Comparing Independent Component Analysis and the Parafac model for artificial multi-subject fMRI data

Alwin Stegeman[1]

University of Groningen

The Netherlands

February 27, 2007

Abstract

Recently, Beckmann and Smith (2005) compared a three-way extension of Independent Component Analysis (ICA) and the Parafac model, as applied to artificial multi-subject fMRI data (voxels $\times$ scans $\times$ subjects). They concluded that the ICA approach yields more accurate estimates of the underlying signal sources and results in less interference between the different sources compared to the Parafac estimates. Moreover, the ICA approach is more robust against overfitting and its computational load is much less than Parafac. In this paper, we offer detailed explanations of the differences between Parafac and the ICA approach and show that the distinction between second-order statistics versus higher-order statistics does not apply to Parafac versus ICA. Using the data of Beckmann and Smith (2005), we show that Parafac performs as well as the ICA approach if the correct number of signal sources is chosen, which is possible by considering Parafac fit values. Moreover, if the fMRI spatial activity maps are well-overlapping, then the ICA approach does not find the correct maps while Parafac does. Additionally, we present and demonstrate a method to decrease the computational load of Parafac.

Keywords: Independent Component Analysis, Parafac, three-way arrays, tensor decompositions, functional Magnetic Resonance Imaging.

[1]Author's address: Alwin Stegeman, Heijmans Institute of Psychological Research, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands, phone: ++31 50 363 6193, fax: ++31 50 363 6304, email: *a.w.stegeman@rug.nl*

1 Introduction

Recently, Beckmann and Smith (2005) presented a three-way extension of the Probabilistic Independent Component Analysis (PICA) model for fMRI data, and named it the Tensor PICA model. Beckmann and Smith (2005) applied the Tensor PICA model to artificial multi-subject fMRI data (voxels $\times$ scans $\times$ subjects) and compared its performance to the Parafac model for three-way component analysis. They concluded that Tensor PICA yields more accurate estimates of the underlying signal sources (i.e. voxel activation maps) and results in less interference between the different sources compared to the Parafac estimates. Moreover, Tensor PICA is more robust against overfitting and its computational load is much less than Parafac.

The explanations given by Beckmann and Smith (2005) for the differences in performance between Tensor PICA and Parafac are as follows. They argue that the higher accuracy, less interference and higher robustness of Tensor PICA with respect to Parafac are due to "the stronger statistical constraints on the spatial domain" (Beckmann & Smith, 2005, p. 309), combined with the sparsity in the spatial domain of typical fMRI activation.

In this paper, we provide detailed theoretical and data analytic explanations for these differences in performance. Also, we assess the differences in performance of Parafac and Tensor PICA for the artificial multi-subject fMRI data of Beckmann and Smith (2005), where the signal-to-noise ratio (SNR) and the level of sparsity of the signal in the spatial domain are varied.

The paper has a theoretical and a data analytic part. In the theoretical part, we show that a comparison between the Parafac model and the Tensor PICA model is different from a comparison between Principal Components Analysis (PCA) and Independent Component Analysis (ICA) for two-way data. The distinction between PCA and ICA is usually characterized by the use of second-order statistics (uncorrelated components in PCA) versus higher-order statistics (statistically independent components in ICA). However, this distinction does not apply to a comparison of Parafac and Tensor PICA.

In the data analytic part, we consider the artificial multi-subject data and show that Parafac recovers the signals as well as Tensor PICA if the correct number of signal sources is chosen. This number can be inferred from Parafac fit values for increasing numbers of signal sources. This conclusion holds for different values of the signal-to-noise ratio (SNR) and different degrees of sparsity of the signals in the spatial domain.

Also, we show that if the artificial spatial fMRI activity maps are well-overlapping, then Tensor PICA does not find the correct maps while Parafac does. This is because Tensor PICA forces the ICA assumption of statistically independent maps on the overlapping fMRI activity maps.

The main advantages of the Tensor PICA method over Parafac are that the run-to-run variability of solutions is much less and interference between signal components does not occur. The latter is common in Parafac solutions if the number of components is larger than the number of signal sources present in the data, i.e. in cases of overfitting. We will explain that the robustness properties of Tensor PICA are due to the robustness of the ICA algorithm used to find the spatial fMRI activation maps. The high run-to-run variability of Parafac solutions is due to the low SNR

of the fMRI dataset.

Regarding the comparison of computational load between the Parafac and Tensor PICA algorithms, we show that the computational load of Parafac can be decreased considerably by compressing the data and imposing a restriction of the spatial components.

This paper is organized as follows. In Section 2 we discuss the component models relevant for our purposes. We first discuss PCA and ICA (and their probabilistic versions) in some detail. Next, we present the Parafac and Tensor PICA models and show how their comparison is different from PCA versus ICA. In Section 3, we compare Parafac and Tensor PICA using the artificial multi-subject fMRI data of Beckmann and Smith (2005) for various SNR levels and various sparsity levels in the spatial domain. Also, we consider several other variations of these datasets. In Section 4, we show how the computational load of Parafac may be reduced and use this faster algorithm on the artificial fMRI data. In Section 5, we present a discussion of our results.

2 Model descriptions

2.1 Principal Component Analysis

Principal Component Analysis (PCA) is an exploratory data analysis tool for extracting uncorrelated components and their loadings from an $n \times m$ data matrix $\mathbf{X}$ containing for instance the scores of n subjects on m variables. The matrix PCA model is

$$\mathbf{X} = \mathbf{A}\,\mathbf{B}^T + \mathbf{E} = \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{b}_r + \mathbf{E}\,, \tag{2.1}$$

where the columns $\mathbf{a}_r$ of $\mathbf{A}$ $(n \times R)$ are the latent components, the columns $\mathbf{b}_r$ of $\mathbf{B}$ $(m \times R)$ contain the loadings of the m variables on the R components, $\circ$ denotes the outer product, and $\mathbf{E}$ is a residual term, see Pearson (1901). It is usually assumed that $m \leq n$, the columns of $\mathbf{X}$ (variables) have mean zero and unit variance, the columns of $\mathbf{A}$ (components) have mean zero and unit variance and are uncorrelated, i.e. $n^{-1}\mathbf{A}^T\mathbf{A} = \mathbf{I}_R$. A PCA solution is found by minimizing the sum-of-squares of $\mathbf{E}$, i.e. the square of the Frobenius norm $\|\underline{\mathbf{E}}\|$.

Eckart and Young (1936) show that a PCA solution $(\mathbf{A}, \mathbf{B})$ can be obtained from the Singular Value Decomposition (SVD) of $\mathbf{X}$. Indeed, let the SVD of $\mathbf{X}$ be given by $\mathbf{X} = \mathbf{U}\,\mathbf{S}\,\mathbf{V}^T$, where $\mathbf{U}$ $(n \times m)$ has orthogonal columns, $\mathbf{S}$ is the diagonal matrix containing the singular values in decreasing order, and $\mathbf{V}$ $(m \times m)$ is orthogonal. Then a PCA solution satisfying the assumptions above, is

$$\mathbf{A} = n^{1/2}\,\mathbf{U}_R\,, \qquad \mathbf{B}^T = n^{-1/2}\,\mathbf{S}_R\,\mathbf{V}_R^T\,, \tag{2.2}$$

where $\mathbf{U}_R$ and $\mathbf{V}_R$ contain the first R columns of $\mathbf{U}$ and $\mathbf{V}$, respectively, and $\mathbf{S}_R$ is the diagonal matrix containing the first R singular values of $\mathbf{X}$. Hence, $\mathbf{U}_R\,\mathbf{S}_R\,\mathbf{V}_R^T$ is a truncated SVD of $\mathbf{X}$ using only the first (and largest) R singular values and vectors. Eckart and Young (1936) show that this PCA solution $\mathbf{A}\,\mathbf{B}^T$ is a best rank-R approximation to $\mathbf{X}$. Note that for this PCA solution the first component explains the most variance in the data, the second component explains the most

variance in the data after the first component has been subtracted, etcetera. This is due to the ordering of the singular values and the orthogonality of the left- and right singular vectors.

The PCA solution above is unique (up to sign changes) in this ordering if the first R singular values are distinct. However, for any $R \times R$ orthogonal matrix $\mathbf{Q}$, the solution $(\mathbf{AQ}, \mathbf{BQ})$ has the same residuals $\mathbf{E}$, since $\mathbf{A}\,\mathbf{B}^T = (\mathbf{AQ})\,(\mathbf{BQ})^T$. Hence, only the space of the PCA components $\mathbf{A}$ is uniquely determined (ignoring the ordering). Within this space, any set of basis vectors can be taken as components. In psychological applications of PCA, the ordering of the components is usually not meaningful and a rotation $\mathbf{Q}$ is calculated which yields interpretable components, see Browne (2001).

Probabilistic PCA

Tipping and Bishop (1999) present a Probabilistic PCA (PPCA) model in which the PCA model (2.1) is given a stochastic interpretation. The data matrix $\mathbf{X}$ is interpreted as consisting of n samples or observations of m random variables. We denote the $m \times 1$ vector of these m random variables as $\mathbf{x}$. Analogously, the component matrix $\mathbf{A}$ is interpreted as consisting of n samples of R random variables, which are also called *source signals* or *sources*. We denote the $R \times 1$ vector of these R source random variables as $\mathbf{a}$. The PPCA model is

$$\mathbf{x} = \mathbf{B}\,\mathbf{a} + \mathbf{e}\,, \tag{2.3}$$

where the $m \times R$ matrix $\mathbf{B}$ is called the *mixing matrix* and $\mathbf{e}$ denotes the $m \times 1$ vector of noise random variables. The random vector $\mathbf{a}$ is assumed to have a Gaussian distribution with mean zero and variance $\mathbf{I}_R$, i.e. $\mathbf{a} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_R)$. Also, it is assumed that $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_m)$, the source variables $\mathbf{a}$ and the errors $\mathbf{e}$ are statistically independent and $\mathrm{rank}(\mathbf{B}) = R$. Note that due to Gaussianity, the sources $\mathbf{a}$ are not only uncorrelated but also statistically independent. The goal is to estimate $\mathbf{B}$ and n realisations of the random vector $\mathbf{a}$, which are the rows of $\mathbf{A}$ $(n \times R)$.

Tipping and Bishop (1999) show that the Maximum Likelihood estimators of $\mathbf{A}$, $\mathbf{B}$ and σ^2 are given by

$$\widehat{\mathbf{A}}_{\mathrm{ML}} = \mathbf{U}_R\,\mathbf{S}_R\,(n^{-1}\mathbf{S}_R^2 - \sigma^2\,\mathbf{I}_R)^{-1/2}\,\mathbf{Q}\,, \tag{2.4}$$

$$\widehat{\mathbf{B}}_{\mathrm{ML}} = \mathbf{V}_R\,(n^{-1}\mathbf{S}_R^2 - \sigma^2\,\mathbf{I}_R)^{1/2}\,\mathbf{Q}\,, \tag{2.5}$$

$$\hat{\sigma}^2_{\mathrm{ML}} = \frac{1}{(m-R)\,n}\sum_{j=R+1}^{m} s_{jj}^2\,, \tag{2.6}$$

where $\mathbf{Q}$ is any $R \times R$ orthogonal rotation matrix, $\mathbf{U}_R\,\mathbf{S}_R\,\mathbf{V}_R$ is the truncated SVD of $\mathbf{X}$, and s_{jj}^2 is the j-th diagonal element of $\mathbf{S}^2$. It can be seen that the difference between the PPCA estimates (2.4) and (2.5) and the PCA solution (2.2) is that the noise variance is taken into account in the PPCA estimates. Recall that also the PCA solution (2.2) can be rotated by $\mathbf{Q}$ without loss of fit.

The PPCA model (2.3) and the assumptions above imply that

$$\mathrm{Cov}(\mathbf{x}) = E(\mathbf{x}\mathbf{x}^T) = \mathbf{B}\, E(\mathbf{a}\mathbf{a}^T)\, \mathbf{B}^T + E(\mathbf{e}\mathbf{e}^T) = \mathbf{B}\mathbf{B}^T + \sigma^2\, \mathbf{I}_m\,. \tag{2.7}$$

Replacing $E(\mathbf{x}\mathbf{x}^T)$ by its estimator $n^{-1}\mathbf{X}^T\mathbf{X}$ and using the SVD of $\mathbf{X}$ it follows from (2.7) that

$$\mathbf{B}\mathbf{B}^T = \mathbf{V}\,(n^{-1}\mathbf{S}^2 - \sigma^2\, \mathbf{I}_m)\, \mathbf{V}^T\,. \tag{2.8}$$

Since $\mathbf{B}\mathbf{B}^T$ is non-negative definite, $\mathrm{rank}(\mathbf{B}\mathbf{B}^T) = \mathrm{rank}(\mathbf{B}) = R$ and (2.8) is an eigendecomposition of $\mathbf{B}\mathbf{B}^T$, it follows that the last (and smallest) $m - R$ diagonal elements of $(n^{-1}\mathbf{S}^2 - \sigma^2\, \mathbf{I}_m)$ are zero. By considering the eigenvalues of $n^{-1}\mathbf{X}^T\mathbf{X}$, this fact can be used to obtain a rough estimate of the number R of source signals present.

