

Improving CCA based fMRI Analysis by Covariance Pooling - Using the GPU for Statistical Inference

Anders Eklund, Mats Andersson and Hans Knutsson

Linköping University Post Print

N.B.: When citing this work, cite the original article.

Original Publication:

Anders Eklund, Mats Andersson and Hans Knutsson, Improving CCA based fMRI Analysis by Covariance Pooling - Using the GPU for Statistical Inference, 2011, presented at Joint MICCAI Workshop on High Performance and Distributed Computing for Medical Imaging, HP-MICCAI, September 22nd, Toronto, Canada .

Postprint available at: Linköping University Electronic Press

<http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-71281>

Improving CCA based fMRI Analysis by Covariance Pooling

-

Using the GPU for Statistical Inference

Anders Eklund^{1,2}, Mats Andersson^{1,2}, Hans Knutsson^{1,2}

¹ Division of Medical Informatics, Department of Biomedical Engineering

² Center for Medical Image Science and Visualization (CMIV)

Linköping University, Linköping, Sweden

Abstract. Canonical correlation analysis (CCA) is a statistical method that can be preferable to the general linear model (GLM) for analysis of functional magnetic resonance imaging (fMRI) data. There are, however, two problems with CCA based fMRI analysis. First, it is not feasible to use a parametric approach to calculate an activity threshold for a certain significance level. Second, two covariance matrices need to be estimated in each voxel, from a rather small number of time samples. We recently solved the first problem by doing random permutation tests on the graphics processing unit (GPU), such that the null distribution of any maximum test statistics can be estimated in the order of minutes. In this paper we consider the second problem. We extend the idea of variance pooling, that previously has been used for the GLM, to covariance pooling to improve the estimates of the covariance matrices. Our GPU implementation of random permutation tests is used to calculate significance thresholds, which are needed to compare the different activity maps in an objective way. The covariance pooling results in more robust estimates of the covariance matrices. The number of significantly active voxels that are detected (thresholded at $p = 0.05$, corrected for multiple comparisons) is increased with 40 - 120% (if 8 mm smoothing is applied to the covariance estimates). Too much covariance pooling can however result in a loss of small activity clusters, 7-10 mm of smoothing gives the best results. The calculations that were made in order to generate the results in this paper would have taken a total of about 65 days with a Matlab implementation and about 10 days with a multithreaded C implementation, with our multi-GPU implementation they took about 2 hours. By using fast random permutation tests, suggested improvements of existing methods for fMRI analysis can be evaluated in an objective way.

1 Introduction

Functional magnetic resonance imaging (fMRI) is a modality that is commonly used in neuroscience and clinic for investigating brain activity patterns and for

planning brain surgery. The most common approach to analyze the fMRI data is to apply the general linear model (GLM) separately to each voxel time series, calculate a t-test or a F-test value and then use a parametric threshold in order to classify each voxel as active or inactive [11]. While this approach is straight forward and easy to understand, its main drawback is that it considers each voxel time series separately. The only thing that is done in order to use information from neighbouring voxels is to apply a Gaussian smoothing of each volume. The problem is that the optimal amount, and orientation, of smoothing that should be applied is not constant but varies in the brain. To this end, canonical correlation analysis (CCA) [14] can be used instead of the GLM, since CCA can handle *two* multidimensional variables (e.g. temporal and spatial basis functions [9]) while the GLM only can handle *one* multidimensional variable (e.g. temporal basis functions [11]). The temporal basis functions for CCA are the same as for the GLM. The spatial basis functions can for example be neighbouring pixels [10, 16] or a number of anisotropic filters [9], which can be linearly 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. Even if the CCA approach has been proven to have a superior detection performance compared to the GLM [10, 9, 16, 18], its use has been rather limited. One reason for this is that it is much harder to calculate a threshold for a certain significance level, as the distribution of the canonical correlation coefficients is rather complicated. It is thereby difficult to objectively compare fMRI activity maps from GLM and CCA.

