Signal identification in ERP data by decorrelated Higher Criticism Thresholding

Emeline Perthame, Ching-Fan Sheu, David Causeur

Signal identification in ERP data by decorrelated Higher Criticism Thresholding

Emeline Perthame, Ching-Fan Sheu, David Causeur

May 3, 2016

Abstract: Event-related potentials (ERPs) are intensive recordings of electrical activity along the scalp time-locked to motor, sensory, or cognitive events. A main objective in ERP studies is to select (rare) time points at which (weak) ERP amplitudes (features) are significantly associated with experimental variable of interest. The Higher Criticism Thresholding (HCT), as an optimal signal detection procedure in the "rare-and-weak" paradigm, appears to be ideally suited for identifying ERP features. However, ERPs exhibit complex temporal dependence patterns violating the assumption under which signal identification can be achieved efficiently for HCT. This article first highlights this impact of dependence in terms of instability of signal estimation by HCT. A factor modeling for the covariance in HCT is then introduced to decorrelate test statistics and to restore stability in estimation. The detection boundary under factor-analytic dependence is derived and the phase diagram is correspondingly extended. Using simulations and a real data analysis example, the proposed method is shown to estimate more efficiently the support of signals compared with standard HCT and other HCT approaches based on a shrinkage estimation of the covariance matrix.

1 Introduction

Event-related potentials (ERP) are voltage changes along the scalp time-locked to some physical or mental occurrence in the ongoing electrical brain activity recorded as electroencephalogram (EEG). Like functional magnetic resonance imaging (fMRI), ERPs are noninvasive instruments that directly reflect cortical neuronal activity. Unlike fMRI, ERPs studies provide better temporal resolution to chart the time course of mental processes and are less expensive to conduct. In basic research, ERPs offer a psychophysiological method for studying attentional processes, language, and memory functions, yielding information not available from behavioral studies alone. In clinical research, ERPs are one of several noninvasive biomarkers that have been proposed for evaluating neurological and psychiatric disorders such as Alzheimer's disease, amnestic mild cognitive impairment, attention deficit, hyperactivity disorder, among others.

For a typical trial in a study, ERP amplitudes are measured in milliseconds (ms) for up to one or more seconds with reference to the onset of an external event. ERP waveforms are notoriously noisy and highly variable, both within and between subjects, which explains why ERPs are usually averaged across trials of the same condition for the same subjects. To identify time points at which ERP amplitudes can reliably be linked to either stimulus (or response) events, researchers must shift, simultaneously, through thousands of features for significance testing. A balance must be struck between keeping a low false positive error rate while maintaining sufficient power for correct signal identification. How to achieve this objective for ERP data exhibiting arbitrarily strong temporal dependence is the focus of the present paper.

Searching for time points at which ERP amplitudes are significantly associated with a response variable can be seen as a signal identification issue in the "Rare and Weak" (RW) paradigm introduced by [8]. Large-scale significance analysis of ERP waveforms is indeed based on a T-vector $\mathcal{T} = (\mathcal{T}_1, \ldots, \mathcal{T}_T)$ of test statistics for the collection of corresponding null hypotheses $H_{0,t}$ of no association between the ERP measured at time t and the target variable. The RW paradigm provides a simple yet useful framework to investigate test procedures for detecting the presence of a nonzero signal. In the multiple testing setting, [8] proposed Higher Criticism Thresholding (HCT) inspired by an idea of [27] based on the significance of an overall body of tests. HCT is known to be effective to detect signals under independence. Indeed, closed-form detection bounds can be derived analytically and [8] demonstrated that HCT attains the theoretically optimal decision limits. For the more challenging objective of selecting non-null features for classification or prediction, [9] also demonstrated the superior performance of HCT with respect to multiple testing procedures controlling False Discovery Rate (FDR).

As reported in [6], the pronounced dependence observed in ERP waveforms can induce a long-range regularity for the test statistics resulting in spuriously small p-values outside of the support of the signal and causing the non-null features to be misidentified. This instability in the ranking of p-values due to dependence has also been reported in high-dimensional genomic data analysis (see for example [11], [1], [20]). Although HCT is known to be effective even when tests are weakly correlated (see [13]), its performance can still be improved when dependence is accounted for ([1] and [14]). For example, [14] reported that the theoretical detection bounds derived in the RW framework are markedly affected by a strong dependence among the test statistics. Therefore, [14] extended the RW framework by introducing the so-called innovated HCT (iHCT) and showed that iHCT restores the effectiveness of the HCT procedure under strong dependence.

Elsewhere, for feature selection in the Linear Discriminant Analysis (LDA) context, [1] applied HCT to Correlated Adjusted T-scores (CAT-scores), which are tests statistics decorrelated by a James-Stein shrinkage estimator of the covariance matrix (see [29]). [1] showed that performance of HCT is improved by this decorrelation using an inverse-square root of the covariance matrix. Alternatively, [20] extended the decorrelation method introduced in [11] based on a regression

factor model, to generalized linear modeling for a variable selection issue in supervised classification. [20] devised an adaptive algorithm to jointly estimate signal, covariance matrix and the class probabilities in order to efficiently decorrelate data for classification.

As demonstrated by [6], the complex dependence pattern observed in the correlation structure of test statistics derived from ERP data can be well approximated using a factor decomposition, often with a moderate number of factors. The procedure proposed in this paper is expected to better select ERP features which are associated with treatment variable of interest. Moreover, it also takes advantage of simple and efficient algebraic tools to derive an inverse-square root of the covariance matrix. This article is organized as follows. Section 2 describes the oddball auditory ERP paradigm to motivate the subsequent modeling and data analysis issues. It also introduces signal detection methods in a multivariate linear settings for ERP experiments. A brief review of actual methods dealing with HCT under independence and dependence is presented in Section 3. The factor model framework is introduced in Section 4, leading to an extension of detection boundary under dependence and a Factor-innovated HCT method. The properties of the method are demonstrated through simulations and a real data study in Section 5. We conclude with a discussion in Section 6.

2 Signal detection in ERP experiments

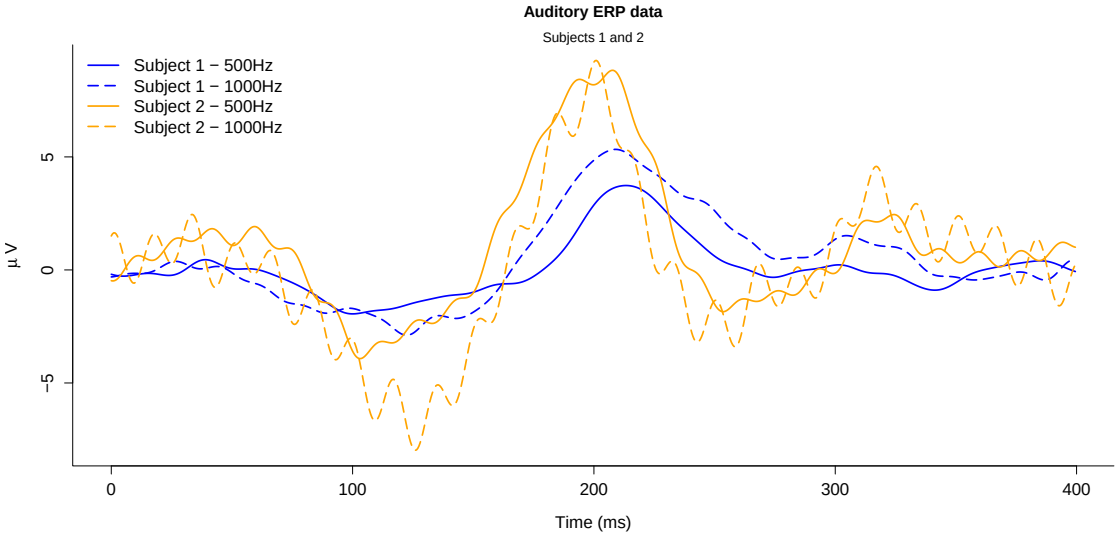
We first consider hereafter the statistical testing for the mean comparison of I groups of ERP curves, with n_i curves in group i. The $n = n_1 + \ldots + n_I$ ERP curves are observed at timepoints in $\{t_1, \ldots, t_T\}$ on J subjects. The former signal detection issue is first motivated by an auditory oddball ERP study.

2.1 The auditory oddball experiment

In ERP studies, perhaps the most commonly used experimental task is the oddball paradigm ([21]). In this paradigm, typically two classes of stimuli are presented, one occurring frequently (standard) and the other occurring infrequently (target). The subject is required to distinguish between the two stimuli and to respond to the stimuli that are designated as targets.

An auditory ERP study was performed at Kaohshung Medical University in Taiwan to provide an illustrative data set for the present investigation. The task uses two pure tones of 500 Hz and 1,000 Hz. The former is presented 120 out of 150 trials, whereas the latter (target) is presented only for 30 trials. The order of tone presentation is random and the subject is asked to (silently) count the number of targets. At each of 4 electrode locations (FZ, C3, C4, & O1), ERP waveform was obtained from each of the two tone conditions for each of the $J = 15$ participants, each curve

Figure 1: ERP curves for subjects 1 (blue lines) and 2 (orange lines) of the auditory oddball experiment in conditions Hz500 (plain) and Hz1000 (dashed)

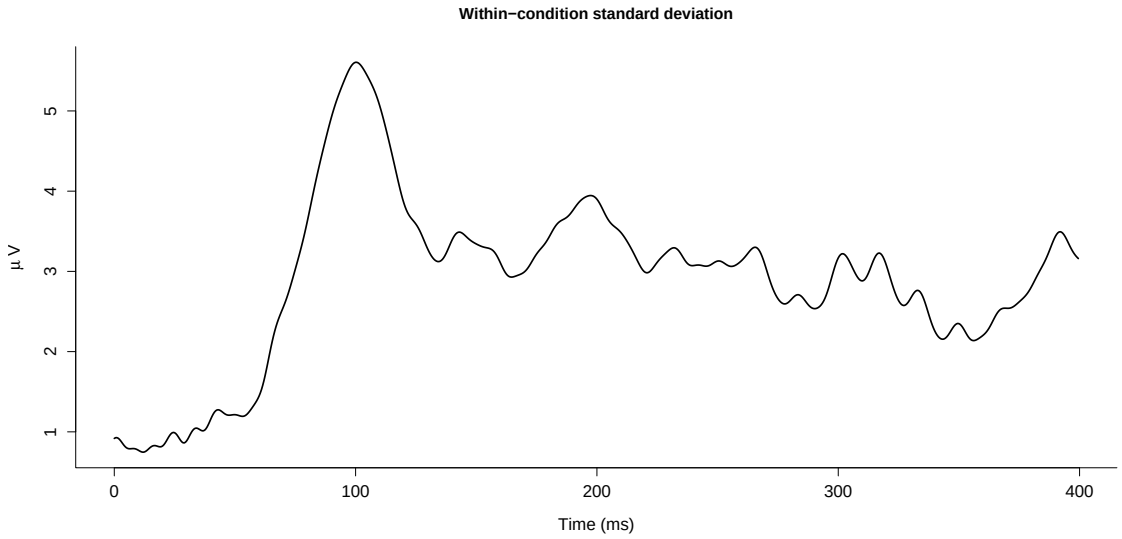


begins at -100 milliseconds (ms) and terminates at 399.5 milliseconds (ms) with two records per 1 ms. The stimulus onset is at 0 ms. For subsequent analysis, only the ERPs from the electrode location FZ will be used.

[28] and many other studies have demonstrated that an ERP waveform across the parieto-central area of the skull is usually observed around 300 ms (the so-called P300 component) and is larger after the target event. The question to be addressed is whether it is possible to select time points at which ERP features can reliably detect which one of the two tones was presented to the subject and whether these time points are indeed around 300 ms as expected. This verification is of fundamental importance if the P300 component is to be considered as an electrophysiological marker for further assessment of psychiatric and neurological disorders.

It is conjectured that brain activations would be different over different time points depending on whether participants listen tones of 500Hz or 1000Hz. The data consist of $T = 799$ time points measured for $J = 15$ subjects and $I = 2$ conditions. Figure 1 shows ERP curves for subjects 1 (blue) and 2 (orange) for condition Hz500 (plain) and Hz1000 (dashed) as an example. This figure illustrates the large variability among subjects.

Figure 2: Within-condition standard errors in the auditory ERP dataset



2.2 Multivariate linear settings

One-way analysis of variance Let Y_{ijt} denote the ERP at time t, t in $\{t_1, \ldots, t_T\}$, in condition i, with i in $[1; I]$, for subject j, with j in $[1; n_i]$. The total number of ERP curves is $n = n_1 + \ldots + n_I$. The following multivariate one-way analysis of variance model is first assumed:

$$Y_{ijt} \quad = \quad x'_{0ij}\mu_t + a_{it} + \varepsilon_{ijt}, \tag{1}$$

where x_{0ij} is a r-profile of baseline covariates for subject j, which does not depend on condition i, and ε_{ijt} are random errors, normally distributed with mean 0, standard deviations $\sigma = (\sigma_{t_1}, \ldots, \sigma_{t_T})'$ and correlation matrix R. In the above auditory oddball experiment, $x_{0ij} = (1, \delta_j)$ where δ_j is just the $0-1$ variable which takes the value 1 if the subject is j. As shown in Figure 2, the within-condition standard errors $s = (s_{t_1}, \ldots, s_{t_T})'$ vary quite a lot along time, where s_t^2 is the degree-of-freedom corrected mean squared error at time t.

