Bias in Resampling-Based Thresholding of Statistical Maps in fMRI

Ola Friman and Carl-Fredrik Westin

Laboratory of Mathematics in Imaging, Department of Radiology
Brigham and Women's Hospital, Harvard Medical School

Abstract. Selecting a threshold for the statistical parameter maps in functional MRI (fMRI) is a delicate matter. The use of advanced test statistics and/or the complex dependence structure of the noise may preclude parametric statistical methods for finding appropriate thresholds. Non-parametric statistical methodology has been presented as a feasible alternative. In this paper we discuss resampling-based methods for finding thresholds and show that proposed non-parametric approaches can lead to severely biased results.

1 Introduction

Selecting a threshold for the statistical parameter maps in fMRI is a challenging and important problem. The challenge lies in the fact that employed test statistics and/or the dependence structure of fMRI noise may not conform with classical statistical procedures and assumptions. Nonetheless it is important to assess the statistical significance provided by a specific threshold. The statistical significance is customarily measured by the p-value, which in fMRI context translates to the probability of declaring voxels active when in fact they are not. In order to find a threshold that provides a desired p-value, knowledge about the null-distribution of test statistic used for forming the statistical maps is required, see Fig. 1. Under certain assumptions about the noise structure and for certain test statistics, an analytic expression for this distribution is known. Examples

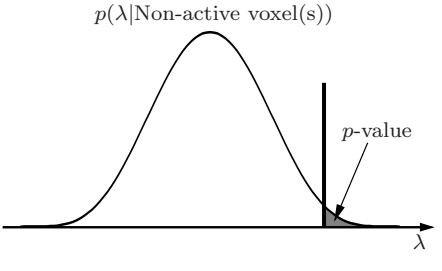

Fig. 1. The null-distribution is the distribution of a test statistic λ given that there is no activity in the examined voxel(s). The p-value is the probability of observing a value of the test statistic exceeding the threshold.

C. Barillot, D.R. Haynor, and P. Hellier (Eds.): MICCAI 2004, LNCS 3217, pp. 711–718, 2004.

are the t and F statistics encountered in the widely employed General Linear Model analysis [1]. However, as soon as the test statistic or the dependence structure of the noise depart from those afforded by classical statistical theory, the analytic expression for the null-distribution is in general intractable. The difficulties in deriving analytic expressions for the test statistic's null-distribution have spawned a number of alternative non-parametric ways to finding thresholds for the statistical maps [2,3,4,5,6,7]. Instead of assuming a parameterized form of the null-distribution, non-parametric approaches estimate it by analyzing data sets synthesized to mimic real fMRI data. Since no distributional assumptions about the data are required, non-parametric thresholds can be more accurate than those found by parametric methods [8]. The accuracy of the non-parametric thresholds is, however, strongly dependent on our ability to generate data with characteristics similar to real fMRI data. For this purpose various resampling techniques, for example whitening resampling [3,5], Fourier resampling [9,7] and wavelet resampling [6], have been applied.

Even though non-parametric procedures assume less about the nature of the test statistic and noise dependence structure, there exist pitfalls which may lead to severely biased thresholds. In this paper we point on such pitfalls and show that Fourier and wavelet resampling methods are not suitable for finding thresholds for fMRI statistical parameter maps. In the following sections we present the data, methods and results that underpin this conclusion. Finally, under Discussion we provide a theoretical explanation of the results.

2 Material

To prove our point we make use of simulated and real fMRI data sets. The simulated data set consists of 500 Gaussian white noise time series, each 128 samples long. In 20 % of the time series a synthetic smooth blocked design Blood Oxygen Level Dependent (BOLD) response has been embedded, i.e. 20 % of the time series correspond to 'active' voxels. Even though this data do not have the same characteristic as real fMRI data, it serve an illustrative purpose since we will be able to compare estimated null-distributions with the theoretically correct null-distribution. The real fMRI data set is a blocked design mental calculation test acquired using a 1.5 T GE scanner with imaging parameters: TR 2 s, TE 60 ms, FOV 24 cm, slice thickness 3 mm, image size 128×128 voxels and 180 time points. The images were realigned and spatially smoothed with a 4 mm FWHM Gaussian filter prior to the analysis described below.