Non-parametric tests, such as permutation tests, have previously been applied to *multi subject* fMRI [13, 17] while they are more rare for *single subject* fMRI [3]. We recently solved the problem of calculating significance thresholds, corrected for multiple testing, for arbitrary test statistics for single subject fMRI analysis, by taking advantage of the graphics processing unit (GPU) [7]. Speeding up random permutation tests by using the GPU has previously been done in biostatistics [19, 15]. We can thereby make a fair comparison of different methods for fMRI analysis, by for example looking at the number of significantly active voxels that are detected. Fast random permutation tests also make it practically possible to evaluate improvements of existing methods, in particular improvements for which the effect on the null distribution is hard to predict.

In this paper we therefore focus on another difficulty with CCA, namely that two covariance matrices have to be estimated in each voxel, from a rather low number of time samples (normally 100 - 200). One way to improve variance estimates is to increase the degrees of freedom by a statistical technique known as variance pooling. In fMRI this is done by spatially smoothing the variance estimates, in order to use information from neighbouring voxels. This idea has previously been used for the GLM [21] and also in order to improve the estimates of auto regressive (AR) parameters [20] that are used for whitening of the time series. To our knowledge, variance pooling has not been proposed for CCA even though CCA is likely to benefit more from variance pooling than the GLM, as a high number of covariances need to be estimated in each voxel.

1.1 fMRI analysis on the GPU

The idea of taking advantage of the GPU for fMRI analysis is quite new. The main advantage of GPUs compared to CPUs is the much higher degree of parallelism. As the time series are normally regarded as independent in fMRI analysis, it is perfectly suited for parallel implementations.

The first work about fMRI analysis on the GPU is the work by Gembris et al. [12] that used the GPU to speedup the calculation of correlations between voxel time series, a technique that commonly is used in resting state fMRI [2] for identifying functional brain networks. We extended the work by Gembris et al. to a GPU accelerated interactive interface, with 3D visualization, for exploratory functional connectivity analysis [6]. Another example is the work by da Silva [1] that used the GPU to speedup the simulation of a Bayesian multilevel model. We recently described how to perform both preprocessing and statistical analysis of fMRI data on the GPU [8] by using the CUDA (Compute Unified Device Architecture) programming language by Nvidia. Our work was then extended to random permutation tests [4, 7], in order to for example compare activity maps from GLM and CCA at the same significance level.

In this work we take further advantage of our fMRI GPU implementation to improve CCA based fMRI analysis.

2 Methods

2.1 Canonical correlation analysis

Ordinary correlation between two one-dimensional variables x and y with zero mean can be written as

$$\rho = \text{Corr}(x, y) = \frac{E[xy]}{\sqrt{E[x^2]E[y^2]}} \quad (1)$$

This expression can easily be extended to multidimensional variables. The GLM calculates the correlation between one multidimensional variable \mathbf{x} and one one-dimensional variable y according to

$$\rho = \text{Corr}(\beta^T \mathbf{x}, y) \quad (2)$$

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

$$\rho = \text{Corr}(\beta^T \mathbf{x}, \gamma^T \mathbf{y}) = \frac{\beta^T \mathbf{C}_{\mathbf{x}\mathbf{y}} \gamma}{\sqrt{\beta^T \mathbf{C}_{\mathbf{x}\mathbf{x}} \beta \gamma^T \mathbf{C}_{\mathbf{y}\mathbf{y}} \gamma}} \quad (3)$$

where $\mathbf{C}_{\mathbf{x}\mathbf{y}}$ is the covariance matrix between \mathbf{x} and \mathbf{y} , $\mathbf{C}_{\mathbf{x}\mathbf{x}}$ is the covariance matrix for \mathbf{x} and $\mathbf{C}_{\mathbf{y}\mathbf{y}}$ is the covariance matrix for \mathbf{y} . The temporal and spatial weight vectors, β and γ , that give the highest correlation are calculated as the

eigenvectors of two eigenvalue problems. The canonical correlation is the square root of the corresponding eigenvalue. The two eigenvalue problems can be written as