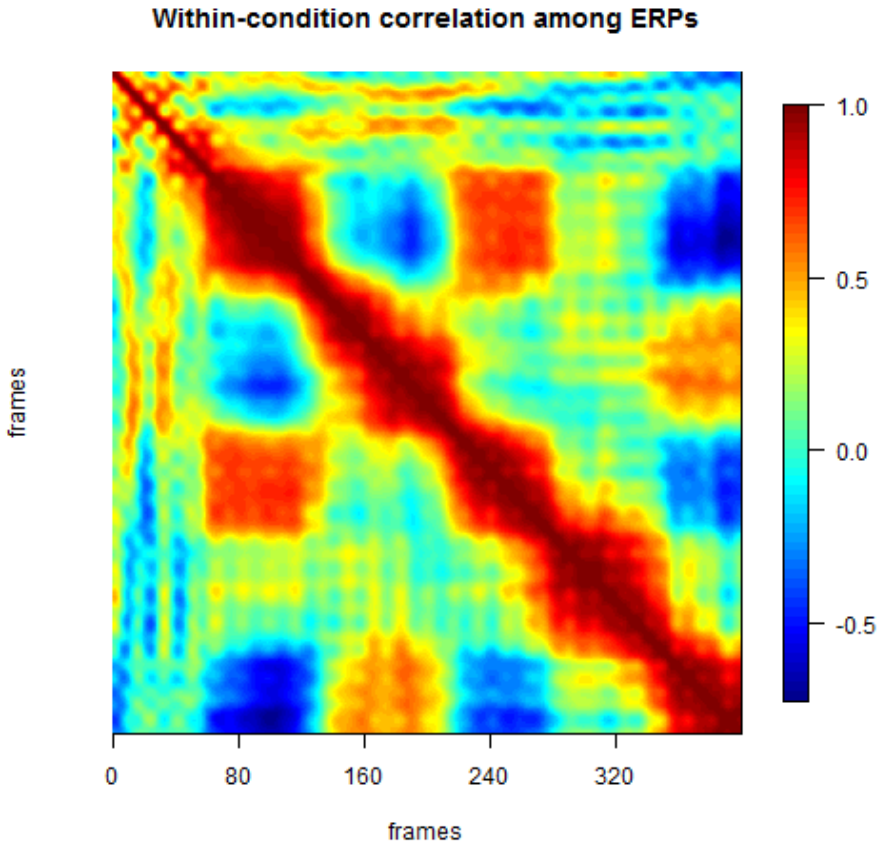
Moreover, the image plot of the within-condition correlation matrix among ERPs in Figure 3 shows that it is both structured by a strong-autocorrelation component and a block pattern.

Model (1) is similarly expressed as follows in matrix notations:

$$Y_t \quad = \quad X_0\mu_t + Xa_t + \varepsilon_t,$$

where $Y_t = (Y_{11t}, \ldots, Y_{1,n_1,t}, Y_{21t}, \ldots, Y_{2,n_2,t}, \ldots, Y_{I1t}, \ldots, Y_{I,n_I,t})'$, X_0 is the $n \times r$ matrix whose rows are x'_{0ij}, $a_t = (a_{2t}, \ldots, a_{It})'$ is the $(I-1)-$vector of condition effect parameters, X

Figure 3: Within-condition correlations in the auditory ERP dataset



is a $n \times (I-1)$ matrix whose kth column X_k is 0, except the entries between indices $n_1+\ldots+n_k+1$ and $n_1+\ldots+n_{k+1}$ which are equal to 1, and $\varepsilon_t = (\varepsilon_{11t}, \ldots, \varepsilon_{1,n_1,t}, \varepsilon_{21t}, \ldots, \varepsilon_{2,n_2,t}, \ldots, \varepsilon_{I1t}, \ldots, \varepsilon_{I,n_I,t})'$.

Under the within-subject independence assumption introduced above, for all heteroscedasticity and correlation pattern (σ, R), the generalized least-squares (GLS) estimate of a_t coincides with the maximum likelihood (ML) estimator under normality:

$$\hat{a}_t \quad = \quad S_{xx}^{-1} S_{xy_t}, \tag{2}$$

where S_{xx} is the $(I-1) \times (I-1)$ empirical variance matrix derived from the design matrix of model (1):

$$S_{xx} \quad = \quad \frac{1}{n} X' P_0 X,$$

where $P_0 = \mathbb{I}_n - X_0 (X_0' X_0)^{-1} X_0'$. Similarly, S_{xy_t} is the $(I-1)-$ empirical covariance vector:

$$S_{xy_t} \quad = \quad \frac{1}{n} X' P_0 Y_t.$$

Finally, the residual variance σ_t^2 is estimated by the residual degree of freedom-corrected mean square of residuals errors:

$$s_t^2 \quad = \quad \frac{Y_t'(P_0 - P)Y_t}{n - (r + I - 1)}, \tag{3}$$

where P is the $n \times n$ orthogonal projection matrix:

$$P = X P_0 \Big[X' P_0 X \Big]^{-1} P_0 X'.$$

Signal detection Signal detection can be expressed as the statistical test of the whole-signal null hypothesis

$$\mathrm{H}_0 \quad : \quad \text{for all } t,\ a_t = 0.$$

The type-I error rate α for the former test is also the probability of declaring erroneously that, for at least one time t, $a_t \neq 0$. Controlling the type-I error rate for the global test on the whole curves can therefore be viewed as equivalent to controlling the Family-Wise Error Rate (FWER), namely the probability of at least one false rejection, in the simultaneous testing of the collection of null hypothesis

$$\mathrm{H}_{0t} \quad : \quad a_t = 0.$$

For each of the former tests, the following F-tests F_t are appropriate:

$$F_t \quad = \quad \frac{n}{I-1}\frac{\hat{a}_t' S_{xx} \hat{a}_t}{s_t^2}.$$

Let $p_t = 1 - G_{I-1,n-(r+I-1)}(F_t)$, where $G_{I-1,n-(r+I-1)}(.)$ is the probability distribution function of the Fisher distribution with $(I-1)$ and $(n-(r+I-1))$ degrees of freedom, denote the corresponding p-value. Many multiple testing methods can be used to determine a rejection threshold p^* on the p-values for the simultaneous tests of H_{0t}, $t = \{t_1, \ldots, t_T\}$. In the present situation of independent tests, one of the most famous method is the Bonferroni correction: choosing $p^* = \alpha/T$ ensures that the Family-Wise Error Rate (FWER) is lower than α:

$$\mathbb{P}_{H_0}\Big(\bigcup_t \big[\frac{n}{I-1}\frac{\hat{a}_t' S_{xx} \hat{a}_t}{s_t^2} \geq f^*\big]\Big) \leq \alpha,$$

where $f^* = G^{-1}_{I-1,n-(r+I-1)}(1-p^*)$. However, the FWER-controlling multiple testing procedures are known to be conservative and the Bonferroni correction is not adapted to dependence among tests.

Multivariate analysis of variance An alternative approach is to derive a single test statistics for H_0 by gathering the curves into a single vector $Y = (Y_{t_1}', Y_{t_2}', \ldots, Y_{t_T})'$. The multivariate analysis of variance model (1) is now expressed in the following expanded form:

$$\begin{aligned} Y \quad &= \quad [\mathbb{I}_T \otimes X_0]\mu + [\mathbb{I}_T \otimes X]a + \varepsilon, \\ &= \quad \tilde{X}_0 \mu + \tilde{X} a + \varepsilon, \end{aligned} \tag{4}$$

where $\otimes$ is the Kronecker matrix product, $\mu = (\mu_{t_1}, \ldots, \mu_{t_T})'$ and $a = (a_{t_1}', \ldots, a_{t_T}')'$. The above expanded linear framework is not homoscedastic. Indeed,

$$\text{Var}(\varepsilon) \quad = \quad [D_\sigma R D_\sigma] \otimes \mathbb{I}_n = V_\varepsilon,$$

where D_σ is the $T \times T$ diagonal matrix which diagonal terms are $\sigma_{t_1}, \ldots, \sigma_{t_T}$.

Therefore,

$$\text{Var}(\hat{a}_{\text{gls}}) \quad = \quad \frac{1}{n}[D_\sigma R D_\sigma] \otimes S_{xx}^{-1}. \tag{5}$$

Recent papers (see [4, 24, 25]) suggest analysis of variance F-tests based on the above presentation of the ERP curves. Furthermore, [4, 24, 25] take advantage of this framework to introduce a smooth non-parametric model for μ and a, using B-splines or wavelets. This modifies the design matrix of the model and correspondingly reduces the number of regression coefficients. Indeed, if φ stands for the $T \times S$ matrix associated to the basis functions, e.g. B-splines, (such that $\varphi_{is} = \phi_s(t_i),\ s =$

$1, \ldots, S;\ i = 1, \ldots, T$), then $\tilde{X}_0 = \varphi \otimes X_0$ and $\tilde{X} = \varphi \otimes X$ in expression (4) and the parameters μ and a are respectively $Sr-$ and $S(I-1)$ vectors of regression coefficients on the basis functions. In the former functional analysis of variance approach, the handling of dependence in V_ε remains the same as in the multivariate linear model (4).

The Generalized Least Squares (GLS) method is here appropriate since it provides the minimum variance linear unbiased estimator of the regression parameters provided (σ, R) is known. The corresponding F-statistics for $H_0 :\ a = 0$ is given by the following expression:

$$F_{\text{gls}} \quad = \quad n\hat{a}'_{\text{gls}}([D_{1/\sigma}R^{-1}D_{1/\sigma}] \otimes S_{xx})\hat{a}_{\text{gls}}. \tag{6}$$

which null distribution is $\chi^2_{(I-1)T}$. However, F_{gls} cannot be calculated because its expression depends on the unknown variance parameters. To handle this issue, the analysis of variance F-tests in [4, 24, 25] are either based on the usual homoscedastic assumption $V_\varepsilon = \sigma^2 I_{nT}$ or an assumption of a lag-1 autoregressive process for R:

$$R_\rho \quad = \quad \begin{pmatrix} 1 & \rho & \rho^2 & \ldots & \rho^{T-1} \\ \rho & 1 & \rho & \ldots & \rho^{T-2} \\ \vdots & \vdots & & & \vdots \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \ldots & 1 \end{pmatrix},$$

where ρ is the lag-1 autocorrelation.

Homoscedasticity assumption Let $\hat{Y}$ denote the vector of fitted values of Y, using the above GLS estimates of the model parameters. The corresponding F-test for the null hypothesis $H_0 :\ a = 0$ under the homoscedasticity and independence assumption is given by the following expression:

$$F_{\text{ols}} \quad = \quad n\frac{\hat{a}'_{\text{gls}}(I_T \otimes S_{xx})\hat{a}_{\text{gls}}}{\hat{\sigma}^2}, \tag{7}$$

where

$$\hat{\sigma}^2 \quad = \quad \frac{(Y - \hat{Y})'(Y - \hat{Y})}{T(n - (r + I - 1))}. \tag{8}$$

Heteroscedasticity assumption Note that, if the above homoscedasticity assumption is relaxed, then $V_\varepsilon = D_{\sigma^2} \otimes I_n$ and the corresponding F-statistics:

$$F_s \quad = \quad n\hat{a}'_{\text{gls}}(D_{1/s^2} \otimes S_{xx})\hat{a}_{\text{gls}}, \tag{9}$$

where $s^2 = (s^2_{t_1}, \ldots, s^2_{t_T})'$ is given by expression (3), is explicitly derived from the sum of the individual F_t statistics:

$$F_s \quad = \quad (I-1)\sum_{i=1}^{T} F_{t_i}. \tag{10}$$

Therefore, in the present situation, heteroscedasticity can be straightforward accounted for by considering that the null distribution of F_s is the distribution of $(I-1)\times$ the sum of T independent $\mathcal{F}_{I-1,n-(r+I-1)}$ variables. In the following, the former null distribution is denoted $\bar{\mathcal{F}}_{I-1,n-(r+I-1)}$.

Autoregressive covariance assumption Under the lag-1 autoregressive covariance assumption, let us introduce the matrix $L_{\rho,\sigma}$ defined as follows:

$$L_{\rho,\sigma} \quad = \quad \begin{pmatrix} -\frac{\rho}{\sigma_{t_1}} & \frac{1}{\sigma_{t_2}} & 0 & \ldots & 0 \\ 0 & -\frac{\rho}{\sigma_{t_2}} & \frac{1}{\sigma_{t_3}} & \ldots & 0 \\ \vdots & \vdots & & & \vdots \\ 0 & 0 & \ldots & -\frac{\rho}{\sigma_{t_{T-1}}} & \frac{1}{\sigma_{t_T}} \end{pmatrix}.$$

If $\varepsilon^* = (L_{\rho,\sigma} \otimes I_n)\varepsilon$ stands for the $(n-1)T-$ vector of innovations, then it is straightforward checked that $\text{Var}(\varepsilon^*) = (1-\rho^2)I_{(n-1)T}$. Therefore, starting from a non-zero initial estimate $\hat{\rho}_0$ of ρ to derive $L_{\rho,\sigma}$, an update is obtained by the variance of the corresponding ε^*:

$$\hat{\rho}_1^2 = 1 - \frac{\sum_{i=1}^{n}\sum_{j=1}^{T} \varepsilon^{*2}_{it_j}}{(n-(r+I-1))(T-1)}.$$

Plugging-in this updated estimate of ρ in $L_{\rho,\sigma}$ provides new decorrelated residuals ε^* and in turn a new estimate $\hat{\rho}_2$ of ρ. This defines an iterative algorithm which converges to the so-called Feasible GLS estimator $\hat{\rho}$ of ρ.