3 Methods

Our objective is to detect active voxels in the data sets described above with thresholds that provide prespecified p-values. Below we describe the procedures for calculating the statistical maps, resampling the data sets and finding suitable thresholds.

3.1 Statistical Maps

For analyzing the synthetic data set, we simply use the correlation coefficient between each time series and the known embedded BOLD response shape. In this simplified synthetic setting, we know that the correlation coefficient is the c optimal test statistic for detecting 'active' voxels [10]. In the the real data case we do not know the exact shape of the BOLD response. In a traditional GLM analysis fashion [1], we produce a BOLD response model by convolving the binary on/off paradigm with a canonical impulse response function, and augment this model with its temporal derivative in order to account for unknown delays. With this BOLD response model we calculate F-maps, i.e. statistical maps consisting of F-statistics [11].

3.2 Resampling

Resampling is the process of producing artificial null-data sets with a statistical dependence structure similar to an original data set. From such resampled data sets we can estimate the null-distribution of the employed test statistic and consequently find a threshold that provides the desired p-value. In its simplest form, resampling boils down to a random reshuffling or permutation of the samples in the fMRI time series [2,4]. However, since fMRI data are serially correlated [12], such an approach does not preserve the temporal dependence structure, leading to erroneous threshold estimates. Instead we need to transform the data to a domain where a reshuffling does not alter the statistical structure, randomly permute the data, and then apply an inverse transform. To this end, whitening, Fourier and wavelet transforms have been proposed for resampling fMRI data [3, 5,6,7]. The resampling schemes based on these transforms are described in more detail below.

Whitening resampling. By assuming a particular model for the serial correlation structure we can apply a whitening transform to the time series, after which a random reshuffling of the samples is allowed. AR(1) and ARMA noise models have been proposed for this purpose [3,5]. In this paper, the whitening resampling was implemented by first fitting an AR(1) noise model to each time series. The time series was then whitened and the samples in the resulting time series permuted. Lastly, the permuted time series was passed through the fitted AR(1) process in order to form a resampled time series.

Fourier resampling. The intrinsic whitening property of the Fourier transform makes it potentially useful for resampling. Fourier resampling was carried out by taking the Fourier transform of the time series, keeping the magnitude of each frequency but permuting the phase components, and then applying the inverse Fourier transform.

Wavelet resampling. The wavelet resampling works similarly to the Fourier resampling. As devised by Bullmore et al. [6], a 4:th order Daubechies wavelet was used for wavelet transforming the time series. The wavelet coefficients within each scale were then permuted before applying the inverse wavelet transform.

Since fMRI data sets are spatio-temporal, in addition to preserving the serial correlation, it is also important to preserve the spatial autocorrelation. This is easily accomplished by applying the same random permutation to every time series in the fMRI data set [7,9]. Finally, as will be evident in the Results section, resampling the original data set as it is or resampling the residual data set obtained after regressing out the BOLD response model (and other deterministic components such as drifts and trends) yield very different results. The standard way would be to resample the residual data, but here we examine both alternatives as it provides insight into the biases we are about to see.

3.3 Finding Thresholds

The inference in fMRI analysis is usually carried out either at a voxel-wise level or at a family-wise level. The former means that we have the probability of a single voxel falsely being declared active under control and the latter implies that we have the probability of seeing any false positive activations over an entire set of voxels under control [13,14]. By producing a large number of resampled data sets, both voxel-wise and family-wise thresholds can be found. Having a large number of observations of the test statistic, calculated using resampled data, it is an easy task to determine the threshold for the quantile implied by the p-value. A family-wise threshold is found in a similar manner, but instead using only the maximum statistics recorded from the resampled data sets. Hence, it is computationally more demanding to find family-wise thresholds compared to voxel-wise thresholds.