$$\mathbf{C}_{xx}^{-1/2} \mathbf{C}_{xy} \mathbf{C}_{yy}^{-1} \mathbf{C}_{yx} \mathbf{C}_{xx}^{-1/2} \mathbf{a} = \lambda^2 \mathbf{a} \quad (4)$$

$$\mathbf{C}_{yy}^{-1/2} \mathbf{C}_{yx} \mathbf{C}_{xx}^{-1} \mathbf{C}_{xy} \mathbf{C}_{yy}^{-1/2} \mathbf{b} = \lambda^2 \mathbf{b} \quad (5)$$

To get the weight vectors from \mathbf{a} and \mathbf{b} a change of base is necessary, according to

$$\boldsymbol{\beta} = \mathbf{C}_{xx}^{-1/2} \mathbf{a}, \quad \boldsymbol{\gamma} = \mathbf{C}_{yy}^{-1/2} \mathbf{b} \quad (6)$$

It is sufficient to solve one of the problems, since the second weight vector can be calculated from the first. If \mathbf{x} is defined to be the temporal basis functions, it is sufficient to estimate its covariance matrix \mathbf{C}_{xx} once prior to the fMRI analysis, since the temporal basis functions are the same for each voxel. The other covariance matrices, \mathbf{C}_{yy} and \mathbf{C}_{xy} , have to be estimated for each voxel time series, according to

$$\mathbf{C}_{yy} = \frac{1}{N-1} \sum_{t=1}^N \mathbf{y}(t) \mathbf{y}(t)^T, \quad \mathbf{C}_{xy} = \frac{1}{N-1} \sum_{t=1}^N \mathbf{x}(t) \mathbf{y}(t)^T \quad (7)$$

where N is the number of time samples. The size of \mathbf{C}_{xx} is $n \times n$ where n is the number of temporal basis functions, the size of \mathbf{C}_{yy} is $m \times m$ where m is the number of spatial basis functions and \mathbf{C}_{xy} is of size $n \times m$. Typical numbers of n and m are 2-7 and 2-6 respectively.

The CCA implementation that is used in this paper is based on the work in [9] but slightly modified in order to be able to implement it on the GPU. In our case 4 spatial basis functions (one isotropic lowpass filter and three anisotropic lowpass filters) and 2 temporal basis functions (the stimulus paradigm convolved with the hemodynamic response function and its temporal derivative) are used.

To accurately estimate *one* covariance from 80 time samples is quite difficult. In our case 18 covariances (10 for \mathbf{C}_{yy} (the matrix contains 16 values but the matrix is symmetric) and 8 for \mathbf{C}_{xy}) need to be estimated from 80 time samples. By extending the idea of variance pooling to *covariance* pooling, the estimates of the covariance matrices become more robust. The covariance matrices are first estimated for each voxel time series separately and then a spatial smoothing with a Gaussian lowpass filter is applied to each element of the covariance matrices separately.

The difficult part is to figure out how the null distribution of the maximum canonical correlation coefficient is affected by this covariance pooling. Fortunately, this is in our case implicitly solved in a very easy way by doing the covariance pooling in each permutation of the random permutation test.

2.2 Preprocessing & the random permutation test

Before the fMRI data was statistically analyzed, slice timing correction, motion compensation [5] and cubic detrending was applied.

Our multi-GPU implementation of random permutation tests [7] is the key component that makes it practically possible to estimate the null distribution of any maximum test statistics. Prior to the permutations an AR(4) model is estimated for each voxel time series of the preprocessed fMRI data. A spatial Gaussian smoothing (8 mm FWHM) is then applied to improve the estimates of the AR parameters [20] and then a whitening is applied with the smoothed parameters. The whitening has to be done prior to the permutations, since the random permutation test requires that the time samples are exchangeable under the null hypothesis. In each permutation a new null dataset is then generated by applying an inverse whitening transform with the permuted whitened time series as innovations. This null dataset is then analyzed by smoothing with the four filters, applying cubic detrending and finally the canonical correlation is calculated for each voxel.

