Comparing fMRI Activity Maps from GLM and CCA at the Same Significance Level by Fast Random Permutation Tests on the GPU

Anders Eklund*†, Ola Friman‡, Mats Andersson*† and Hans Knutsson*†
*Division of Medical Informatics, Department of Biomedical Engineering, Linköping University
†Center for Medical Image Science and Visualization (CMIV)
‡Swedish Defence Research Agency, Linköping

***Abstract*—Parametric statistical methods are traditionally employed in functional magnetic resonance imaging (fMRI) for identifying areas in the brain that are active with a certain degree of statistical significance. These parametric methods, however, have two major drawbacks. First, it is assumed that the observed data are Gaussian distributed and independent; assumptions that generally are not valid for fMRI data. Second, the statistical test distribution can be derived theoretically only for very simple linear detection statistics. In this work it is shown how the computational power of the Graphics Processing Unit (GPU) can be used to speedup non-parametric tests, such as random permutation tests. With random permutation tests it is possible to calculate significance thresholds for any test statistics. As an example, fMRI activity maps from the General Linear Model (GLM) and Canonical Correlation Analysis (CCA) are compared at the same significance level.**

I. Introduction

Functional magnetic resonance imaging (fMRI) is used in neuroscience and clinic for investigating brain activity patterns and for planning brain surgery. Activity is detected by fitting an activity model to each observed fMRI voxel time series and then testing whether the null hypothesis of no activity can be rejected or not based on the model parameters. Specifically, this test is performed by subjecting a test statistic calculated from the model parameters to a threshold. To control the randomness due to noise in this test procedure, it is desirable to find the statistical significance associated with the detection threshold, i.e., how likely it is that a voxel is declared active by chance. When the statistical distribution of the data is known *and* when the probability (null-)distribution of the test statistic can be derived, parametric statistics can be used to this end. This is for example the case for the commonly used General Linear Model (GLM), for which the well known t-test and F-test can be derived when the input data are independently Gaussian distributed. However, when the data distribution is not known or the distribution of the test statistic cannot be derived, parametric statistical tests can only yield approximate thresholds or cannot be applied at all. This is generally the case in fMRI analysis as the noise in fMRI data is not Gaussian and independent. Furthermore, more advanced detection approaches often adaptively utilize the spatial context of fMRI activation patterns to improve the detection, or they perform other operations that make the derivation of the test statistic distribution mathematically intractable. Said otherwise, only for the very simplest test statistics, such as the GLM, can a parametric test distribution be derived theoretically. On top of the problems described above, the multiple testing problem must be solved since more than 20 000 voxels normally are tested for activity. This complicates the derivation of the test statistic distribution even further. To conclude, the parametric statistical approach is applicable only to a very limited set of tests and is subject to many sources of error.

An alternative is to use non-parametric statistics. The major drawback of non-parametric statistical approaches for single subject fMRI analysis is the computational complexity, requiring hours or days of processing time on regular computer hardware. In this work, it is shown how random permutation tests can be made practical for fMRI analysis by using the parallel processing power of the Graphics Processing Unit (GPU), making it possible to estimate the null-distribution of a test statistic, corrected for multiple testing, in the order of minutes. This has significant implications on the way fMRI analysis can be carried out as it opens the possibility to routinely apply more powerful detection methods than the GLM. As an example, the results of the standard GLM detection is in this work compared with a restricted Canonical Correlation Analysis (CCA) method [1] that adaptively incorporates spatial context in the detection.

II. Methods

A. Basics of random permutation tests

One subset of non-parametric tests is permutation tests where the statistical analysis is done for all the possible permutations of the data. Complete permutation tests are however not feasible if the number of possible permutations is very large. For a voxel time series with 80 samples, there exists $7.16 \cdot 10^{118}$ possible permutations. It is therefore common to instead do *random* permutation tests [2], also called Monte Carlo permutation tests, where the statistical analysis is made for a sufficiently large number of random permutations, for example 10 000, of the data. The main idea is to estimate the null distribution of the test statistics, by generating and analysing surrogate data that is similar to the original data. The surrogate data is generated by permuting, or reshuffling, the data between the different groups to be compared.