A corresponding feasible version $F_{s,\hat{\rho}}$ of the F_{gls}-statistics is obtained by plugging-in the estimates of σ_t and R_ρ in expression (6):

$$F_{s,\hat{\rho}} \quad = \quad n\hat{a}'_{\text{gls}}(D_{1/s}R^{-1}_{\hat{\rho}}D_{1/s} \otimes S_{xx})\hat{a}_{\text{gls}}. \tag{11}$$

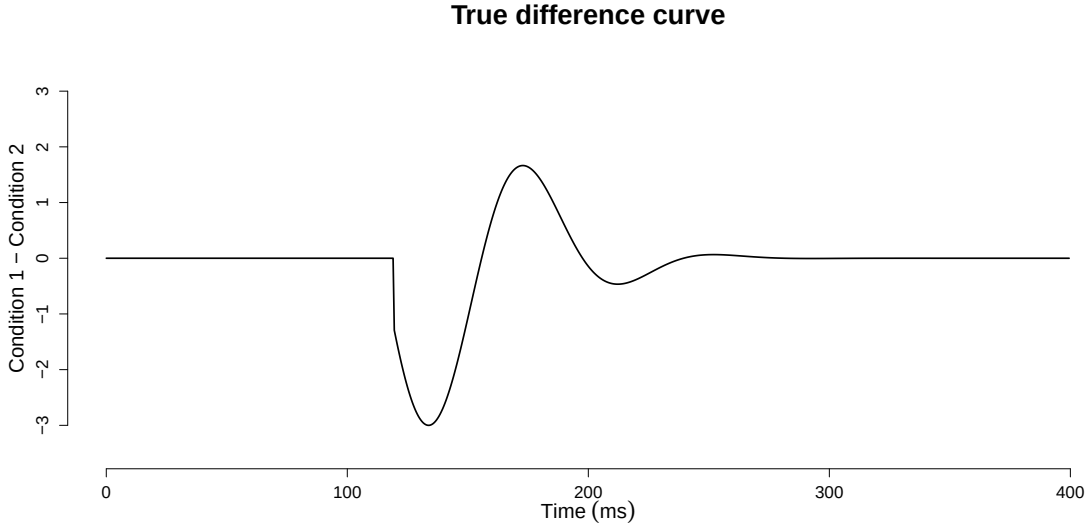
The exact null distribution of $F_{s,\hat{\rho}}$ is untractable. The analysis of variance F-tests are usually based on the asymptotic approximation of the null distribution by a $\mathcal{F}_{(I-1)T,(n-(r+I-1))T}$ degrees of freedom.

Finally, we suggest here to consider the F-test based on F_s (see expression (9)). Noticing that

$$\text{Cor}(F_{t_i}, F_{t_j}) \quad \approx_{n\to+\infty} \quad \rho^{2|i-j|},$$

the null distribution of F_s under the lag-1 autoregressive covariance assumption can be approximated by the distribution of $(I-1)\times$ the sum of lag-1 autocorrelated $\mathcal{F}_{I-1,n-(r+I-1)}$ variables, with auto-correlation ρ^2. In the following, this null distribution is denoted $\bar{\mathcal{F}}_{I-1,n-(r+I-1)}(\rho)$.

Figure 4: Expected difference ERP curve in the simulation study



2.3 Impact of a model misspecification on signal detection

In order to compare the different signal detection strategies described in the present section, we simulate datasets which dimensions and distribution mimic those obtained from the real auditory oddball dataset. The original dataset contains 30 ERP curves observed in two conditions, 15 curves in each condition on the time interval [0;400ms], at the frequency of 1 observation every half millisecond. We focus here on the significance of the difference curve. In the present simulation settings, it is assumed under H_1 that the expected ERP curve are constantly zero in condition 1, whereas it describes the waveform plotted in Figure 4 in condition 2. In the next section, we will show that, in the Rare-and-Weak paradigm of [8], the present situation falls into the so-called estimable region in terms of sparsity and weakness of the signal in the phase diagram under independence.

1000 datasets are generated, under H_0 and under H_1, with $n = 30$ rows and $T = 800$ columns, with standard deviation profiles s and a lag-1 autoregressive correlation matrix, with autocorrelation $\rho = 0.99$. Note that the estimated auto-correlation in the auditory oddball experiment is 0.997. On each simulated dataset, the following F-tests are implemented:

1. The F-test F_{ols} (see expression (7)) based on the homoscedasticity and independence assumption, with the null distribution $\mathcal{F}_{T,(n-2)T}$;
2. The same F-test as above calculated in the model including a B-spline smoothing of the regression coefficients. The corresponding F-test is implemented using the functions `bam` and

`anova.gam` in the `R` package `mgcv` with the default option for the choice of the smoothing parameter (minimizing the Generalized Cross Validation criterion). Note that the null distribution of the F-statistics is still a Fisher distribution, whose degrees-of-freedom accounts for the smoothness of the fit by means of the trace of the smoothing matrix;

3. The F-test $F_{s,\hat{\rho}}$ (see expression (11)) with the null distribution $\mathcal{F}_{T,(n-2)T}$;
4. The same F-test as above calculated in the model including a B-spline smoothing of the regression coefficients. The corresponding F-test is also implemented in the function `bam`, using arguments especially designed to introduce a lag-1 autocorrelation prior;
5. The F test F_s (see expression 9) with the null distribution $\bar{\mathcal{F}}_{I-1,n-(r+I-1)}(\hat{\rho})$. The calculation of the p-values are based on Monte-Carlo estimation of the former null distribution.

Figure 5 displays the empirical null probability distribution function of the p-values for five F-statistics. The plots confirms that the main impact of the model misspecification here is a too liberal F_{ols}-test, not controlling the type-I error rate. The concern is even more obvious when the fit is based on regression spline smoothing. This is consistent with the observation reported in many papers, including [4], that the confidence bands based on the homoscedasticity and independence assumption are strikingly too narrow. Another conclusion is that the approximate null distribution of $F_{s,\hat{\rho}}$ is obviously wrong, leading to a too liberal test, both without and with spline smoothing. The F-test based on the sum of individual F-tests, with an ad-hoc correction of the null distribution, turns out to control the type-I error rate at the desired level. Correspondingly, Figure 6, which shows the empirical non-null probability distribution functions of the same five F-tests, demonstrates the lesser ability of F_s to detect the signal.

Finally, the above simulation study shows that, provided we can efficiently derive the null distribution of the F-test under the dependence assumption, a signal detection strategy based on the individual F_t statistics can help achieving a correct control of the type-I error rate. The Higher Criticism Thresholding method presented hereafter is also exclusively based on the vector of F_t statistics. Note that, in the above simulation study, the ERP time dependence is supposed to be described by a lag-1 autoregression dependence structure, which is far from true (see Figure 3 for the image plot of the correlation matrix calculated with the auditory oddball dataset).

In the following, to be consistent with the framework defined by [8], we will focus on the situation where $I = 2$. Consequently, we will prefer dealing with the T-tests $\mathcal{T} = (T_{t_1}, \ldots, T_{t_T})'$ deduced by taking the signed square root of the F-tests $\mathcal{F} = (F_{t_1}, \ldots, F_{t_T})'$.

Figure 5: Empirical null probability distribution function of the F-test p-values under lag-1 autoregressive correlation

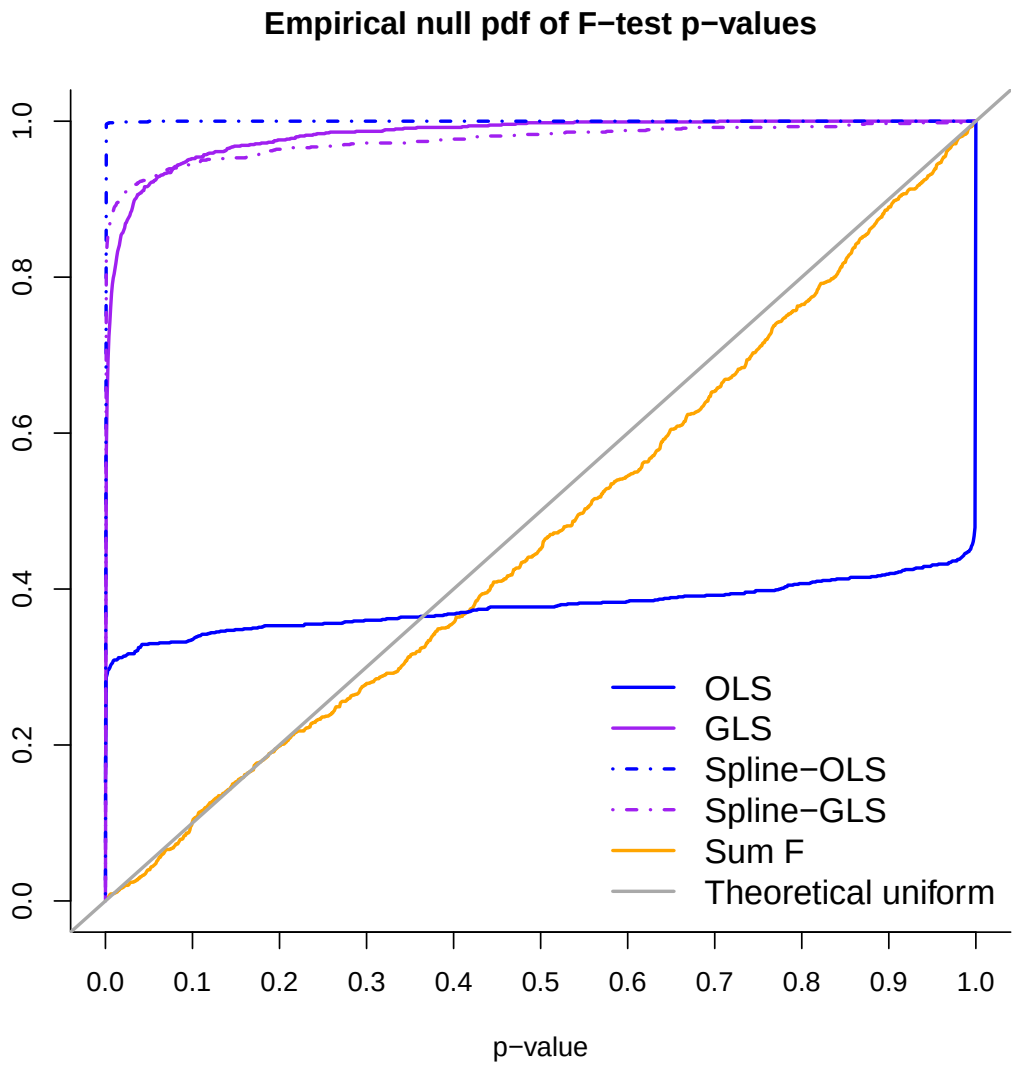
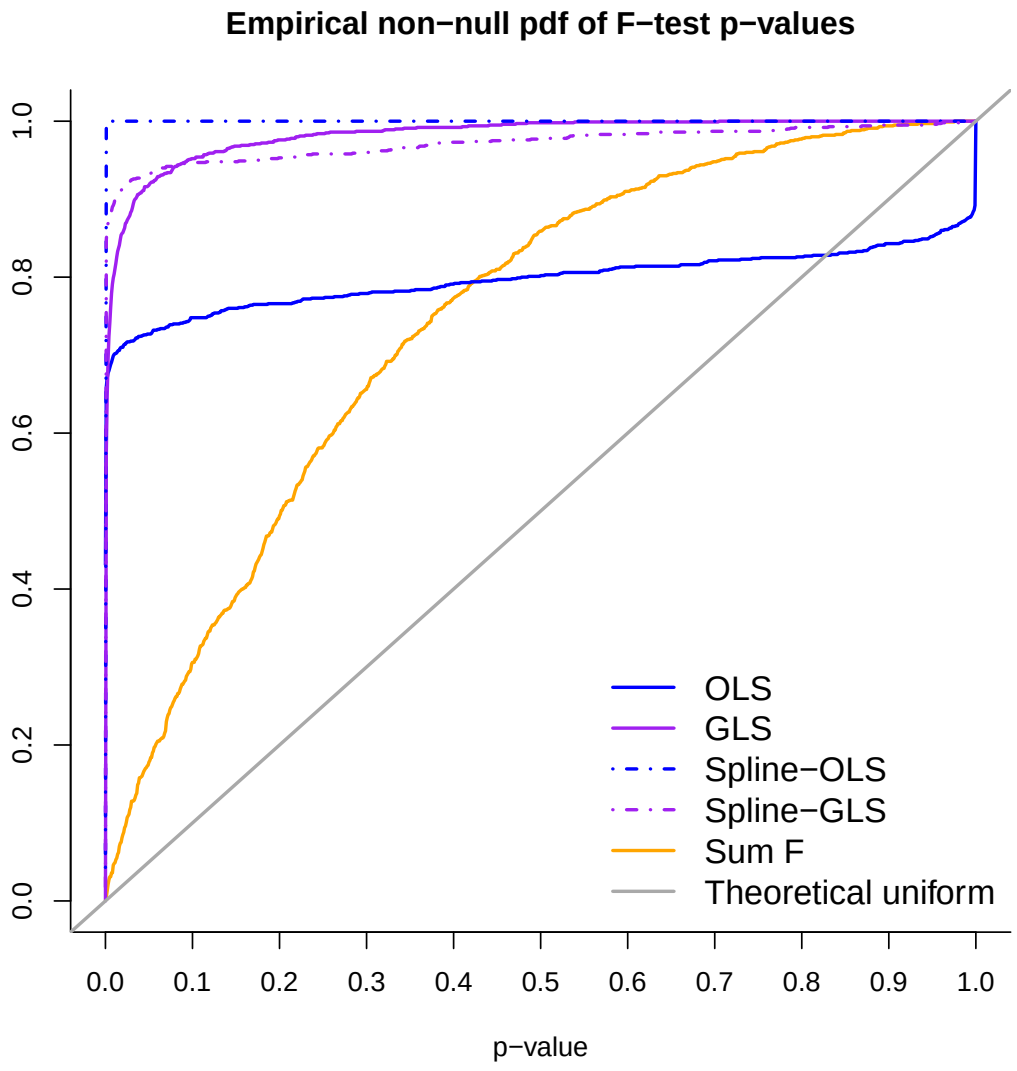