4 Results

Using the methods described above, the synthetic and real data sets were resampled 1000 times each. The estimated voxel-wise and family-wise null-distributions, together with the theoretically correct null-distribution for the correlation coefficient[1], obtained by resampling the synthetic data set are shown in Fig. 2. The important observation here is that the Fourier and wavelet methods produce null-distribution estimates far from the true null-distribution. While the threshold obtained by resampling original fMRI data is severely overestimated (Fig. 2ab), the threshold estimated by resampling residual data is severely underestimated (Fig. 2cd). The whitening resampling method produces null-distributions with less dramatic, though still significant, errors. The behavior of the different resampling methods applied to this simplified synthetic data set is

[1] $f(r) = \frac{1}{\sqrt{\pi}} \frac{\Gamma\left(\frac{N-1}{2}\right)}{\Gamma\left(\frac{N-2}{2}\right)} \left(1-r^2\right)^{\frac{N-4}{2}}$, where N is the number of samples in the time series.

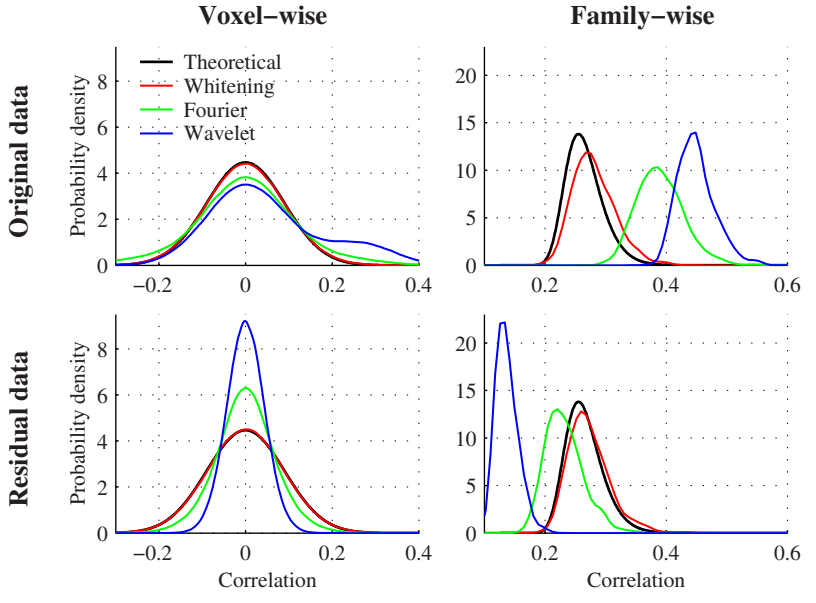

Fig. 2. Theoretical and estimated null-distributions for the correlation coefficient obtained by resampling the synthetic data set. Estimated voxel-wise and family-wise thresholds for a desired p-value can be determined from these distributions, cf. Fig. 1.

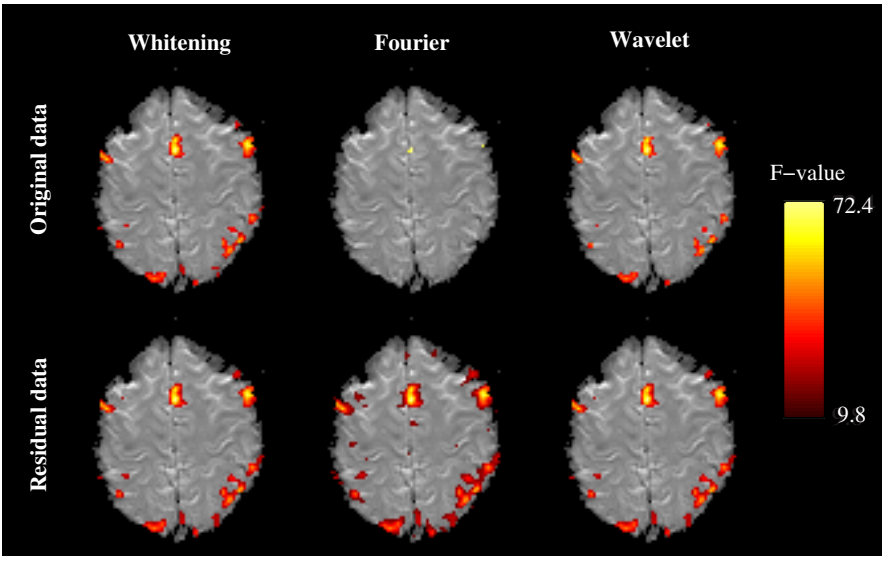

Fig. 3. A slice in the real fMRI data set subjected to the thresholds in Table 1. Note the difference between using the thresholds obtained by resampling original data and those obtained by resampling residual data.