The main principle of the GPU implementation is that each GPU thread works on the time series of one voxel. For the multi-GPU implementation, each GPU performs one third of the permutations. This means that the processing time scales linearly with the number of GPUs.

As an fMRI dataset normally contains more than 20 000 brain voxels, it is necessary to calculate a significance threshold that is *corrected* for the multiple testing, otherwise there will be a lot of false positives. This is done by only saving the maximum canonical correlation from each permutation, such that the *maximum* null distribution is estimated. For this reason, it is necessary to apply at least 10 000 random permutations in order to get a good estimate of the significance threshold.

3 Results

3.1 Data

Three single subject datasets have been used to test our algorithms, the test subject was a 50 year old healthy male. The data was collected with a 1.5 T Philips Achieva MR scanner. The following settings were used: repetition time 2 s, echo time 40 ms, flip angle 90 degrees, isotropic voxel size 3.75 mm. A field of view of 240 mm thereby resulted in slices with 64 x 64 pixels, a total of 22 slices were collected every other second. The experiments were 160 s long, resulting in 80 volumes to be processed. The datasets contain about 20 000 within-brain voxels. For the *Motor 1* dataset the subject periodically activated the left hand (20 s activity, 20 s rest), for the *Motor 2* dataset the subject periodically activated the right hand. For the *Language* dataset the subject periodically performed a reading task (20 s activity, 20 s rest). The task was to read sentences and determine if they were reasonable or not.

3.2 Processing time

By using our multi-GPU implementation with three Nvidia GTX 480 GPUs (giving a total of 1440 processor cores), significance thresholds, corrected for

multiple testing, for CCA can be calculated from a random permutation test with 10 000 permutations in about 80 seconds, compared to about 20 hours with a Matlab implementation and 3 hours with a multi-threaded C implementation. To objectively evaluate the effect of the covariance pooling is clearly not possible without the computational power of the GPU.

3.3 The effect of covariance pooling on the null distribution

To see how the covariance pooling affects the null distribution of the maximum canonical correlation coefficient, the null distribution was estimated when different amounts of smoothing (0, 8, 15 mm FWHM) were applied to the elements of the covariance matrices, 100 000 permutations were used to estimate each distribution. The resulting null distributions are given in Figure 1. The canonical correlation thresholds, for corrected $p = 0.05$, are 0.63, 0.37 and 0.23 respectively. With the covariance pooling the canonical correlations are generally lower. This might at first seem to be bad, but it is not the strength of a correlation *per se* that is interesting, but rather how strong a correlation is compared to the null distribution.

