Available online at www.sciencedirect.com

ScienceDirect

Journal of Magnetic Resonance 185 (2007) 138–151

JMR
Journal of
Magnetic Resonance

www.elsevier.com/locate/jmr

fMRI bold signal analysis using a novel nonparametric statistical method

Patrick A. De Mazière *, Marc M. Van Hulle

K.U.Leuven, Laboratorium voor Neuro- en Psychofysiologie, Herestraat 49–bus 1021, B–3000 Leuven, Belgium

Received 30 June 2006; revised 13 November 2006
Available online 28 December 2006

Abstract

We present in this article a novel analytical method that enables the application of nonparametric rank-order statistics to fMRI data analysis, since it takes the omnipresent serial correlations (temporal autocorrelations) properly into account. Comparative simulations, using the common General Linear Model and the permutation test, confirm the validity and usefulness of our approach. Our simulations, which are performed with both synthetic and real fMRI data, show that our method requires significantly less computation time than permutation-based methods, while offering the same order of robustness and returning more information about the evoked response when combined with/compared to the results obtained with the common General Lineal Model approach.

Keywords: fMRI; Statistics; Nonparametrics; Signal-analysis

1. Introduction

The tools that are employed for fMRI data analysis can be divided in two main categories: model-based and model-free tools. The best known representative of the former is the General Linear Model (GLM) [1]. A GLM represents every effect assumed to be present in the recorded fMRI signal by a single regressor that, in addition, is convolved with a so-called haemodynamic response function (HRF) to model the haemodynamic delay of the brain [2]. This GLM can be solved with the ordinary least squares (OLS) method to obtain the regression coefficients for each voxel which, when properly combined with the regression error, returns a single t-value that expresses the responsiveness of the corresponding voxel.

While it has proven very successful and useful, this GLM approach should be applied with care given the assumptions behind it. One such point of concern is whether the non-linear relationship between the fMRI measurements and the neuronal activity is correctly mapped by the assumed linear transform model, on which the GLM relies [3,4]. Another point of interest concerns the question whether the recorded data can be analysed using Gaussian tests [5,6]: the GLM's residue must have a Gaussian distribution to obtain a valid analysis [7]. Also the commonly applied Gaussian Random Field theory (GRF, [8]) or the recently developed Discrete Local Maxima (DLM, [9]), which are used to assess the statistical significance, rely on assumptions of Gaussianity. To meet this assumption, one often uses a (Gaussian) smoothing as a pre-processing step (see, *e.g.*, [10]). In addition, it is proven that the OLS renders too optimistic results [11].

Given these concerns, and the fact that nonparametric tests are the only kind of tests that are invariably valid and exact when the nature of the data is unknown [5], we explore the application of nonparametric tests. Albeit that there already exists a fair amount of literature on the application of nonparametric tests to fMRI, the largest share deals with nonparametric alternatives to the GRF. Only a small fraction deals with the detection of activation.

[*] Corresponding author. Fax: +32 16 345960.
E-mail addresses: patrick.demaziere@med.kuleuven.be (P.A. De Mazière), marc.vanhulle@med.kuleuven.be (M.M. Van Hulle).

1090-7807/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmr.2006.12.001

The Kolmogorov–Smirnov test (*KS*) and the Mann–Whitney test (*MW*) are often mentioned within this context [6,12], and are included in some fMRI analysis tools, *e.g.*, Lyngby [13] and AFNI [14]. However, the presence of serial correlations (temporal autocorrelations) invalidates the straightforward application of these nonparametric tests to fMRI data [6].

We propose in this article an extension to rank-order tests for handling the serial correlations issue properly, so that these tests can be applied to fMRI. We investigate the following rank-order tests in this article: the Mann–Whitney test (*MW*), the Kolmogorov–Smirnov test (*KS*), and the Cramér–von Mises test (*CvM*). The article's structure is as follows: we first introduce the data sets used for the validation and comparison of our novel extension with some existing analysis methods, followed by a brief review of the properties of the implemented tests, and how nonparametric statistics are to be applied to fMRI data. Next, we introduce the proposed extension and evaluate it using the permutation test as a reference, together with the popular GLM technique. Finally, we consider a number of analyses of real fMRI data, discuss our findings and formulate our conclusions.

2. Materials and methods

2.1. Synthetic and real fMRI data sets

The synthetic data sets are created by autocorrelating Gaussian noise using a first order autoregressive model (*AR(1)*) that very closely resembles the autocorrelation structure of real fMRI data. The scheme used to create these data sets is analogous to that presented in Gautama and Van Hulle [11]: $N = 10,000$ time series $x(t)$ of length $T = 420$ are generated using an *AR(1)* model by implementation of the formula $x(t) = \rho_1 * x(t-1) + u(t)$ where $u(t)$ represents a white Gaussian noise (WGN) source.[1] The value ρ_1 is set to 0.4, a value which is very representative for fMRI signals [15].

We use also two real world fMRI data sets. One is the publicly available fMRI null data set created by the Brain Mapping Unit (University of Cambridge, UK), and is commonly referred to as the "BAMM" data set. The other one is data coming from a real activation study performed by Vanduffel and colleagues as described in Vanduffel et al. [16]. This study investigated which monkey brain areas were involved during three-dimensional structure-from-motion processing (3D-SFM). Each functional scan (time series) consisted of gradient echo-planar whole-brain images (EPI; TR 2.3 s; TE = 32 ms; $2 \times 2 \times 2$ mm^3 voxels; 64×64 matrix; 32 sagittal slices).[2] We applied non-linear realignment procedures [17,18] to these monkey images. For the purpose of validation, no additional smoothing was applied, except for the one introduced during the realignment.

We use only the synthetic and BAMM data sets for the quantitative evaluation. To obtain useful fMRI signals, we first realigned the BAMM images using the SPM99 software [19]. Next, we applied the brain extraction tool of FSL [20] to obtain grey matter images from which the time series are extracted in random order. For the BAMM data set, pre-processed time series were concatenated to obtain noise time series of the right length, *i.e.*, 420 values. For the synthetic data set, we set TR = 3 s, a value identical to that of the BAMM data set. These time series—both the ones extracted from the BAMM images and the synthetic ones—are detrended using a second order polynomial and standardised, *i.e.*, with zero mean and unit variance. Active time series are then simulated by adding an on–off block-pulse to these signals (synthetic/BAMM) in a ratio *1:noise-level*. We varied the noise-level in our experiments from 2 to 8 in steps of 0.5. The used block-pulse contains 14 alternating blocks of activity (on) and inactivity (off). Each block contains 30 values. Activity is represented by ones and inactivity by zeroes. We thus mimic 14 epochs of 30 scans each, summing up to a total length of 420 values. Whenever HRF modelling is applied, a haemodynamic delay (HD) of 7 s was chosen. The HRF employed to perform this modelling, and with which the block-pulse is convolved, is the one used in SPM99 (parameter values for this HRF can be found in, *e.g.*, Worsley et al. [15]).

2.2. GLM-based fMRI data analysis

To properly express the brain activity using the standard GLM approach we need accurate estimates of the GLM parameters to obtain an accurate statistical *t*-value, and in turn a proper marking of the active brain regions as fully explained in, *e.g.*, [15]. In the remainder of this paper, we represent a GLM mathematically by $\mathbf{V} = \mathbf{X}\beta + \mathbf{e}$, with $\mathbf{V}$ the fMRI signal, $\mathbf{X}$ the design matrix containing the regressors, $\mathbf{e}$ the error which is assumed to have a normal distribution, and β the regression coefficients to be estimated.

To obtain such accurate estimates, one must correct for the presence of serial correlations in fMRI signals, either by "colouring" the data with a band-pass filter [21], or by using a two-step pre-whitening method: first, one applies an OLS to obtain β and $\mathbf{e}$, of which the serial correlation is estimated (see Section 2.3). Second, one uses this serial correlation estimate to perform a decorrelation of both the fMRI signal $\mathbf{V}$, and (column-wise) of the GLM's design matrix $\mathbf{X}$, prior to the final calculation of β and $\mathbf{e}$. This two-step method is referred to as the Durbin–Watson method [7,22]. In case one applies this method iteratively to reduce the remaining serial correlation, it is referred to as the Cochrane–Orcutt method (OLS-CO) [7,23]. In fMRI, the Durbin–Watson method is frequently applied since Bullmore et al. noted that the improvements gained by

[1] We performed the experiments using Matlab, Mathworks Inc. As WGN source we used Matlab's built-in random number generator.

[2] TR represents the repetition time or the time between successive whole brain scans; TE represents the time delay between excitation and echo maximum.

the Cochrane-Orcutt method are rather limited [24]. The pre-whitening method is advised whenever the serial correlation is accurately known, since it returns the best linear unbiased estimates (BLUE) of the GLM parameters [25] and thus a more accurate estimation of the brain activity.

The t-value is then calculated as the product of β and the contrast **c** divided by a term that is function of **e** [21]. The contrast vector **c** specifies which stimuli (represented by 1s) are compared to which other stimuli (represented by -1s), and which stimuli are not considered at all (represented by 0s). We further refer to this method as OLS-(CO)/t.

2.3. Serial correlations and (rank-order) statistical tests

Obviously, the efficiency of the pre-whitening method is determined by the accuracy of the serial correlation estimation. For this reason, and since such estimation is fundamental to our novel method as well, we first review here the estimation of serial correlations, a topic which has attracted quite some attention in fMRI literature: *e.g.*, [11,21,25,26].