the main result in this paper. However, before discussing the reasons underlying the biases seen in Fig. 2, we briefly show that similar results are obtained with real fMRI data too. In Table 1, the estimated family-wise thresholds for the real data set at $p = 0.05$ are listed. Note that while we cannot assess the accuracy of these thresholds, also here we observe substantially higher thresholds when resampling the original data set compared to the residual data set. Finally, in Fig. 3 the effect of the different thresholds are visualized.

Table 1. Estimated family-wise thresholds at significance level $p = 0.05$ for the F-maps calculated using the real data set.

	Whitening	Fourier	Wavelet
Original data	20.6	72.4	25.3
Residual data	16.9	9.8	17.0

5 Discussion

The results in the previous section show that Fourier- and wavelet-based resampling methods provide greatly biased thresholds while the whitening resampling approach seems to yield more accurate thresholds. There are two factors explaining this behavior. The first source of bias pertains to the fact that there are two classes of voxels in fMRI data, namely those containing a BOLD response and those who do not. In the blocked experimental design case, in original fMRI data voxels from these two classes have rather different spectra/autocorrelation functions, see top panel of Fig. 4. The second factor contributing to the bias is the number of degrees of freedom the resampling method has available for imitating the serial correlation structure in time series to be resampled. Both the Fourier and wavelet resampling schemes have large freedom in generating a time series with a spectrum matching that of the original time series. Therefore, when resampling original data with the Fourier and wavelet methods, if the time series is active also the resampled time series will have have a strong variation in pace with the BOLD response we are looking for. Hence, in the resampling process we will unproportionally often get time series that correlate well BOLD response model, leading to the biased null-distributions seen in Fig. 2a and Fig. 2c. If we try to circumvent this problem by removing the expected BOLD response from all voxels (i.e. create what we here denote residual data) prior to the resampling, we arrive in the situation shown in the bottom panel in Fig. 4. In this case the Fourier and wavelet methods produce resampled time series with too little power in the BOLD response frequencies. We will therefore instead find unproportionally small correlations between the BOLD response model and the resampled time series, as was seen in Fig. 2bd, leading to underestimated thresholds.

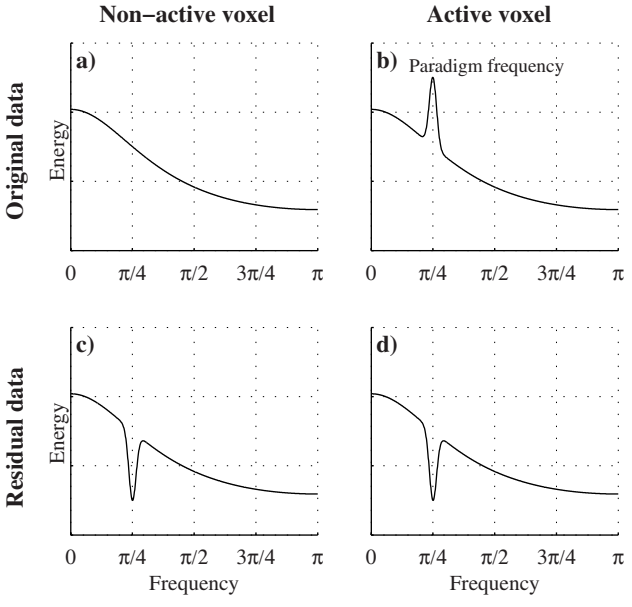

Fig. 4. Schematic spectra of time series in original and residual data (i.e. with a blocked BOLD response model removed from the time series).

In contrast to the Fourier and wavelet methods, the whitening resampling approach utilizes a specific model for the noise spectrum. In this paper an AR(1) model with only one degree of freedom is employed. When fitting this model to the observed spectrum of a time series, the absence or presence of power in the BOLD response frequencies has lesser impact on the resulting fit. Hence, the resampling process is regularized by the prior information provided by the noise model. Nevertheless, to some extent the presence of a BOLD response biases also the whitening approach, as was seen in Fig. 2b.