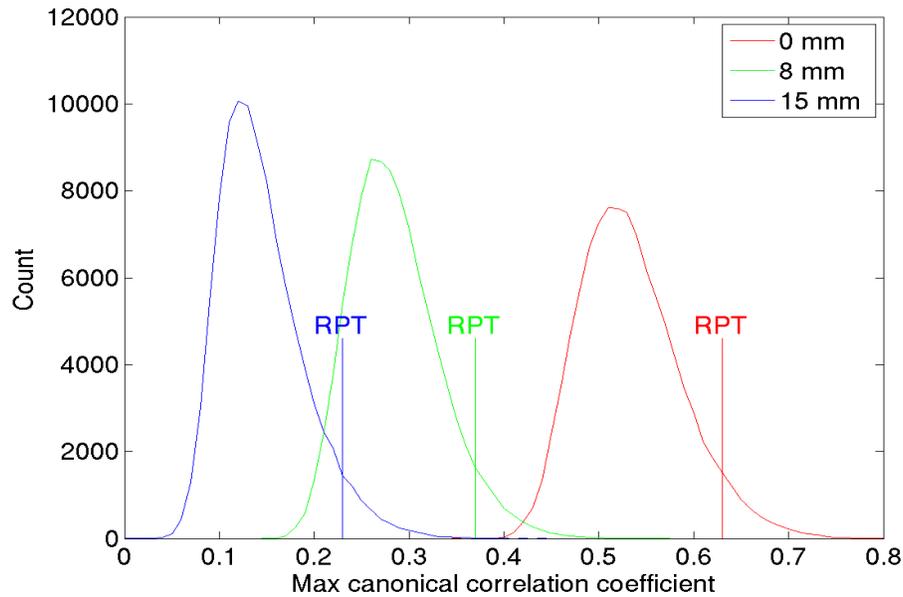


Fig. 1. The null distribution of the maximum canonical correlation coefficient, when different amounts of covariance pooling were applied (from right to left, 0 mm (red), 8 mm (green), 15 mm (blue)). The thresholds for corrected $p = 0.05$ are marked with RPT (random permutation test). The Motor 1 dataset was used with 8 mm of smoothing of the fMRI volumes.

3.4 The effect of covariance pooling on the activity map

To see how the covariance pooling affects the activity map, the activity map was calculated when different amounts of smoothing (0, 8, 15 mm FWHM) were applied to the elements of the covariance matrices (the fMRI volumes were smoothed with 8 mm in all three cases). The original and thresholded activity maps for one slice are given in Figure 2. The activity maps were thresholded at the same significance level, corrected $p = 0.05$. No smoothing was applied to the activity map, only to the covariance estimates.

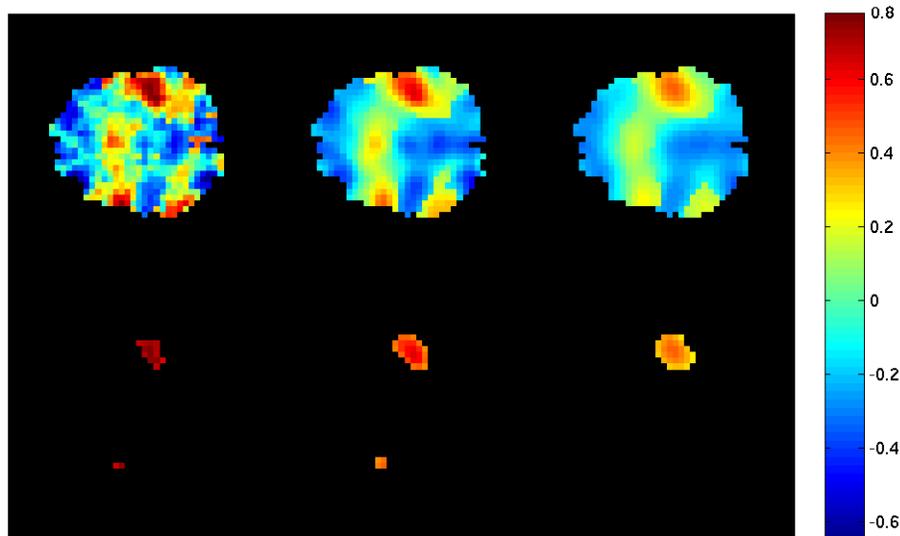


Fig. 2. The resulting original activity maps (top) and thresholded activity maps (bottom) for different amounts of covariance pooling. The Motor 1 dataset was used with 8 mm of smoothing of the fMRI volumes. **Left:** The resulting activity map when no covariance pooling was used. Note that the activity map is very rough, due to bad estimates of the covariances. **Middle:** The resulting activity map when 8 mm covariance pooling was used. Note that the activity map is much smoother than without the covariance pooling. The number of significantly voxels clearly increases. **Right:** The resulting activity map when 15 mm covariance pooling was used. The number of active voxels increases even more, but at the price of a decreased spatial locality and a loss of small activity clusters.

3.5 Number of activity clusters and significantly active voxels

The number of voxels that are classified as significantly active is one way to objectively evaluate different methods for detecting brain activity. The number of voxels that were classified as significantly active (at $p = 0.05$, corrected for

multiple testing), as function of the amount of covariance pooling, are given in Figure 3. For each case, 10 000 permutations were used and 8 mm of smoothing was applied to the fMRI volumes. Smoothing of the covariance estimates clearly increases the number of significantly active voxels.