B. The problem of multiple testing

By applying a threshold to the activity map, each voxel can be classified as active or inactive. The threshold is normally selected as a level of significance, one may for example want that only voxels that with at least 95% significance are to be considered as active. If a statistical test is repeated and a family wise error rate α is desired, the error rate for each test must be smaller than α. This is known as the problem of multiple testing. If *Bonferroni* correction is used, the error rate for each comparison becomes α/N, where N is the number of tests. This is a correct solution, if the tests are independent. In fMRI it is common to perform the statistical analysis for more than 20 000 brain voxels, if a threshold of p = 0.05 is used to consider the voxel as active, the p-value becomes 0.05/20000 with Bonferroni correction. The assumptions that are made about the behaviour of the tail of the distribution are thereby critical.

The non-parametric approach can be used to solve the problem of multiple testing as well. This is done by estimating the null distribution of the *maximum* test statistic by only saving the maximum test value from each permutation, to get a *corrected* threshold. This means that something like 10 000 permutations have to be used, while as little as 10 permutations can be enough if an *uncorrected* threshold is sufficient.

C. Preprocessing of fMRI time series

As fMRI time series are temporally correlated [3], the time series have to be preprocessed before they are permuted, otherwise the exchangeability criteria is not satisfied and the temporal structure is destroyed. A lot of the temporal correlations originate from different kinds of trends, like scanner imperfections and physiological noise. In our case these trends are removed by a cubic detrending, such that the mean and any polynomial trend up to the third order is removed, but more advanced detrending is possible.

Several approaches have been proposed for the resampling, the most common being whitening transforms, wavelet transforms and Fourier transforms. A comparison of these approaches [4] indicates that whitening performs best, at least for fMRI data that is collected during block based stimuli paradigms. The whitening transform is done by first estimating an auto regressive (AR) model for each time series. An AR model, of order p, states that a time series $x(t)$ is generated as

$$x(t) = \alpha_1 x(t-1) + ... + \alpha_p x(t-p) + e(t) \quad (1)$$

where $e(t)$ is white noise. The parameters $\alpha_1, ..., \alpha_p$, the AR parameters for the different time lags, can be estimated by solving the equation system that is given by the Yule-Walker equations. The whitened time series $w(t)$ are then calculated as

$$w(t) = x(t) - \sum_{i=1}^{p} \hat{\alpha}_i x(t-i) \quad (2)$$

where p is the order of the AR model and $\hat{\alpha}_i$ are the estimated AR parameters. An AR model of order 4 was used in our case.

D. Statistical analysis, GLM & t-test

The general linear model (GLM) [5] is the most used approach for statistical analysis of fMRI data. For each voxel time series, a linear model is fitted according to

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad (3)$$

where $\boldsymbol{Y}$ is the time series, $\boldsymbol{X}$ are the regressors that model brain activity, $\boldsymbol{\beta}$ are the parameters to estimate and $\boldsymbol{\epsilon}$ are the residuals.

The regressors were created by convolving the stimulus paradigm with the hemodynamic response function (HRF) (difference of gammas) and its temporal derivative [3]. The regression weights are estimated as

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \quad (4)$$

and the t-test value is then calculated as

$$t = \frac{\boldsymbol{c}^T\hat{\boldsymbol{\beta}}}{\sqrt{\mathrm{var}(\boldsymbol{\epsilon})\boldsymbol{c}^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{c}}} \quad (5)$$

where $\boldsymbol{c}$ is the contrast vector ($[1\ 0]^T$).

Prior to the GLM the time series were whitened by using the same AR(1) model for all the voxels [3].

E. Statistical analysis, CCA

One statistical approach for fMRI analysis that provides more adaptivity to the data is canonical correlation analysis (CCA) [6]. While the GLM works with *one* multidimensional variable (temporal basis functions), CCA works with *two* multidimensional variables (temporal and spatial basis functions). Ordinary correlation between two one-dimensional variables x and y can be written as

$$\rho = \mathrm{Corr}(\mathrm{x}, \mathrm{y}) = \frac{\mathrm{E[xy]}}{\sqrt{\mathrm{E[x^2]E[y^2]}}} \quad (6)$$