Hitherto, we have only discussed blocked experimental designs, as opposed to more rapid event-related designs. Due to the higher entropy, i.e. randomness, of event-related designs they tend to be more similar to noise when resampled. Thus, the bias effects discussed above are less pronounced, but still valid, when resampling fMRI data sets acquired during event-related experimental conditions.

6 Conclusions

Resampling-based methods based on Fourier and wavelet transforms have previously been proposed as appropriate for finding accurate thresholds for fMRI statistical maps. However, we have shown that even under simplified and controlled conditions, Fourier and wavelet resampling methods fail badly in this task. What ultimately makes Fourier and wavelet resampling unsuitable is the

many degrees of freedom they have available for mimicking the serial correlation in the fMRI time series. We have also shown that due to the regularizing effect of a serial correlation model, a whitening resampling approach has the potential to provide accurate thresholds. We therefore conclude that whitening resampling is the preferred non-parametric method for finding thresholds for statistical maps in fMRI.

Acknowledgments. The authors acknowledge the NIH for research grant P41 RR 13218.

References

1. Friston, K., Jezzard, P., Turner, R.: Analysis of functional MRI time-series. Human Brain Mapping **1** (1994) 153–171
2. Holmes, A., Blair, R., Watson, J., Ford, I.: Nonparametric analysis of statistic images from functional mapping experiments. Journal of Cerebral Blood Flow and Metabolism **16** (1996) 7–22
3. Bullmore, E., Brammer, M., Williams, S., Rabe-Hesketh, S., Janot, N., David, A., Mellers, J., Howard, R., Sham, P.: Statistical methods of estimation and inference for functional MR image analysis. Magnetic Resonance in Medicine **35** (1996) 261–277
4. Brammer, M., Bullmore, E., Simmons, A., Williams, S., Grasby, P., Howard, R., Woodruff, P., Rabe-Hesketh, S.: Generic brain activation mapping in functional magnetic resonance imaging: A nonparametric approach. Magnetic Resonance Imaging **15** (1997) 763–770
5. Locascio, J., Jennings, P., Moore, C., Corkin, S.: Time series analysis in the time domain and resampling methods for studies of functional magnetic resonance brain imaging. Human Brain Mapping **5** (1997) 168–193
6. Bullmore, E., Long, C., Suckling, J., Fadili, J., Calvert, G., Zelaya, F., Carpenter, T., Brammer, M.: Colored noise and computational inference in neurophysiological (fMRI) time series analysis: resampling methods in time and wavelet domains. Human Brain Mapping **12** (2001) 61–78
7. Laird, A., Rogers, B., Meyerand, M.: Comparison of Fourier and wavelet resampling methods. Magnetic Resonance in Medicine **51** (2004) 418–422
8. Nichols, T., Holmes, A.: Nonparametric permutation tests for functional neuroimaging: A primer with examples. Human Brain Mapping **15** (2001) 1–25
9. Prichard, D., Theiler, J.: Generating surrogate data for time series with several simultaneously measured variables. Physical Review Letters **73** (1994) 951–954
10. van Trees, H.: Detection, Estimation, and Modulation Theory, Part I. John Wiley & Sons (1968)
11. Friston, K., Fletcher, P., Josephs, O., Holmes, A., Rugg, M., Turner, R.: Event-related fMRI: Characterizing differential responses. NeuroImage **7** (1998) 30–40
12. Woolrich, M., Ripley, B., Brady, M., Smith, S.: Temporal autocorrelation in univariate linear modeling of FMRI data. NeuroImage **14** (2001) 1370–1386
13. Worsley, K.: Local maxima and the expected Euler characteristic of excursion sets of χ^2, F and t fields. Advances in Applied Probability **26** (1994) 13–42
14. Nichols, T., Hayasaka, S., Wager, T.: Controlling the familywise error rate in functional neuroimaging: A comparative review. Statistical Methods in Medical Research **12** (2003) 419–446