Figure 6: Empirical non-null probability distribution function of the F-test p-values under lag-1 autoregressive correlation



3 Higher Criticism for signal detection

3.1 Variants of the Higher Criticism method

The Higher Criticism (HC) method is designed to detect signals, namely to provide a single p-value for the significance of a signal, based on the vector of significance tests for the individual coordinates of this signal. It was initially proposed by [27] in his course notes at Princeton University and brought up to date by [8]. HC is designed to be optimal in a rare and weak model for the signal, defined in [8] as a sparse normal mixture model for z-scores. In this paper, we use the equivalent notation of [14]

$$\mathcal{T} \quad = \quad \mu + \mathcal{E}\,, \mathcal{E} \sim \mathcal{N}(0, R), \tag{12}$$

where R is set to the identity matrix and μ is a sparse T-vector with a proportion of non-null features $0 \leq \varepsilon_T \leq 1$ and non zero entries set to the same signal amplitude $A_T \geq 0$. Note that the normality assumption introduced above holds for most ERP studies in which the tests for the association between the ERPs and the response variable can be handled by t-tests for the significance of a single parameter ([6], [12]). The RW model supposes a subtle situation in which the signal is rare, as $\varepsilon_T = T^{-\beta}$ with $\beta \in (\frac{1}{2}, 1)$ and weak $A_T = \sqrt{2r\log(T)}$, $0 < r < 1$. When parameters (β, r) are known, explicit detection boundaries can be calculated for the RW setup under independence, which separate the space (β, r) into undetectable, detectable ([16]) and estimable ([8]) regions and are summarized in a so-called phase diagram (see Figure 7): if $r > \rho_D^*(\beta)$, the signal is detectable which means that the sum of type I and type II errors of the Neymann-Pearson Likelihood Ratio test (LRT) tends to 0 when the number of tests tends to infinity. If $r > \rho_E^*(\beta)$, the signal is not only detectable but the support of non null coordinates is identifiable. If (β, r) are out of these bounds, the signal is undetectable so that the sum of type I and type II errors for any test tends to 1 as the number of tests tends to infinity. The bounds ρ_D^* of detectability and ρ_E^* of estimability have the following expressions:

$$\begin{aligned} \rho_D^*(\beta) \quad &= \quad \begin{cases} \beta - \frac{1}{2} & \text{if } \frac{1}{2} < \beta \leq \frac{3}{4} \\ (1 - \sqrt{1-\beta})^2 & \text{if } \frac{3}{4} < \beta < 1, \end{cases} \\ \rho_E^*(\beta) \quad &= \quad \beta. \end{aligned}$$

In the example introduced in the simulation study of the previous section, the difference curve between two conditions shows a waveform. Therefore, the corresponding curve of r coefficients is not constant on intervals of nonzero signals as in the above RW setup, as shown in Figure 8. However, note that a proportion of 3% ($\beta = 0.52$) of the signal has a r coefficient larger than 0.7. The corresponding RW situation ($\beta = 0.52, r = 0.7$) falls into the estimable region in the phase diagram displayed in Figure 7.

Figure 7: Phase diagram of undetectable, detectable and estimable regions under independence

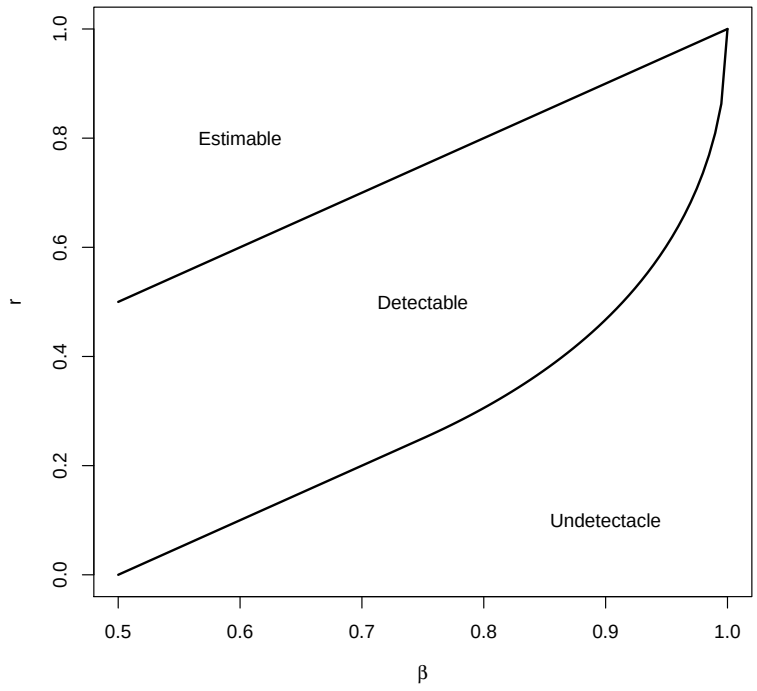
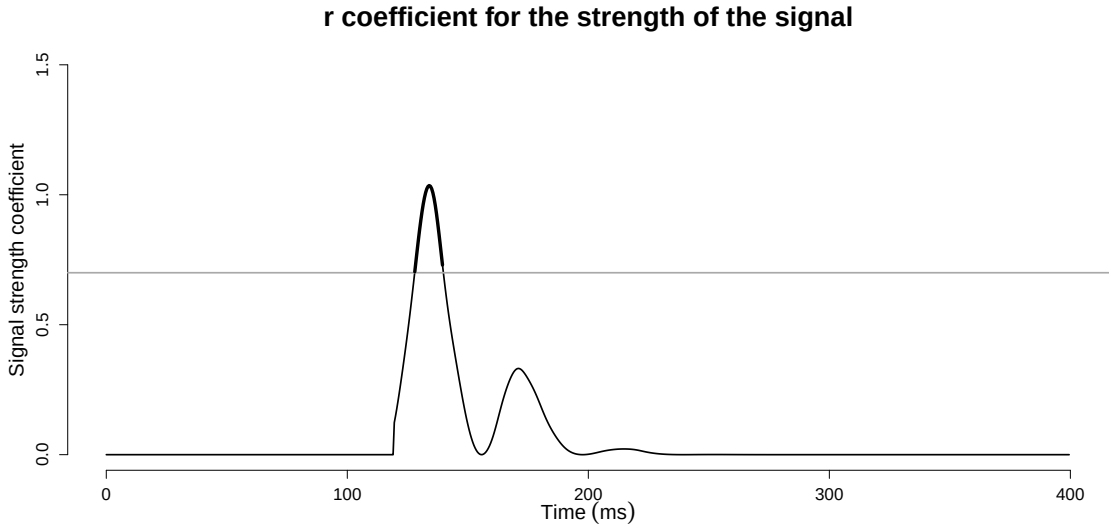


Figure 8: r coefficient curve in the simulation setting of section 2. A proportion of 3% ($\beta = 0.52$) of the signal has an amplitude larger than 0.7.



Signal detection by HC starts from the observation that the presence of a signal generates a discrepancy between the empirical distribution of the p-values and the theoretical null uniform distribution $\mathcal{U}[0;1]$. Indeed, it is expected that, for non-zero signal coordinates, the corresponding t-th increasingly ordered p-value $p_{(t)}$ is such that $p_{(t)} \ll t/T$. This discrepancy is measured by a Higher Criticism function $HC(p_{(t)}, t)$. The various HC methods are based on different definitions of the HC objective function: [9] and [17] standardize the difference $t/T - p_{(t)}$ by the standard-deviation of ordered p-values under the null, $Var(p_{(t)}) = \frac{t}{T}(1 - \frac{t}{T})$ and define the HC objective function as:

$$HC(p_{(t)}, t) \quad = \quad \sqrt{T}\frac{t/T - p_{(t)}}{\sqrt{t/T(1 - t/T)}}.$$

Alternatively, [8] and [14] suggest to standardize the difference $t/T - p_{(t)}$ by $p_{(t)}(1 - p_{(t)})$ and define the HC objective function as:

$$HC(p_{(t)}, t) \quad = \quad \sqrt{T}\frac{t/T - p_{(t)}}{\sqrt{p_{(t)}(1 - p_{(t)})}}.$$

Using arguments from the theory of empirical processes, [8] shows that

$$\frac{HC^*}{\sqrt{2 \log \log T}} \to 1, \text{ in Probability},$$

where HC^* is the maximum of $HC(p_{(t)}, t)$. It is deduced that the type-I error rate of any signal detection rule of the form $HC^* \geq (1+a)\sqrt{2\log\log T}$, for $a > 0$, tends to 0 as T tends to infinity. Moreover, [8] shows that the former tests are optimal in the sense that, for any Rare-Weak situation (r, β) within the detectable region, the type-II error rate of the tests based on HC^* also tends to 0 when T tends to infinity. This property is known as the optimal adaptivity of Higher Criticism. In order to obtain a testing strategy which controls the type-I error rate, [5] suggest to derive the value of a from a Monte-Carlo estimation of the null distribution of HC^*. We have implemented the former method in the simulation study introduced in the previous section. Figures 9 and 10 display the empirical null (resp. non-null) probability distribution function of the p-values for the HC statistics and the $F_{\hat{\sigma}}$-statistics, which accounts for dependence. The plots show the two methods have similar performance.

Note that, in practice, the HC function is often only maximized over a subset $\mathcal{I}$ of features with small p-values:

$$HC^* \quad = \quad \max_{t \in \mathcal{I}} HC(p_{(t)}, t).$$

Thus, another major variation among the different HCT methods is on the definition of the subset $\mathcal{I}$: for [8] and [9], $\mathcal{I} = \{t, 1/T \leq t/T \leq \alpha_0\}$ where $\alpha_0 \in [0, 1]$ is a preset proportion of small p-values. Alternatively, [14] proposes to avoid the extremely small p-values by taking $\mathcal{I} = \{t, 1/T \leq p_{(t)} \leq \alpha_0\}$. At last [17] proposes a global maximization $\mathcal{I} = \{1, \ldots, T\}$. In the above simulation of the HC statistics, the value $\alpha_0 = 0.1$ has been chosen, to be consistent with the recommendation in [8].

In the next sections, the standard HCT refers to the following criterion:

$$HC^* \quad = \quad \max_{1/T \leq t/T \leq \alpha_0} \sqrt{T} \frac{t/T - p_{(t)}}{\sqrt{t/T(1-t/T)}} \tag{13}$$

which appears to be the most widely used in the literature.

In a feature selection objective for supervised classification, [9] highlight the link between maximizing the HC function and minimizing the error rate of the classifier implemented on the features selected as follows: if the maximum value HC^* of the HC function is reached at index t^*, then the subset of selected features is given by $\{t, p_t \leq p_{(t^*)}\}$. [9] demonstrates that the former Higher Criticism Threholding (HCT) method outperforms methods based on the False Discovery Rate (FDR) thresholding. However, [17] shows the equivalence between HCT and the definition of a cutoff on the local FDR.