2.2 Independent Component Analysis

In the Independent Component Analysis (ICA) framework, the PCA model (2.1) is given a stochastic interpretation and the assumption of uncorrelated components $\mathbf{a}$ is strengthened to statistically independent components, i.e.

$$P(a_1 \leq \tau_1, \ldots, a_R \leq \tau_R) = P(a_1 \leq \tau_1) \cdots P(a_R \leq \tau_R)\,, \tag{2.9}$$

where a_j denotes the j-th source random variable and τ_j are real numbers. This approach is the same as PPCA discussed above. However, in the ICA framework, the source signals are assumed to have a non-Gaussian distribution. As we will see below, this assumption fixes the rotational ambiguity of PCA and PPCA in the presence of Gaussian noise. For an introduction to ICA we refer to De Lathauwer, De Moor, and Vandewalle (2000) and the references therein. For a rigorous mathematical treatment of ICA, see Comon (1994a). The principles of ICA will be made clear in a discussion of the Probabilistic ICA model.

Probabilistic ICA

Next, we discuss the Probabilistic ICA (PICA) model of Penny, Roberts, and Everson (2001), which is adapted by Beckmann and Smith (2004). The PICA model is analogous to the PPCA model (2.3) discussed above. The only difference is that the source signals $\mathbf{a}$ have a non-Gaussian distribution. Note that they are still assumed statistically independent with mean zero and unit variance. Since the form of the distribution of $\mathbf{a}$ is usually not known exactly, the Maximum Likelihood estimates of $\mathbf{A}$, $\mathbf{B}$ and σ^2 cannot be calculated exactly. However, the estimates (2.4)-(2.6) can still be used. Indeed, for the PICA model, equations (2.7) and (2.8) still hold. As argued above, we may replace (2.8) by

$$\mathbf{B}\mathbf{B}^T = \mathbf{V}_R\,(n^{-1}\mathbf{S}_R^2 - \sigma^2\, \mathbf{I}_R)\, \mathbf{V}_R^T\,, \tag{2.10}$$

and it follows that the SVD of $\mathbf{B}$ is given by (2.5). For a fixed mixing matrix $\mathbf{B}$, we obtain the sources in $\mathbf{A}$ as the parameters of the regression problem $\mathbf{X}^T = \mathbf{B}\,\mathbf{A}^T + \mathbf{E}^T$. Combining the standard estimate $\mathbf{A}^T = (\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{X}^T$, the SVD of $\mathbf{X}$ and the expression for $\mathbf{B}$ in (2.5), yields

the estimate of $\mathbf{A}$ in (2.4). As for σ^2, under the PICA model it still holds that the last (and smallest) $m - R$ diagonal elements of $(n^{-1}\mathbf{S}^2 - \sigma^2\,\mathbf{I}_m)$ are zero. Hence, a logical estimate for σ^2 is (2.6).

The PICA estimation procedure can be considered as a two-step algorithm. The first step has been described above and yields approximately uncorrelated PPCA sources and fixes $\mathbf{A}$ and $\mathbf{B}$ up to an orthogonal rotation. The second step then finds an optimal rotation $\mathbf{Q}$ such that the sources are approximately statistically independent, using some measure of independence. The optimal $\mathbf{Q}$ is usually found by considering the whitened data $\widetilde{\mathbf{X}}^T = (n^{-1}\mathbf{S}_R^2 - \sigma^2\,\mathbf{I}_R)^{-1/2}\,\mathbf{V}_R^T\,\mathbf{X}^T$, which satisfies

$$\widetilde{\mathbf{X}}^T = \mathbf{Q}\,\mathbf{A}^T + \widetilde{\mathbf{E}}^T\,, \tag{2.11}$$

where $\widetilde{\mathbf{E}}^T = (n^{-1}\mathbf{S}_R^2 - \sigma^2\,\mathbf{I}_R)^{-1/2}\,\mathbf{V}_R^T\,\mathbf{E}^T$. Note that (2.11) follows from $\mathbf{X}^T = \mathbf{B}\,\mathbf{A}^T + \mathbf{E}^T$ and (2.5).

ICA algorithms

The two steps of the PICA estimation procedure are common to a large class of ICA algorithms. Next, we describe two different ICA algorithms to obtain an optimal rotation $\mathbf{Q}$ of the components (2.4). Let the whitened ICA model be given by $\tilde{\mathbf{x}} = \mathbf{Q}\,\mathbf{a} + \tilde{\mathbf{e}}$, see (2.11), with non-Gaussian sources $\mathbf{a}$ and Gaussian noise $\tilde{\mathbf{e}}$. Then $\mathbf{a} = \mathbf{Q}^T\tilde{\mathbf{x}} - \mathbf{Q}^T\tilde{\mathbf{e}}$. The goal is to find $\mathbf{Q}$ such that $\mathbf{a}$ are approximately statistically independent.

The ICA algorithm of Comon (1994a) considers the K-th order cumulants of the sources $\mathbf{a}$, which are given by

$$\mathrm{Cum}^{(K)}(a_{j_1}, \ldots, a_{j_K}) = \sum (-1)^{k-1}(k-1)!\; E\left(\prod_{j\in S_1} a_j\right)\cdots E\left(\prod_{j\in S_k} a_j\right)\,, \tag{2.12}$$

where $a_{j_1}, \ldots, a_{j_K}$ are elements of the random vector $\mathbf{a}$ and the summation involves all possible partitions $\{S_1, \ldots, S_k\}$, $1 \le k \le K$, of the integers $\{j_1, \ldots, j_K\}$. For $K \ge 3$, the cumulants of the sources $\mathbf{a}$ are equal to the cumulants of $\mathbf{Q}^T\,\tilde{\mathbf{x}}$. This is due to the multilinearity of the cumulant function and the fact that for $K \ge 3$ the cumulants of Gaussian random variables are zero, see De Lathauwer et al. (2000). To obtain statistically independent sources, the following property of cumulants is used. If the mean-zero sources $\mathbf{a}$ are statistically independent, then $\mathrm{Cum}^{(K)}(a_{j_1}, \ldots, a_{j_K}) \neq 0$ only if $j_1 = j_2 = \ldots = j_K$, $K \ge 1$. The algorithm of Comon (1994a) makes the K-way array of (the sample estimates of) the K-th order cumulants of $\mathbf{Q}^T\,\tilde{\mathbf{x}}$ as diagonal as possible by choosing the rotation $\mathbf{Q}$. An algorithm for $K = 4$ can be obtained from Comon (1994b). A more efficient algorithm for $K = 4$ is called JADE and is due to Cardoso and Souloumiac (1993). For other algorithms using cumulants see De Lathauwer et al. (2000) and the references therein. Comon (1994a) shows that approximate diagonalisation of the (estimate of the) cumulant K-way array of $\mathbf{a}$ is a natural ICA algorithm in view of results from information theory. In this paper we will denote the K-th order cumulant algorithm of Comon (1994a) as Comon-K.

A different ICA algorithm is due to Hyvärinen (1999) which uses as a measure of non-Gaussianity

$$[E(G(a)) - E(G(z))]^2 \ , \tag{2.13}$$

where a is one source variable of $\mathbf{a}$, z is a standard Gaussian variable and G is some non-quadratic function. The (sample estimate of the) objective function (2.13) is maximized over $\mathbf{Q}$ using approximative Newton iterations and this procedure is used for each source variable separately. This algorithm is called the FastICA algorithm (Hyvärinen & Oja, 2000; Hyvärinen, 2005). For remarks on the influence of the choice of G on the performance of the FastICA algorithm, see Hyvärinen (1999). In Hyvärinen and Oja (2000) it is shown that the ICA objective function (2.13) is natural from an information theoretical point of view.

General remarks on ICA

The criterion used to find the statistically independent sources $\mathbf{a}$ is often referred to as *maximizing non-Gaussianity*. Under Gaussian noise and non-Gaussian sources, the orthogonal linear combinations of the sources are to be found which differ from Gaussian random variables as much as possible. The underlying reason is that it follows from the Central Limit Theorem that any non-trivial linear combination of non-Gaussian variables becomes "more Gaussian". An ICA algorithm incorporating this approach maximizes some measure of non-Gaussianity over the orthogonal linear combinations $\mathbf{Q}$ of the sources. In Section 3, we describe two different ICA algorithms used in this paper. Under the assumption of Gaussian noise and non-Gaussian sources, and all other assumptions above, the ICA solution $(\mathbf{A}, \mathbf{B})$ is unique up to sign changes and a permutation of the source signals (i.e. a permutation of the columns of $\mathbf{A}$ and $\mathbf{B}$ simultaneously), see De Lathauwer et al. (2000).

The first step of the ICA estimation procedure, yielding approximately uncorrelated yet rotationally ambiguous sources, is called the second-order step since it finds approximately uncorrelated sources by using second-order statistics. The second step of the PICA algorithm then uses higher-order statistics to obtain an optimal rotation $\mathbf{Q}$ and approximately statistically independent sources. ICA is sometimes presented as a "higher-order fine tuning of PCA", since it considers higher-order statistics of the sources instead of only the order-2 restriction of uncorrelatedness in PCA and it fixes the non-unique PCA components by finding an optimal rotation.

It should be noted that ICA solutions may depend on the particular objective function that is used to quantify the notion of "maximally non-Gaussian sources".

Apart from the ICA framework decribed above, there are also other ICA approaches. For example, stationary sources which are individually correlated in time can be separated by a joint diagonalization of a set of covariance matrices (Belouchrani, Abed-Meraim, Cardoso, & Moulines, 1997). In this case only second-order statistics are used. ICA algorithms for non-stationary sources are presented in Pham and Cardoso (2001).

The PICA model for single-subject fMRI data

Beckmann and Smith (2004) apply the PICA model to single-subject fMRI data. The $n \times m$ data matrix $\mathbf{X}$ consists of m fMRI brain scans of the same subject, where the fMRI signal is measured in n voxels (volume elements of the brain). Hence, $\mathbf{x}$ is the vector of m scan variables having n realisations (one for each voxel) which constitute the rows of $\mathbf{X}$. The rows of $\mathbf{A}$ contain n realisations of the R independent source processes and may be considered as R independent ICA spatial maps of voxel activity. the mixing matrix $\mathbf{B}$ then contains the R associated time courses as columns. For each voxel, the noise $\mathbf{e}$ is Gaussian distributed, which is a good approximation of the theoretical Rician distribution of magnitudes in MRI images, see Gudbjartsson and Patz (1995) and Wink and Roerdink (2006). Of course, this assumes that physiological artifacts due to respiration, heart beat, subject movement, etc, have either been removed or are considered as signals. For single-subject fMRI data, Thomas, Harshman, and Menon (2002) conclude that ICA is a suitable technique for the recognition and removal of physiological artifacts. For details on artifact recognition and removal using ICA, we refer to Thomas et al. (2002), the discussion in McKeown, Hansen, and Sejnowski (2003) and the references therein. Apart from being Gaussian, the noise $\mathbf{e}$ is also assumed to be statistically independent over the m scans and to have the same variance σ^2 for each voxel, which is not likely to hold for fMRI data. Therefore, prior to calculating the ICA estimates for the mixing matrix $\mathbf{B}$ and the spatial maps $\mathbf{A}$, the fMRI data should be temporally pre-whitened for each voxel. For details, we refer to Beckmann and Smith (2004) and the references therein. Also, the average scan should be subtracted to make the data mean zero.

Note that the PICA model assumes statistically independent spatial maps of voxel activity as ICA components, while the temporal patterns form the mixing matrix. Another ICA approach is known in the literature in which the statistically independent ICA components are the temporal patterns and the spatial patterns for the mixing matrix. For a discussion, see Calhoun, Adali, Hansen, Larsen, and Pekar (2003) and the references therein. Spatially independent ICA for single-subject fMRI data was first applied by McKeown, Makeig, Brown, Jung, Kindermann, Bell, and Sejnowski (1998) and was found to outperform PCA in recovering voxel activation patterns. For an overview of ICA for fMRI data we refer to Calhoun et al. (2003).

2.3 The Parafac model

Carroll and Chang (1970) and Harshman (1970) have independently proposed the same method for component analysis of three-way data arrays, and named it Candecomp and Parafac, respectively. Since the term "Parafac model" is more widely known than "Candecomp", we use the former term. We denote the Parafac model as

$$\underline{\mathbf{X}} = \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r + \underline{\mathbf{E}}\,, \tag{2.14}$$

where $\underline{\mathbf{X}}$ is an $n \times m \times p$ data array and the vectors $\mathbf{a}_r$, $\mathbf{b}_r$ and $\mathbf{c}_r$ form the columns of the component matrices $\mathbf{A}$ $(n \times R)$, $\mathbf{B}$ $(m \times R)$ and $\mathbf{C}$ $(p \times R)$, respectively. The sum-of-squares of the residual

array $\underline{\mathbf{E}}$ (i.e. $\|\underline{\mathbf{E}}\|^2$) is minimized to find the R components $\mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r$. The can be done using an Alternating Least Squares algorithm, in which each component matrix is sequentially optimized, given the other two component matrices, see Carroll and Chang (1970).. For an overview and comparison of several Parafac algorithms, see Tomasi and Bro (2006).

