Dealing with long-time range dependence in large-scale multiple testing of Event-Related Potentials data

Émeline Perthame [1] & Ching-Fan Sheu [2] & Yuh-Shiow Lee [3] & David Causeur [4]

[1] *Agrocampus Ouest, IRMAR UMR 6625 CNRS, 65 rue de St-Brieuc, CS 84215, 35042 Rennes cedex, France - emeline.perthame@agrocampus-ouest.fr*
[2] *National Cheng-Kung University, Intitute of Education, Social Science Building, North B419, 1 University Road, Tainan 701, Taiwan - csheu@mail.ncku.edu.tw*
[3] *National Chung-Cheng University, Department of Psychology, Chiayi 621, Taiwan - psyysl@ccu.edu.tw*
[4] *Agrocampus Ouest, IRMAR UMR 6625 CNRS, 65 rue de St-Brieuc, CS 84215, 35042 Rennes cedex, France - david.causeur@agrocampus-ouest.fr*

Résumé.

Les potentiels évoqués cognitifs (ERPs) sont utilisés dans le domaine de la recherche en psychologie pour décrire par électro-encéphalographie (EEG) l'évolution dans le temps de l'activité cérébrale induite par des évocations. Sur les courts intervalles de temps au cours desquels les variations d'ERPs peuvent être liées à des conditions expérimentales, le signal psychologique est le plus souvent faible, au regard de la forte variablité inter-individuelle des courbes d'EEG.

Groppe *et al.* (2011a, 2011b) proposent une revue des procédures de tests simultanés pour les données haut-débit de potentiels. Toutefois, ils se limitent à la comparaison de méthodes classiques mais ne mentionnent pas les problèmes liés à la forte dépendance entre les statistiques de tests au cours du temps. Pourtant, il ressort d'articles récents sur les tests multiples pour données à haut-débit (voir par exemple Efron, 2007) qu'une forte corrélation entre les tests réduit considérablement la précision et la stabilité des procédures.

Cette dépendance est ici fortement structurée dans le temps, avec une composante autorégressive et une structure en blocs. Nous proposons une méthode basée sur la modélisation jointe du signal et de la dépendance entre les statistiques de tests au cours du temps, dans le but d'améliorer les procédures de tests multiples telles que celles de Benjamini et Hochberg (1995), de Guthrie et Buchwald (1991), conçue pour l'analyse de données d'ERP, et les plus récentes approches de décorrélation proposées par Leek et Storey (SVA, voir Leek and Storey, 2008) et Sun *et al.* (2012).

Mots-clés. Dépendance, Données d'ERP, Grande dimension, Tests multiples.

Abstract.

Event-related potentials (ERPs) are now widely collected in psychological research to determine the time courses of mental events. In the significant analysis of the relationships between event-related potentials and experimental covariates, the psychological signal is often both rare, since it only occurs on short intervals and weak, regarding the huge between-subject variability of ERP curves. Testing simultaneously for differences over the entire digitized time intervals creates a serious multiple comparison problem in which the probability of false positive errors must be controlled, while maintaining reasonable power for correct detection.

A pair of papers published recently summarized current status on mass univariate analysis of event-related brain potentials/fields (Groppe *et al.*, 2011a, 2011b). These authors focused on comparing a variety of the False Discovery Rate (FDR) controlling procedures. Missing conspicuously from the two articles is any reference to the problem of dependent tests generated by the strong temporal dependence in the ERPs. It is yet now well known from the literature on large-scale significance analysis (Efron, 2007) that highly correlated data can severely affect the accuracy of simultaneous testing.

In the present situation, dependence is obviously structured over time with both a strong autocorrelation component and a block pattern. We propose a joint modeling of the signal and time-dependence among test statistics to improve the properties of multiple testing procedures, regarding standard methods such as the Benjamini and Hochberg (1995) false discovery rate procedure, the Benjamini and Yekutieli (2001) procedure designed for dependent test statistics, the Guthrie and Buchwald (1991) method for ERP analysis and the more recent decorrelation approaches by Leek and Storey (SVA, see Leek and Storey, 2008) and Sun *et al.* (2012).

Keywords. Dependence, ERP data, High-dimensional data, Multiple testing.

1 Introduction

High-throughput instrumental data such as Event-Related Potentials (ERPs) and functional magnetic resonance imaging (fMRI) data have increasingly become common in psychological research. 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 amount of data from ERP or fMRI studies, researchers must face the challenge of multiple comparison corrections: in shifting, simultaneously, through thousands or tens of 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 data exhibiting arbitrarily strong temporal dependence is the focus of this presentation.