If Eq. (1) represents a time series $x(t)$ with $t = 1, \ldots, T$, whose autocorrelation is modelled by an autoregressive model of order $k(AR(k), k < T)$, the raw autocorrelation estimate, $\hat{\rho}_\tau$, at lag τ is given by Eq. (2):

$$x(t) = \rho_1 x(t-1) + \rho_2 x(t-2) + \cdots \rho_k x(t-k) + u(t) \quad (1)$$

where $u(t)$ is normally distributed with

$$\bar{u}(t) = 0, \quad \sigma^2_{u(t)} = \text{constant } \forall t,$$

$$\sigma_{u(t)u(t-s)} = 0 \quad \forall t, \ \forall s \neq 0$$

$$\hat{\rho}_\tau = \frac{\sum_{t=\tau+1}^{\mathrm{T}} x(t)\ x(t-\tau)}{\sum_{t=1}^{\mathrm{T}} x^2(t)}, \quad \tau \in \mathbb{N},$$

$$x \text{ unit standardised.} \quad (2)$$

While τ expresses the time over which the value of a time series **x** at time t is influenced by another one, ρ_j expresses the amount of that influence at lag j. With respect to fMRI, ρ and the lag for which the autocorrelation value is the highest depend, among other factors, on the value TR, which varies between 2 and 7 s for human fMRI studies. Higher TR values are typical for older data sets, while smaller TR values are used more recently due to the advanced MRI scanner technology. Serial correlations cannot be neglected whenever TR drops below 5 s. Bullmore et al. has stated that *AR(1)* models are appropriate for modelling these autocorrelations [27]. Recently, some authors have proposed methods to correct for serial correlations using *AR* models up to an order of four [15]. The need for such corrections has been called into question by Woolrich et al., who demonstrated that the order of valuable *AR* models varies between 0 and 5, but with only a few voxels requiring an order greater than 2 [25].

Several authors [11,25,27] have suggested additional routines to increase the accuracy delivered by the raw serial correlation estimate. Examined methods embrace single and multi-tapering in combination with high-pass filters, non-parametric estimation techniques, usage of the partial autocorrelation coefficient (PACF, [28,29]) and autoregressive parametric models with $k > 1$. Woolrich et al. found that a high-pass filter in combination with a single Tukey taper performed best [25].

Given the fact that we are considering only box car designs in this paper, we prefer the autoregressive model estimator given its simplicity. In addition, we use the PACF to determine the best value for k. Indeed, when fitting an *AR(k)* model to the time series, the last partial coefficient, α_k, measures the excess correlation at lag k not accounted for by an *AR(k − 1)* model. α_k plotted for all k is the PACF. The lowest value of k for which α_k is not significantly different from zero (using Bartlett's 95% confidence limits of approximately $\pm 2/\sqrt{T}$) then specifies the order to be used. We prefer Bartlett's test [30] since it can test every individual (partial) autocorrelation coefficient for being significantly different from zero or not. Conversely, the better known Box–Pierce Q-statistic tests whether all (partial) autocorrelation coefficients are derived from a white noise process. For Bartlett's test, the confidence intervals are calculated as follows: if a series of length T is generated by a white noise process, the estimates of the (partial) autocorrelation coefficients are approximately normally distributed random variables with zero mean and variance $1/T$. The confidence limits are then equal to $\pm z_{1-\alpha/2}/\sqrt{T}$, with α the desired significance level and z the percent point function of the standard normal distribution. For a 95% confidence interval these limits approximate $\pm 2/\sqrt{T}$.

As an illustration, we show the amount of autocorrelation present in the BAMM data set as a function of the lag τ assuming an *AR(1)* model. For this purpose, we selected at random 1000 signals from the BAMM data set, which is pre-processed as explained in Section 2.1. Next, Eq. (2) is used to calculate $\hat{\rho}_\tau$ for $\tau = 0, \ldots, 14$ for each signal. The average of the 1000 obtained $\hat{\rho}_\tau$ values is then displayed in Fig. 1. Necessarily, a value of one is obtained for lag zero. From the literature (*e.g.*, [15]), and confirmed by Fig. 1, it is clear that in general the amount of autocorrelation does not exceed $\hat{\rho} = 0.4$ at lag 1, while the serial correlations at higher lags decrease rapidly when using an *AR(1)* model. The autocorrelation structure depicted in our figure, including the strange effect that the autocorrelation becomes negative for higher lags, corresponds very well with previous findings for fMRI data as reported by, *e.g.*, Bullmore et al. [27].

2.4. Nonparametric tests for analysing fMRI data

We opt for rank tests that are based on an *empirical distribution function* (EDF), since they are known to be the most powerful nonparametric ones [31], and since they can be easily extended to handle serial correlations (cf. infra). Both the *KS*- and *CvM*-test are EDF tests [31]. The *MW*-test is not, but its statistical values are obtained

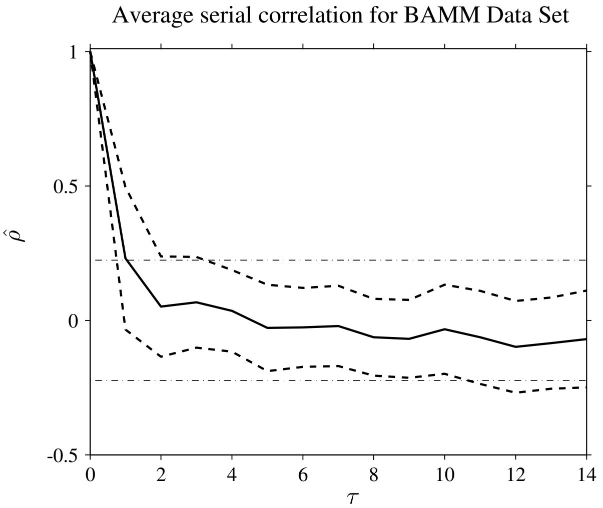

Fig. 1. Autocorrelation plot of the BAMM fMRI null data set. The figure is created by randomly drawing 1000 signals from this data set and plotting the average $\hat{\rho}$ as a function of τ using an *AR(1)* model (*see text*). The dashed lines represent one standard deviation. The horizontal dash dotted lines are Bartlett's approximate 95% confidence interval ($\pm 2/\sqrt{T}$).

in a very similar manner, and we therefore include that one too. The theoretical formulae behind these tests are given in Appendix A.

In order to apply nonparametric tests to fMRI data, the data points must be reorganised in a specific way whenever multi-condition ($\geqslant 2$) paradigms or contrasts are used. The reorganisation we propose here is suited for any nonparametric test that is able to compare two samples of data points ($\{X_i\}$ and $\{Y_i\}$) that possibly contain an unequal number of data points. To explain the method, we use the following example: a study contains six experimental conditions labelled A through F, and the desired contrast equals: $A + B + D - E$. In order to test for a significant difference between the evoked response of conditions A, B and D versus the response of condition E, we put in sample $\{X_i\}$ all data points recorded under conditions A, B and D, while $\{Y_i\}$ contains those recorded under condition E. The fact that one sample can contain more data points than the other one is correctly dealt with by the respective statistical tests, thus omitting the need to balance the samples. In the case where HRF-convolved time series are used, we have opted for a very simple approach to cope with the haemodynamic delay: we skipped the transitional scans from the actual statistical analysis, *i.e.* the first $\lceil \mathrm{HD}/\mathrm{TR} \rceil$ data points[3] of every epoch.

2.4.1. Assumptions with rank-order tests

Three assumptions need to be satisfied when applying the *KS*-, *CvM*- and *MW*-tests [31]: the measurement scale should be ordinal, the random variables should be continuous, and the data points should be exchangeable. Due to serial correlations within fMRI signals, all assumptions but the third are fulfilled: since the theoretical significance thresholds for EDF tests are calculated under the assumption of white noise, these tests fail when applied to autocorrelated data. In the next section we therefore introduce our novel extension for handling correlated data.

2.4.2. The permutation test

The permutation test, which is used as a reference for our method, was introduced by Holmes et al. for PET data [5], and later on extended for fMRI. A discussion of permutation tests for fMRI data can be found in Nichols and Holmes and references therein [32]. The permutation test we adopted is the one proposed by Liu et al. [33]. It uses the data itself to extract a proper null distribution, which is used to calculate the proper significance thresholds. For this reason, the permutation test is a better reference than the overly optimistic OLS/t-test [11]. This permutation test in fact randomises the labels (conditions), instead of the data values themselves, to ensure that the temporal autocorrelation structure is preserved within each permuted time series. To obtain a reliable null distribution, we opted to draw 1000 permutations for each synthetic/BAMM fMRI signal. The statistical significance value for a permutation test is obtained as follows: the statistical value (OLS/t, *MW*, *KS*, or *CvM*) is calculated for every permuted fMRI signal using the identical experimental paradigm and contrast. The p-value for the original, non-permuted signal is then the proportion of the distribution that is at least as extreme as the observed test. Given the layout of this algorithm, the time complexity is $P = 1000$ times that of the statistical test itself, with P being the number of permutations performed. The time complexity for a whole brain analysis equals then $N \times P \times Q$, with N the number of voxels, and Q the statistical test's complexity.