Figure 9: Empirical null probability distribution function of the p-values of the HC and $F_{\hat{\sigma}}$ statistics under lag-1 autoregressive correlation

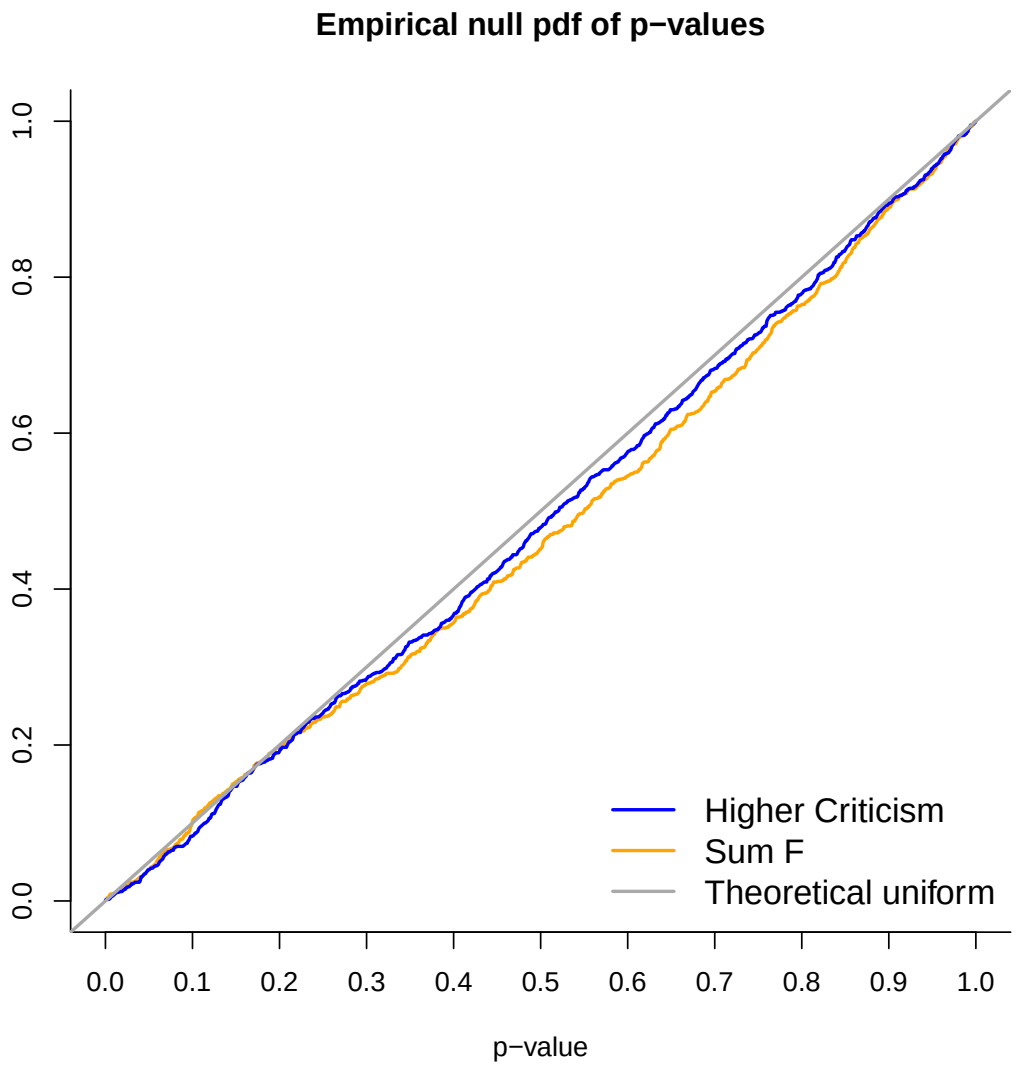
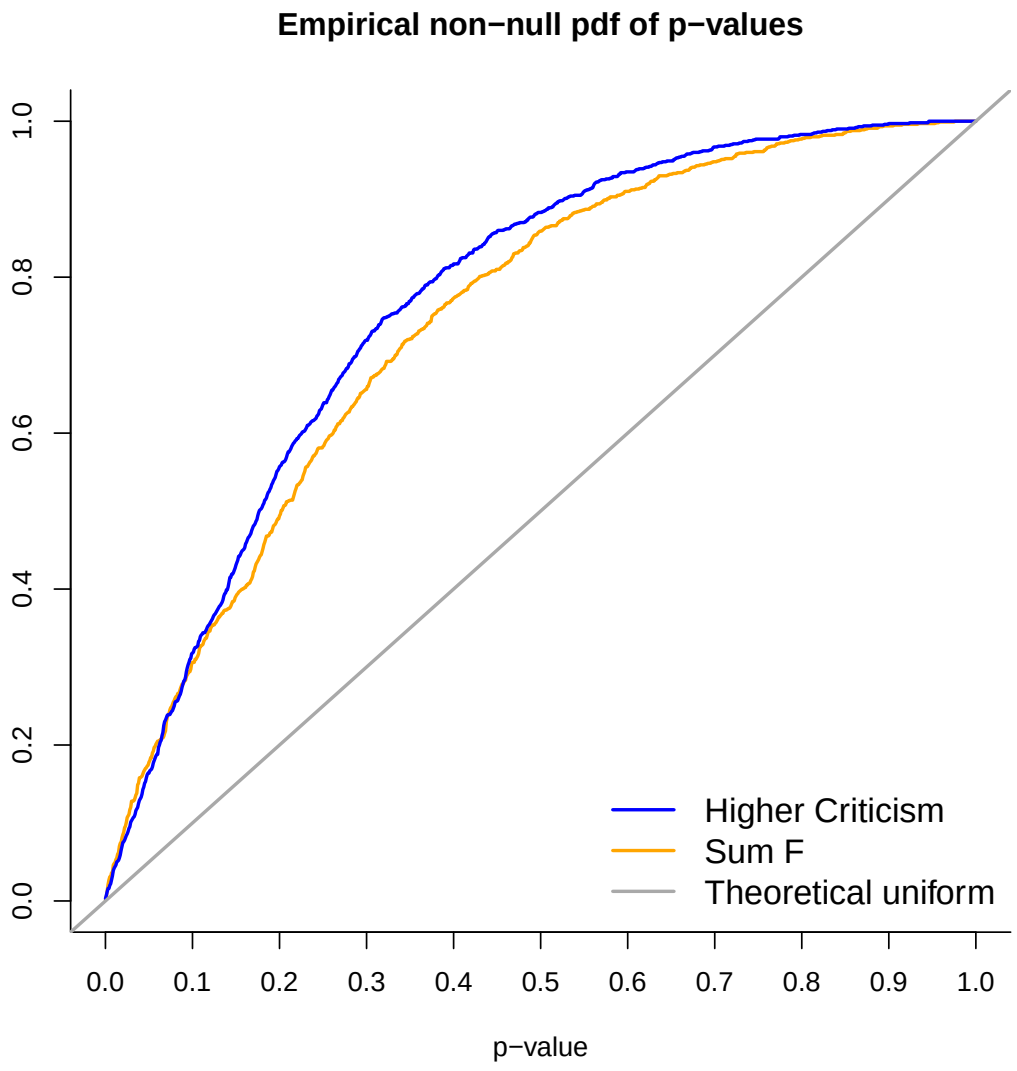


Figure 10: Empirical non-null probability distribution function of the p-values of the HC and $F_{\hat{\sigma}}$ statistics and under lag-1 autoregressive correlation



3.2 HCT under dependence

Even if the optimality of HCT is known to be robust to weak dependence, the simulation study above suggests that it would be improved by explicitly taking into account dependence. The existing variants of HC accounting for dependence are mainly based on a decorrelation step of the vector $\mathcal{T}$ of test statistics. Initially proposed by [29] in a multiple testing framework but also used as a variable selection procedure in a supervised classification issue in [1], the Correlation-Adjusted T-scores are defined as:

$$\tau^{adj} \quad = \quad R^{-1/2}\tau,$$

where τ are standard t-scores arising from a multiple testing procedure. The correlation matrix R is estimated by a James-Stein estimator $R_\gamma = \gamma \mathbb{I}_m + (1-\gamma)R$ as proposed by [22]. Interestingly, [22] provide an analytical expression for the estimation of γ. The inverse square root of R_γ is computationally feasible in high dimension by observing that the matrix $Z = \frac{1}{\gamma}R_\gamma$ can be decomposed as $Z = \mathbb{I}_m + UMU'$ where M is a definite positive symmetric matrix and U an orthonormal basis matrix. Hence,

$$Z^\alpha \quad = \quad \mathbb{I}_m - U(\mathbb{I}_r - (\mathbb{I}_r + M)^\alpha)U' \tag{14}$$

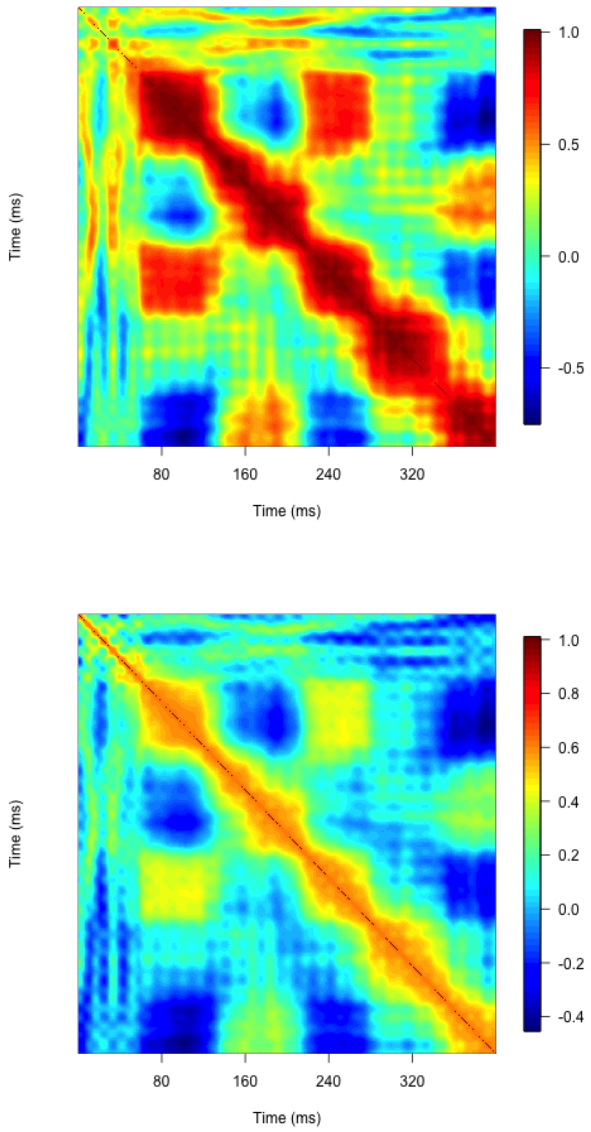
where r is the rank of the empirical correlation matrix. Even if it is efficient in practice, especially for many applications in genomic data analysis, the estimator proposed by [22] seems to fail to catch the dependence structure of ERP data. Indeed, the right panel of Figure 11 displays an image of the residual correlation matrix presented in Figure 3, which shows that the blocks of auto-correlation are underestimated.

Similarly, [14] propose the innovated HCT (iHCT) based on a preliminary whitening of $\mathcal{T}$. To this end, they introduce a matrix U_m such as $U_m R U_m' = \mathbb{I}_m$ and apply standard HCT to decorrelated test statistics in the manner of [29]. They recommend to use the inverse of the Cholesky factorization of the matrix R, which is computationally unstable in high dimension. As the support of the signal is slightly shifted by this linear transformation, the authors propose to restore the support by smoothing techniques applied on U_m, in the context of an auto-regressive dependence pattern.

4 Factor-Innovated Higher Criticism Thresholding

The proposed method is conceptually similar to innovated HCT and CAT-scores except that the decorrelation is based on a factor model. In many areas in which high-throughput devices are used to observe complex systems, such as neuroscience (see [6]) or genomics (see [3], [11], [19], [26]), this model is now widely used to implement data reduction techniques. In these papers, the authors suggest to cope with dependence by adjusting data from the effects of a moderate number of latent variables, which catch the dependence structure.

Figure 11: Estimation of within-condition correlation matrix by AFA algorithm ([23]) under a factor model assumption with 6 factors (left panel) and by the shrinkage estimator of [22] (right panel)



4.1 Whitening using latent factors

Factor model assumes the conditional independence of the variables given latent factors. Indeed, conditionally on latent factors, Expression (1) becomes:

$$Y_{ijt} \quad = \quad x'_{0ij}\mu_t + a_{it} + b'_t z_{ij} + e_{ijt}, \tag{15}$$

where $z_{ij} = (z_{ij1}, \ldots, z_{ijq})$ is normally distributed with mean 0_q and variance $\mathbb{I}_q$ and $e_{ij} = (e_{ij,t_1}, \ldots, e_{ij,t_T})'$ are independent random error vectors such that $e_{ij} \sim \mathcal{N}_T(0_T, \Psi_\Sigma)$ where Ψ_Σ is a diagonal matrix of specific variances described hereafter. q is the number of latent factors, $q \ll T$. The algorithms developed in a large number of papers (see [11], [6] [26], [18], [20]) are designed for the estimation of the latent factors in order to adjust data from their effects. In this paper, we intentionally focus on another implication induced by the rank-reduced factor structure of the covariance matrix Σ. Indeed, factor model assumes equivalently that the covariance matrix Σ can be decomposed as follows:

$$\Sigma \quad = \quad \Psi_\Sigma + B_\Sigma B'_\Sigma,$$

where B_Σ is a $T \times q$ matrix of loadings, which describes the common dependence shared by the variables and $\Psi_\Sigma = \mathrm{diag}(\Psi_{\Sigma,1}, \ldots, \Psi_{\Sigma,\mathrm{T}})$ is a diagonal matrix, which defines the specific variance. A similar factor structure holds for residual correlation matrix:

$$R \quad = \quad \Psi + BB',$$

where $\Psi_\Sigma = D_{\sigma^2}\Psi$, $\Psi = \mathrm{diag}(\Psi_1, \ldots, \Psi_{\mathrm{T}})$ and $B_\Sigma = D_\sigma B$.