A matrix notation of the Parafac model (2.14) is as follows. Let $\mathbf{X}_k$ $(n \times m)$ and $\mathbf{E}_k$ $(n \times m)$ denote the k-th slices of $\underline{\mathbf{X}}$ and $\underline{\mathbf{E}}$, respectively. Then (2.14) can be written as

$$\mathbf{X}_k = \mathbf{A}\,\mathbf{C}_k\,\mathbf{B}^T + \mathbf{E}_k\,, \qquad k = 1, \ldots p\,, \tag{2.15}$$

where $\mathbf{C}_k$ $(R \times R)$ is the diagonal matrix with the k-th row of $\mathbf{C}$ as diagonal.

The Parafac model (2.14) can be seen as a three-way extension of the PCA model (2.1) for matrices. For example, if $\mathbf{A}$ is interpreted as the components in the first mode, then $\mathbf{B}$ and $\mathbf{C}$ are the loadings on these components for the second and third modes, respectively. However, essentially the Parafac model is symmetric in its three modes. The real-valued Parafac model, i.e. where $\underline{\mathbf{X}}$ and the model parameters are real-valued, is used in a majority of applications in psychometrics and chemometrics; see Kroonenberg (1983) and Smilde, Bro, and Geladi (2004). Complex-valued applications of Parafac occur in e.g. signal processing and telecommunications research; see Sidiropoulos (2004). In this paper, we only consider the real-valued Parafac model.

Usually, the three-way data $\underline{\mathbf{X}}$ are centered and normalized before a Parafac algorithm is applied. Although more options for centering and normalizing exist than in the matrix case (PCA for example), only the following choices leave the Parafac model structure intact. Centering should be done across only one mode at a time, e.g. $x_{ijk} - x_{i\bullet k}$, where $\bullet$ implies that the average is taken over all indices in this mode. Centering the data in this way yields a Parafac model with $\mathbf{B}$ having mean zero columns. Normalizing should be done within one mode at a time, e.g. x_{ijk}/σ_i, where σ_i denotes the standard deviation of all elements with first index equal to i. This yields a Parafac model where the rows of $\mathbf{A}$ are normalized by σ_i. For more details on centering and normalization we refer to Bro and Smilde (2003).

The most attractive feature of Parafac is its uniqueness property. Kruskal (1977) has shown that, for fixed residuals $\underline{\mathbf{E}}$, the vectors $\mathbf{a}_r$, $\mathbf{b}_r$ and $\mathbf{c}_r$ are unique up to rescaling/counterscaling and a reordering of the summands in (2.14) if

$$k_{\mathbf{A}} + k_{\mathbf{B}} + k_{\mathbf{C}} \geq 2\,R + 2\,, \tag{2.16}$$

where $k_{\mathbf{A}}$, $k_{\mathbf{B}}$, $k_{\mathbf{C}}$ denote the k-ranks of the component matrices. The k-rank of a matrix is the largest number x such that every subset of x columns of the matrix is linearly independent. Hence, contrary to the matrix PCA model, the Parafac components are rotationally unique if (2.16) holds. Note that by fixing the columns of two of the matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ to length 1, a Parafac solution is unique up to sign changes and a reordering of the components if (2.16) holds.

The rank of a three-way array $\underline{\mathbf{Y}}$ is usually defined as the minimum number of rank-1 arrays whose sum equals $\underline{\mathbf{Y}}$. A three-way array has rank-1 if it is the outer product of three vectors. Hence, the components $\mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r$ all have rank 1. In this sense, the Parafac model (2.14) tries to find a best rank-R approximation to $\underline{\mathbf{X}}$. Moreover, the smallest R for which $\underline{\mathbf{X}}$ has a full Parafac

decomposition equals the rank of $\underline{\mathbf{X}}$. Unfortunately, a best rank-R approximation does not exist for all three-way arrays $\underline{\mathbf{X}}$, see De Silva and Lim (2006). For a variety of cases, Stegeman (2006abc) shows that if no best rank-R approximation exists, then the sequence of Parafac updates will exhibit diverging components.

The Parafac model and algorithms can be easily extended to multi-way arrays. For a multi-way extension of Kruskal's uniqueness condition (2.16) see Sidiropoulos and Bro (2000).

Data compression

When large datasets are fitted with Parafac, it is useful to compress the data before calculating the Parafac solution. Below, we show how the data can be compressed if the size of one mode is at least at large as the product of the other two sizes. Consider the matrix form (2.15) of the Parafac model. Let $\mathbf{X}^{(n\times mp)}$ be the matrix with the slices $\mathbf{X}_k$ stacked next to each other. Analogously, let $\mathbf{E}^{(n\times mp)}$ be the matrix with the slices $\mathbf{E}_k$ stacked next to each other. Then

$$\mathbf{X}^{(n\times mp)} = \mathbf{A}\,[\mathbf{C}_1\,\mathbf{B}^T|\ldots|\mathbf{C}_p\,\mathbf{B}^T] + \mathbf{E}^{(n\times mp)} = \mathbf{A}\,(\mathbf{C}\odot\mathbf{B})^T + \mathbf{E}^{(n\times mp)}\,, \tag{2.17}$$

where $\odot$ denotes the Khatri-Rao product, i.e. the column-wise Kronecker product. Assume that $n > mp$. Let the QR-decomposition of $\mathbf{X}^{(n\times mp)}$ be given by $\mathbf{Q}_x\mathbf{R}_x$, where $\mathbf{Q}_x$ is $n \times mp$ with orthonormal columns and $\mathbf{R}_x$ is $mp \times mp$ upper triangular. If $\mathbf{R}_x$ has an optimal Parafac solution $(\widetilde{\mathbf{A}}, \mathbf{B}, \mathbf{C})$ with R components, then $\mathbf{X}^{(n\times mp)} = \mathbf{Q}_x\mathbf{R}_x$ has an optimal Parafac solution $(\mathbf{Q}_x\widetilde{\mathbf{A}}, \mathbf{B}, \mathbf{C})$ with R components. This is because $\|\mathbf{R}_x - \widetilde{\mathbf{A}}\,(\mathbf{C}\odot\mathbf{B})^T\|^2$ equals $\|\mathbf{Q}_x\mathbf{R}_x - \mathbf{Q}_x\widetilde{\mathbf{A}}\,(\mathbf{C}\odot\mathbf{B})^T\|^2$ due to $\mathbf{Q}_x$ being column-wise orthonormal. Hence, it suffices to calculate a Parafac solution for the $mp \times mp$ matrix $\mathbf{R}_x$ instead of the larger $n \times mp$ matrix $\mathbf{Q}_x\mathbf{R}_x$. Note that this only works if $n > mp$. A more general form of the procedure above is described in detail in Kiers and Harshman (1997).

Parafac for multi-subject fMRI data

Let the $n \times m \times p$ data array $\underline{\mathbf{X}}$ consist of m fMRI brain scans of p subjects, where the fMRI signal is mapped into a reference brain containing n voxels. Andersen and Rayens (2004) have compared the Parafac model on $\underline{\mathbf{X}}$ to PCA on $\mathbf{X}^{(n\times mp)}$. Parafac then yields R voxel activation maps as the columns of $\mathbf{A}$, the R associated time courses as the columns of $\mathbf{B}$, and the activation strengths for each subject in $\mathbf{C}$. A PCA on $\mathbf{X}^{(n\times mp)}$ yields R voxel activation patterns in $\mathbf{A}$ and all combinations of scan$*$subject loadings in $\mathbf{B}$. Andersen and Rayens (2004) found that Parafac is to be preferred because of its uniqueness and the preservation of multilinear linkages and interactions. Moreover, a PCA on matricized multi-way data yields an unnecessarily complex solution.

Note that for fMRI data, the number of voxels n is often larger than the product mp. Hence, the compression method described above can be used in this case.

Concerning the uniqueness of Parafac, we would like to remark the following. This convenient property should not be mistaken as evidence that the true underlying structure of the data has been found. Indeed, consider the Parafac solution $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ for an array $\underline{\mathbf{X}}$ filled with random

numbers drawn independently from a continuous distribution. Then Kruskal's condition (2.16) will hold for R small enough. This implies a unique Parafac solution, while the data $\underline{\mathbf{X}}$ has no trilinear structure at all.

2.4 The Tensor PICA model for multi-subject fMRI data

The Tensor PICA model of Beckmann and Smith (2005) is a three-way extension of the PICA model described above, for the case of multi-subject fMRI data. As above, the three-way data array $\underline{\mathbf{X}}$ $(n \times m \times p)$ contains m fMRI scans of p subjects, where the scans have been mapped to a reference brain with n voxels. It is assumed that $\underline{\mathbf{X}}$ obeys the structure of the Parafac model. The Tensor PICA model is the PICA model corresponding to the matricized Parafac model (2.17), i.e.

$$\mathbf{x} = (\mathbf{C} \odot \mathbf{B})\,\mathbf{a} + \mathbf{e}\,. \tag{2.18}$$

Here, $\mathbf{x}$ is a vector of mp random variables having n realisations in the rows of $\mathbf{X}^{(n \times mp)}$, the vector $\mathbf{a}$ of R statistically independent sources has n realisations in the rows of $\mathbf{A}$, and the mixing matrix $(\mathbf{C} \odot \mathbf{B})$ has a particular structure. All assumptions of the PICA model above also apply to the Tensor PICA model (2.18). Hence, $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{mp})$, which implies that the noise variance σ^2 is the same for each subject and each voxel and the scans of the different subjects are statistically independent. As in the PICA model above, prior to estimating $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$, the fMRI data should be temporally pre-whitened for each voxel and each subject separately. Also, the mean activation maps (averaged over scans for each subject separately) should be subtracted to obtain zero mean data.

Using trilinear models such as Parafac or Tensor PICA to analyse multi-subject fMRI data implies that it is assumed that the underlying signal sources are proportional for the different subjects. The physiological artifacts in multi-subject fMRI data are not likely to satisfy this assumption. Therefore, these artifacts are better removed prior to a trilinear analysis of multi-subject fMRI.

The estimation of the Tensor PICA model is a two-stage procedure. In the first step, the structure of the mixing matrix in (2.18) is ignored and the PICA estimates of $\mathbf{A}$ and the compound mixing matrix, which we denote by $\mathbf{M}$, are obtained. In the second step, the matrices $\mathbf{B}$ and $\mathbf{C}$ are estimated from $\mathbf{M}$ as follows. Each column r of $\mathbf{M}$ is mapped to an $m \times p$ matrix $\mathbf{M}^{(r)}$ which should have rank 1 according to the Tensor PICA model (2.18). From the SVD of $\mathbf{M}^{(r)}$ the best rank-1 approximation of $\mathbf{M}^{(r)}$ is obtained as well as the estimates of the r-th columns of $\mathbf{C}$ and $\mathbf{B}$ (these are the first left- and right singular vectors appropriately scaled). Moreover, the relative size of the first singular value can be used as a goodness-of-fit measure of the rank-1 approximation of $\mathbf{M}^{(r)}$. When the solution $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ has been obtained, the value of $(\mathbf{C} \odot \mathbf{B})$ can be used as initial value for the mixing matrix in another run of the Tensor PICA estimation procedure. This yields an iterative procedure which ends when two consecutive estimates of $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ are sufficiently alike. As we will see below, it is our experience that we only need two of these iterations in most cases, i.e. the algorithm yields the same estimates in different runs.

When only two Tensor PICA iterations are needed, the Tensor PICA model is equivalent to the two-way PICA model applied to $\mathbf{X}^{(n\times mp)}$, followed by an estimation of the temporal and subject modes from the compound PICA mixing matrix. As mentioned above, Andersen and Rayens (2004) found a PCA on $\mathbf{X}^{(n\times mp)}$ unsatisfying due to its non-uniqueness and complicated solution structure. In the Tensor PICA model, the latter problems is circumvented by estimating the temporal and subject modes from the compound mixing matrix in the second step of the Tensor PICA estimation procedure. Moreover, the non-uniqueness of PCA is replaced with the uniqueness of ICA (under non-Gaussian sources and Gaussian noise).

2.5 Parafac versus Tensor PICA is different from PCA versus ICA

As mentioned above, an ICA model avoids the rotational ambiguity of the PCA components by finding an optimal rotation such that the ICA components are approximately statistically independent, with respect to some measure of statistical independence. The second-order constraint of uncorrelated sources in PCA is strengthened to statistically independent sources in ICA, by making use of higher-order statistics. Below, we argue that this second-order versus higher-order distinction cannot be applied to a comparison of Parafac versus Tensor PICA.

A common property of PCA and Parafac is that they both have a sum-of-squares objective function. A difference is that PCA assumes uncorrelated components and unconstrained loadings (apart from scaling), while Parafac treats all three modes equally and does not assume uncorrelated (i.e. orthogonal) columns in any mode. A second difference is that a PCA solution has rotational ambiguity, while there is (usually) not any rotational ambiguity in a Parafac solution.