A pair of papers published recently summarized current status on mass univariate analysis of event-related brain potentials/fields (Groppe *et al.*, 2011a, 2011b). These authors focused on comparing a variety of the False Discovery Rate (FDR) control procedures (Benjamini and Hochberg, 1995) and permutation tests (e.g., Blair and Karinski, 1993). Instead of controlling for family-wise error rate, the FDR procedure controls for expected proportion of incorrectly rejected null hypotheses (false discoveries) over the total number of rejections made. Permutation tests are nonparametric and within-subject data structure is preserved. Missing conspicuously from the two articles is any reference to the problem of dependent tests generated by the strong temporal dependence in the ERPs. It is yet now well known from the literature on large-scale significance analysis (Efron, 2007) that highly correlated data can severely affect the accuracy of FDR estimation and the stability of simultaneous testing (i.e., variances of discovery proportions). Ignoring dependence among test statistics also reduces the detection of true positives (Leek and Storey, 2008).

One approach is to model the dependence structure in the data by a hidden Markov model (see Sun and Cai, 2009). Another more general approach is to account for the multivariate dependence by some data reduction techniques involving latent variables (see Leek and Storey, 2008 or more recently Sun *et al.*, 2012). A notable example of the latter approach in genomic data analysis is the powerful factor analysis multiple testing procedure proposed by Friguet *et al.* (2009) under the assumption that the conditional covariance of the responses given the treatment variables can be well approximated by its factor components. The former method is adapted to a dynamic factor adjusted modeling of ERPs arising from standard analysis of variance designs in Causeur *et al.* (2012). Results of a simulation study showed that the new procedure performed well against those standard multiple testing procedures: both in power gains and in stabilizing true discoveries. We propose to extend the former method by a joint modelling of the signal and the time-dependence structure among test statistics.

2 Settings for significance analysis of ERP data

For $i = 1, \ldots, n$, let Y_{it} denote the ith ERP measurement at time t, with $t = 1, \ldots, T$, where T is the number of frames. The following multivariate linear model is assumed as a general framework to study the relationship between the ERPs and covariates $x_i = (x_{i1}, \ldots, x_{ip})'$, eventually adjusted from the effect of other covariates $z_i = (z_{i1}, \ldots, z_{ir})$:

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

where μ_t is the intercept coefficient at time t, b_t and β_t are the $r-$ and $p-$vectors of slope coefficients relating the ERP at time t with z and x respectively and ε_{it} is the random error term, normally distributed with mean 0 and standard deviation σ_t. Typically, in basic multiple testing procedures, no time dependence is assumed among the residual errors ε_{it}: for each subject i, the vector $\varepsilon_i = (\varepsilon_{i,1}, \varepsilon_{i,2}, \ldots, \varepsilon_{i,T})'$, where T is the number of time frames, is then assumed to be normally and independently distributed with mean

0 and variance $D_\sigma = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_T^2)$, where diag(.) stands for the matrix operator which transforms a T-vector into the $T \times T$ diagonal matrix which diagonal elements are given by this vector. Hereafter, this assumption is relaxed to account for time-dependence: $\mathrm{Var}(\varepsilon) = \Sigma = D_\sigma^{1/2} R D_\sigma^{1/2}$, where R is a $T \times T$ residual correlation matrix.

For ERP data, the signal at each channel is usually both rare and weak: rare because for most t, the null hypothesis $H_{0,t}: \ \beta_t = 0$ is true, 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 often not in favor for successful detection of time points for which $H_{0,t}$ is not true. According to the general linear model theory, the selection of significant time points is based on the observed signal $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_T)'$ obtained by ordinary least-squares estimation of model (1). For $p = 1$,

$$\hat{\beta} \quad = \quad \frac{x' P_z Y}{x' P_z x},$$

where $P_z = I_n - Z(Z'Z)^{-1}Z'$, Z is the $n \times (r+1)$ matrix which ith row is $(1, z_i')$, Y is the $n \times T$ matrix which generic (i,t) element is Y_{it} and $x = (x_1, \ldots, x_n)$ is the $n-$vector collecting the measurements of the covariate x. The corresponding vector $\mathcal{T}$ of $t-$statistics for the set of null hypotheses $H_{0,t}$ is given by the following expression:

$$\mathcal{T} \quad = \quad \sqrt{x' P_z x}\, \mathrm{diag}(\hat{\sigma})^{-1} \hat{\beta}, \tag{2}$$

where $\mathrm{diag}(\hat{\sigma})$ is the $T \times T$ diagonal matrix which t-th diagonal element is the standard degree-of-freedom corrected estimate $\hat{\sigma}_t$ of model (1). Under the null hypothesis $H_{0,t}$, each component $\mathcal{T}_t$ of $\mathcal{T}$ is distributed according to a Student distribution with $d = n - p - r - 1$ degrees of freedom. In the following, p_t stands for the p-value associated to $\mathcal{T}_t$.

FDR-controlling Multiple testing procedures

The collection of p-values $(p_t)_{t=1,\ldots,T}$ is generally the only input for multiple testing procedures. Indeed, most methods consist in rejecting the null H_{0t} if $p_t \leq p^*$, where the threshold p^* is chosen in order to guarantee that the corresponding number V of erroneous rejections of the null is controlled.