2.5. Novel extension for EDF-like tests to handle serial correlations

At least to the best of our knowledge, we developed a completely novel extension for EDF-based tests to correct for serial correlations. It depends on the value $\tau_{\max}$, which represents the maximum lag one wants to correct for. The value of $\tau_{\max}$ can be determined using the PACF and Bartlett's approximate 95% confidence interval: first, the PACF (Section 2.3) is applied to find the best value for k, with $AR(k)$ the model used to fit the fMRI signal. Next, we again use Bartlett's approximate confidence interval to define $\tau_{\max}$ as the highest lag for which the amount of autocorrelation is still significantly different from zero [27]. This confidence interval is represented in Fig. 1 by horizontal dash dotted lines. Our extension is applicable to any value of $\tau_{\max}$. For the sake of clarity, and without loss of generality, we use a contrast equal to $A - B$, and apply a unit lag autocorrelation correction ($\tau_{max} = 1$). According to the method of Section 2.4, the fMRI signal $\mathbf{V}$ is split into the samples $\{X_i\}$ and $\{Y_i\}$. We consider only $\{X_i\}$, since

[3] $\lceil x \rceil \equiv$ smallest integer $\geqslant x$.

$\{Y_i\}$ can be treated analogously. We divide $\{X_i\}$ into $\tau_{\max} + 1 = 2$ parts, labelled $\{X_i^{1*}\}$ and $\{X_i^{2*}\}$:

$$\begin{aligned} X_i^{1*} &= X_{2k}, \quad k = 1, \ldots, \lfloor N_X/(\tau_{\max}+1) \rfloor, \\ X_i^{2*} &= X_{2k+1}, \end{aligned} \tag{3}$$

with $\lfloor x \rfloor \equiv$ largest integer $\leqslant x$. This partitioning sees to it that $\{X_i^{1*}\}$ (or $\{X_i^{2*}\}$) no longer has the original unit lag correlation, and becomes thus exchangeable. For both $\{X_i^{1*}\}$ and $\{X_i^{2*}\}$ (and $\{Y_i^{1*}\}$ and $\{Y_i^{2*}\}$), one can calculate the *p*-values using the *MW*-, *KS*-, or *CvM*-test. To obtain a single *p*-value for the complete fMRI signal, we need to combine the different $(\tau_{\max} + 1)$ *p*-values. Two possible solutions are discussed in this article: (a) a simple multiple comparison correction (MCC) method and (b) an algorithm from the *meta-analysis* domain. Meta-analysis, "the analysis of analyses", which is quite often used in the field of experimental psychology, is the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings. In fact, even *p*-values obtained by applying different statistical tests can be combined, as long as the hypothesis is identical. We first discuss both methods theoretically and then show the results obtained with each of them.

2.5.1. MCC method to combine p-values

Given the idea that the different *p*-values are obtained by performing identical tests, the evidence for using MCC is rather clear. Personal communication with Benjamini and Yekutieli—the authors of the False Discovery Rate (FDR) technique [34]— confirmed that a simple ”Simes FDR test for the intersection hypothesis” is valid for obtaining a single corrected p^*-value:

(a) Order the *p*-values as follows: $p_1 \leqslant p_2 \leqslant \cdots \leqslant p_{(\tau_{\max}+1)}$.
(b) $\forall i, \exists j | p^* = p_j \times (\tau_{\max} + 1)/j$ and $p_j \geqslant \max_i \left(p_i \times \binom{j}{i}\right)$.

2.5.2. Meta-analysis to combine p-values

A good introduction to meta-analysis can be found in Wolf [35]. We selected the Stouffer combined test [35,36] given its simplicity and rather straightforward implementation, which avoids logarithmic transformations or adjustments of the degrees of freedom. The Stouffer combined test converts each "partial" p_i-value into a *z*-value; since every p_i-value has an identical probability for occurring, the p_i-values are uniformly distributed and can be transformed into *z*-values under the null hypothesis. Compared to the classic meta-analysis approach, where *t*-values are summed, this approach is slightly more powerful [35]. If we denote the *z*-value derived from the p_i-value by z_i, we can calculate the global *z*-value, denoted by z^*, as:

$$z^* = \sum_{i=1}^{\tau_{\max}+1} \frac{z_i}{\sqrt{\tau_{\max}+1}}, \quad i = 1, \ldots, (\tau_{\max}+1), \tag{4}$$

where $\tau_{\max} + 1$ equals the number of combined tests, thus the number of partial samples examined. The global *p*-value, *p**, can easily be derived from this z^*. Besides its computational simplicity, the Stouffer combined test offers another, more conceptual advantage: the *t*-maps, as created by the GLM-based analysis tools, are easier to compare to the calculated *z*-values, than to the statistical values obtained by the FDR extension (*MW*, *KS*, or *CvM* values). However, we still have to verify whether the different p_i-values, extracted from the different partial time series, are exchangeable, as required for a valid meta-analysis. Experiments carried out in the next section will verify this assumption.

3. Results of the validation and comparison experiments

In order to validate and compare the proposed extensions to existing tools like the GLM and the permutation test, we calculate and display both the true and false positive rates (FPR and TPR, respectively) and the receiver–operator characteristic (ROC) curve: for the FPR scenario, a bare noise signal (synthetic/BAMM) is used to which *no* block-pulse (see Section 2.1) is added but which is analysed as if a block-pulse were present (using noise-level = 1 and $\tau_{\max}$ values in the discrete range [0, 5], with $\tau_{\max} = 0$ representing no serial correlation correction at all). A rejection of the null hypothesis, which states that no activation is present, is therefore a false positive. For the TPR scenario, a block-pulse is added to noise (synthetic/BAMM) in a ratio 1:noise-level (with noise-level varying between 2 and 8 in steps of 0.5) and examined as such. A rejection of the same null hypothesis now refers to a true positive. Every statistical value (or the thereof derived values and conclusions) is based on 10,000 time series or iterations, which allows us to apply the statistically common threshold of 0.01. This nominal α guarantees that, at least theoretically, 100 cases should pass the test, which is a significantly perceptible amount. Given the exactly known properties of the synthetic data set, we use this data set to examine the validity of both extensions; we use the BAMM data set to illustrate the extension's capabilities with respect to real fMRI data, *i.e.*, fMRI noise. In order to introduce all tests and demonstrate the serial correlation problem, we first display the uncorrected statistical tests using white Gaussian noise (WGN), and the synthetic and the BAMM data sets.

In Fig. 2(a) and (b), we show the TPR curves for all tests without application of any serial correlation correction for the nonparametric tests for WGN and the synthetic data set, respectively. ROC curves for the *MW*- and *CvM*-tests using both extensions are shown in Fig. 5 for a noise-level equal to 5 and for $\tau_{\max} = 0, 1, 2, 3$, with $\tau_{\max} = 0$ meaning no correction. As was to be expected according to [31,37], the difference in power between the nonparametric and parametric GLM-based tests (OLS and OLS-CO) is larger for Gaussian data (Fig. 2(a) and (b)), than for non-Gauss-

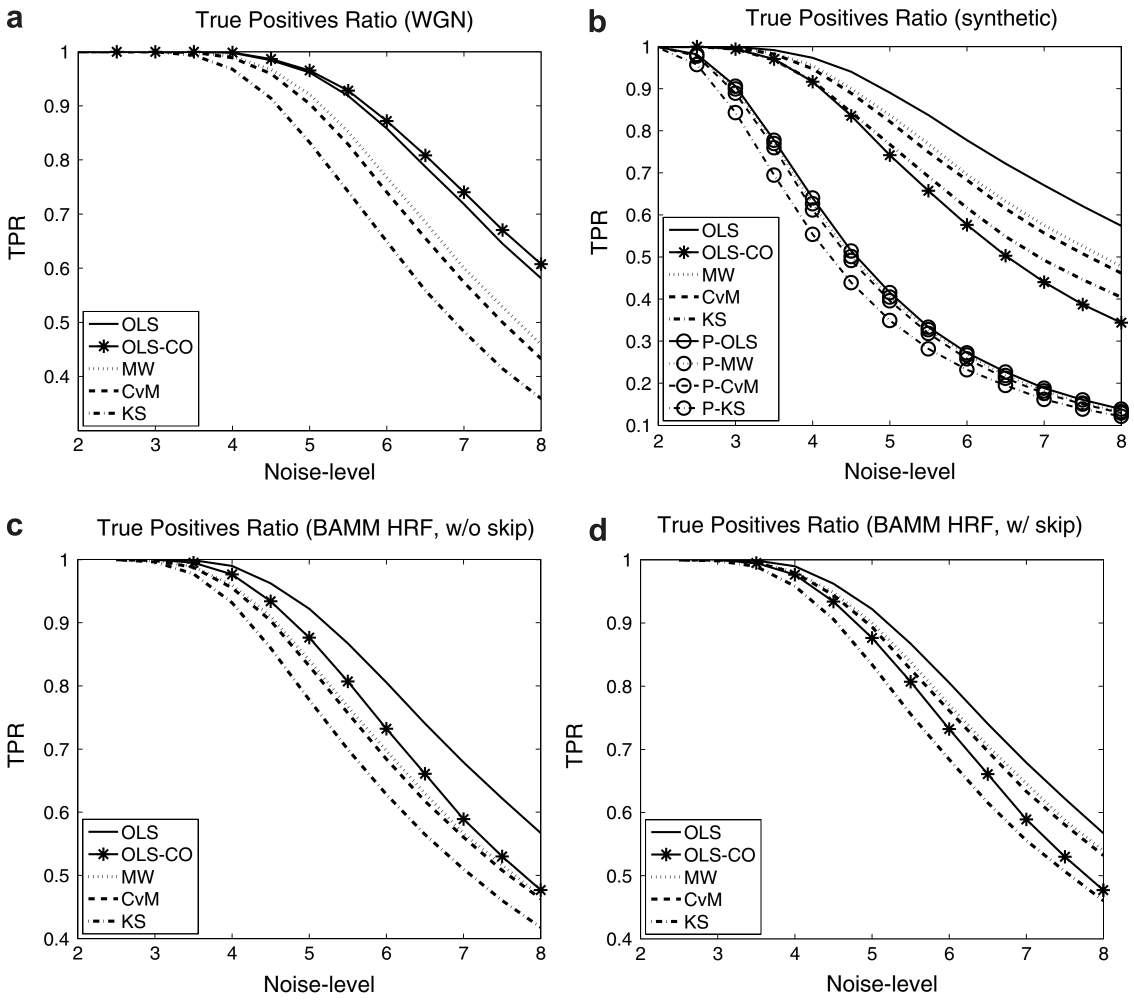

Fig. 2. Power of the OLS/*t*, *MW*-, *KS*-, and *CvM*-test as a function of the amount of noise (noise-level) for different kinds of noise: (a) WGN, (b) synthetic data set, and (c) and (d) the BAMM data set with a HRF convolved block-pulse. (c) Displays the outcome when no HRF correction is applied to the nonparametric tests, while (d) displays the outcome with a HRF correction (see text). In addition, permutation tests are referred to in the legend of (b) by a '*P*-' that precedes the name of the used statistical test (TPR = 1 corresponds to 100% true activations).