The parameters (Ψ, B) are estimated by an EM-algorithm proposed by [23] and implemented in the `R` package `ERP` available online on the `CRAN` ([7]). The authors provide an adaptive estimation of the dependence structure under the factor model (15) based on a simultaneous estimation of the signal and covariance matrix. Moreover, the method is designed to take advantage of a prior knowledge on time intervals where the signal does not occur in order to denoise the least-squares estimation of the signal. The number of factors q is estimated by the variance inflation criterion proposed by [11].

The left panel of Figure 11 illustrates that factor modeling is flexible enough to recover the dependence pattern observed in the auditory oddball ERP dataset. The Figure shows an image plot of the estimation of correlation matrix presented in Figure 3 by a factor model with $\hat{q} = 6$ factors.

4.2 Detection boundary

Deriving exact boundaries in an arbitrary dependence pattern is impossible since the possibly non-homogeneous dependence along time can generate a different detectability and estimability of the

signal for some subset of features. By the way, [14] do not propose a closed-form phase diagram under dependence but explicit lower and upper bounds for the detection boundaries under some dependence patterns. In this section, we also propose to extend the phase diagram presented in Section 3 to the dependent case in a factor model framework.

Provided the test statistics are linear transformations of the response variables, which is generally the case for t-tests, the covariance structure of the test statistics is inherited from the residual correlation of those response variables. If $\mathcal{Z}$ stands for the same linear transformation applied to the $n \times q$ matrix Z, which rows are z_{ij}, then model (12) for the test statistics $\mathcal{T}$ becomes:

$$\mathcal{T} \quad = \quad \mu + B\mathcal{Z}' + E \, , E \sim \mathcal{N}(0, \Psi).$$

Equivalently,

$$\Psi^{-1/2}(\mathcal{T} - B\mathcal{Z}') \quad = \quad \Psi^{-1/2}\mu + E^* \, , E^* \sim \mathcal{N}(0, \mathbb{I}_T). \tag{16}$$

The above equation offers a RW setup similar to model (12), in which the decorrelated test statistics $\Psi^{-1/2}(\mathcal{T} - B\mathcal{Z}')$ are independent with equal variance 1, given the factors $\mathcal{Z}$. Upper and lower bounds for the detection boundary of a signal with constant amplitude $A_T = \sqrt{2r\log(T)}$ for a proportion ε_T of signal coordinates are deduced. Indeed, depending on the distribution of the specific variances Ψ_t along the features, the amplitude of the transformed signal $\Psi^{-1/2}\mu$ can take any values between $\underline{\gamma_0} = A_T/\sqrt{\Psi_{\max}} = \sqrt{2\underline{r}\log(T)}$ and $\overline{\gamma_0} = A_T/\sqrt{\Psi_{\min}} = \sqrt{2\overline{r}\log(T)}$, where $\Psi_{\min} = \min_t(\Psi_t)$, $\Psi_{\max} = \max_t(\Psi_t)$, $\underline{r} = r/\Psi_{\min}$ and $\overline{r} = r/\Psi_{\max}$. The conditions on (r, β) for detectability established under independence can be invoked here, with two different values $\underline{r}$ and $\overline{r}$ for the strength parameter r, providing respectively the lower and upper bounds for detectability. First, if r satisfies the following condition:

$$r > \Psi_{\max}\rho_D^*(\beta), \tag{17}$$

then all the signal is detectable. Moreover, if r does not satisfy the following condition,

$$r > \Psi_{\min}\rho_D^*(\beta), \tag{18}$$

then the signal can not be detected. Between these two bounds, detection of the signal is uncertain, depending on the correspondence between the profile of specific variances and the support of the signal.

The same reasoning for estimability gives similar forms for estimation boundaries. At last, we propose to define a *factor RW setup* in which the sparsity of the signal is given by

$$\varepsilon_T \quad = \quad T^{-\beta} \text{ with } \beta \in (1/2, 1),$$

and the strength of the signal by

$$A_T \quad = \quad \sqrt{2rlog(T)} \text{ with } 0 < r < \Psi_{max}.$$

Under independence, the signal strength is set to $\sqrt{2r\log(T)}$ with $0 < r < 1$ to make signal detection challenging. Indeed, $\sqrt{2\log(T)}$ is the expectation of the maximum of T independent normal variables with mean 0 and variance 1. Note that signals with amplitudes $\sqrt{2r\log(T)}$, with $\Psi\max < r \leq 1$, are considered as weak under independence but not under dependence. Consistently, this observation is also reported by [14], who explains that correlated designs are actually favorable for signal detection with procedures accounting for dependence. Moreover, this is illustrated by Figure 12, which shows the detectability and estimability phase diagrams obtained with the estimated specific variances in the auditory oddball experiment. Indeed, the plot confirms that both the detectability and the estimability regions are wider under dependence.

The bounds in Expressions (17) and (18) are consistent with the ones established by [14]. Indeed, Theorems 3.1 and 4.2 of the former paper give the same expression as above for the detection bounds, where $\underline{\gamma}_0$ and $\overline{\gamma}_0$ are defined as follows:

$$\underline{\gamma_0} = \lim \inf_{T\to+\infty} \max_{\sqrt{T}\leq k\leq T-\sqrt{T}} R_{kk}^{-1},$$
$$\overline{\gamma_0} = \lim \sup_{T\to+\infty} \max_{\sqrt{T}\leq k\leq T-\sqrt{T}} R_{kk}^{-1},$$

where $(R_{11},\ldots,R_{kk},\ldots,R_{TT})$ stands for the diagonal of matrix R. In the present situation, the following sandwich inequality holds:

$$\underline{\gamma_0} = \frac{1}{\max(\Psi)} \quad \leq \quad \text{diag}(R^{-1}) \leq \frac{1}{\min(\Psi)} = \overline{\gamma_0},$$

which gives expressions (17) and (18).

4.3 Factor innovated HCT

As in [29] and [14], we propose a factor-innovated HCT (F-iHCT) which consists in applying standard HCT to decorrelated test-statistics. In the factor RW setup defined above, we define the following square root L of the correlation matrix R, such as $R = LL'$:

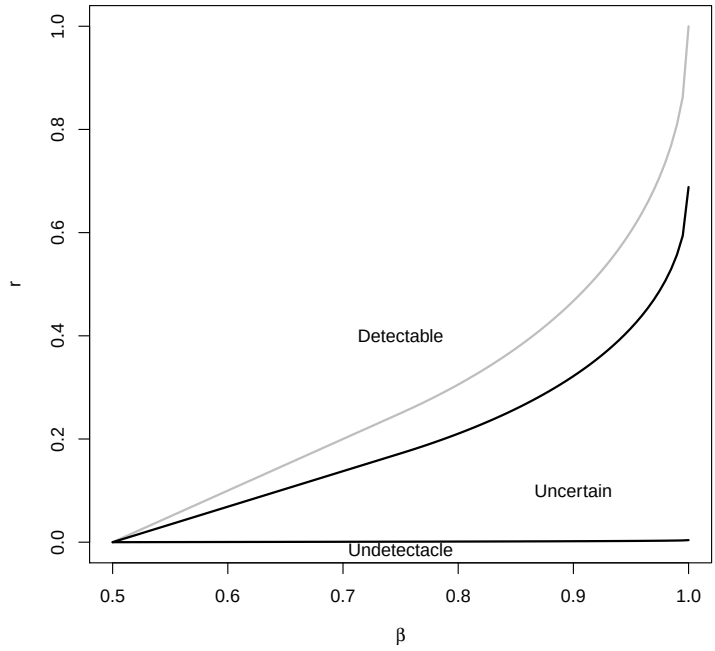
$$L \quad = \quad (\mathbb{I}_m - U[(\mathbb{I}_q + [\mathbb{I}_q + D^2]^{1/2})^{-1} + \mathbb{I}_q]^{-1}U')\Psi^{-1/2},$$

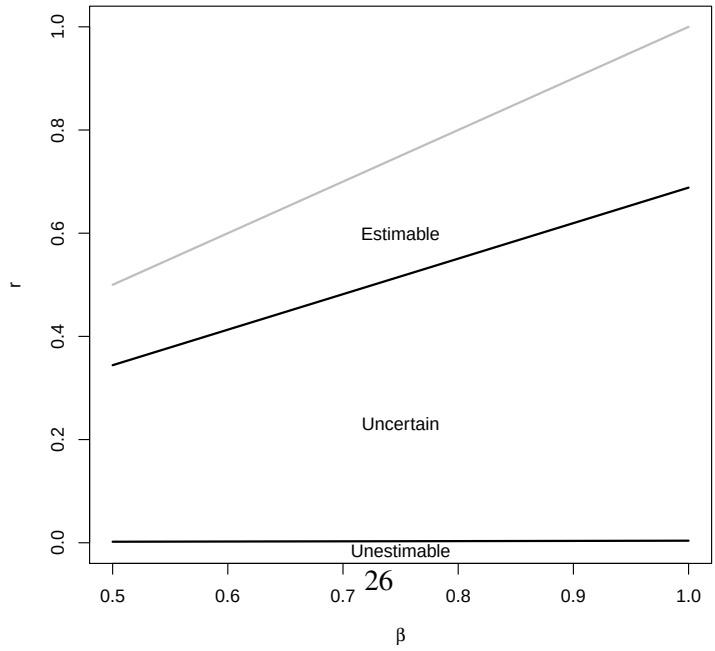
where U and D comes from the singular value decomposition of the standardized loadings $\Psi^{-1/2}B = UDV$. Note that this formula only requires inversion and square root computation of diagonal matrices. The decorrelated test statistics are therefore deduced:

$$\mathcal{T}^* \quad = \quad L'\mathcal{T}$$

and the corresponding p-values are denoted p_t^*. HCT is then applied on the collection of p-values $p^* = (p_1^*,\ldots,p_T^*)$.

Figure 12: Detection (upper plot) and identification (lower plot) boundaries for the auditory oddball ERP data. The grey line is the boundary under independence and black lines are upper and lower bounds of detectability (top) and identification (bottom) of a signal under dependence.





5 Simulation results and real data analysis

The properties of the proposed method are now compared to standard HCT and to some other methods based on decorrelation by innovations or by adjustment for effect of latent variables. The comparison is made through a simulation study and an application on the auditory oddball experiment.

5.1 Simulation study

Simulation settings The properties of F-iHCT are now investigated through simulations. 1,000 datasets with dimensions 30×799 are generated according to a multivariate normal distribution. Both the correlation structure and the within-condition variances are estimated from the auditory oddball ERP data introduced in section 2 (see Figure 3 and Figure 2). This simulation plan mimics the observed data on the oddball experiment by dimensions and covariance structure except that the true signal is known. Each dataset is split into two balanced groups. The normal distribution has expectation zero for the first 15 subjects (group 1) and the expectation for the 15 last subjects (group 2) is plotted on Figure 13. The difference curve is therefore a waveform with various amplitudes and the indices of non null features are in $[150ms, 200ms]$. $1,000$ training datasets are generated for each signal strength. Eight corresponding testing data of size 1000×799 with two balanced groups are also generated according to the same simulation plan for a prediction purpose. The RW model parameters for this simulation plan are $\varepsilon_T = 12\%$ and the maximum amplitude of signals is expressed as $A_T = \sqrt{2rlog(T)}$ with r taking 8 equally distributed values in $[0.004; 0.688]$. According to the RW setup, the present combination of r and β characterizes a not very sparse signal, with a weak to strong strength.

Methods Four methods are compared in this simulation study and are described hereafter. As in [9] and [10] the variable selection step by different versions of HCT, with two possible values for α_0 ($\alpha_0 = 0.1$ or 0.5), is followed by a supervised classification by Diagonal Discriminant Analysis on the subset of selected variables. The following methods are compared:

1. Variable selection by standard HCT on raw p-values, classification by Naives Bayes (see [2]) denoted by *Standard HCT*;
2. Variable selection by HCT on CAT-scores ([29, 1]), classification by diagonal Shrinkage Discriminant Analysis (SDA, see [1]) denoted by *CAT-scores*;
3. Variable selection by Factor-innovated HCT, classification by conditional Bayes classifier (proposed by [20]) denoted by *F-iHCT*;
4. Variable selection by standard HCT performed on p-values adjusted for effects of latent factors as returned by the AFA ([23]) procedure using `erpfatest` function of `R` package

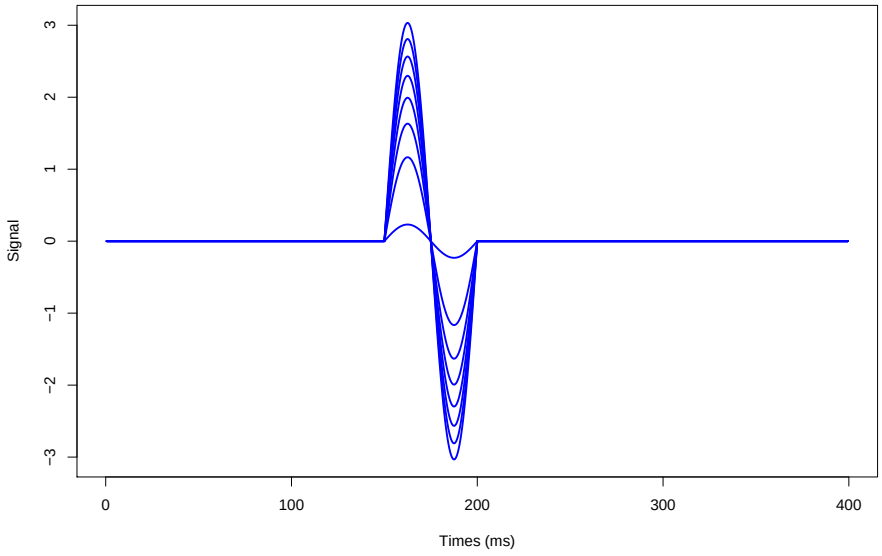


Figure 13: Simulation study - Signal amplitude curves along time

`ERP` ([7]), classification by conditional Bayes classifier (see [20]) denoted by *AFA*.