Basically, these methods can first be divided into two families according to the overall type-I error rate they wish to control: the so-called Family-Wise Error Rate (FWER) defined as $\mathrm{FWER} = \mathbb{P}(V \geq 1)$, and the False Discovery Rate (FDR), defined as the expected proportion of erroneous rejections of the null among the positive tests: $\mathrm{FDR} = \mathbb{E}(\mathrm{FDP})$, where the False Discovery Proportion FDP is 0 if the number R of rejections is itself 0 and $\mathrm{FDP} = V/R$ if $R > 0$. The renewal of large-scale multiple testing induced by this change of objective dates from Benjamini and Hochberg (1995), which introduces the so-called Benjamini-Hochberg (BH) procedure: p_* is here defined as the largest $p_{(k)}$, where $p_{(k)}$ is the kth increasingly ordered p-value, such that $p_{(k)} \leq k\alpha/T$. Under an assumption of independence among tests, Benjamini and Hochberg (1995) show that the former thresholding method guarantees that FDR $\leq \pi_0\alpha \leq \alpha$, where π_0 is the unknown proportion of true nulls. Among the many refinements of the seminal BH procedure, some have focused on the control of the FDR by the BH procedure in situations where tests are correlated (Benjamini and Yekutieli, 2001).

More recent papers investigate the negative impact of dependence on the accuracy of multiple testing procedures, especially due to the instability of ranking. It shall be noted that the former approaches do not consist in improvements of the thresholding procedure but in modifications of the calculation of p-values. For example, in the context of genomic data analysis, Leek and Storey (2008), Friguet *et al.* (2009) and Sun *et al.* (2012) propose to model the dependence among tests using a latent factor model, which is used to decorrelate the test statistics and consequently restore the consistency of p-values ranking.

3 Time-dependence among test statistics

It is first important to note that, in the present multivariate linear context, the dependence pattern of the t-statistics is directly inherited from the residual correlation R introduced in model (1): under the family-wise null hypothesis $H_0 = \cap_t H_{0,t}$, $\mathcal{T}$ is indeed distributed according to a multivariate Student distribution with correlation R and d degrees of freedom.

The method proposed by Guthrie and Buchwald (1991), hereafter referred to as the GB procedure, is noticeably the first one addressing this dependence issue by assuming an autoregression correlation structure of order 1 for the t-tests. In order to better account for the complexity of the dependence pattern, we propose to model the residual correlation of model (1) using a more flexible factor model.

It is now assumed that there exists q latent factors $f = (f_1, \ldots, f_q)'$, normally distributed with mean 0 and variance I_q, such that, conditionally on z_i, x_i and f_i,

$$Y_{it} = \mu_t + b_t' z_i + \beta_t' x_i + \lambda_t' f_i + e_{it}, \tag{3}$$

where λ_t is the $q-$vector of factor loadings for the ERP measurement at time 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 variance matrix Σ:

$$\Sigma = \Psi + \Lambda\Lambda', \tag{4}$$

where $\Psi = \text{diag}(\psi_1^2, \ldots, \psi_T^2)$ and Λ is the $T \times q$ matrix which $t-$th row is λ_t'.

In other words, the factors are introduced in the model in order to capture linearly the time-dependence among residuals of model (1). The same model with different estimating strategies can be found in Leek and Storey (2008), Friguet *et al.* (2009) and Sun *et al.* (2012) for multiple testing issues in high-dimension.

The talk will present an iterative estimation procedure which alternates the estimation of the factor model parameters and the signal. We will show that it improves the stability of error rates and the overall true discovery proportions.

Bibliography

[1] Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, Ser. B, **57**, 289–300.

[2] Benjamini, Y., Yekutieli, D. (2001). *The control of the false discovery rate in multiple testing under dependency.* Annals of Statistics **29**, 1165–1188.

[3] Blair, R., Karinski, W. (1993). An alternative method for significance testing of waveform difference potentials. *Psychophysiology*, **30**, 518–524.

[4] Causeur, D., Chu, M.C., Hsieh, S., Sheu, C.F. 2012. A factor-adjusted multiple testing procedure for ERP data analysis. *Behavior Research Methods*, **44**, 635–643.

[5] Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association*, **102**(477), 93–103.

[6] 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.

[7] 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**, 17111725

[8] Groppe, D. M., Urbach, T. P., Kutas, M. (2011b). Mass univariate analysis of event-related brain potentials/fields: II. Simulation studies. *Psychophysiology*, **48**, 17261737.

[9] Guthrie, D., Buchwald, J.S. (1991). Significance testing of difference potentials. *Psychophysiology*, **28**, 240–244.

[10] Leek, J.T. and Storey, J. (2008). A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences*,**vol. 105 no. 48**, 18718–18723.

[11] Sun, W., Cai, T.T. (2009). Large-scale multiple testing under dependence. *Journal of the Royal Statistical Society, B.*, **71**(2), 1–32.

[12] Sun, Y., Zhang, N.R., Owen, A.B. (2012). *Multiple hypothesis testing adjusted for latent variables, with an application to the AGEMAP gene expression data.* The Annals of Applied Statistics **6**, no. 4, 1664–1688.