ian data, *i.e.*, the BAMM data set (Fig. 2(d)). Next, Fig. 2(b) indicates that the *MW*- and *CvM*-tests have more power than the *KS*-test. The corresponding FPR values for this setup are: 0.016 (OLS-CO), 0.061 (OLS), 0.065 (*KS*), 0.073 (*CvM*), and 0.073 (*MW*). Based on these values and Fig. 2(b), we can state that the *MW*- and *CvM*-tests turned out to be more powerful than the *KS*-test, with the *MW*-test slightly ahead of the *CvM*-test, since the TPR gained is larger than the FPR lost. Given these results and the fact that the *KS*- and the *CvM*-test have an identical hypothesis, we further omit the *KS*-test. This is also the reason why we omitted ROC curves for the *KS*-test. Finally, the bottom row of Fig. 2 demonstrates that the TPR curves are higher for all nonparametric tests when the transitional scans are skipped (Fig. 2(d)), than when no haemodynamic delay correction is applied (Fig. 2(c)). Leaving the transitional scans out of any nonparametric analysis is thus recommended in a case where we deal with real fMRI signals.

3.1. Validation of the FDR extension using synthetic data

To verify the validity of our extensions, we display the TPR curves and FPR curves using the synthetic data set. The TPR and FPR curves for the *MW*- and *CvM*-tests are displayed from right to left in Fig. 3. The lag we have corrected for in the TPR figures, *i.e.*, the value of $\tau_{\max}$, is represented by the number behind the *MW*/*CvM* notation in the legend. If no serial correlation correction is applied ($\tau_{\max} = 0$), no number is displayed behind the test's abbreviated name in the legend of the figure. The permutation test (represented in the legend by a '*P*-' that precedes the test's abbreviated name) and the GLM-based tests are displayed for the sake of comparison.

Inspection of the FPR values for the uncorrected case ($\tau_{\max} = 0$ at Fig. 3(c)), teaches us that the OLS/*t*-test fails for the nominal size of rejections (0.01), as expected, while the OLS-CO/*t*-test better controls the FPR. The FPR values for the uncorrected nonparametric statistical tests

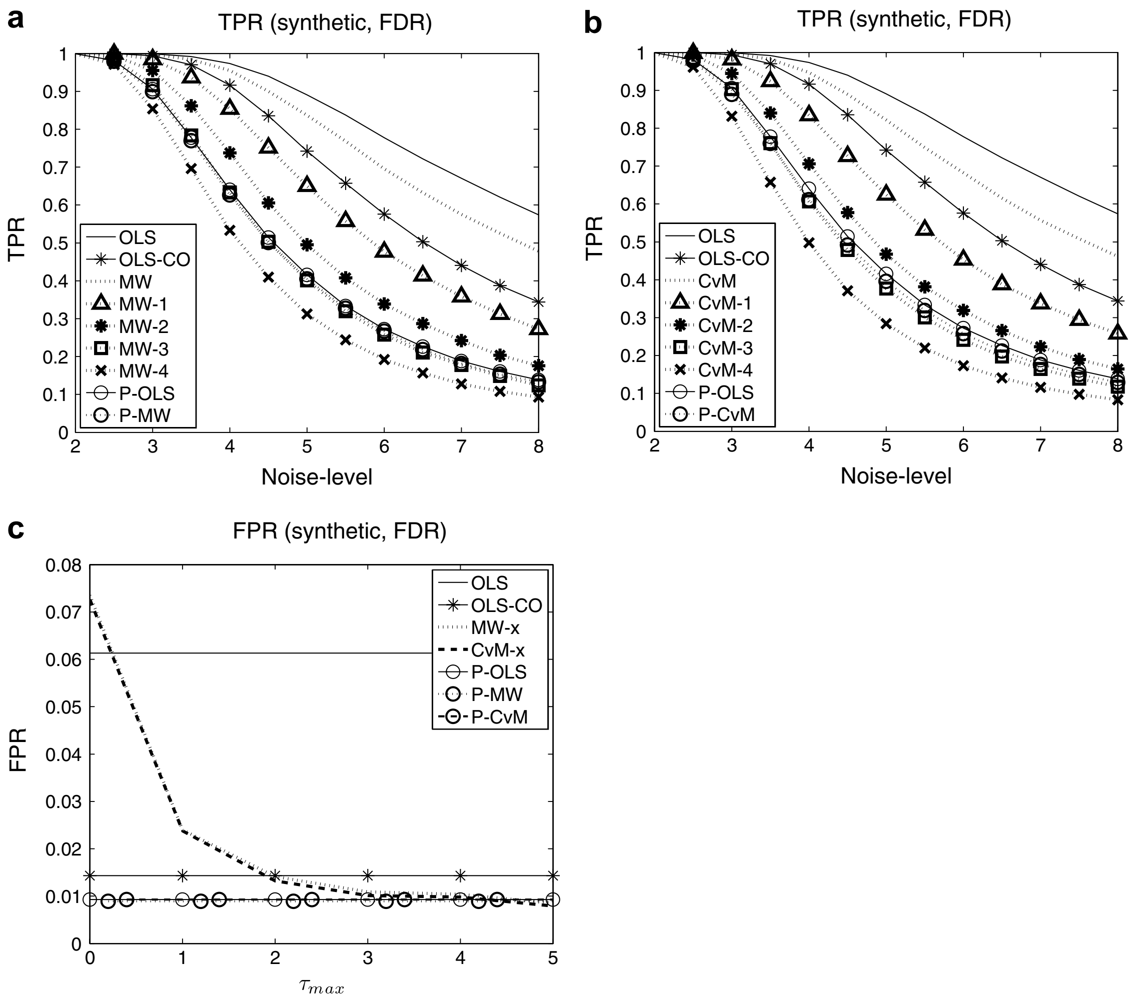

Fig. 3. TPR and FPR curves for the OLS-CO/*t*-, *MW*-, and *CvM*-tests using the synthetic data set. The FDR extension is used to correct for serial correlations. (a) and (b) depict the TPR curves as a function of noise-level. The lag we have corrected for is represented by the number behind the *MW*/*CvM* notation in the legend. If no number is given, *i.e.*, $\tau_{\max} = 0$, no correction is applied. (c) Displays the FPR values for all tests (y-axis, $1 \equiv 100\%$).

deviate markedly from the nominal size. This is in agreement with the statement that the number of false positives (for the *KS*-test, [6]) is higher than that of the *t*-test. Our simulations confirm this and extend this finding to the *MW*- and *CvM*-tests. Considering the $\tau_{\max} \neq 0$ case, a significant decrease in FPR is already visible for a unit lag correction. For a lag three correction, we see that, for both the TPR and FPR curves, our correction method coincides with the values obtained with the corresponding permutation tests. This implies that, using a lag three correction, our method has the same level of performance as the permutation test. Furthermore, our simulations demonstrate that the OLS-CO/*t*'s FPR is slightly larger than the nominal size. This is not surprising given the fact that the OLS is rather optimistic [11].

3.2. Validation of the meta-analysis extension using synthetic data

We show in Fig. 4 the results obtained when applying the meta-analysis extension instead of the FDR extension. As could be expected, the TPR curves for the nonparametric tests are now higher than those obtained with the FDR extension. However, the FPR curves for these tests are (dramatically) higher too: this extension clearly fails to control the FPR appropriately, even for a $\tau_{\max} = 5$ correction. Furthermore, the TPR curves no longer coincide with those of the corresponding permutation tests. This confirms that the meta-analysis extension is far too optimistic, and consequently, not a recommended procedure for dealing with serial correlations.