It is however not only the number of active voxels that is interesting, but also the spatial locality of the activity map. Spatial locality is however rather hard to measure. A simple related measure is the number of clusters in the thresholded activity maps. The number of clusters for each amount of covariance pooling are given in Figure 4. It is clear that small activity clusters can be missed if too much smoothing is applied to the covariance estimates. The number of clusters is however not a perfect measure of the spatial locality, since two neighbouring active voxels can count as two clusters if a voxel between them is inactive, but as one cluster if the inactive voxel becomes active as well.

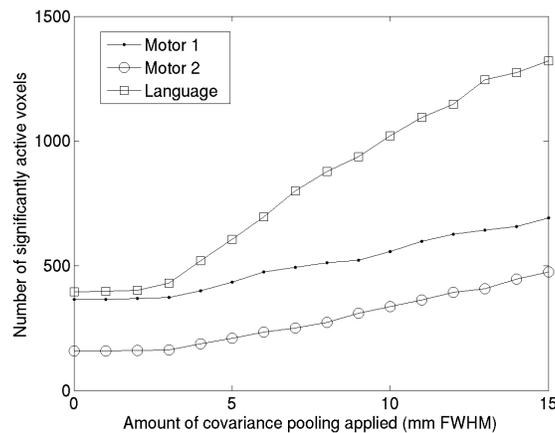


Fig. 3. The number of significantly active voxels, for the three datasets, as function of the amount of smoothing applied to the covariance estimates.

4 Discussion

We have presented how to apply pooling of covariance matrices, in order to improve CCA based fMRI analysis. The number of voxels that can be classified as significantly active increase with 40 - 120% (for 8 mm of smoothing). The covariance estimates are improved as the amount of smoothing increases, but at the price of decreased spatial locality in the activity map. A visual inspection of Figures 2, 3 and 4 indicates that a smoothing of 7 - 10 mm FWHM seems to give the best compromise between an increase in the number of significantly active voxels and a decrease in the number of activity clusters.

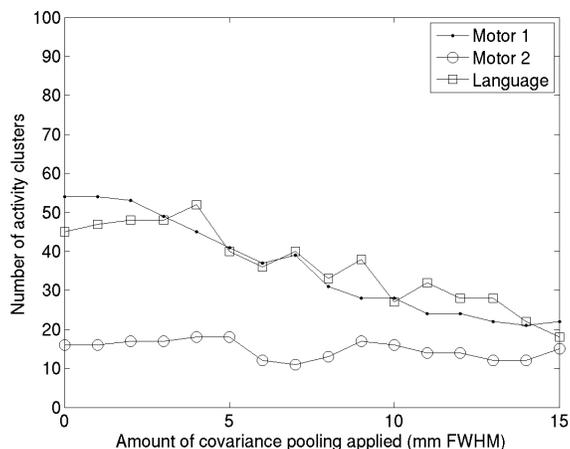


Fig. 4. The number of activity clusters, for the three datasets, as function of the amount of smoothing applied to the covariance estimates.

Without the random permutation test it would be impossible to predict exactly how the covariance pooling affects the null distribution of the maximum canonical correlation coefficient and thereby the resulting activity maps can not be compared in an objective way. The calculations that were made in order to generate the results in this paper would have taken a total of about 65 days with a Matlab implementation and 10 days with a multi-threaded C implementation, with our multi-GPU implementation they took about 2 hours. We have thereby presented the first step towards a new era of fMRI analysis, where suggested improvements of existing methods easily can be evaluated in an objective way.