For all the methods described above, the proportion of signal recovery, called precision, the false discovery rate (FDR), the number of selected features and the prediction error rate are computed. For all datasets, variable selection and estimation of classification rule are performed on training data (including the optimization of meta-parameters) and prediction error is computed on testing data.

Results When α_0 is well specified regarding to the proportion of null hypothesis ($\alpha_0 = 0.1$), Figure 14 shows that selection by CAT-scores appears to be the most efficient to catch weak signals, with both the smallest FDR and the largest precision for small amplitudes of signal. Even if CAT-scores does not achieve the best performance for large signal strengths, the FDR, precision and number of selected variables are remarkably stable. Standard HCT seems robust to dependence as the method performs well in term of FDR but its precision is small regarding methods based on decorrelation. Moreover, the number of selected variables is also small, which suggests that HCT is conservative under dependence. Lastly, classification by Naive Bayes fails as the error rates are the largest for weak to moderate strengths of signal. Variable selection and classification procedures based on the factor model assumption (AFA and F-iHCT) provide the best results both in terms of false positive, recovery of the signal and prediction error. FDR turns out to be small for moderate

Table 1: Real data study - Number of selected time points for ERP auditory experiment

α_0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
Number of variables	40	80	120	160	200	240	280	320	360	400

to high signal strengths and a correct power of signal identification is achieved.

As shown on Figure 15, all methods are affected by a misspecification of α_0 parameter ($\alpha_0 = 0.5$). CAT-scores method selects too many variables so it achieves a good precision at the cost of large FDR. F-iHCT and AFA perform well in classification despite the number of false positives. Classification and precision rates of standard HCT are also affected by this wrong choice of α_0.

5.2 ERP data analysis

The 4 methods compared in the simulation study are now applied on the auditory oddball ERP data presented in Section 2. It is reminded that the purpose associated to the oddball experiment is to predict a new label from ERP curves. For each method, the number of selected features and the prediction error are computed. As the number of observations is small, the classification error is computed by leave-one-out cross-validation (CV).

Table 1 shows the number of selected features by the 4 compared selection methods for different values of α_0. Whatever the value of α_0, the 4 methods select the same number of features. The difference between all methods lies on the indices of selected features as they do not selected the same time points, as shown as an example on Figure 17.

Figure 16 presents the cross-validated error rates for several values of α_0. For values of α_0 larger than 0.15, standard HCT is stable and performs rather well. For more sparse models, standard HCT reaches larger error rates and is improved by decorrelation methods based on a factor model assumption (AFA and F-iHCT). The performance of the CAT-scores method varies slightly depending on α_0. The curves of F-iHCT and AFA are erratic for values of α_0 smaller than 0.125 but they stabilize when α_0 increases. For values of α_0 larger than 0.275, F-iHCT and AFA appear to be the most effective methods as they perfectly classify data. Nevertheless, one can notice that for equal error rates, the two methods do not select the same features as shown on the bottom of Figure 17.

Figure 17 shows the curve of the mean difference among the two groups and the time points selected by the 4 compared methods for $\alpha_0 = 0.125$ (top) and $\alpha_0 = 0.275$ (bottom). These values of α_0 are chosen because they provide two levels of sparsity but comparably small CV error. As expected, time points after 300ms are selected by all methods which is consistent with the literature but time points around 100 ms also appear to be significant.

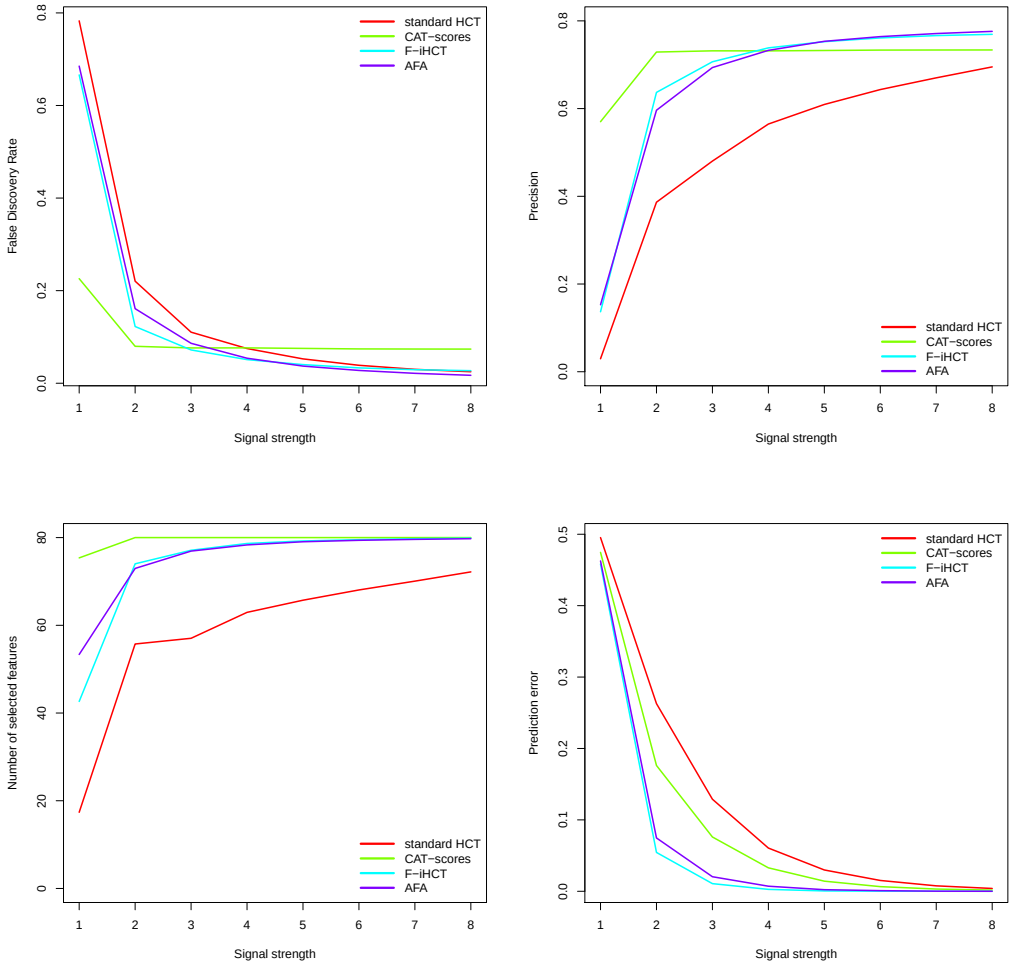


Figure 14: Results of the simulation study depending on signal strength and $\alpha_0 = 0.1$: False Discovery Rate (top left), Precision (top right), Number of selected features (bottom left), Prediction error (bottom right).

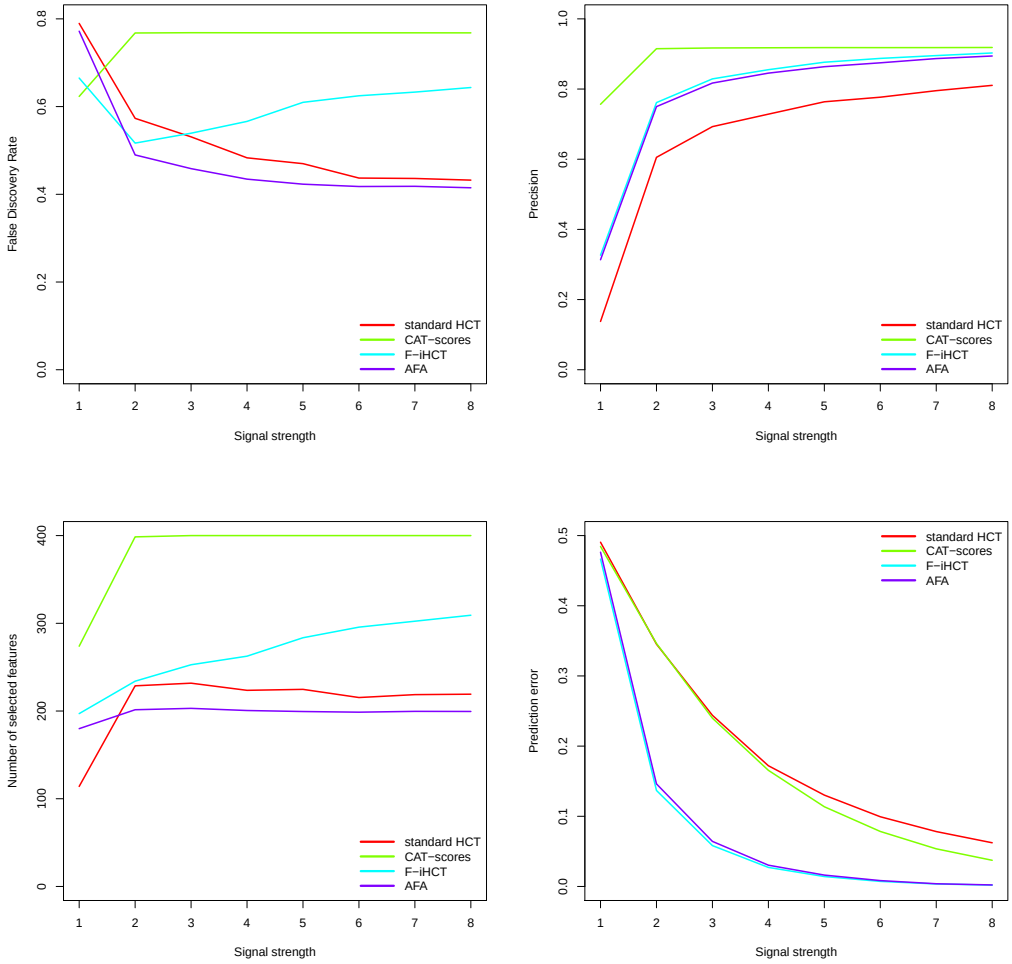


Figure 15: Results of the simulation study depending on signal strength and $\alpha_0 = 0.5$: False Discovery Rate (top left), Precision (top right), Number of selected features (bottom left), Prediction error (bottom right).

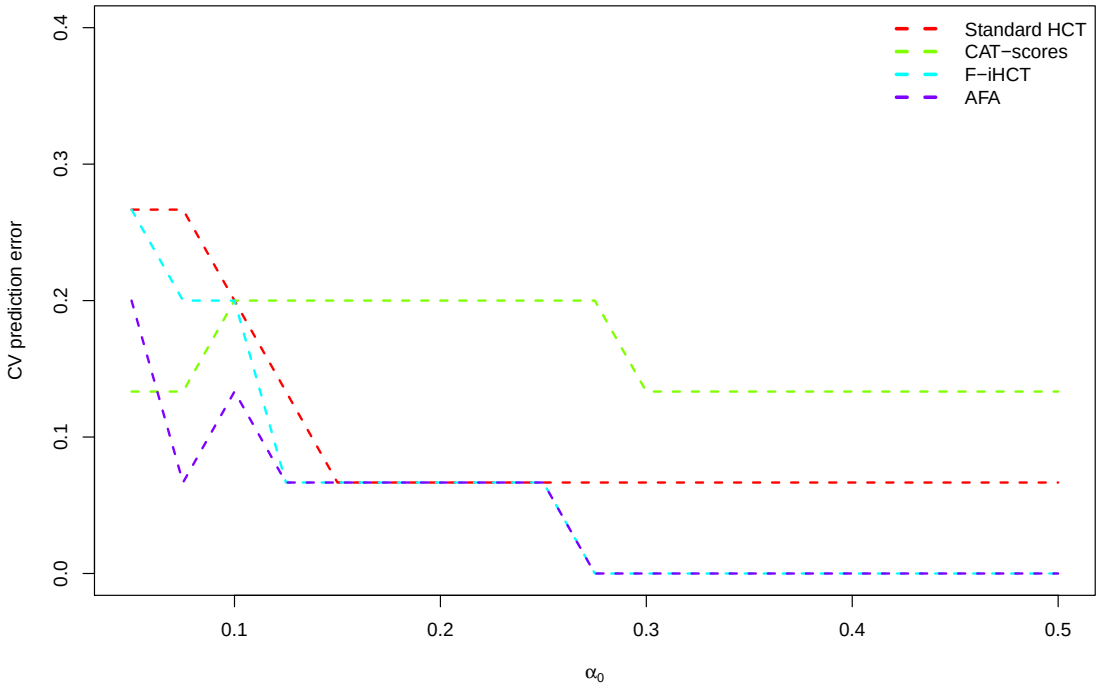


Figure 16: Real data study - Cross-validated prediction error of standard HCT, HCT performed on CAT-scores, factor innovated HCT and HCT performed on pvalues provided by AFA on auditory ERP experiment for several values of α_0

To conclude, this study highlights the importance of the choice of α_0 in HCT, since it has a significant impact on the sparsity and classification rates of the models for all methods. The method denoted by AFA which performs decorrelation by adjustment of covariates for the effect of latent factors seems to be the most suitable method in this example, even if F-iHCT also achieves interesting classification rates.