Separate TPR and FPR curves can be combined into a single receiver operator characteristic (ROC, [38]). An ROC curve is easy to interpret since a larger area beneath the curve indicates that the considered test better detects the true activity in the signal. The area can be maximally equal to one (a perfect test) and should not be less than 0.5 (the performance of a random guess). An ROC curve is obtained by plotting the TPR and FPR values for nominal α values in the range $[0, 1]$. The TPR and FPR values are calculated using their respective signals as explained in the first paragraph of

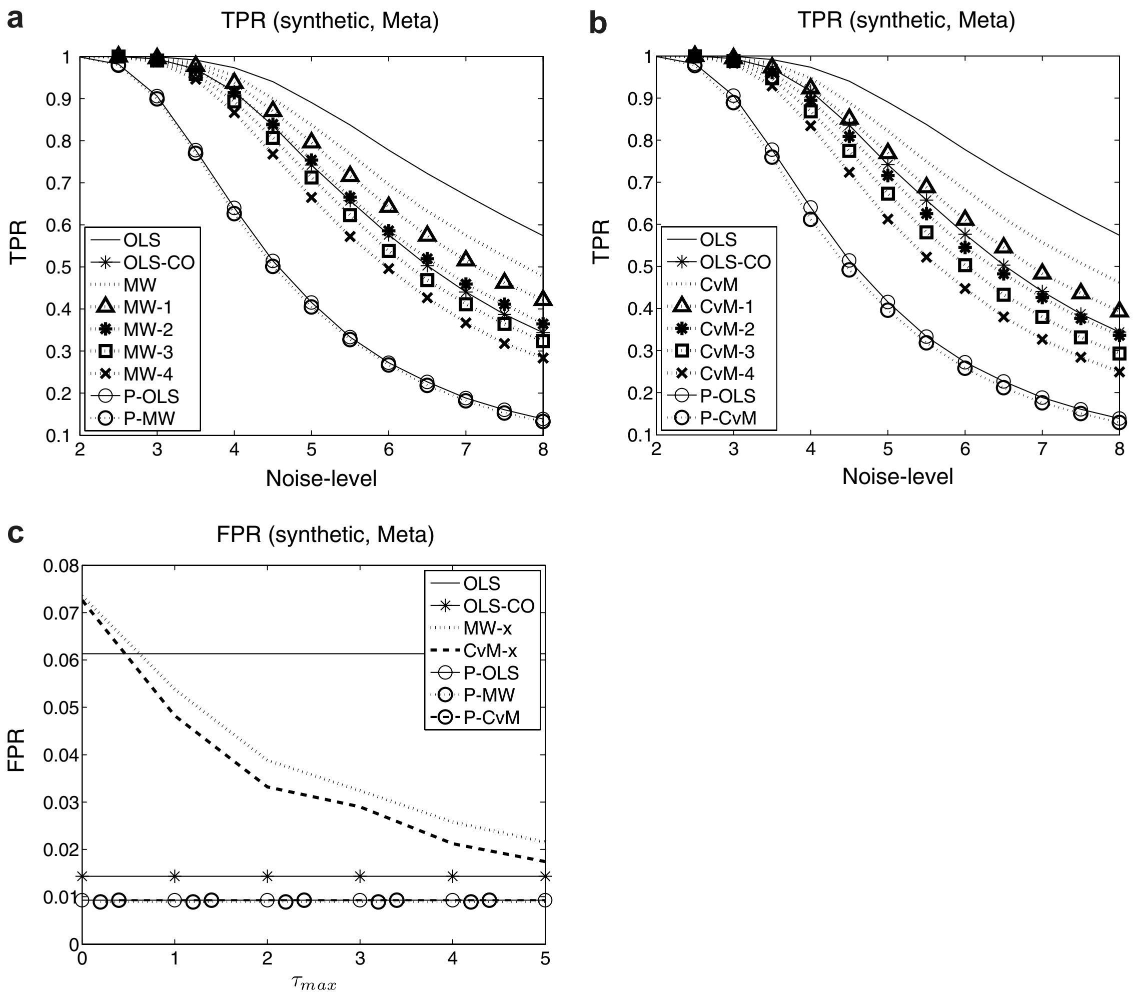

Fig. 4. TPR and FPR curves for the OLS-CO/t-, MW-, and CvM-tests using the synthetic data set. The meta-analysis extension is used to correct for serial correlations. (a) and (b) depict the TPR curves as a function of noise-level, while (c) displays the FPR values. The same conventions are used as in Fig. 3.

this section. To preserve a clear display, we show in Fig. 5 only the ROC curves for the MW-test (top row) and CvM-test (bottom row), as obtained for synthetic signals with a noise-level equal to 5. Again, we see that the nonparametric statistics, which are based on the FDR extension, have less power than the OLS-CO/t-test, but show a performance similar to that of their corresponding permutation test. This coincidence is better than the coincidence between the OLS-CO/t-test and its corresponding permutation test. In addition, these curves confirm the difference between the FDR (Fig. 5, right column) and meta-analysis extension (left column) as already deducible from the separate TPR and FPR curves: the ROC curves for the meta-extension are higher on the y-axis than those for the FDR extension (higher sensitivity), but they are also shifted to the right (lower specificity). This shift to the right is visible by tracking the leftmost mark of the ROC curves for the nonparametric statistics: with respect to the meta-analysis extension, this mark is more to the right than the one of the FDR extension.

3.3. Summary

Considering both extensions, we can state that the proposed signal splitting scheme clearly decreases the FPR rates, but that only the FDR extension guarantees that the nominal size, 0.01, is achieved for reasonable $\tau_{\max}$ values, and that the obtained ROC curves better match the ones of the corresponding permutation tests. Higher $\tau_{\max}$ corrections are to be avoided given the rather limited length of epochs in practice. We can summarise (for this data set and using the FPR curves as primary constraint) that only the FDR extension is suitable for fMRI analyses, that for a $\tau_{\max} = 1$ correction a better false positive control is achieved than for the standard OLS/t-test, and that a $\tau_{\max} = 2$ correction suffices to obtain a FPR equal to that of the OLS-CO/t-test, but that only the $\tau_{\max} = 3$ correction returns a reasonable false positives control. Considering also the TPR curves (or ROC curves) we see that the performance of the FDR extension very well coincides with that of the corresponding permutation test, and even with that of the GLM-based permutation test. Moreover, taking

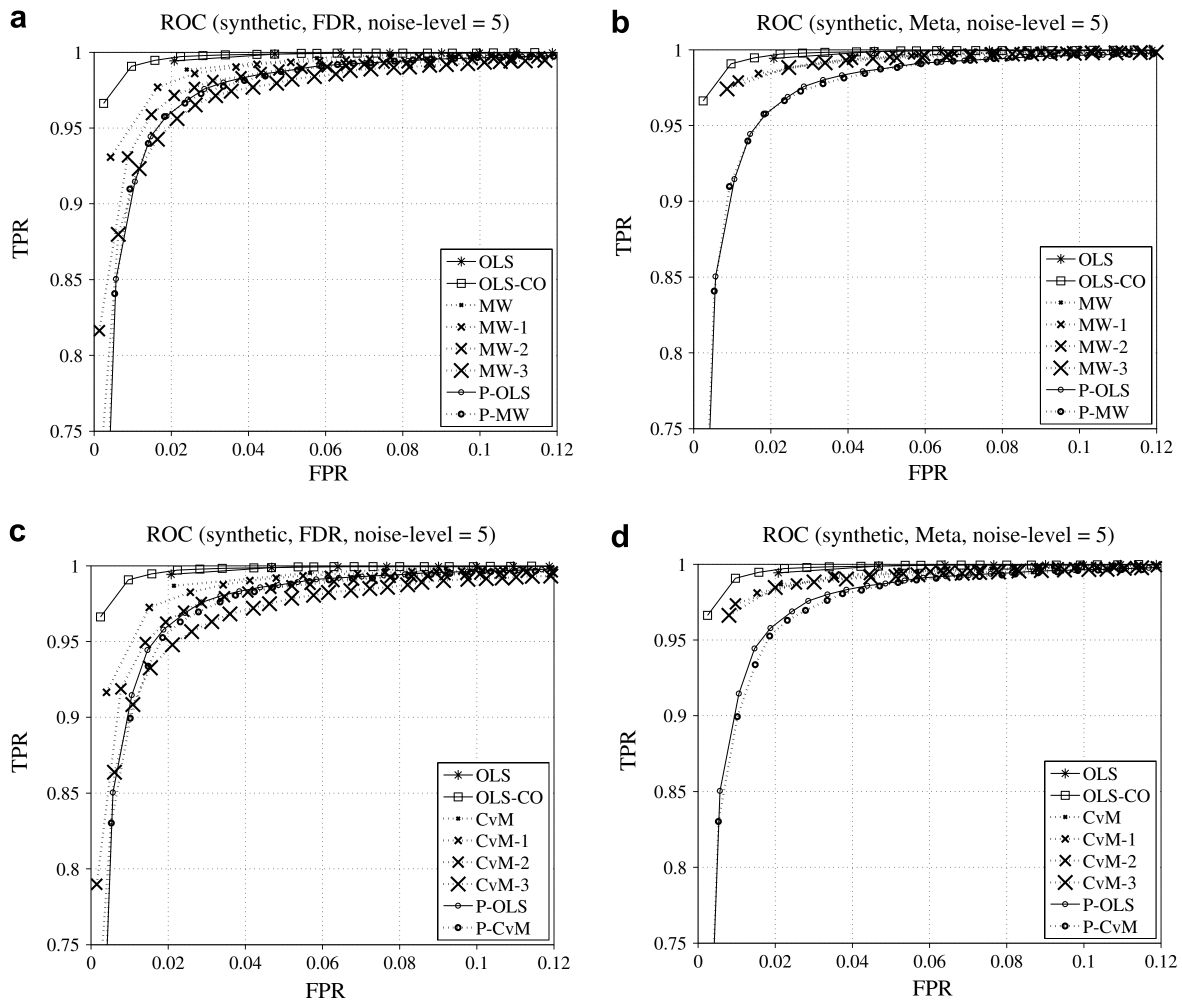

Fig. 5. ROC curves for the OLS-CO/t-, MW-, CvM-, and their corresponding permutation tests using synthetic signals with a noise-level equal to 5. The top row ((a) and (b)) shows ROC curves for the MW-test, while the bottom row ((c) and (d)) contains those for the CvM-test. The ROC curves in the right column ((a) and (c)) are obtained with the FDR extension, those of the left column ((b) and (d)) with the meta-analysis extension (see text for interpretation).

the time complexity into account, which is $N \times P \times Q$ for the permutation test (see Section 3), the time complexity of a full brain analysis for our extension is only $N \times \tau_{\max} \times Q$ with $\tau_{\max} \lll P$. This renders our extension much faster.