The GLM calculates the correlation between one multidimensional variable $\boldsymbol{x}$ and one one-dimensional variable y according to

$$\rho = \mathrm{Corr}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}, \mathrm{y}) \quad (7)$$

where $\boldsymbol{\beta}$ is the weight vector that determines the linear combination of $\boldsymbol{x}$. Canonical correlation analysis is a further generalization of the GLM, such that both the variables are multidimensional. The canonical correlation is defined as

$$\begin{aligned} \rho &= \mathrm{Corr}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}, \boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{y}) \\ &= \frac{\boldsymbol{\beta}^T\boldsymbol{C_{xy}}\boldsymbol{\gamma}}{\sqrt{\boldsymbol{\beta}^T\boldsymbol{C_{xx}}\boldsymbol{\beta}\,\boldsymbol{\gamma}^T\boldsymbol{C_{yy}}\boldsymbol{\gamma}}} \end{aligned} \quad (8)$$

where $\boldsymbol{C_{xy}}$ is the covariance matrix between $\boldsymbol{x}$ and $\boldsymbol{y}$, $\boldsymbol{C_{xx}}$ is the covariance matrix for $\boldsymbol{x}$ and $\boldsymbol{C_{yy}}$ is the covariance matrix for $\boldsymbol{y}$. The temporal and spatial

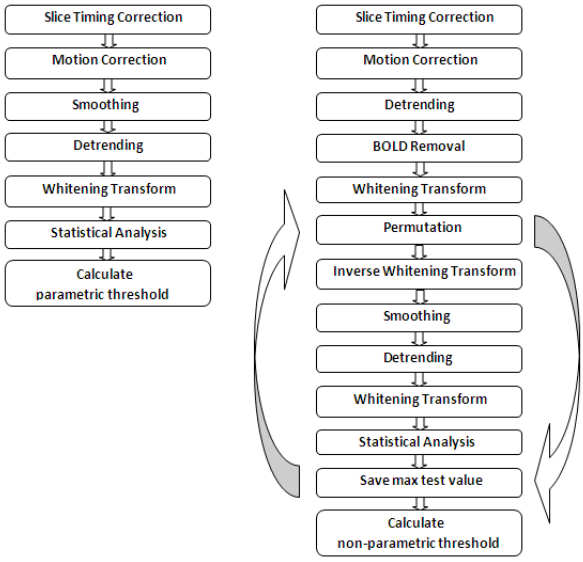


Fig. 1. ***Left:*** *Flowchart for conventional parametric analysis of fMRI data.* ***Right:*** *Flowchart for non-parametric analysis of fMRI data. In each permutation a new null dataset is generated and analyzed.*

weight vectors, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, that give the highest correlation are calculated as the eigen vectors of two eigen value problems. The canonical correlation is the square root of the corresponding eigen value. The two eigen value problems can be written as

$$\boldsymbol{C_{xx}}^{-1/2}\,\boldsymbol{C_{xy}}\,\boldsymbol{C_{yy}}^{-1}\,\boldsymbol{C_{yx}}\,\boldsymbol{C_{xx}}^{-1/2}\,\boldsymbol{a} = \lambda^2\,\boldsymbol{a} \quad (9)$$

$$\boldsymbol{C_{yy}}^{-1/2}\,\boldsymbol{C_{yx}}\,\boldsymbol{C_{xx}}^{-1}\,\boldsymbol{C_{xy}}\,\boldsymbol{C_{yy}}^{-1/2}\,\boldsymbol{b} = \lambda^2\,\boldsymbol{b} \quad (10)$$

It is sufficient to solve one of the problems, since the second weight vector can be calculated from the first. The temporal basis functions for CCA are the same as for the GLM. The spatial basis functions can for example be neighbouring pixels [7], [8] or a number of anisotropic filters [1], that linearly can be combined to a lowpass filter with arbitrary orientation, to prevent unnecessary smoothing. In contrast to the GLM, an adaptive anisotropic smoothing is obtained, instead of a fix isotropic smoothing.

One disadvantage with CCA is that it is much harder to calculate the threshold for a certain significance level, since the distribution of the canonical correlation coefficients is rather complicated.

F. The complete algorithm

The complete algorithm for the random permutation test is given in Figure 1. As reference, the algorithm for a conventional parametric fMRI analysis is also included.