Acknowledgement

This work was supported by the Linnaeus center CADICS, funded by the Swedish research council. 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. A.R. Ferreira da Silva: A Bayesian multilevel model for fMRI data analysis. *Computer Methods and Programs in Biomedicine* 102, 238–252 (2010)
2. Biswal, B., Yetkin, F., Haughton, V., Hyde, J.: Functional connectivity in the motor cortex of resting state human brain using echo-planar MRI. *Magnetic Resonance in Medicine* 34, 537–541 (1995)

3. Brammer, M.J., Bullmore, E.T., Simmons, A., Williams, S.C.R., Grasby, P.M., Howard, R.J., R.Woodruff, P., Rabe-Hesketh, S.: Generic brain activation mapping in functional magnetic resonance imaging: A nonparametric approach. *Magnetic Resonance Imaging* 15, 763–770 (1997)
4. Dwass, M.: Modified randomization tests for nonparametric hypotheses. *The Annals of Mathematical Statistics* 28, 181–187 (1957)
5. Eklund, A., Andersson, M., Knutsson, H.: Phase based volume registration using CUDA. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010. pp. 658–661 (2010)
6. Eklund, A., Friman, O., Andersson, M., Knutsson, H.: A GPU accelerated interactive interface for exploratory functional connectivity analysis of fMRI data. In: *IEEE International Conference on Image Processing (ICIP)* (2011)
7. Eklund, A., Andersson, M., Knutsson, H.: Fast random permutation tests enable objective evaluation of methods for single subject fMRI analysis. *International Journal of Biomedical Imaging*, Article ID 627947, In Press (2011)
8. Eklund, A., Andersson, M., Knutsson, H.: fMRI analysis on the GPU - possibilities and challenges. *Computer Methods and Programs in Biomedicine*, <http://dx.doi.org/10.1016/j.cmpb.2011.07.007> (2011)
9. Friman, O., Borga, M., Lundberg, P., Knutsson, H.: Adaptive analysis of fMRI data. *NeuroImage* 19, 837–845 (2003)
10. Friman, O., Carlsson, J., Lundberg, P., Borga, M., Knutsson, H.: Detection of neural activity in functional MRI using canonical correlation analysis. *Magnetic Resonance in Medicine* 45(2), 323–330 (2001)
11. Friston, K., Holmes, A., Worsley, K., Poline, J., Frith, C., Frackowiak, R.: Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping* 2, 189–210 (1995)
12. Gembris, D., Neeb, M., Gipp, M., Kugel, A., Männer, R.: Correlation analysis on GPU systems using NVIDIA's CUDA. *Journal of real-time image processing* pp. 1–6 (2010)
13. Holmes, A., Blair, R., Watson, J., Ford, I.: Nonparametric analysis of statistic images from functional mapping experiments. *Journal of Cerebral Blood Flow & Metabolism* 16, 7–22 (1996)
14. Hotelling, H.: Relation between two sets of variates. *Biometrika* 28, 322–377 (1936)
15. John L. Van Hemert, Julie A. Dickerson: Monte Carlo randomization tests for large-scale abundance datasets on the GPU. *Computer Methods and Programs in Biomedicine* 101, 80–86 (2011)
16. Nandy, R., Cordes, D.: A novel nonparametric approach to canonical correlation analysis with applications to low CNR functional MRI data. *Magnetic Resonance in Medicine* 49, 1152–1162 (2003)
17. Nichols, T.E., Holmes, A.P.: Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping* 15, 1–25 (2001)
18. Ragnehed, M., Engström, M., Knutsson, H., Söderfeldt, B., Lundberg, P.: Restricted canonical correlation analysis in functional MRI - validation and a novel thresholding technique. *Journal of Magnetic Resonance Imaging* 29, 146–154 (2009)
19. Shterev, I., Jung, S.H., George, S., Owzar, K.: permGPU: Using graphics processing units in RNA microarray association studies. *BMC Bioinformatics* 11, 329 (2010)
20. Worsley, K., Liao, C., Aston, J., Petre, V., Duncan, G., Morales, F., Evans, A.: A general statistics analysis for fMRI data. *NeuroImage* 15, 1–15 (2002)
21. Worsley, K., Marrett, S., Neelin, P., Friston, K., Evans, A.: A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping* 4, 58–73 (1996)