3.4. The FDR extension and the BAMM data set

Using the BAMM data set, we verify how well the above statements hold for real world fMRI data, *i.e.*, we examine whether a $\tau_{\max} = 3$ correction using the FDR extension is sufficient to control the FPR. To mimic as best as possible the real world situation, we convolved the block-pulse with the HRF, and performed the analysis accordingly. The results of this experiment are shown in Fig. 6. We omitted the curves for the permutation tests from these figures for the sake of clarity. Identical to the experiment with the synthetic data set, a $\tau_{\max} = 3$ correction delivers again an appropriate FPR, at least as good as the OLS-CO/t-test (CvM), or even better (MW). Furthermore, we detect again a decrease in power as was the case for the synthetic data set. However, in this way we are now confident to have performed a valid analysis given the nonparametric tests that are used.

3.5. The FDR extension and real fMRI data

In this section we examine the FDR extension's behaviour and properties with respect to real world fMRI data sets containing real activity. Albeit that we do not have any ground truth for such data, we can demonstrate some advantages when we compare the obtained results using FDR-extended nonparametric statistics, with those obtained using a classical GLM-based approach.

We use the earlier described monkey data, select a single run, and analyse it with SPM99 as described in Vanduffel et al. [16], but now starting from the non-linear realigned fMRI signals, which are detrended using a first order poly-

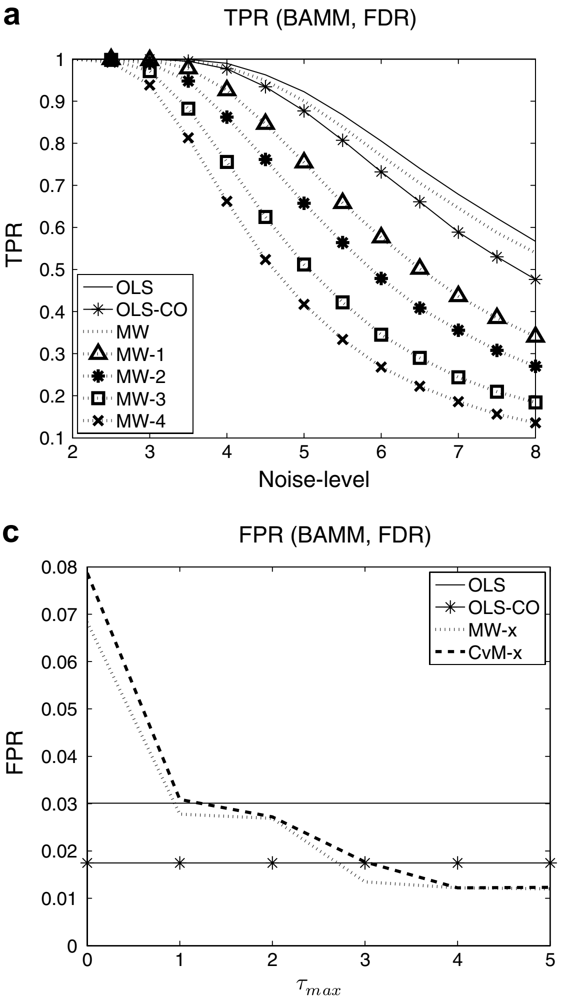

Fig. 6. TPR and FPR curves for the OLS-CO/*t*-, *MW*-, and *CvM*-tests using the BAMM data set. The FDR extension is used to perform the serial correlation correction. (a) and (b) depict the TPR curves as a function of noise-level, while (c) displays the FPR values. The same conventions are used as in Fig. 3.

nomial. Volumes are resliced to 1 mm^3 voxels. The haemodynamic delay for monkeys is estimated at 8 s [39]. The *MW*- and *CvM*-tests are applied to the same fMRI signals used for the SPM99 analysis. When analysing the grey matter fMRI signals using an *AR(1)* model, we found that a lag 2 FDR correction was mandatory. For the application of our novel method, we replaced in the SPM analysis procedure only the GLM/*t* calculation by the FDR-based *MW*/*CvM* calculation.

For the sake of clarity, the results shown are limited to a small set of axial slices. As contrast, we use 3D versus 2D motion and a threshold of $p < 0.001$ for either method. Uncorrected *p*-values are used here to avoid any influence from a statistical inference method. Looking at Fig. 7, identical regions are detected in general by either analysis. However, the size and shape of the regions does differ. The regions obtained using nonparametric statistical tests are smaller and are better delineating the anatomical/functional areas due to the absence of smoothing. This smoothing was introduced by SPM99 (GLM approach) to obtain valid results. Using nonparametric statistics we are no longer forced to smooth our data by a given factor as required with GLM-based analyses, but we can now omit smoothing or specify the amount of smoothing to optimise the signal-to-noise ratio (SNR) without being bothered by other constraints. Small regions are also no longer fused together, but clearly separated, thereby indicating the existence of separate functional regions as was to be expected [16].

In addition to this *qualitative* comparison, we performed a small *quantitative* comparison by using the coordinates of the local maxima for some well-known functional areas. Table 1 gives an overview of these maxima as found by SPM99-, *MW*-, and *CvM*-based analyses. The areas and SPM99 coordinates are adapted from Vanduffel et al. [16] since we applied a different realignment procedure. The corresponding coordinates for the nonparametric methods are found by searching the local maximum in the neighbourhood of the SPM99 coordinates. This comparison shows the advantage offered by applying different kinds

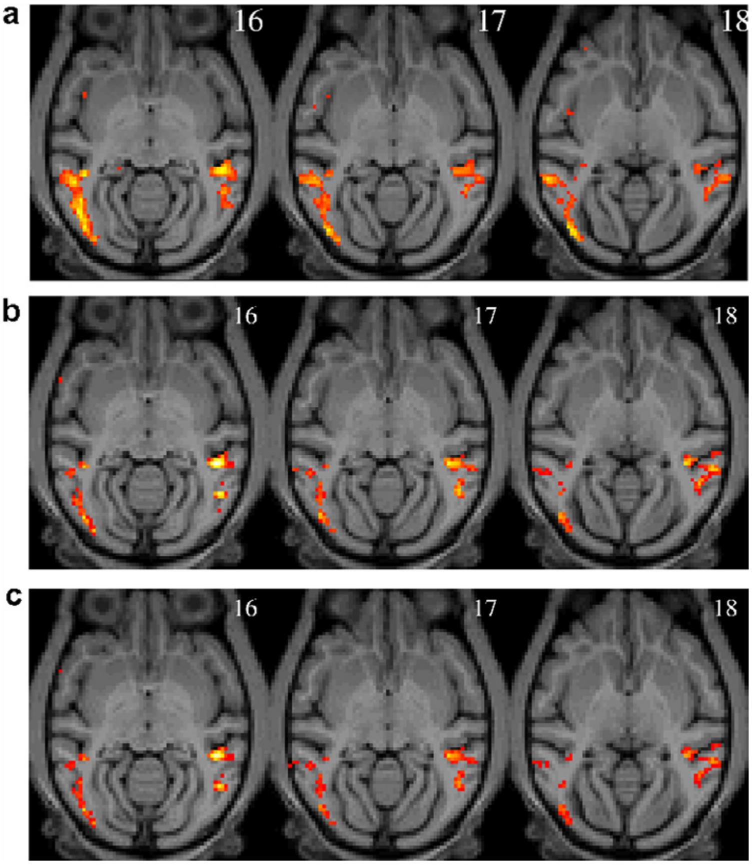

Fig. 7. Qualitative comparison of the results obtained using (a) SPM99, (b) the *MW*, and (c) the *CvM* method for the monkey 3D-SFM data set. For all methods a threshold of $p < 0.001$ on uncorrected p-values was used. The numbers in the upper right corner of the figures represent the coordinate of that slice in mm. Identical slices are displayed for each method. One can notice that the regions as detected by the non-parametric statistical analyses better match with the anatomy than the results obtained with the parametric analysis where some regions are fused together, *e.g.*, in the right posterior part.