G. Implementation

The random permutation test was implemented with the CUDA (compute unified device architecture) programming language by Nvidia. The used graphics cards were three Nvidia GTX 480, each equipped with 480 processor cores and 1.5 GB of memory, giving a total of 1440 processor cores.

III. Results

With our multi-GPU implementation, 10 000 null datasets can be analyzed in about 50 seconds, compared to about 4 hours for the GLM and 16 hours for CCA with a standard C implementation.

With the random permutation test it is possible to calculate corrected p-values for fMRI analysis by CCA, and thereby activity maps from GLM and CCA can finally be compared at the same significance level. The activity maps generated by using 2D smoothing are given in Figure 2, the activity maps generated by using 3D smoothing are given in Figure 3. For these comparisons 10 000 permutations were used both for GLM and CCA. For the fMRI dataset used the test subject periodically activated the left hand.

With 8 mm of 2D smoothing, GLM detects 302 significantly active voxels while CCA detects 344 significantly active voxels. With 8 mm of 3D smoothing, GLM detects 475 significantly active voxels while CCA detects 684 significantly active voxels.

IV. Conclusions

We have applied random permutation tests for single subject analysis of fMRI data using the GPU. Our work enables objective evaluation of arbitrary methods for single subject fMRI analysis. As a pleasant side effect, the problem of multiple testing is solved in a way that significantly reduces the number of necessary assumptions.

Acknowledgment

This work was supported the Linnaeus center CADICS, funded by the Swedish research council. The fMRI data was collected at the Center for medical image science and visualization (CMIV).

The authors would like to thank the Novamedtech project at Linköping university for financial support of our GPU hardware and Johan Wiklund for support with the CUDA installations.

References

[1] O. Friman, M. Borga, P. Lundberg, and H. Knutsson, "Adaptive analysis of fMRI data," *NeuroImage*, vol. 19, pp. 837–845, 2003.

[2] M. Dwass, "Modified randomization tests for nonparametric hypotheses," *The Annals of Mathematical Statistics*, vol. 28, pp. 181–187, 1957.

[3] K. Friston, O. Josephs, E. Zarahn, A. Holmes, S. Rouquette, and J. Poline, "To smooth or not to smooth - bias and efficiency in fMRI time-series analysis," *Neuroimage*, vol. 12, pp. 196–208, 2000.

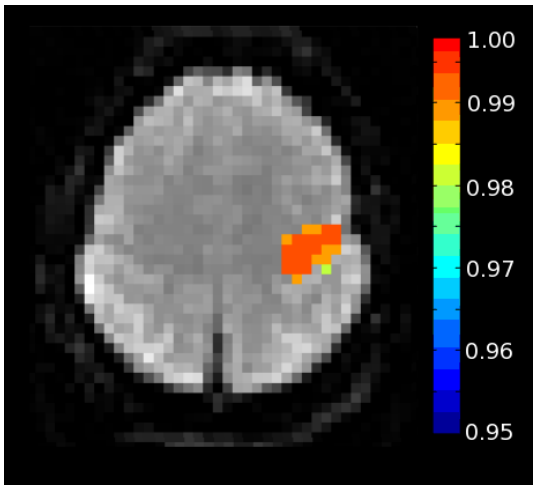
[4] O. Friman and C.-F. Westin, "Resampling fMRI time series," *NeuroImage*, vol. 25, pp. 859–867, 2005.

[5] K. Friston, A. Holmes, K. Worsley, J. Poline, C. Frith, and R. Frackowiak, "Statistical parametric maps in functional imaging: A general linear approach," *Human Brain Mapping*, vol. 2, pp. 189–210, 1995.

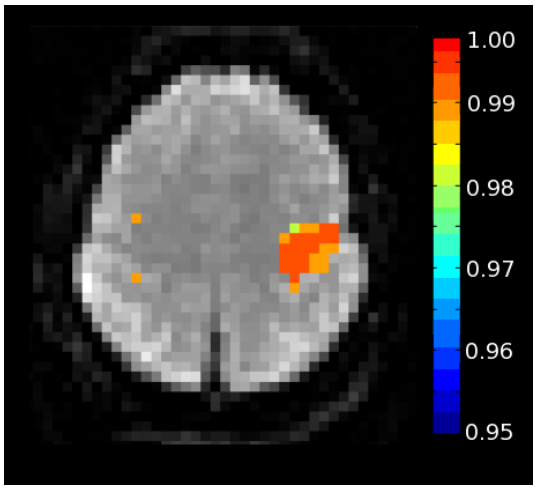
[6] H. Hotelling, "Relation between two sets of variates," *Biometrika*, vol. 28, pp. 322–377, 1936.

