$$\hat{\beta}_t = (x'P_zx)^{-1}x'P_zY_t,$$

(2)

where $P_z = I_n - Z(Z'Z)^{-1}Z'$, $Z$ is the $n \times (r+1)$ matrix whose $i$th row is $(1, z_i')$, $Y_t = (Y_{1t}, \ldots, Y_{nt})'$ and $x$ is the $n \times p$ matrix whose $i$th 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:

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

(3)

where $\hat{\sigma}^2_t$ is the standard degree-of-freedom corrected estimate of the residual variance in model (1).

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 $\mathcal{T}$. 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 (1): under the family-wise null hypothesis $H_0 = \cap_t H_{0,t}$, the components of $\mathcal{T}$ are indeed F-statistics with the following correlation structure:

$$\text{Cor}(\mathcal{T}_t, \mathcal{T}_{t'}) = \left(1 + \frac{1}{d} \right) \frac{p(d-4)}{p + d - 2} \approx_{d \to +\infty} r^2_{t't'}.$$
where \( r_{tt} \) is the generic term of the matrix \( R \).

**Multiple testing**

The collection of p-values \( (p_t)_{t=1,...,T} \) is generally the only input for multiple testing procedures. Among them, the method proposed by [Guthrie & 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 \( \alpha \). 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 [van der Laan & Dudoit, 2007; Efron, 2010] for a review of the popular procedures and [Groppe et al., 2011a,b; 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 \( \text{FDP} \) is 0 if the number \( R \) of rejections is itself 0 and \( \text{FDP} = V / R \) if \( R > 0 \) [Benjamini & 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 & 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 ??.

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 et al., 2009; Leek & Storey, 2008; Sun et al., 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.

**Time-dependence among test statistics**

First, we propose a flexible factor modeling of the residual correlations in model (1) 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 \( t \) is:

\[
Y_{it} = \mu_t + b_t'z_i + \beta_t'x_i + \lambda_t'f_i + e_{it}, \tag{4}
\]

where \( \lambda_t \) 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 \( \psi^2_t \). Moreover, it is assumed that the specific errors \( e_{it} \) are mutually independent, which induces the following decomposition of the residual covariance matrix \( \Sigma \):

\[
\Sigma = \Lambda \Lambda' + \Psi, \tag{5}
\]

where \( \Lambda \) is the \( T \times q \) matrix whose \( t \)-th row is \( \lambda'_t \) and \( \Psi \) is the diagonal matrix whose \( t \)-th diagonal element is \( \psi^2_t \). In other words, latent factors are intro-
duced to capture linearly the time dependence among residuals of model (1) [Causeur et al., 2012].

To illustrate the ability of model (4) to fit the complex dependence pattern observed in Figure ??, models with 1, 5 and 10 factors, respectively, are estimated for the residual correlations of model (??) at channel CZ using the EM algorithm described in [Friguet et al., 2009]. The results are compared in Figure 1, 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.

**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 $T_0$, the process of estimation errors outside of $T_0$ is upgraded by making use of its correlation with the counterpart in $T_0$. 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 $t$–th row is $\delta_t = (x'P_zx)^{-1}x'P_z\epsilon_t$, and $\epsilon_t = (\epsilon_1,t, \ldots, \epsilon_{nT})'$ is the $n$–vector of residual errors in model (1). Let $\text{vec}(\cdot)$ be the matrix operator transforming a matrix into a vector by concatenating its rows. The $pT$–vector, $\text{vec}(\Delta)$, is distributed according to a normal distribution with mean 0 and covariance $V_\delta = \Sigma \otimes (x'P_zx)^{-1}$, 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 signal is unlikely to be present, i.e., $\beta_t = 0$ for $t$ belonging
Figure 1: Image plots of the fitted correlation matrix of the residuals of model (??) 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 ??.

Thus, the estimation error $\delta_t$, for $t \in T_0$, is not confounded with the true signal $\beta_t$: $\Delta_0 = \hat{\beta}_0$, where $\Delta_0$ (resp. $\hat{\beta}_0$) is the submatrix of $\Delta$ (resp. $\hat{\beta}$) restricted to $t \in T_0$. This allows us to partition $\Delta$ into two submatrices:

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

(6)

where $\Delta_{-0}$ is the submatrix of $\Delta$ with rows $\delta_t$, $t \notin T_0$, and rearrange $V_0$ corre-
spondingly:

\[ \hat{\varphi}_t = \left( \begin{array}{cc} \Sigma_{0,0} & \Sigma'_{0,0} \\ 0,0 & \Sigma_{0,0} \end{array} \right) \otimes (x'P_z x)^{-1}, \]

where \( \Sigma_{0,0} \) (resp. \( \Sigma_{0,0}^{-1} \)) is the submatrix of \( \Sigma \) restricted to rows and columns corresponding to time points \( t \in T_0 \) (resp. \( t \notin T_0 \)) and \( \Sigma_{0,0}^{-1} \) is the submatrix of \( \Sigma \) restricted to rows corresponding to \( t \notin T_0 \) and columns corresponding to \( t \in T_0 \).