Concerning the "second-order" nature of Parafac the following can be said. First of all, the Parafac model is simply trying to fit a trilinear structure to the data as well as possible, measured in a sum-of-squares sense. However, the Parafac model can also be interpreted as a Gaussian Maximum Likelihood model, where the residuals are uncorrelated standard Gaussian and $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ are the parameters to be determined by maximizing the likelihood function of the data; see Vega-Montoto and Wentzell (2003). Under the assumptions above, the Maximum Likelihood estimators for $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ are the same as the optimal Parafac solution. However, although Parafac is equivalent to Gaussian Maximum Likelihood (which only considers second-order statistics of the Gaussian data), the component matrices $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ are mere parameters of this Gaussian distribution and they are not considered as realisations of random variables themselves. Hence, also in this Maximum Likelihood model, the second-order statistics of the residuals are considered and not those of the components matrices $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ themselves. The second-order versus higher-order distinction, which applies to a comparison of PCA and ICA, is applicable to the components, and, hence, does not directly apply to a comparison of Parafac versus Tensor PICA.

As we will see in the following sections, what is important in the comparison of Parafac versus Tensor PICA, as applied to data satisfying the Tensor PICA assumptions, is the signal-to-noise ratio (SNR). Since Parafac minimizes the sum-of-squares of the residual term, it may not find all signal components if the SNR is low. Indeed, the sum-of-squares objective function may be lower if

a noise component instead of a signal component is included in the Parafc trilinear model part. In this case, a different criterion is needed to distinguish between signal and noise. And if the noise is Gaussian and the signal sources are non-Gaussian, the Tensor PICA model delivers exactly such a criterion. However, it should be emphasized that, if the SNR is sufficiently high, then Parafac will find the signal components regardless of their distribution. It may even be Gaussian, which is not allowed in the Tensor PICA model.

3 Parafac and Tensor PICA for artificial multi-subject fMRI data

In this section, we compare Parafac and Tensor PICA in their ability to extract the activation patterns in the artificial multi-subject fMRI data of Beckmann and Smith (2005). In Section 3.1, we describe the data in detail, discuss the preprocessing procedure and define several useful measures of signal-to-noise ratios. In Section 3.2, we compare Parafac and Tensor PICA on the original dataset. In Section 3.3, we consider several variations of the dataset in which the strength of the signal and the sparsity of the signal in the spatial domain have been changed. Again, Parafac and Tensor PICA are compared for these datasets.

3.1 Data description, preprocessing and signal-to-noise ratios

The artificial multi-subject data of Beckmann and Smith (2005) was kindly provided by Christian Beckmann. The signal part of this data consists of artificial voxel activation maps, artificial time patterns and activation strengths for three subjects. Random Gaussian noise is added to this signal part, where for each voxel the noise mean and variance are estimated from real single-subject resting state fMRI measurements (for details, see Beckmann and Smith, 2005). For different voxels, the noise is uncorrelated. The voxel-wise noise mean and variance are the same for each of the three subjects.

Detailed data description

Beckmann and Smith (2005) consider five different artificial multi-subject fMRI datasets, named (A)-(E), which differ only in their signal part. First, we consider dataset (A) in detail. Let $\underline{\mathbf{X}}$ be the three-way array of this dataset, which has size 2489 voxels $\times$ 196 time points $\times$ 3 subjects. The voxels form three brain slabs and each slab contains a different spatial activation pattern (crosses, vertical stripes and horizontal stripes, respectively). The total number of voxels is actually 12288 (64 times 64 for each of the three brain slabs), but as Beckmann and Smith (2005) we only consider the 2489 intra-cranial voxels. The number of active voxels in each slab is 45 out of 962 for slab 1, 90 out of 838 for slab 2 and 54 out of 689 for slab 3. Hence, only approximately 8 percent of all voxels is active. Each of the three voxel activation patterns has a different associated time course. The time courses correspond to a simple block design, a single-event design with fixed interstimulus interval, and a single-event with random interstimulus interval. The three time courses are convolved with a canonical hemodynamic response function (Gamma variate with 3 seconds standard deviation and

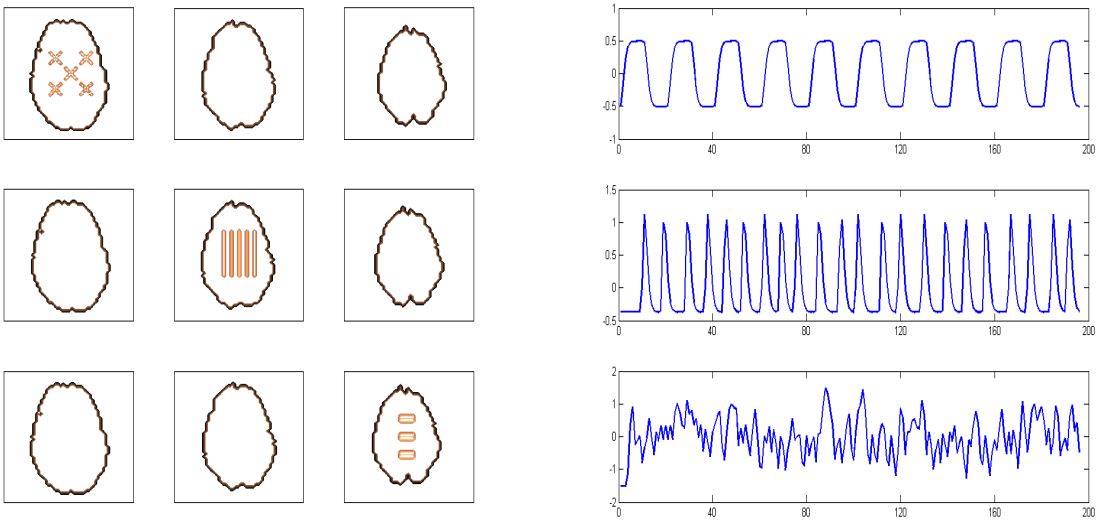

Figure 1: Artificial voxel activation maps and time courses in dataset (A) of Beckmann and Smith (2005). Each row shows the activation pattern in the three brain slabs and the associated time course.

6 seconds lag). The time courses are mean zero. In Figure 1 the voxel activation maps and their associated time courses are depicted.
Let the 2489×3 matrix $\mathbf{A}_0$ contain the voxel activation maps and let the 196×3 matrix $\mathbf{B}_0$ contain the associated time courses. The subject activation strengths are the elements of the 3×3 matrix $\mathbf{C}_0$, with

$$\mathbf{C}_0 = \begin{bmatrix} 3 & 4 & 5 \\ 2 & 3 & 4 \\ 2 & 2 & 3 \end{bmatrix} . \tag{3.1}$$

Hence, for subject 1 the three activation patterns have strengths 3, 4 and 5, etcetera. The data $\underline{\underline{\mathbf{X}}}$ satisfies the Parafac model with $R = 3$, component matrices $(\mathbf{A}_0, \mathbf{B}_0, \mathbf{C}_0\mathbf{\Lambda})$ and Gaussian residuals. Here, $\mathbf{\Lambda}$ is a 3×3 diagonal matrix with weight coefficients such that the SNR equals 2, which is defined as the signal Z-statistics being twice as large as the noise standard deviation, on average over the active voxels. This has been confirmed to the author by C.F. Beckmann in a personal communication.

Preprocessing

The definition of SNR above applies to the appropriately centered and normalized data. Centering of the data is done by subtracting the average activation map for each subject separately, i.e. $x_{ijk} - x_{i\bullet k}$. This does not change the signal part of the data (since the columns of $\mathbf{B}_0$ are already mean zero), but the noise means will be approximately zero. The centered data satisfies almost all

assumptions of the Tensor PICA model. We only need to normalize the data by estimates of the voxel-wise noise standard deviations, such that the noise variances become approximately identical for all voxels and subjects. Beckmann and Smith (2005) obtain noise variance estimates $\hat{\sigma}^2_{ik}$ (i.e. for voxel i and subject k) from the residuals of a PCA decomposition of the matricized time $\times$ (voxels $*$ subjects) data. The number of components in the PCA is equal to the model order R. Note that a normalisation x_{ijk}/σ_{ik} does not leave the trilinear model structure intact, while a normalisation x_{ijk}/σ_i does.

Beckmann and Smith (2005) estimate the model order R from an approximation to the model order of the PPCA model placed in a Bayesian framework, see also Beckmann and Smith (2004). Using this method, they obtain a model order estimate of $R = 13$ for dataset (A). The reason why this estimate is much higher than the correct order of 3, is because there is a lot of noise with respect to signal in the data. Consequently, large noise "components" (e.g. for voxels with a high noise variance) are treated as signal. In Figure 2 the true voxel-wise noise standard deviations are plotted for the three slabs. As can be seen, there are large noise standard deviations in slabs 1 and 2, at the back of the head. When the voxel-wise noise variances are estimated from the residuals of a PCA with 13 components, the large noise peaks caused by large noise standard deviations are included as components of the PCA. Hence, the large noise variances are systematically underestimated using this approach. This can be seen in Figure 3, where the esimates $\hat{\sigma}_{ik}$ are plotted against the true values σ_i for each subject k. For comparison, we also estimate the voxel-wise noise variances from the residuals of a PCA with 3 components. As can be seen from Figure 4, even if the correct model order is used, still one noise peak is considered as signal. Note that the large voxel-wise noise standard deviations do not occur for voxels included in the three spatial patterns, see Figure 1.

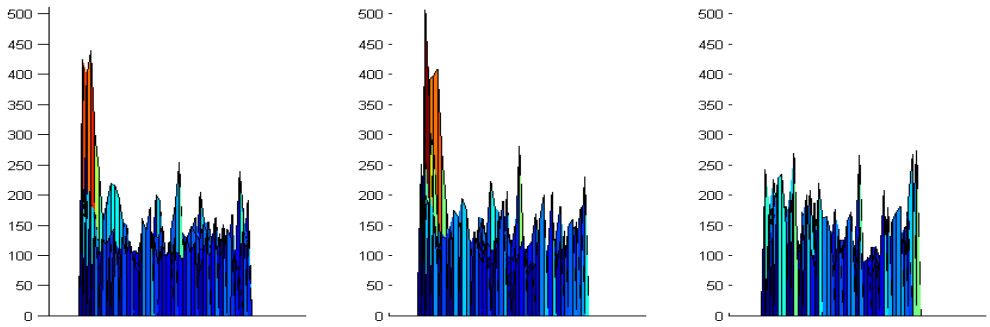

Figure 2: Side view of the true voxel-wise noise standard deviations for the three slabs of the brain. As can be seen, large standard deviations are present in slabs 1 and 2, at the back of the head.

Signal-to-noise ratios

Next, we quantify the signal-to-noise chasracteristics of the dataset. As mentioned above, the SNR of the artificial fMRI datasets in Beckmann and Smith (2005) equals 2, which is defined as

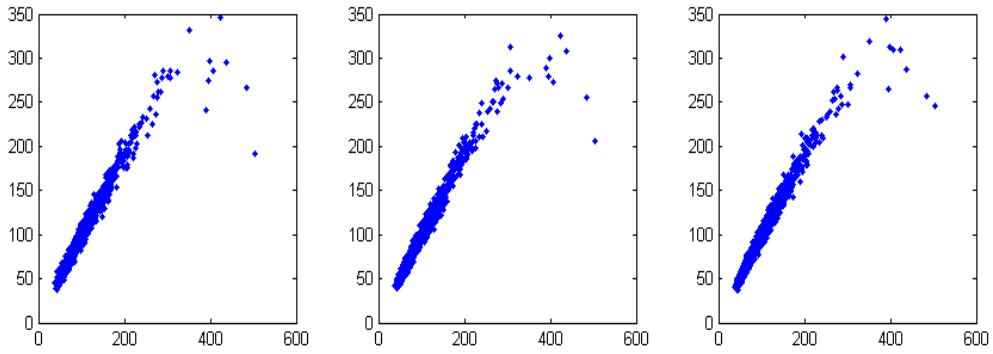

Figure 3: Voxel-wise standard deviation estimates $\hat{\sigma}_{ik}$ (y-axis) plotted against the true values σ_i (x-axis) for each of the three subjects k. The $\hat{\sigma}_{ik}$ are estimated from the residuals of a PCA with 13 components. As can be seen, the large values of σ_i are underestimated.