[7] O. Friman, J. Carlsson, P. Lundberg, M. Borga, and H. Knutsson, "Detection of neural activity in functional MRI using canonical correlation analysis," *Magnetic Resonance in Medicine*, vol. 45, no. 2, pp. 323–330, 2001.

[8] R. Nandy and D. Cordes, "A novel nonparametric approach to canonical correlation analysis with applications to low CNR functional MRI data," *Magnetic Resonance in Medicine*, vol. 49, pp. 1152–1162, 2003.

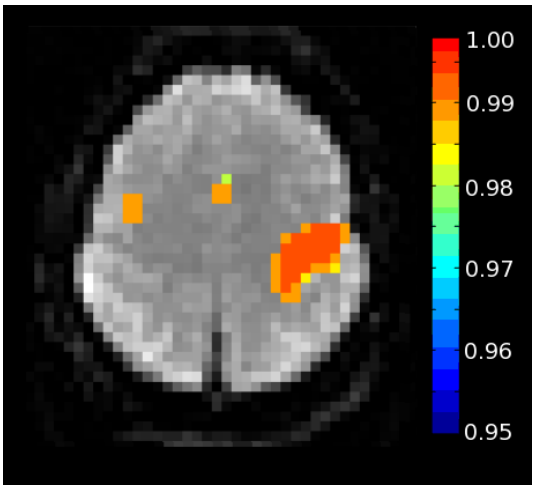


(a) *Activity map generated by using 2D GLM.*

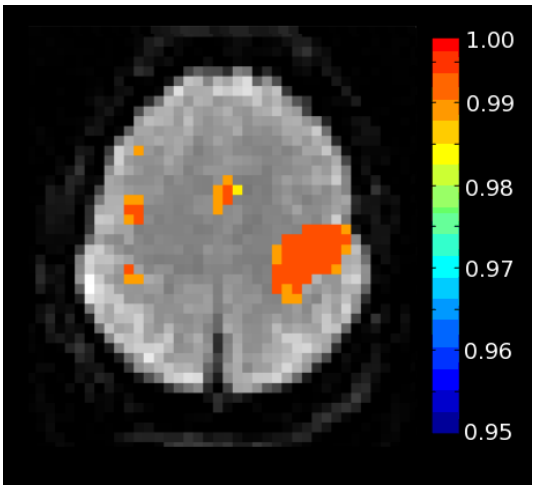


(b) *Activity map generated by using 2D CCA.*

Fig. 2. *A comparison between corrected p-values from 2D GLM and 2D CCA, calculated from a random permutation test with 10 000 permutations. The activity maps are thresholded at the same significance level (corrected p = 0.05). The GLM used one isotropic 8 mm FWHM 2D Gaussian smoothing kernel while CCA used one isotropic 2D Gaussian kernel and 3 anisotropic 2D Gaussian kernels, designed such that the largest possible filter that CCA can create has a FWHM of 8 mm. The neurological display convention is used (left is left), 1 - p is shown instead of p. Note that CCA detects active voxels in the left motor cortex and left somatosensory cortex that not are detected with the GLM.*



(a) *Activity map generated by using 3D GLM.*



(b) *Activity map generated by using 3D CCA.*

Fig. 3. *A comparison between corrected p-values from 3D GLM and 3D CCA, calculated from a random permutation test with 10 000 permutations. The activity maps are thresholded at the same significance level (corrected p = 0.05). The GLM used one isotropic 8 mm FWHM 3D Gaussian smoothing kernel while CCA used one isotropic 3D Gaussian kernel and its derivative, designed such that the largest possible filter that CCA can create has a FWHM of 8 mm. The neurological display convention is used (left is left), 1 - p is shown instead of p. Note that CCA detects active voxels in the left somatosensory cortex that not are detected with the GLM.*