Table 1
Quantitative comparison of the SPM99-, *MW*-, and *CvM*-analysis: the coordinates (in mm^3) of local maxima and corresponding statistical values are shown for some of the functional areas found in the 3D-SFM monkey data set using the 3D versus 2D motion contrast (*MTr*/*MTl* = right and left part of the middle temporal area; $V4r/V4l$ = right and left part of the fourth visual area)

Area	SPM99		*MW*		*CvM*	
	$[x,y,z]$	t	$[x,y,z]$	T_1	$[x,y,z]$	T_2
MTr	21, −2, 15	9.69	21, −3,15	5.75	21, −3, 15	3.69
MTl	−19, −4, 16	7.09	−19, −4, 16	5.17	−19, −4, 16	2.86
V4r	25, −4, 22	10.72	24, −4, 21	5.75	24, −4, 21	2.79
V4l	−23, −2, 21	3.49	−18, −6, 21	3.99	−18, −6, 22	2.17

of (nonparametric) statistical tests. With respect to the local maxima of areas *MTr* (the right part of the middle temporal area) and *V4r* (the right part of the fourth visual area), which are both known to be active for this contrast [16], we see that the *MW*-test declares them equally active, while the *CvM*-test declares them differently active. This indicates that the ratio of the average activation for perceiving 2D stimuli and 3D stimuli is similar for both areas, while the ratio of the distribution of the activation is different. In such cases, a more detailed look at the raw time series can reveal additional information about the exact behaviour of the respective areas with respect to the used stimuli. Without our extension, a statistical analysis using distribution-based test would not be possible and such differences in brain-responsiveness over time could go unnoticed.

4. Discussion

The experiments performed demonstrate that the method of splitting the time series to deal with temporal autocorrelations or serial correlations decreases the false positive rate, independently of the applied extension. Using the true and false positive rates (ROC curves), the synthetic data and the permutation test as a reference, we showed that only the FDR extension controls the false positive rate in an adequate manner and for reasonable values of $\tau_{\max}$. In general, our simulations show and/or confirm that (a) in correspondence to what theory predicts [31], the classic OLS-based test outperforms the nonparametric tests with respect to its power whenever the data is derived from a Gaussian process (Fig. 2(b)); (b) the TPR values obtained with the nonparametric statistical tests better match those obtained with the OLS(-CO)/t values in cases where the data is derived from a non-Gaussian process, like the BAMM data set; and (c) that, independently from the data set used, a lag $\tau_{\max,} = 3$ FDR correction has an identical or even better FPR control than the OLS-CO/t-test, as can be deduced from the TPR and FPR curves.

We proposed also an objective method based on the PACF, to determine the $\tau_{\max}$ value using an *AR(k)* model of the right order k. fMRI noise signals or the error series obtained by using a GLM model are used as input for this *AR(k)* model. Albeit that a GLM/OLS is used in the latter case to find this error, a possible improper application of this method, due to the fact that some assumptions are not met, has no influence on the final results since this GLM is only used to estimate the *integer* $\tau_{\max}$. For both the synthetic *AR(1)* autocorrelated Gaussian noise and the BAMM data set, a $\tau_{\max} = 3$ correction controls the false positive rate appropriately.

With respect to the lower sensitivity (TPR values) of the nonparametric tests (even the permutation test) in comparison to those obtained with the OLS-CO/t approach, we can mention two arguments in defence of them:

- Application of an identical HRF model for both the creation of the synthetic signals and their analysis clearly favours the GLM-based methods. In practice, the exact HRF is not known and an estimate is used, if not a template.
- As mentioned before, the OLS-based tests are rather optimistic [11]. This behaviour is confirmed by the OLS/t permutation test that clearly has less power than the OLS(-CO)/t-test (Fig. 3(a) and (b)).

In addition to the decreased sensitivity, the proposed extension has two other, albeit minor, disadvantages. First, the extension is only applicable for block design fMRI studies and not for event-related fMRI studies, since our method splits each epoch into (several) parts. Second, contrary to the GLM-based methods that employ a design matrix to model all known effects, nonparametric statistical tests do not allow the modelling of additional effects like eye movements or cardio-respiratory activity. A possible solution with respect to the nonparametric tests is to apply a statistical test that checks for any relationship between the selected time series and the assumed effects. A warning for the researcher can then be issued in case a given threshold is exceeded. Nonparametric equivalents, such as a Cochran test or a Friedman test [31], are a possible solution.

We also stress that the proposed extension is used on single runs. Random or fixed effects analyses are not directly possible using the proposed methods. The GLM-based approach allows to design random/fixed effects analyses in a hierarchical way without the need for large and time consuming calculations [15]. The here proposed nonparametric statistics cannot calculate intra-run or intra-subject variability. We think that the use of confidence intervals for the difference between two means [31] (*i.e.*, the means of $\{X_i\}$ and $\{Y_i\}$ as defined in Section 2.4) can offer a variability estimate. More research needs to be done to find, examine, and verify such methods. However, this research falls outside the scope of this article.

Finally, a current line of research exists that might render the FDR technique even more promising: Yekutieli and Benjamini are developing hierarchical extensions to the basic FDR principle (personal communication) that allows one to include information, gathered while investigating part of the problem (*i.e.*, the serial correlation correction method), into the procedure that calculates the adjusted p-values for the complete problem (*i.e.*, the spatial multiple comparison correction). This path of research might return a solution yielding a higher sensitivity while still keeping the FPR within bounds.

5. Conclusion

Traditionally, a GLM is used to analyse fMRI data. However, the question remains whether a GLM approach is valid given the underlying Gaussian and linear assumptions. Specific pre-processing operations like data-smoothing [10,15] are necessary to help the data meet the required assumptions. We started investigating nonparametric statistical tests, to avoid such pre-processing operations since they could affect the final results. Moreover, from a formal perspective, nonparametric tests are the only kind of statistical tests that are guaranteed to be valid and exact in cases where the nature of the distribution is unknown [5].

We have developed an FDR-based extension that enables the application of EDF-like nonparametric tests (Kolmogorov–Smirnov, Cramér–von Mises, and Mann–Whitney) to fMRI data by appropriately handling the serial correlations within fMRI signals. In comparison with the GLM-based tests that are solved using an OLS, we spotted a significant decrease in statistical power. However, these OLS-based approaches are known to return too optimistic results [11], and they should also be applied with great care given their assumptions that require, *e.g.*, a Gaussian distribution of the regression error. Using nonparametric tests we do not have to consider these assumptions, nor do we have to know the exact nature of the distribution. Furthermore, our experiments show that the performance of an FDR extended nonparametric test very well coincides with that of the corresponding permutation test, and even with that of the GLM-based permutation test. Whenever a statistically valid analysis is aspired, we could therefore consider nonparametric tests.

In comparison to other nonparametric tests like the Bayesian approaches or the permutation test, our extension takes only a few additional seconds with respect to the unextended tests. Our method allows thus for a major speedup of the whole analysis when nonparametric tests are selected for analysing fMRI data. In addition, the proposed extension enables the application of EDF tests (Kolmogorov–Smirnov and Cramér–von Mises) to fMRI data. These EDF tests return more information from a fMRI signal as shown in Table 1, since they compare the distributions of the data points recorded under different stimuli, rather than the mean/median of those data points (GLM-approach/Mann–Whitney, respectively). Lange and coworkers have already mentioned that the application of a range of statistical procedures, parametric and data-driven, linear and nonlinear, would be most useful [12]. Regions might show an equal average level of activity, but a different distribution of the observed activation (see Table 1). The application of *e.g.*, the Cramér–von Mises in addition to a Mann–Whitney test is therefore certainly a source of valuable, additional information for the researcher.

Acknowledgments

The authors thank Dr. Walter Van Assche (Department of Mathematics, K.U. Leuven) and Dr. Steve Raiguel (Laboratory of Neuro- & Psychophysiology, K.U. Leuven) for their helpful suggestions on this manuscript. We thank also F. Deleus for his valuable scientific comments, and Dr. W. Vanduffel and the Brain Mapping Unit for the use of the monkey and BAMM null data set, respectively. Finally, we thank the editor and the anonymous reviewers for their useful remarks that significantly improved this paper.

P.A. De Mazière is currently supported by a scholarship from the European Commission (NEST-2003-012963). M.M. Van Hulle is supported by research grants received from the Belgian Fund for Scientific Research—Flanders (G.0248.03 and G.0234.04), the Interuniversity Attraction Poles Programme—Belgian Science Policy (IUAP P5/04), the Flemish Regional Ministry of Education (Belgium) (GOA 2000/11), and the European Commission (NEST-2003-012963, IST-2002-016276, and IST-2004-027017).

Appendix A. Rank-order tests

A.1. Mann–Whitney test

The hypothesis of the Mann–Whitney test (*MW*) states that both samples of data points have an identical median. Given $\{X_i\}$ and $\{Y_i\}$, containing N_x and N_y data points respectively, a set of data points $\{Z_i\} = \{X_i\}\bigcup\{Y_i\}$ is created and a rank assigned to the respective data points of $\{X_i\}$ and $\{Y_i\}$. The statistical value is calculated using Eq. (5), where $N = N_x + N_y$ and $\sum_{i=1}^{N} R_i^2$ represents the sum of squares of *all N* ranks [31]. The significance values (*p*-values) are easy to calculate since T_1 is approximately a standard normal random variable [31], to which well-known Gaussian formulae are applicable.

$$T_1 = \frac{T - N_x\frac{N+1}{2}}{\sqrt{\frac{N_xN_y}{N(N-1)}\sum_{i=1}^{N} R_i^2 - \frac{N_xN_y(N+1)^2}{4(N-1)}}}, \quad T = \sum_{k=1}^{N_x} R(X_k). \tag{5}$$

A.2. Kolmogorov–Smirnov and Cramér–von Mises test

The Kolmogorov–Smirnov (*KS*) and Cramér–von Mises test (*CvM*) have an identical null hypothesis, namely that both samples have an identical distribution. Using the same notation, we represent $\{X_i\}$ and $\{Y_i\}$ by their respective EDF: $S_1(x)$ and $S_2(x)$. The EDF $S(x)$ represents the fraction of X_is that are less than or equal to x [31]. The hypothesis is verified using the deviations between these EDFs: $d_k = S_1(x_k) - S_2(x_k)$, for $k = 1, \ldots, (N_x + N_y)$. The difference between the *KS*- and *CvM*-test lies in the way the deviations, d_ks, are interpreted: the statistical value for the *KS*-test is simply $\sup(|d_k|)$, while the statistical value for the *CvM*-test, T_2, is based on all d_ks, as shown in Eq. (6):

$$T_2 = \frac{N_xN_y}{(N_x + N_y)^2} \sum_{x_k \in \{X_i\}\cup\{Y_i\}} (S_1(x_k) - S_2(x_k))^2. \tag{6}$$

The difference in definition causes also a difference in the range of the statistical values: $[0, 1]$ for the *KS*-test, and $[0, \infty)$ for the *CvM*-test. These ranges do not contain any *absolute* information about the significance. The calculation of the significance value *p* is rather complex [40–42] and can be obtained upon request from the authors.