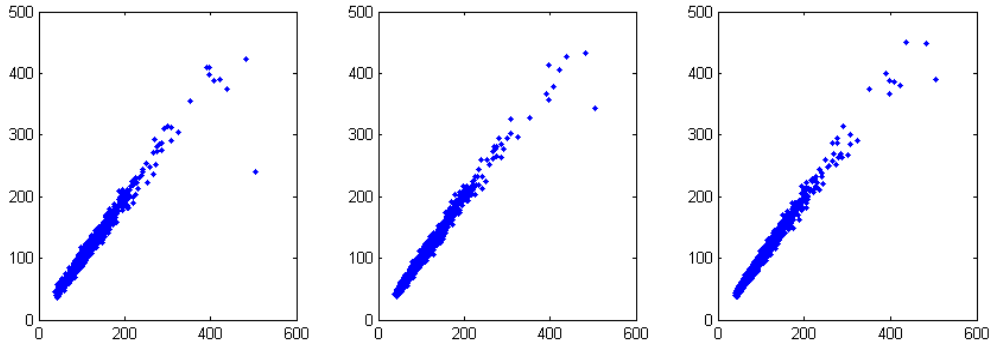

Figure 4: Voxel-wise standard deviation estimates $\hat{\sigma}_{ik}$ (y-axis) plotted against the true values σ_i (x-axis) for each of the three subjects k. The $\hat{\sigma}_{ik}$ are estimated from the residuals of a PCA with 3 components. As can be seen, one large value of σ_i is underestimated.

the signal Z-statistics being twice the noise standard deviation, on average over the active voxels. An SNR of 2 suggests that signal is twice as strong as the noise. However, this only holds (on average) for the active voxels and only approximately 8 percent of all voxels are active in the dataset. We will use a different measure of SNR, which is more suitable for our purposes. We define the SNR as the Frobenius norm of the signal divided by the Frobenius norm of the noise, i.e. $\text{SNR}_{\text{total}} = \|\underline{\mathbf{X}} - \underline{\mathbf{E}}\|/\|\underline{\mathbf{E}}\|$. Analogously, we define $\text{SNR}_{\text{active}}$ where we only take into account the active voxels. And $\text{SNR}_{\text{act:}s}$ is defined for the active voxels of pattern s. These SNR measures are comparable to the Parafac objective function and can be used to explain why some signals are found as Parafac components and others are not. In Table 1 these SNR values are given for the centered dataset without normalization, and for the centered and normalized dataset with two different normalizations, namely normalizing by the voxel-wise standard deviation estimates

	Centered data	Centered & PCA norm.	Centered & true norm.
$\text{SNR}_{\text{total}}$	0.05	0.07	0.08
$\text{SNR}_{\text{active}}$	0.23	0.27	0.28
$\text{SNR}_{\text{act:1}}$	0.17	0.19	0.19
$\text{SNR}_{\text{act:2}}$	0.20	0.26	0.26
$\text{SNR}_{\text{act:3}}$	0.29	0.34	0.35

Table 1: Various signal-to-noise ratios for dataset (A) of Beckmann and Smith (2005). The columns correspond to the centered data, the data centered and normalized by estimated voxel-wise standard deviations from PCA residuals (using 13 components), and the data centered and normalized using true voxel-wise standard deviations.

from the residuals of a PCA with 13 components and normalizing by the true voxel-wise standard deviations. In all three cases, the same noise instance is used as we will use throughout the paper. As can be seen, the amount of noise with respect to signal is very high and the SNR values do not differ very much for the two different normalizations. Also, the signal-to-noise ratios for artificial signal patterns 1, 2 and 3 satisfy $\text{SNR}_{\text{act:1}} < \text{SNR}_{\text{act:2}} < \text{SNR}_{\text{act:3}}$.

3.2 Comparison of Parafac and Tensor PICA on dataset (A)

Here, we compare the performance of Parafac and Tensor PICA in recovering the artificial activation patterns in the dataset (A) described in Section 3.1. We use the following measures of recovery of the artificial signals. For the spatial activation maps, we consider the correlation between the estimated maps and the maps $\widetilde{\mathbf{A}}_0$ obtained from an OLS regression

$$\mathbf{X}^{(mp\times n)} = (\mathbf{C}_0\mathbf{\Lambda} \odot \mathbf{B}_0)\, \mathbf{A}^T + \mathbf{E}^{(mp\times n)}, \tag{3.2}$$

where $n = 2489$ voxels, $m = 196$ time points, $p = 3$ subjects and $\mathbf{X}^{(mp\times n)}$ is the centered and normalized matricized data. In a personal communication to the author, it has been confirmed by C.F. Beckmann that this correlation measure of recovery in the spatial domain is also used in Beckmann and Smith (2005). For the estimated time patterns we consider the correlations with $\mathbf{B}_0$. The estimated subject activation strengths will be compared to $\mathbf{C}_0$.

For later use, we give the correlations of the maps $\widetilde{\mathbf{A}}_0$ and the time courses $\mathbf{B}_0$ with themselves.

$$\text{Corr}(\widetilde{\mathbf{A}}_0) = \begin{bmatrix} 1 & 0.01 & -0.12 \\ 0.01 & 1 & -0.04 \\ -0.12 & -0.04 & 1 \end{bmatrix} \qquad \text{Corr}(\mathbf{B}_0) = \begin{bmatrix} 1 & -0.02 & 0.16 \\ -0.02 & 1 & 0.05 \\ 0.16 & 0.05 & 1 \end{bmatrix}. \tag{3.3}$$

Parafac - results and discussion

We consider the dataset (A) described in Section 3.1 with the same noise instance as used throughout the paper. For $R = 3, 4, 6, 10, 13, 20$ we estimate the voxel-wise noise standard deviations from the residuals of a PCA with R components on the matricized and centered data, then normalise the data by these estimates, and next use Parafac with R components on the normalized data to estimate the artificial fMRI signals. We use the standard Alternating Least Squares algorithm for Parafac. The size of the data array $\underline{\mathbf{X}}$ is $2489 \times 196 \times 3$ and, hence, we may use the compression method described in Section 2.3. This means that the size of the data array in the Parafac algorithm is $588 \times 196 \times 3$.

For each R we ran the Parafac algorithm 10 times with a random starting position, and out of 10 runs we picked the solution with the highest Parafac objective function. In Table 7 on p. 29 the results are presented. For each R, Table 7 contains the Parafac components which had the highest correlation to the three artificial fMRI activation maps. For each such Parafac component, the correlations of the voxel mode with the spatial maps $\widetilde{\mathbf{A}}_0$, the correlations of the temporal mode with the time courses $\mathbf{B}_0$, and the estimates of the subject activation strengths $\mathbf{C}_0$ in (3.1) are given.

As can be seen, the second and third artificial fMRI signals are nearly always found, while the first signal is found with less precision. This is consistent with the SNR values in Table 7, which satisfy $\mathrm{SNR}_{\mathrm{act:1}} < \mathrm{SNR}_{\mathrm{act:2}} < \mathrm{SNR}_{\mathrm{act:3}}$. For $R = 3$ the first fMRI signal is not found at all. Apparently, one noise peak is larger than the first fMRI signal. This is probably due to one large noise standard deviation being underestimated, see Figure 4 and our discussion above.

There are many Parafac solutions with nearly the same objective value and run-to-run variability of the Parafac solutions is considerable. While the second and third artificial fMRI signals are nearly always well recovered, the correlation with the first fMRI signal varies. We suspect that this is due to the low SNR values and the fact that the voxel-wise noise is spatially and temporally independent and approximately isotropic, which implies that it "points in all directions". For low values of R, a higher fit value tends to indicate higher correlations of the Parafac solution with the artificial fMRI signals. However, for higher values of R this is not true. For example, for $R = 13$ the best solution of 10 runs (with a fit percentage of 3.95) has a correlation of 0.52 with the first artificial fMRI signal, while a solution with a slightly worse fit (3.94 percent) has a correlation to the same signal of 0.71. Beckmann and Smith (2005) use Parafac with $R = 13$ and report the latter correlation value with the first fMRI signal.

The standard method to get an idea of the model order R is by considering the graph of additionally explained variance for increasing values of R. The fit percentage values of the Parafac solutions tend to increase by the same amount of 0.25-0.30 for $R = 2$ up to $R = 40$. One could infer a model order of $R = 2$, since each increase of R beyond 2 seems to add a noise component of equal strength. Due to the low SNR values of the dataset, the first signal component is not noticed in this way. However, a plot of this Parafac component clearly revels its structure if its correlation with the first fMRI signal is above 0.6. In Table 4 the fit percentages are given for $R = 1, \ldots, 6$.

The Parafac fit percentage is very low, e.g. 3.95 percent for $R = 13$. One could argue that this is because only 8 percent of the voxels contain signal which implies that the Parafac model is valid for only 8 percent of the voxels. However, the Parafac fit is low due to the low SNR values and not due to sparsity of the signal. This is illustrated in Section 3.3, where we increase the signal strength.

In the Parafac solutions for $R \geq 4$, nearly always there are only three components containing signal. However, from Table 7 it can be seen that the Parafac component with the highest correlation with one of the artificial fMRI signals, sometimes also has considerable correlation with another artificial fMRI signal. In Beckmann and Smith (2005) this phenomenon is called "cross-talk". As can be seen from (3.3), these cross-talk correlations cannot be justified by the correlations in the maps $\widetilde{\mathbf{A}}_0$ or the time courses $\mathbf{B}_0$. If R is significantly larger than the number of signal components present in the data, cross-talk is a common phenomenon in Parafac solutions. Then it becomes profitable to mix the signals such that as much of the noise variation as possible is included in the Parafac solution. In the fMRI dataset at hand, the SNR values are low which cause some cross-talk to occur also for smaller R. The largest noise peaks are also included in the Parafac solution at the cost of distinguishing between the signal components.

Imposing orthogonality restrictions on the Parafac voxel mode matrix $\mathbf{A}$ does not reduce the cross-talk. Although the spatial maps $\mathbf{A}_0$ are orthogonal, their best OLS estimates $\widetilde{\mathbf{A}}_0$ are not, and neither are their unrestricted Parafac estimates. Since only 8 percent of all voxels are active and the signals $\mathbf{A}_0$ are orthogonal, the orthogonality restriction affects the estimation of the largest noise peaks more than the estimation of the signals. Also in this case, the low SNR causes the estimation of the largest noise peaks to occur at the cost of distinguishing between the signal components.

However, for $R = 3$ the cross-talk can be reduced by choosing a more restrictive convergence criterion for the Parafac algorithm. The solutions in Table 7 are obtained with a convergence criterion of 10^{-6}. If we change this to 10^{-9} the best solution of 10 Parafac runs with $R = 3$ is as in Table 2. The estimates of $\mathbf{C}_0$ are almost identical to the solution in Table 7 and have been omitted. As can be seen, the cross-talk of the above solution is less than the cross-talk of the solution with $R = 3$ in Table 7. In Section 3.3 we will see that for higher SNR values there is less cross-talk for $R = 3$.

Tensor PICA - results and discussion

The Tensor PICA model is fitted to the same centered and normalized dataset (A) as is Parafac, for $R = 3, 4, 6, 10, 13, 20$. As ICA algorithms, we try FastICA, Comon-4 and JADE. In a personal communication to the author, it has been confirmed by C.F. Beckmann that a version of the FastICA algorithm is used in Beckmann and Smith (2005), with $G(x) = x^3$.

In Table 8 on p. 30 the Tensor PICA solutions are presented for the Comon-4 algorithm. For $R = 3, 4, 6, 10, 13$ the solutions for FastICA were nearly identical, and for all $R = 3, 4, 6, 10$ the solutions for JADE are also nearly identical. For $R = 13, 20$ the solutions for JADE are comparable to Table 8. For $R = 20$, the FastICA algorithm suffered from convergence problems. The FastICA

	1	2	3
Map 1	*	*	*
Map 2	*	0.96	*
Map 3	*	-0.18	0.96
Time 1	*	*	0.18
Time 2	0.12	0.94	0.19
Time 3	*	*	0.94

Table 2: Best Parafac solution of 10 runs for dataset (A) with $R = 3$ and convergence criterion 10^{-9}.

algorithm can compute the independent components all together (symmetric approach) or one after the other (deflation approach). The symmetric approach had convergence problems for R larger than 5, and also the deflation approach had difficulty to converge. This problem was fixed by running FastICA in the "stabilization" mode. Recall that the Tensor PICA procedure uses the obtained estimate $(\mathbf{C} \odot \mathbf{B})$ of the compound mixing matrix as initial value in the next run, and stops when two consecutive estimates of $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ are sufficiently alike. With FastICA, the Tensor PICA procedure needed only two such iterations for $R = 3, 4, 6, 10$ and up to 6 for $R = 13$. For $R = 20$ the procedure ran prohibitively long. This is probably due to the slow convergence of FastICA, which causes consecutive estimates to be insufficiently alike. However, setting a higher maximum number of iterations for FastICA did not seem to help. Contrary to FastICA, Comon-4 and JADE had no convergence problems and always needed only two iterations of the Tensor PICA procedure.

As can be seen from Table 8, the Tensor PICA solutions do not differ much for different values of R. Also, there is almost no cross-talk between the components, i.e. the cross correlations present are due to (3.3). These advantages are also observed by Beckmann and Smith (2005). Like Parafac, the Tensor PICA method recovers the second and third fMRI signals well. However, the first fMRI signal is recovered much less clearly than with Parafac. The run-to-run variability of Tensor PICA is much less than for Parafac, i.e. if the Tensor PICA procedure is applied several times (with a random starting point for the first iteration) then the obtained solutions do not differ much.