For each \( t \notin T_0 \), we predict \( \delta_t \) from \( \Delta_0 \) by its best linear predictor:

\[
\begin{align*}
\text{vec}(\hat{\Delta}_{-0}) &= [\hat{\Sigma}_{-0,0} \otimes (x'P_z x)^{-1}] [\hat{\Sigma}_{0,0}^{-1} \otimes (x'P_z x)^{-1}]^{-1} \text{vec}(\Delta_0), \\
&= [\hat{\Sigma}_{-0,0} \otimes (x'P_z x)^{-1}] [\hat{\Sigma}_{0,0}^{-1} \otimes (x'P_z x)] \text{vec}(\Delta_0), \\
&= [\hat{\Sigma}_{-0,0} \hat{\Sigma}_{0,0}^{-1}] \otimes I_p \text{vec}(\Delta_0),
\end{align*}
\]

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:

\[ \hat{\Delta}_{-0} = \hat{\Sigma}_{-0,0} \hat{\Sigma}_{0,0}^{-1} \Delta_0. \] (7)

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 (4) to estimate the whole matrix \( \Sigma \) by \( \hat{\Sigma} = \hat{\Lambda} \hat{\Lambda}' + \hat{\Psi} \), where \( \hat{\Psi} \) is the diagonal matrix of estimated specific variances \( \hat{\Psi}_i^2 \) and \( \hat{\Lambda} \) is the \( T \times q \) matrix of estimated loadings \((q \ll T)\). The partition (6) of \( \Delta \) results in corresponding partitions of \( \Psi \) and \( \Lambda \):

\[
\begin{align*}
\hat{\Lambda} &= \begin{pmatrix} \Lambda_0 \\ \Lambda_{-0} \end{pmatrix}, \\
\hat{\Psi} &= \begin{pmatrix} \Psi_0 & 0 \\ 0 & \Psi_{-0} \end{pmatrix}.
\end{align*}
\]

Estimators of \( \Sigma_{0,0} \) and \( \Sigma_{-0,0} \) are derived as:

\[
\hat{\Sigma}_{-0,0} = \hat{\Lambda}_{-0} \hat{\Lambda}_{0}', \quad \hat{\Sigma}_{0,0} = \hat{\Psi}_0 + \hat{\Lambda}_0 \hat{\Lambda}_0'.
\]
Note that computing $\hat{\Sigma}^{-1}_{0,0}$ in expression (7) involves only the inversion of a $q \times q$ matrix according to Woodbury’s identity [Press et al., 2007]:

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

An estimate $\hat{\Delta}^{(1)}$ for $\Delta$ can be obtained by substituting $\hat{\Delta}_{-0}$, given by expression (7), for $\Delta_{-0}$ in (6). A new estimate for $\beta$ 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 $\hat{\epsilon}$:

$$\hat{\epsilon}^{(1)} = P_{x}(Y - x \hat{\beta}^{(1)'})$$

A new factor decomposition of the covariance of the updated residual errors is again derived, producing a new estimate for $\Delta$ 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 $\beta$.

**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 et al., 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 & Thayer, 1982], presented in detail in [Friguet et al., 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 $\hat{\Psi}^{2}_{r}$ 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 et al., 2009] derive a closed form expression for the variance inflation \( \gamma_k \) of the \( k \)-factor model for this dependence. These authors assess the number of factors by estimating the variance \( \gamma_k \) of the number of false positives when the tests are calculated with the \( k \)-factor adjusted residuals: \( \hat{\epsilon} - \hat{F}_k \hat{\Lambda}_k \) for each \( k \)-factor model \( (\Lambda_k, \Psi_k) \). Finally, the retained number of factors is \( \hat{k} = \text{arg min}_k \gamma_k \). In contrast, the number of factors is determined via parallel analysis [Buja & Eyuboglu, 1992] in surrogate variable analysis [Leek & Storey, 2008] and latent effect adjustment after primary projection [Sun et al., 2012].

Once a factor regression model (4) [Leek & Storey, 2008; Friguet et al., 2009; Causeur et al., 2012; Sun et al., 2012] is fitted to a set of dependent data for multiple testing, the new test statistics \( \tilde{T} \) (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 \( k \)th 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 in 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: \( \hat{\epsilon}^{(k+1)} = \hat{P}_x(y - x \hat{\beta}^{(k+1)'}) \);

- Estimate \( (\hat{\Psi}_{k+1}, \hat{\Lambda}_{k+1}; \hat{F}_{k+1}) \) of the factor model based on \( \hat{\epsilon}^{(k+1)} \). 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 \( \mathcal{F}_0 \), but also with a higher chance of including the support of the signal. We suggest adopting the following rule: \( \mathcal{F}_0^{k+1} = \{ t = 1, \ldots, T, \tilde{p}_t^{(k+1)} \geq 0.2 \} \), where \( \tilde{p}_t^{(k+1)} \) is the current factor-adjusted p-value at time \( t \).

The iteration terminates at step \( k \) such that \( \mathcal{F}_0^{k+1} = \mathcal{F}_0^k \).

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 \( \mathcal{F}_0 \) 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, i.e., 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 & 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.

**Analysis of ERP data from 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 et al., 1994]). Empirical studies of recognition memory using ERPs have indicated 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 400-600 ms after stimulus onset. The finding that the FN400 component increases gradually with recognition confidence [Rugg & 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 $T_0 = [1, 200] \cup [901, 1, 000]$ ms, and, for the posterior locations, $T_0 = [1, 200]$ ms. After examining the variance inflation criterion for each channel, the number of factors was set to 2. The top and bottom panels of Figure 2 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 3 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 400ms, clearer in the TBF condition, implicates the relationship between instruction and the modulation of the FN400 component.
Estimated correlation with recognition performance using OLS at channel CZ

Estimated correlation with recognition performance using AFA at channel CZ

Figure 2: 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.

Results and Discussions

Mass univariate analysis of event-related brain potentials [Groppe et al., 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 [??] 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].

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 & Karniski, 1993; Westfall & 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 & 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 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).

References