References

[1] K. Friston, C. Frith, P. Liddle, R. Frackowiak, Comparing functional; (PET) images: the assessment of significant change, J. Cereb. Blood Flow Metab. 11 (4) (1991) 690–699.

[2] G. Aguirre, E. Zarahn, M. D'Esposito, The variability of human, BOLD hemodynamic responses, Neuroimage 8 (4) (1998) 360–369.

[3] D. Heeger, D. Ress, What does fMRI tell us about neuronal activity? Nat. Rev. Neurosci. 3 (2) (2002) 142–151.

[4] T. Gautama, D. Mandic, M. Van Hulle, Signal nonlinearity in fMRI: a comparison between BOLD and MION, IEEE Trans. Med. Imaging 22 (5) (2003) 636–644.

[5] A. Holmes, R. Blair, J. Watson, I. Ford, Nonparametric analysis of statistic images from functional mapping experiments, J. Cereb. Blood Flow Metab. 16 (1) (1996) 7–22.

[6] G. Aguirre, E. Zarahn, M. D'Esposito, A critique of the use of the Kolmogorov–Smirnov (KS) statistic for the analysis of BOLD fMRI data, Magn. Reson. Med. 39 (3) (1998) 500–505.

[7] R. Thomas, Modern Econometrics, An Introduction, Addison–Wesley, Harlow, UK, 1997.

[8] K. Worsley, A. Evans, S. Marett, P. Neelin, A three-dimensional statistical analysis for cbf activation studies in human brain, J. Cereb. Blood Flow Metab. 12 (6) (1992) 900–918.

[9] K. Worsley, An improved theoretical *P* value for SPMs based on discrete local maxima, Neuroimage 28 (2005) 1056–1062.

[10] T. Nichols, S. Hayasaka, Controlling the familywise error rate in functional neuroimaging: a comparative review, Stat. Meth. Med. Res. 12 (5) (2003) 419–446.

[11] T. Gautama, M. Van Hulle, Optimal spatial regularisation of autocorrelation estimates in fMRI analysis, Neuroimage 23 (2004) 1203–1216.

[12] N. Lange, S. Strother, J. Anderson, F. Nielsen, A. Holmes, T. Kolenda, R. Savoy, L. Hansen, Plurality and resemblance in fMRI data analysis, Neuroimage 10 (3) (1999) 282–303.

[13] L. Hansen, F. Nielsen, P. Toft, M. Liptrot, C. Goutte, S. Strother, N. Lange, A. Gade, D. Rottenberg, O. Paulson, Lyngby—a modeler's Matlab toolbox for spatio-temporal analysis of functional neuroimages, in: Neuroimage, vol. 9, 1999, p. S241.

[14] R. Cox, AFNI: software for analysis and visualization of functional magnetic resonance neuroimages, Comput. Biomed. Res. 29 (3) (1996) 162–173.

[15] K. Worsley, C. Liao, J. Aston, V. Petre, G. Duncan, F. Morales, A. Evans, A general statistical analysis for fMRI data, Neuroimage 15 (1) (2002) 1–15.

[16] W. Vanduffel, D. Fize, H. Peuskens, K. Denys, S. Sunnaert, J. Todd, G. Orban, Extracting 3D from motion: differences in human and monkey intraparietal cortex, Science 298 (5592) (2002) 413–415.

[17] C. Chef d'Hotel, G. Hermosillo, O. Faugeras, Flows of diffeomorphisms for multimodal image registration, Proc. IEEE Int. Symp. Bio. Im. 7–8 (2002) 21–28.

[18] G. Hermosillo, C. Chef d'Hotel, O. Faugeras, Variational methods for multimodal image matching, Int. J. Comput. Vis. 50 (2002) 329–343.

[19] K. Friston, P. Jezzard, R. Turner, The analysis of functional MRI time-series, Hum. Brain Mapp. 1 (1994) 153–171.

[20] S. Smith, M. Jenkinson, M. Woolrich, C. Beckmann, T. Behrens, H. Johansen-Berg, P. Bannister, M. De Luca, I. Drobnjak, D. Flitney, R. Niazy, J. Saunders, J. Vickers, Y. Zhang, N. De Stefano, J. Brady, P. Matthews, Advances in functional and structural MR image analysis and implementation in FSL, Neuroimage 23 (2004) S208–S219.

[21] K. Worsley, K. Friston, Analysis of fMRI time series revisited—again, Neuroimage 2 (3) (1995) 173–181.

[22] J. Johnston, Econometric Methods, third ed., McGraw-Hill, Auckland, 1991.

[23] D. Cochrane, G. Orcutt, Application of least squares regression to relationships containing autocorrelated error terms, J. Am. Stat. Assoc. 44 (1949) 32–61.

[24] E. Bullmore, C. Long, J. Suckling, J. Fadili, G. Calvert, F. Zelaya, T. Carpenter, M. Brammer, Colored noise and computational inference in neurophysiological (fMRI) time series analysis: Resampling methods in time and wavelet domains, Hum. Brain Mapp. 12 (2) (2001) 61–78.

[25] M. Woolrich, B. Ripley, M. Brady, S. Smith, Temporal autocorrelation in univariate linear modeling of fMRI data, Neuroimage 14 (6) (2001) 1370–1386.

[26] P. Purdon, R. Weisskoff, Effect of temporal autocorrelation due to physiological noise and stimulus paradigm on voxel-level false-positive rates in fMRI, Hum. Brain Mapp. 6 (4) (1998) 239–249.

[27] E. Bullmore, M. Brammer, S. Williams, S. Rabe-Hesketh, N. Janot, A. David, J. Mellers, R. Howard, P. Sham, Statistical methods of estimation and inference for functional MR image analysis, Magn. Reson. Med. 35 (2) (1996) 261–277.

[28] Chatfield, The Analysis of Time Series: an Introduction, fifth ed., Chapman and Hall Texts in Statistical Science Series, Chapman & Hall, London, 1996.

[29] J. Fan, Q. Yao, Nonlinear Time Series, Nonparametric and Parametric Methods, Springer Series in Statistics, Springer-Verlag, New York, 2003.

[30] M. Bartlett, On the theoretical specification and sampling properties of autocorrelated time-series, J. Roy. Stat. Soc. Suppl. 8 (1) (1946) 27–41.

[31] W. Conover, Practical Nonparametric Statistics, third ed., John Wiley & Sons, New York, 1999.

[32] T. Nichols, A. Holmes, Nonparametric permutation tests for functional neuroimaging: a primer with examples, Hum. Brain Mapp. 15 (1) (2001) 1–25.

[33] C. Liu, J. Raz, B. Turetsky, An estimator and permutation test for single-trial fMRI data, in: Abstracts of ENAR meeting of the International Biometric Society, 1998.

[34] Y. Benjamini, D. Yekutieli, The control of the false discovery rate in multiple testing under dependency, Am. Stat. 29 (4) (2001) 1165–1188.

[35] F. Wolf, Meta-Analysis: Quantitative Methods for Research Synthesis, Quantitative Applications in the Social Sciences, Sara Miller McCune, Sage Publications Inc., Newbury Park, 1990.

[36] S. Stouffer, E. Suchman, L. De Vinney, S. Star, R. WilliamsThe Amercian Soldier: Adjustment During Army Life, vol. 1, Princeton University Press, Princeton, NJ, 1949.

[37] D. Montgomery, G. Runger, Applied Statistics and Probability for Engineers, second ed., John Wiley & Sons, New York, 1999.

[38] D. Green, J. Swets, Signal Detection Theory and Psychophysics, Wiley, New York, 1966.

[39] D. Malonek, U. Drinagl, U. Lindauer, K. Yamada, I. Kanno, A. Grinvald, Vascular imprints of neuronal activity: relationships between the dynamics of cortical blood flow, oxygenation, and volume changes following sensory stimulation, Proc. Natl. Acad. Sci. USA 94 (1997) 14826–14831.

[40] T. Anderson, D. Darling, Asymptotic theory of certain goodness of fit criteria based on stochastic processes, Ann. Math. Stat. 23 (1952) 193–212.

[41] S. Csörgo, J. Faraway, The exact and asymptotic distributions of Cramér–von Mises statistics, J. R. Stat. Soc. (Ser. B) 58 (1) (1996) 221–234.

[42] W. Press, S. Teukolsky, W. Vetterling, B. Flannery, Numerical Recipes in C, The Art of Scientific Computing, second ed., Cambridge University Press, Cambridge, 1997.