The robustness of Tensor PICA is due to the robustness of the ICA algorithms used. These are designed to separate the non-Gaussian signals from the Gaussian noise. The absence of cross-talk can be explained as follows. By the Central Limit Theorem, a mixture of two signals is always "more Gaussian" than one signal alone. Since the ICA algorithms "maximize the non-Gaussianity" of each extracted independent component, one independent component contains a single signal source and not a mixture of two or more signal sources. For R larger than the number of signal sources present in the data, the number of independent components containing signal sources is at

most R. This can be explained as follows. The signal sources are given by $\mathbf{a} = \mathbf{Q}^T\tilde{\mathbf{x}} - \mathbf{Q}^T\tilde{\mathbf{e}}$, where $\mathbf{Q}$ is orthonormal and each column of $\mathbf{Q}$ corresponds to one signal source. Each column of $\mathbf{Q}$ is forced to be orthogonal to the columns of $\mathbf{Q}$ which have already been determined. Hence, once a signal source has been found it will not turn up in other independent components.

Contrary to Parafac, the Tensor PICA method does not treat the three modes of the data equally. After the ICA algorithm yields a solution for the spatial fMRI activation maps and a compound mixing matrix, the associated time courses are then found by the column-wise rank-1 approximation of the compound mixing matrix, as described above. If the spatial maps are well recovered, the time courses and subject strengths are well recovered too.

For their Tensor PICA solution, Beckmann and Smith (2005) report a correlation of approximately 0.90 with the first fMRI signal. This differs significantly from our results. However, in a personal communication to the author, it was explained by C.F. Beckmann that the signal-to-noise characteristics of the dataset may differ from the dataset (A) used in Beckmann and Smith (2005). In Section 3.3 we will see that if the signal strength is increased, then both Parafac and Tensor PICA are able to recover all three artificial signals from the data.

3.3 Increasing signal strength and reducing sparsity

Here, we consider several variations of the signal part of the dataset (A) described in Section 3.1, each with the same noise instance. The time courses $\mathbf{B}_0$ and activation strengths $\mathbf{C}_0$ are left unchanged, while the spatial activation patterns $\mathbf{A}_0$ are changed both in strength and form (i.e. the number of active voxels). In this way, the influence of the SNR values and sparsity in the spatial domain on the abilities of Parafac and Tensor PICA to recover the signals, can be made clear. Although, Beckmann and Smith (2005) use a model order of $R = 13$ we will use $R = 3$ for both Parafac and Tensor PICA. The Tensor PICA solutions do not differ much for different $R \geq 3$. Differences in Parafac solutions for different $R \geq 3$ will be addressed briefly.

We create four altered datasets with stronger signals by multiplying the signal part by 2, 3, 5 and 100, respectively. This increases the SNR values as can be seen from Table 3. The order $\text{SNR}_{\text{act:1}} < \text{SNR}_{\text{act:2}} < \text{SNR}_{\text{act:3}}$ still holds in all cases. We also create two datasets in which the number of active voxels in increased. In the first dataset the numbers of active voxels are 125 for map 1 (the crosses become blocks), 218 for map 2 (the stripes become one block) and 102 for map 3 (the stripes become one block). This amounts to 18 percent of the voxels being active, where in the original dataset only 8 percent of the voxels are active. In the second dataset the numbers of active voxels are 289 for map 1, 456 for map 2 and 432 for map 3. In all three maps the active voxels form one block. In this dataset 47 percent of all voxels are active. The signal strength in both datasets is the same as in the original dataset. In Table 3 the SNR values for these datasets are given. The only significant change with respect to the original dataset is the higher value of $\text{SNR}_{\text{total}}$, because the other SNR values take into account only (subsets of) active voxels.

For later use, we give the correlations of the spatial maps $\widetilde{\mathbf{A}}_0^{(18)}$ and $\widetilde{\mathbf{A}}_0^{(47)}$ (for the datasets with 18 and 47 percent active voxels, respectively) with themselves.

	original	2∗signals	3∗signals	5∗signals	100∗signals	18% active	47% active
$\text{SNR}_{\text{total}}$	0.07	0.14	0.22	0.37	3.71	0.11	0.20
$\text{SNR}_{\text{active}}$	0.27	0.55	0.84	1.38	20.8	0.27	0.28
$\text{SNR}_{\text{act:1}}$	0.19	0.36	0.54	0.90	15.8	0.19	0.21
$\text{SNR}_{\text{act:2}}$	0.26	0.51	0.77	1.28	20.0	0.26	0.27
$\text{SNR}_{\text{act:3}}$	0.34	0.70	1.07	1.78	27.9	0.35	0.33

Table 3: Various signal-to-noise ratios for the original and altered dataset (A). The data are centered and normalized using voxel-wise noise standard deviations estimated from the residuals of a PCA with 3 components on the centered matricized data.

$$\text{Corr}(\widetilde{\mathbf{A}}_0^{(18)}) = \begin{bmatrix} 1 & -0.03 & -0.09 \\ -0.03 & 1 & -0.06 \\ -0.09 & -0.06 & 1 \end{bmatrix} \qquad \text{Corr}(\widetilde{\mathbf{A}}_0^{(47)}) = \begin{bmatrix} 1 & -0.11 & -0.15 \\ -0.11 & 1 & -0.19 \\ -0.15 & -0.19 & 1 \end{bmatrix}. \quad (3.4)$$

Parafac - results and discussion

In Table 9 the Parafac solutions for the six altered datasets are given for $R = 3$. The best solution (i.e. with the highest fit percentage) of 10 Parafac runs is picked. The data are compressed before running the Parafac algorithm, as described in Section 2.3. From Table 9 on p. 31 it can be seen that, contrary to the original dataset, all three artificial fMRI signals are recovered by Parafac. Also, the amount of cross-talk is rather low and cross correlations tend to follow the patterns in (3.3) and (3.4). For values of R larger than 3 more severe cross-talk occurs, but this is due to overfitting as we discussed above. The run-to-run variability of Parafac solutions is still considerable, but for higher SNR values the problem is less severe. Note that for the 100∗signals dataset the estimates of the subject strengths $\mathbf{C}_0$ are less accurate because these are small compared to the order of magnitude of the signals.

As mentioned above, the standard method to get an idea of the model order R is by considering the graph of additionally explained variance for increasing values of R. In Table 4 the fit percentages are given for the six altered datasets and $R = 1, \ldots, 6$. As can be expected, the fit percentages increase as the signal becomes stronger and as more voxels are active. For each dataset, for $R = 4, 5, 6$ the fit percentages increase by the same small amount while the fit increases stronger for $R = 1, 2, 3$. This is more clearly observed as the SNR values become larger. Hence, even though the fit percentages are low they show a pattern which is consistent with a model order of $R = 3$. For $R = 1, 2, 3$ the increase in fit corresponds to an additional signal component, while for $R \geq 4$ the increase in fit corresponds to an additional noise component which results in a near constant

	original	2∗signals	3∗signals	5∗signals	100∗signals	18% active	47% active
1	0.42	0.97	2.52	6.05	59.66	0.82	2.15
2	0.82	1.89	4.64	11.21	85.37	1.45	3.52
3	1.10	2.29	5.21	12.42	93.08	1.83	4.13
4	1.38	2.57	5.47	12.66	93.24	2.11	4.40
5	1.67	2.84	5.74	12.91	93.27	2.38	4.67
6	1.93	3.12	6.01	13.15	93.30	2.66	4.94

Table 4: Parafac fit percentages for the original and altered dataset (A), with $R = 1, \ldots, 6$. The data are centered and normalized using voxel-wise noise standard deviations estimated from the residuals of a PCA with 3 components on the centered matricized data.

increase of fit (over a wide range of $R \geq 4$) due to the noise being approximately isotropic. Hence, it may be concluded that the correct number of components can be inferred from the Parafac fit values. This is an advantage over the Tensor PICA model, for which Beckmann and Smith (2005) obtained a model order of $R = 13$.

Tensor PICA - results and discussion

The Tensor PICA model is fitted to the same centered and normalized altered dataset (A) as is Parafac. To avoid the convergence problems of the FastICA algorithm, we use Comon-4 as ICA algorithm. As for the original dataset, the Tensor PICA procedure needed only two iterations in all cases. In Table 10 on p. 32 the results are presented for $R = 3$. The solutions for JADE are comparable. It can be seen that all three signals are well recovered by the Tensor PICA method and almost no cross-talk occurs. Occurring cross correlations are due to (3.3) and (3.4). As before, also for $R \geq 4$ this is the case. Moreover, the run-to-run variability of the Tensor PICA solutions is rather small. Hence, the only difference with the original dataset is that now map 1 is well recovered. Note that for the 100∗signals dataset the estimates of the subject strengths $\mathbf{C}_0$ are less accurate because these are small compared to the order of magnitude of the signals.

3.4 Other datasets

So far, we have only considered dataset (A) of Beckmann and Smith (2005). Regarding the datasets (B)-(E) in the same article, we can say the following. Dataset (B) is the same as (A) except that there are zeros in the matrix $\mathbf{C}_0$ of subject strengths. Our results for dataset (A) also hold for dataset (B). In dataset (C), the time courses differ per subject due to differences in the parameters of the hemodynamic response functions of the subjects. In this case, the data has a Parafac structure

with nine components (three spatial maps and three groups of three slightly different time courses). Parafac with $R = 9$ is able to revover the signals reasonably well. Tensor PICA assumes statistically independent maps and, hence, finds three spatial maps for $R = 9$. The maps are found slightly more accurately than with Parafac. The three associated time courses are averages of the different time courses taken over the subjects. In this case, the differences in the time courses are judged as not systematic and Tensor PICA can be used. In datasets (D) and (E), time courses are coupled to more than one spatial map. This implies that the data has a Parafac structure in which some spatial maps and time courses occur more than once. When Parafac is run with the correct number of components, the spatial maps and time courses are not well recovered for low SNR values and there is a substantial amount of cross-talk. Tensor PICA recovers the signals well without much cross-talk. For datasets (D) and (E), the data actually satisfies a Tucker model (1966) which allows the coupling of time courses to multiple spatial maps and vice versa. Parafac is a special case of the Tucker model.

The changes in datasets (C)-(E) do not affect the spatial maps. This is to the advantage of Tensor PICA, since its crucial step is finding the correct spatial maps, while Parafac treats all three modes equally. Therefore, it is interesting to compare Tensor PICA and Parafac for the case where the spatial maps differ slightly per subject. For this purpose, we have created a datatset (F). The differences in the spatial maps are such that for each pair of subjects the spatial maps have more than 80 percent of the active voxels in common. The time courses and subject strengths are as in dataset (A). As for dataset (C), dataset (F) has a Parafac structure with nine components. Parafac with $R = 9$ recovers the spatial maps reasonably well, although there is a substantial amount of cross-talk. Tensor PICA with $R = 9$ finds three spatial maps which are averages of the maps taken over the subjects. The time courses are well recovered. Hence, Tensor PICA is also robust against small individual differences in the spatial activation patterns.

It was observed by McKeown, Makeig et al. (1998) and McKeown, Jung, Makeig, Borwn, Kindermann, Lee, and Sejnowski (1998) that overlapping spatial activity maps in fMRI data may violate the assumption in ICA of statistically independent maps. As a result, the obtained ICA spatial maps may not match the overlapping maps. To see whether this is a problem in Tensor PICA as well, we created two datasets with overlapping spatial maps. In dataset (G), the first spatial activity map is the sum of map 1 in dataset (A) and map 2 of the 18% active dataset. The second activity map is the sum of map 2 in dataset (A) and map 1 of the 18% active dataset. Hence, the activity of the first two maps takes place in the first two brain slabs. Maps 1 and 2 of dataset (G) have 51 and 63 percent of their active voxels in common, respectively. Map 3, the time courses $\mathbf{B}_0$, the subject strengths $\mathbf{C}_0$ and the noise instance are the same as in dataset (A). In total, 21 percent of all voxels are active in dataset (G). Dataset (H) is obtained from (G) by removing unique active voxels from maps 1 and 2. These maps now have 68 and 71 percent of their active voxels in common, respectively, and 18 percent of all voxels are active. The correlation between maps 1 and 2 of $\widetilde{\mathbf{A}}_0$ in (3.2) is 0.47 for dataset (G) and 0.58 for dataset (H). In Table 5 the results for Tensor PICA and Parafac are presented for $R = 3$ and the 2$*$signals datasets (G) and (H). The subject strengths were nearly identical for the two methods, and have been omitted.

	dataset (G)						dataset (H)					
	Tensor PICA			Parafac			Tensor PICA			Parafac		
	1	2	3	1	2	3	1	2	3	1	2	3
Map 1	0.82	0.72	*	0.99	0.19	-0.23	0.54	0.85	*	0.97	0.27	-0.17
Map 2	-0.12	0.95	*	0.58	0.95	*	-0.36	0.92	*	0.65	0.94	-0.12
Map 3	*	*	0.99	*	-0.16	0.98	*	*	0.99	0.14	*	0.99
Time 1	0.96	0.55	0.17	0.94	-0.19	0.15	0.79	0.64	0.16	0.93	-0.14	-0.13
Time 2	-0.28	0.82	*	0.29	0.98	0.22	-0.60	0.75	*	0.34	0.98	*
Time 3	0.13	0.13	0.99	0.27	*	0.97	*	0.13	0.99	0.23	*	0.94

Table 5: Tensor PICA and Parafac solutions for $R = 3$ on the 2$*$signals datasets (G) and (H). For Parafac, the best solution of 10 runs was taken. The Tensor PICA procedure needed only 2 iterations for both datasets.