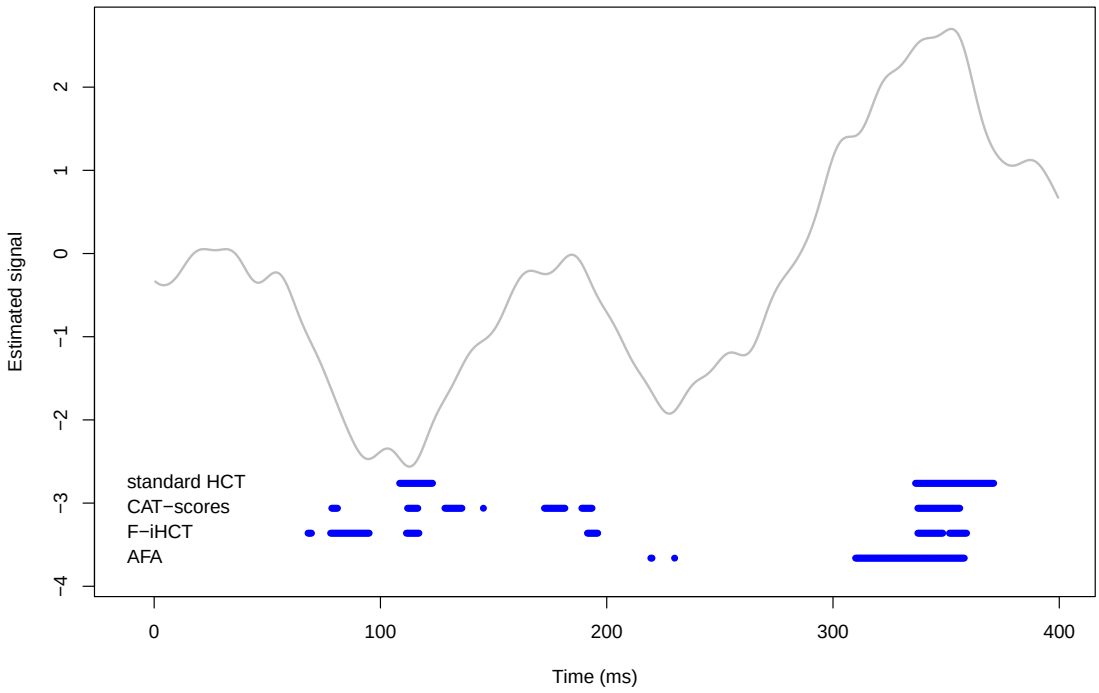
6 Discussion and conclusion

This article addresses signal detection and identification in Event-Related-Potentials (ERP) data. It is motivated by an ERP study in the oddball paradigm, where two classes of stimuli are presented to subjects, one occurring frequently (standard) and the other occurring infrequently (target). The aim is to identify time intervals of ERP curves which could be markers for the rare-frequent difference and to derive a classification rule. When the selection statistics are independent, Higher Criticism is known to be optimal as a signal detection method in the Rare-and-Weak setup initially introduced by [8]. Higher Criticism Thresholding is moreover efficient to estimate the support of the signal. The Rare-and-Weak setup is conceptually adapted to the ERP signal detection issue. However, selection statistics in ERP signal identification issues are characterized by a strong and complex dependence structure.

In ERP data analysis, signal detection is usually handled by F-tests for the overall nullity of the signal on the whole time interval of observation. Under heteroscedasticity and independence, we show that the corresponding Generalized Least Squares (GLS) F-test is expressed as a sum of the individual F-tests for each feature, which gives an explicit expression for the null distribution. Similarly, under a dependence pattern structured by a lag-1 auto-correlation, we show that the former GLS F-test can also be implemented, provided the null distribution accounts for the auto-correlation between the individual F-tests. A simulation study in which dependence among simulated features is structured by a strong lag-1 auto-correlation, demonstrates that the former test controls the type-I error rate, which is far from true for other F-tests, such as those obtained in Functional Analysis of Variance approaches.

The present paper proposes a variant of the HCT procedure which takes advantage of a factor model assumption to decorrelate the test statistics. Indeed, this framework provides algebraical tools to derive an inverse square root of the correlation matrix involved in the computation of innovations. Moreover, similarly to [15] and [8], a phase diagram under a general dependence assumption is deduced and the Factor Innovated HCT, a decorrelated HCT based on innovations, is proposed.

The method is assessed by a more intensive simulation study, in which the dependence structure of simulated datasets mimics the observed correlation pattern in the auditory oddball ERP dataset.



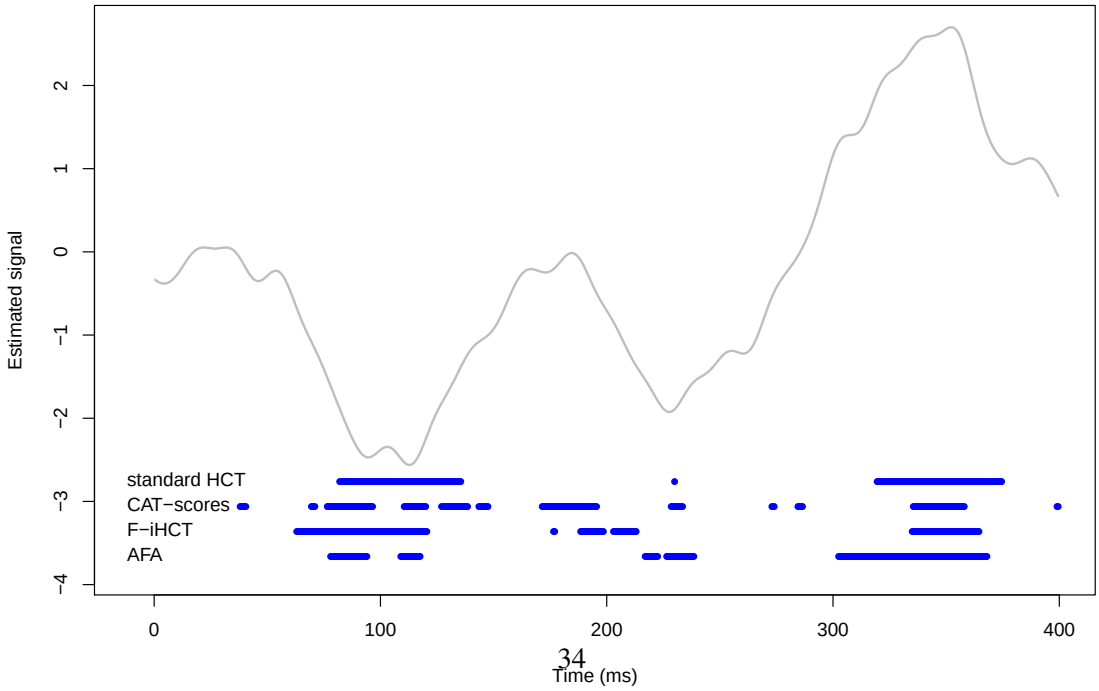


Figure 17: Real data study - Signal estimation (grey line) and significant time points (blue points) selected by standard HCT, HCT performed on CAT-scores, factor innovated HCT and HCT performed on p-values provided by AFA for $\alpha_0 = 0.125$ (top) and $\alpha_0 = 0.275$ (bottom) on auditory ERP experiment

This study illustrates that methods based on decorrelation of HCT under a factor model assumption perform well both in selection and classification. Indeed, F-iHCT and AFA, which is based on data adjustment for latent factors, achieve similar results. When applied on the oddball ERP dataset, all compared methods (standard HCT, CAT-scores, F-iHCT and AFA) select time points around 300 ms as expected but lower cross-validated error rates are achieved by the two methods based on a factor model assumption. Both application on simulations and on real data reveal the sensitivity of HCT to the choice of α_0 parameter, which determines the sparsity of the model.

The above promising results in signal identification should now be exploited in the construction of optimal signal detection strategies in ERP study. Indeed, provided the null distribution of the F-iHC statistics can be estimated, it could be interesting to compare the subsequent testing method to the class of F-tests.

References

[1] M. Ahdesmäki and K. Strimmer. Feature selection in omics prediction problems using cat scores and false non-discovery rate control. *Annals of Applied Statistics*, 4:503–519, 2010.

[2] P.J. Bickel and E. Levina. Some theory for Fisher's Linear Discriminant function, naive Bayes, and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989–1010, 2004.

[3] Y. Blum, G. LeMignon, S. Lagarrigue, and D. Causeur. A factor model to analyze heterogeneity in gene expression. *BMC bioinformatics*, 11:368, 2010.

[4] C. Bugli and P. Lambert. Functional anova with random functional effects: an application to event-related potentials modelling for electroencephalograms analysis. *Statistics in Medicine*, 25:3718–3739, 2006.

[5] T.T. Cai, J. Jeng, and J. Jin. Optimal detection of heterogeneous and heteroscedastic mixtures. *Journal of the Royal Statistical Society, Series B*, 73, 2011.

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

[7] D. Causeur and C.-F. Sheu. *ERP: Significance analysis of Event-Related Potentials data*, 2014. R package version 1.0.1.

[8] D. Donoho and J. Jin. Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32:3:962–994, 2004.

[9] D. Donoho and J. Jin. Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proceedings of the National Academy of Sciences*, 105:39:14790–14795, 2008.

[10] D. Donoho and J. Jin. Feature selection by higher criticism thresholding achieves the optimal phase diagram. *Philosophical Transactions of the Royal Society A*, 367:4449–4470, 2009.

[11] C. Friguet, M. Kloareg, and D. Causeur. A factor model approach to multiple testing under dependence. *Journal of the American Statistical Association*, 104:488:1406–1415, 2009.

[12] D. Guthrie and J.-S. Buchwald. Significance testing of difference potentials. *Psychophysiology*, 28:240–244, 1991.

[13] P. Hall and J. Jin. Properties of higher criticism under strong dependence. *The Annals of Statistics*, 36:1:381–402, 2008.

[14] P. Hall and J. Jin. Innovated higher criticism for detecting sparse signals in correlated noise. *The Annals of Statistics*, 38:3:1686–1732, 2010.

[15] Y.I. Ingster. Some problems of hypothesis testing leading to infinitely divisible distribution. *Mathematical Methods of Statistics*, 6:47–69, 1997.

[16] Y.I. Ingster. Minimax detection of a signal for ℓ_n^p balls. *Mathematical Methods of Statistics*, 7:401–428, 1999.

[17] B. Klaus and K. Strimmer. Signal identification for rare and weak features: higher criticism or false discovery rates? *Biostatistics*, 14:1:129–143, 2013.

[18] J. T. Leek and J. Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9):e161, 2007.

[19] J. T. Leek and J. Storey. A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences*, 105:18718–18723, 2008.

[20] E. Perthame, C. Friguet, and D. Causeur. Stability of feature selection in classification issues for high-dimensional correlated data. *Statistics and Computing*, pages 1–14, 2015.

[21] T. W. Picton. The p300 wave of the human event-related potential. *Journal of Clinical Neurophysiology*, 9(4):456–479, 1992.

[22] J. Schäfer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(32), 2005.

[23] C.F. Sheu, E. Perthame, D. Causeur, and Y.S. Lee. Accounting for time dependence in large-scale multiple testing of event-related potential data. *In revision*, 2015.

[24] N.-J. Smith and M. Kutas. Regression-based estimation of erp waveforms: I. the rerp framework. *Psychophysiology*, 52(2):157–168, 2015.

[25] N.-J. Smith and M. Kutas. Regression-based estimation of erp waveforms: Ii. nonlinear effects, overlap correction, and practical considerations. *Psychophysiology*, 52(2):169–181, 2015.

[26] Y. Sun, N.R. Zhang, and A.B. Owen. Multiple hypothesis testing adjusted for latent variables, with an application to the AGEMAP gene expression data. *The Annals of Applied Statistics*, 6(4):1664–1688, 2012.

[27] J.W. Tukey. T13 N: the higher criticism. *Course Notes*, Princeton University, 1976.

[28] L.M. Williams, E. Simms, C.R. Clark, , and R.H. Paul. The test-retest reliability of a standardized neurocognitive and neurophysiological test battery: neuromarker. *International Journal of Neuroscience*, 115:1605–1630, 2005.

[29] V. Zuber and K. Strimmer. Gene ranking and biomarker discovery under correlation. *Bioinformatics*, 25:2700–2707, 2009.