As can be seen, Tensor PICA finds the spatial maps much less accurate and with more cross-talk than the Parafac maps. Moreover, also the time courses found by Tensor PICA show severe cross-talk and this is not the case for Parafac. It could be argued that in the presence of overlapping maps, Tensor PICA splits the signal into common and unique parts. A visual inspection of the Tensor PICA maps yields that the first map has mostly crosses (and less blocks) in slab 1 and vague stripes in slab 2. Hence, map 1 can be considered the common part of the first two true spatial maps. However, Tensor PICA map 2 contains both crosses and blocks in slab 1 and clear stripes and a (less clear) block in slab 2. This is a mixture of the two unique parts of the first two true maps and their common part. The Tensor PICA solution with $R = 4$ is the same as the solution for $R = 3$ with one noise map added. Hence, Tensor PICA does not clearly split up the signal into common and unique parts.

As suggested in McKeown, Jung et al. (1998) and McKeown, Makeig et al. (1998), this apparently occurs because the overlapping spatial maps violate the independence assumption. Hence, in this situation, Parafac is better able to capture the activation maps and their associated time courses.

In datasets (C)-(F), the Parafac algorithm becomes slower due to collinearity in the columns of the component matrices. Beckmann and Smith (2005) use the Parafac algorithm of Cao, Chen, Mo, Wu, and Yu (2000), in which the Parafac sum-of-squares objective function is adjusted in order to (hopefully) increase the speed of the algorithm in cases where the Parafac solution has several components which are highly correlated. However, it is our experience that the ordinary Parafac Alternating Least Squares algorithm with data compression is still faster than the modified algorithm by Cao et al. (2000). A more promising alternative in this case is to use the Enhanced

Line Search modification of the Parafac Alternating Least Squares algorithm (Rajih & Comon, 2005).

4 Computational times and faster Parafac

The data compression we have used prior to calculating the Parafac solution makes our Parafac algorithm a lot faster than the Parafac algorithm used in Beckmann and Smith (2005). The latter state (p. 306) that for dataset (A) the computational load of Tensor PICA is 15 times less than that of Parafac. In Table 6 our computation times are given for Parafac and Tensor PICA for the 2∗signals dataset and $R = 3$. As can be seen, 10 Parafac runs including data compression and decompression take less than 4 times the computational time for Tensor PICA.

Next, we present a method to reduce the computational time of Parafac even further. In the Tensor PICA framework, the spatial maps are estimated by (2.4), where the rotation $\mathbf{Q}$ is determined by the ICA algorithm. According to (2.4), the spatial maps $\mathbf{A}$ lie in the column span of $\mathbf{U}_R$, which consists of the first R columns of the matrix $\mathbf{U}$ in the SVD $\mathbf{X}^{(n\times mp)} = \mathbf{U}\,\mathbf{S}\,\mathbf{V}^T$. Here, $\mathbf{X}^{(n\times mp)}$ is the matricized voxels × (scans ∗ subjects) data, with $n = 2489$, $m = 196$ and $p = 3$. Consider the matricized form of the Parafac model (2.17). Carroll, Pruzansky, and Kruskal (1980) show that $(\widetilde{\mathbf{A}}, \mathbf{B}, \mathbf{C})$ is an optimal Parafac solution for $\mathbf{U}_R^T\mathbf{X}^{(n\times mp)}$ if and only if $(\mathbf{U}_R\widetilde{\mathbf{A}}, \mathbf{B}, \mathbf{C})$ is an optimal Parafac solution for $\mathbf{X}^{(n\times mp)}$ under the restriction $\mathbf{A} = \mathbf{U}_R\mathbf{D}$ for some $\mathbf{D}$. Hence, it suffices to calculate $\mathbf{U}_R$ and apply the Parafac algorithm to the smaller dataset $\mathbf{U}_R^T\mathbf{X}^{(n\times mp)}$ of size $R \times m \times p$. Moreover, since $m = 196 > pR = 3R$ we can compress this dataset using the QR-decomposition as before and run Parafac on a dataset of size $R \times 3R \times 3$ only. As can be seen from Table 6, the computational time of Parafac is reduced by a factor of almost 3 using this approach.

The incorporation of linear constraints of the type $\mathbf{A} = \mathbf{U}_R\mathbf{D}$ in the Parafac model is known as Candelinc and has been introduced by Carroll et al. (1980).

In Table 11 on p. 33 the results of Parafac Candelinc on the altered dataset (A) are presented for $R = 3$. As can be seen, all three signals are recovered well. However, for low SNR values, there is more cross-talk compared to the ordinary Parafac solutions. Apparently, due to the low SNR values the signals are not well recognized individually in the highly compressed datasets. As for ordinary Parafac, there is cross-talk for $R \geq 4$ due to overfitting. As before, for the 100∗signals dataset the estimates of the subject strengths $\mathbf{C}_0$ are less accurate because these are small compared to the order of magnitude of the signals. For the original dataset (A), Parafac Candelinc recovered only signals 2 and 3 for $R = 3$, as did ordinary Parafac.

5 Discussion

In this paper, we have compared the Parafac and Tensor PICA methods in their ability to recover the three fMRI signals in the artificial multi-subject fMRI data of Beckmann and Smith (2005). We have shown that, theoretically, the comparison of Parafac and Tensor PICA does not boil down

Method	Computational Time
Parafac	93 sec
Parafac - Candelinc	37 sec
Tensor PICA - FastICA	26 sec
Tensor PICA - Comon-4	25 sec
Tensor PICA - JADE	23 sec

Table 6: Computational times for calculating the solution for the centered and normalized 2∗signals dataset (A) with $R = 3$. For Parafac and Candelinc, the time includes data compression and decompression and 10 runs of the Parafac algorithm. For Tensor PICA, the time includes whitening of the data and 2 iterations of the procedure.

to the second-order statistics versus higher-order statistics distinction between PCA and two-way ICA. In PCA, the solution is determined up to an orthogonal rotation and the signal sources are required to be uncorrelated. The two-way ICA method fixes the rotational freedom of the PCA solution by imposing the constraint of statistically independent signal sources, using higher-order statistics of the data. Contrary to PCA, a Parafac solution usually has no rotational freedom and no restriction of uncorrelated components is imposed. The Tensor PICA method iterates the two-way PICA method applied to the matricized three-way data, and a second step to obtain the signals in the remaining two modes from the compound PICA mixing matrix. In our analyses, only two such iterations were needed. Instead of analyzing the matricized three-way data, Parafac finds a trilinear approximation to the data simultaneously for the three modes.

For the original dataset (A) and R up to 6, Parafac clearly outperforms Tensor PICA in recovering the three fMRI signals from the data. Both methods find signals 2 and 3 while Parafac recovers more of signal 1 than Tensor PICA. For R larger than 6, the Parafac solutions suffer from some cross-talk, which is a common phenomenon for Parafac if R is significantly larger than the number of signal components present in the data. The Tensor PICA solutions show almost no cross-talk and do not differ much for different values of R.

For the altered datasets (A) with higher SNR values, both Parafac and Tensor PICA recover the three signals well. For $R = 3$, the Parafac solutions show about as much cross-talk as the Tensor PICA solutions. Hence, if the SNR values are increased by either increasing the strength of the signal or increasing the amount of voxels carrying signal, both methods recover the signals equally well. However, there are two differences between Parafac and Tensor PICA. First, contrary to Parafac the Tensor PICA solutions do not show much cross-talk for R larger than 3. And second, the run-to-run variability of the Tensor PICA solutions is much less than for Parafac. For Tensor PICA, these robustness properties are due to the robustness of the ICA algorithm, as we explained above.

For Parafac, the run-to-run variability can be dealt with as follows. Due to the low SNR values there are many Parafac solutions with near optimal fit (some of which suffer from cross-talk also for $R = 3$). This makes it necessary to choose a small convergence criterion, run the Parafac algorithm at least 10 times and pick the solution with the best fit. It is our experience that this solution recovers the signals well and shows no severe cross-talk for $R = 3$. The occurrence of cross-talk in Parafac solutions for R larger than 3 is a more severe problem which cannot be fixed unless the correct number of signal sources is chosen. However, for the 2$*$signals dataset, the correct number of sources can be inferred from the Parafac fit for increasing values of R. For $R = 1, 2, 3$ the fit increases more than for $R = 4, 5, 6, \ldots$, where the fit increases with a small nearly constant amount. This is due to the noise being approximately isotropic and each increase from $R \geq 3$ to $R + 1$ will fit an additional noise component. Note that for real-life fMRI data the problem of choosing the number of Parafac component is unlikely to be solved by considering explained variances.

The results presented in this paper are obtained using a fixed noise instance. However, trying different noise instances has shown that our conclusions are still true.

Apart from assessing Parafac and Tensor PICA for multi-subject fMRI data, we have also shown that the computational load of the Parafac algorithm can be decreased by compressing the data and using a restriction on the Parafac spatial maps. This makes the computational load of Parafac only 1.5 times larger than that of Tensor PICA.

The artificial multi-subject fMRI datasets we analyzed feature temporally isotropic Gaussian noise. If a trilinear model such as Parafac or Tensor PICA is assumed, then the underlying signal sources are assumed to be proportional for the different subjects. Physiological artifacts (e.g. due to respiration, heart beat, subject movement) are not likely to satisfy this assumption of proportionality. Therefore, these artifacts are better removed prior to the analysis of multi-subject fMRI data using trilinear models such as Parafac and Tensor PICA. Moreover, since the noise is assumed to be temporally isotropic, the data has to be temporally pre-whitened for each voxel (or group of voxels) and subject.

The robustness of Tensor PICA makes this method more suitable for the analysis of real-life multi-subject fMRI datasets. In practice the data will not be as ideal as the datasets (A) analyzed in the current paper, and the determination of the "correct" number of signal sources from Parafac fit values will be more troublesome. However, a serious drawback of Tensor PICA is that it does not find the correct spatial maps if they are well-overlapping. This is due to the assumption of statistical independence of the ICA spatial maps. For Parafac this is not a problem, as can be seen from the results in Table 5. Hence, the Tensor PICA method can be used only if the fMRI activity maps are expected not to be considerably overlapping.

Although our analysis has confirmed the main conclusions of Beckmann and Smith (2005), we have raised some important issues in favor of Parafac. More importantly, we have offered detailed mathematical and statistical explanations for the differences between Parafac and Tensor PICA in the extremely noisy case of fMRI data, and we believe that these explanations are of merit to the community of multi-way, ICA and neurological data analysts.

Acknowledgments. The author would like to thank Christian Beckmann for providing the artificial fMRI datasets and for discussions on preprocessing for fMRI data, Henk Kiers for comments on Parafac data analysis, and Richard Harshman, Lieven De Lathauwer and Maarten De Vos for commenting on an earlier version fo this manuscript. The author is supported by the Dutch Organisation for Scientific Research (NWO), VENI grant 451-04-102.

	$R=3$			$R=4$			$R=6$			$R=10$			$R=13$			$R=20$		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
Map 1	*	*	*	0.66	*	0.10	0.72	0.12	*	0.67	*	*	0.52	0.34	*	0.34	-0.21	*
Map 2	*	0.91	0.20	-0.19	0.94	*	0.18	0.95	*	-0.11	0.83	0.38	-0.21	0.53	-0.44	*	0.89	*
Map 3	*	-0.36	0.94	*	-0.24	0.95	-0.18	-0.16	0.97	0.11	-0.50	0.86	*	0.32	0.76	-0.32	-0.11	0.66
Time 1	*	*	0.17	0.58	*	0.13	0.66	-0.13	0.20	0.60	*	*	0.46	*	*	0.47	-0.16	0.40
Time 2	*	0.92	0.37	-0.22	0.92	0.24	-0.21	0.93	0.14	*	0.83	0.50	-0.50	0.76	-0.41	0.19	0.91	0.22
Time 3	*	-0.18	0.90	-0.14	*	0.93	*	*	0.95	-0.13	-0.37	0.83	-0.33	0.56	0.82	*	*	0.83
Subj 1	-	3.89	4.90	2.41	3.94	4.88	2.93	3.87	4.90	2.48	3.86	4.92	2.76	3.94	5.03	2.73	3.81	5.00
Subj 2	-	3.09	3.87	3.03	3.05	3.86	2.15	3.15	3.84	3.00	3.09	3.93	2.66	2.99	3.82	2.66	3.17	3.95
Subj 3	-	2.07	3.33	1.40	2.04	3.36	1.95	2.04	3.35	1.35	2.14	3.22	1.51	2.13	3.18	1.58	2.11	3.06

Table 7: Results of Parafac for the centered and normalized dataset (A) of Beckmann and Smith (2005). For each R, the data are normalized by the estimated voxel-wise noise standard deviations from the residuals of a PCA with R components on the matricized and centered data. For each R, the Parafac solution with the highest fit of 10 runs was picked. For these best solutions, the table contains the Parafac components which had the highest correlation to the three artificial fMRI activation maps. For each such Parafac component, the correlations of the voxel mode with the spatial maps $\widetilde{\mathbf{A}}_0$, the correlations of the temporal mode with the time courses $\mathbf{B}_0$, and the estimates of the subject activation strengths $\mathbf{C}_0$ in (3.1) are given. An * denotes a correlation less than 0.10 in magnitude. A - indicates that the correlation with the artificial fMRI signal is too low to yield meaningful estimates of the subject strengths.

	$R=3$			$R=4$			$R=6$			$R=10$			$R=13$			$R=20$		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
Map 1	0.17	*	*	0.24	*	*	0.29	*	*	0.25	*	*	0.31	*	*	0.39	*	*
Map 2	*	0.92	*	*	0.92	*	*	0.92	*	*	0.92	*	*	0.92	*	*	0.91	*
Map 3	*	*	0.91	*	*	0.91	*	*	0.91	*	*	0.91	*	*	0.91	*	*	0.91
Time 1	*	*	0.20	0.28	*	0.20	0.29	*	0.19	0.29	*	0.19	0.36	*	0.19	0.44	*	0.20
Time 2	*	0.93	*	*	0.94	*	*	0.94	*	0.12	0.93	*	*	0.94	*	0.10	0.93	*
Time 3	*	*	0.95	*	*	0.95	*	*	0.95	*	*	0.95	*	*	0.95	*	*	0.95
Subj 1	-	3.91	4.74	-	3.91	4.75	-	3.93	4.77	-	3.92	4.77	-	3.91	4.78	3.14	3.88	4.78
Subj 2	-	3.10	3.96	-	3.11	3.96	-	3.08	3.94	-	3.07	3.89	-	3.06	3.89	2.66	3.07	3.87
Subj 3	-	2.02	3.45	-	2.01	3.43	-	2.03	3.43	-	2.04	3.47	-	2.07	3.47	0.20	2.12	3.49

Table 8: Results of Tensor PICA for the centered and normalized dataset (A) of Beckmann and Smith (2005). The Comon-4 ICA algorithm is used (Comon, 1994b). For each R, the data are normalized by the estimated voxel-wise noise standard deviations from the residuals of a PCA with R components on the matricized and centered data. For each R, the Tensor PICA procedure needed only two iterations. For each Tensor PICA solution, the table contains the components which had the highest correlation to the three artificial fMRI activation maps. For each such component, the correlations of the voxel mode with the spatial maps $\widetilde{\mathbf{A}}_0$, the correlations of the temporal mode with the time courses $\mathbf{B}_0$, and the estimates of the subject activation strengths $\mathbf{C}_0$ in (3.1) are given. An * denotes a correlation less than 0.10 in magnitude. A - indicates that the correlation with the artificial fMRI signal is too low to yield meaningful estimates of the subject strengths.

	2∗signals			3∗signals			5∗signals			100∗signals			18% active			47% active		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
Map 1	0.97	*	*	0.99	*	*	1	*	*	1	*	*	0.95	*	0.15	0.99	-0.11	*
Map 2	*	0.99	*	*	1	*	*	1	*	*	1	*	*	0.99	-0.14	*	0.99	-0.11
Map 3	*	*	0.99	-0.13	*	1	-0.11	*	1	*	*	1	-0.25	*	0.96	-0.25	-0.25	0.99
Time 1	0.92	*	0.15	0.96	*	0.20	0.98	*	0.20	1	*	0.16	0.88	*	0.29	0.96	*	0.22
Time 2	-0.13	0.99	*	*	1	*	*	1	*	*	1	*	*	0.99	*	*	0.99	*
Time 3	*	*	0.99	*	*	1	*	*	0.99	0.16	*	1	-0.13	0.11	0.97	*	*	0.99
Subj 1	2.86	3.88	4.75	2.82	3.98	5.02	2.87	3.95	5.08	2.42	2.94	4.37	2.92	3.91	4.93	3.01	3.94	5.06
Subj 2	2.27	3.08	4.03	2.23	3.00	3.92	2.13	3.03	3.88	2.36	3.09	4.12	2.13	3.09	3.87	2.01	3.07	3.93
Subj 3	1.91	2.12	3.34	2.02	2.03	3.08	2.04	2.06	3.02	2.36	3.29	3.73	1.98	2.03	3.27	1.98	2.01	2.99

Table 9: Results of Parafac for the altered dataset (A) with $R = 3$. The data are centered and normalized by the estimated voxel-wise noise standard deviations from the residuals of a PCA with 3 components on the matricized and centered data. For each dataset, the Parafac solution with the highest fit of 10 runs was picked. For these best solutions, the table contains the Parafac components which had the highest correlation to the three artificial fMRI activation maps. For each such Parafac component, the correlations of the voxel mode with the spatial maps $\widetilde{\mathbf{A}}_0$, the correlations of the temporal mode with the time courses $\mathbf{B}_0$, and the estimates of the subject activation strengths $\mathbf{C}_0$ in (3.1) are given. An * denotes a correlation less than 0.10 in magnitude.

	2∗signals			3∗signals			5∗signals			100∗signals			18% active			47% active		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
Map 1	0.92	*	*	0.98	*	*	0.99	*	*	0.99	*	*	0.89	*	*	0.98	-0.13	*
Map 2	*	0.99	*	*	1	*	*	1	*	*	1	*	*	0.99	*	-0.12	0.99	-0.18
Map 3	*	*	0.99	*	*	1	*	*	1	*	*	1	*	*	0.98	*	-0.15	0.99
Time 1	0.94	*	*	0.98	*	*	0.99	*	0.16	1	*	0.16	0.92	*	0.19	0.98	*	0.17
Time 2	*	0.99	*	*	0.99	*	*	1	*	*	1	*	*	0.99	*	*	0.99	*
Time 3	*	*	0.99	*	*	0.99	0.14	*	1	0.16	*	1	*	*	0.98	0.14	*	0.99
Subj 1	2.83	3.88	4.75	2.82	3.98	5.02	2.88	3.95	5.08	2.42	2.94	4.37	2.95	3.91	4.90	3.02	3.94	5.05
Subj 2	2.26	3.07	4.04	2.22	3.00	3.92	2.13	3.03	3.88	2.36	3.09	4.12	2.07	3.11	3.88	1.99	3.07	3.95
Subj 3	1.97	2.11	3.34	2.03	2.03	3.07	2.04	2.06	3.02	2.36	3.29	3.73	2.01	2.02	3.31	1.98	2.01	2.99

Table 10: Results of Tensor PICA for the altered dataset (A) with $R = 3$. The Comon-4 ICA algorithm is used (Comon, 1994b). The data are normalized by the estimated voxel-wise noise standard deviations from the residuals of a PCA with 3 components on the matricized and centered data. In all cases, the Tensor PICA procedure needed only two iterations. For each Tensor PICA solution, the table contains the components which had the highest correlation to the three artificial fMRI activation maps. For each such component, the correlations of the voxel mode with the spatial maps $\widetilde{\mathbf{A}}_0$, the correlations of the temporal mode with the time courses $\mathbf{B}_0$, and the estimates of the subject activation strengths $\mathbf{C}_0$ in (3.1) are given. An * denotes a correlation less than 0.10 in magnitude.

	2∗signals			3∗signals			5∗signals			100∗signals			18% active			47% active		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
Map 1	0.87	*	-0.34	0.98	*	-0.18	1	*	*	1	*	*	0.79	*	*	0.98	-0.15	-0.23
Map 2	*	0.99	*	*	1	-0.12	*	1	-0.12	*	1	*	*	0.99	*	*	0.99	-0.16
Map 3	0.27	*	0.94	*	*	0.98	*	*	1	*	*	1	-0.50	*	0.98	-0.10	*	0.99
Time 1	0.85	*	*	0.94	*	0.17	0.99	*	0.18	1	*	0.16	0.92	*	0.48	0.97	*	0.14
Time 2	*	0.99	*	*	0.99	*	*	0.99	*	*	1	*	*	0.99	*	*	0.99	*
Time 3	0.57	*	0.97	0.43	0.14	1	0.26	0.14	1	0.17	*	1	0.21	*	0.92	0.32	*	0.99
Subj 1	2.81	3.88	4.76	2.83	3.98	5.02	2.88	2.95	5.08	2.42	2.94	4.37	2.94	3.91	4.92	3.02	3.94	5.05
Subj 2	2.26	3.07	4.05	2.22	3.00	3.92	2.13	3.03	3.88	2.36	3.09	4.12	2.08	3.11	3.85	2.00	3.07	3.95
Subj 3	2.00	2.12	3.31	2.02	2.04	3.07	2.04	2.06	3.02	2.36	3.29	3.73	2.01	2.01	3.32	1.98	2.01	2.99

Table 11: Results of Parafac Candelinc for the altered dataset (A) with $R = 3$. The data are normalized by the estimated voxel-wise noise standard deviations from the residuals of a PCA with 3 components on the matricized and centered data. For each dataset, the Parafac solution with the highest fit of 10 runs was picked. For these best solutions, the table contains the components which had the highest correlation to the three artificial fMRI activation maps. For each such component, the correlations of the voxel mode with the spatial maps $\widetilde{\mathbf{A}}_0$, the correlations of the temporal mode with the time courses $\mathbf{B}_0$, and the estimates of the subject activation strengths $\mathbf{C}_0$ in (3.1) are given. An * denotes a correlation less than 0.10 in magnitude. A - indicates that the correlation with the artificial fMRI signal is too low to yield meaningful estimates of the subject strengths.

References

Andersen, A.H., & Rayens, W.S. (2004). Structure-seeking multilinear methods for the analysis of fMRI data. *NeuroImage*, *22*, 728–739.

Beckmann, C.F., & Smith, S.M. (2004). Probabilistic Independent Component Analysis for Functional Magnetic Resonance Imaging. *IEEE Transactions on Medical Imaging*, *24*, 137–152.

Beckmann, C.F., & Smith, S.M. (2005). Tensorial extensions of independent component analysis for multisubject FMRI analysis. *NeuroImage*, *25*, 294–311.

Belouchrani, A., Abed-Meraim, K., Cardoso, J.-F., & Moulines, E. (1997). A blind source separation technique using second order statistics. *IEEE Transactions on Signal Processing*, *45*, 434–444.

Bro, R., & Smilde, A.K. (2003). Centering and scaling in component analysis. *Journal of Chemometrics*, *17*, 16–33.

Browne, M.W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*, 111–150.

Calhoun, V.D., Adali, T., Hansen, L.K., Larsen, J., & Pekar, J.J. (2003). ICA of functional fMRI data: an overview. In *Proceedings of the 4th International Symposium on Independent Component Analysis and Blind Signal Separation* (pp. 281–288). Nara, Japan.

Cao, Y.Z., Chen, Z.P., Mo, C.Y., Wu, H.L., & Yu, R.Q. (2000). A Parafac algorithm using penalty diagonalization error (PDE) for three-way data array resolution. *The Analyst*, *125*, 2303–2310.

Cardoso, J.-F., & Souloumiac, A. (1993). Blind beamforming for non Gaussian signals. In *IEE Proceedings-F*, *140*, 362–370.

Carroll, J.D., & Chang, J.J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of Eckart-Young decomposition. *Psychometrika*, *35*, 283–319.

Carroll, J.D., Pruzansky, S., & Kruskal, J.B. (1980). Candelinc: a general approach to multidimensional analysis of many-way arrays with linear constraints on parameters. *Psychometrika 45*, 3–24.

Comon, P. (1994a). Independent component analysis, a new concept? *Signal Processing*, *36*, 287–314.

Comon. P. (1994b). Fourth order cumulants ICA algorithm. Available online at http://www.i3s.unice.fr/∼comon/codesICA.txt

De Lathauwer, L., De Moor, B., & Vandewalle, J. (2000). An introduction to independent component analysis. *Journal of Chemometrics*, *14*, 123–149.

De Silva, V., & Lim, L.-H. (2006). Tensor rank and the ill-posedness of the best low-rank approximation problem. SCCM Technical Report *06-06*. Available online at http://www-sccm.stanford.edu/nf-publications-tech.html

Eckart, C., & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, *1*, 211–218.

Gudbjartsson, H., & Patz, S. (1995). The Rician distribution of noisy MRI data. *Magnetic Resonance in Medicine*, *34*, 910–914.

Harshman, R.A. (1970). Foundations of the Parafac procedure: models and conditions for an "explanatory" multimodal factor analysis. *UCLA Working Papers in Phonetics*, *16*, 1–84.

Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, *10*, 626–634.

Hyvärinen, A., & Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural Networks*, *13*, 411–430.

Hyvärinen, A. (2005). FastICA routine. Available online at http://www.cis.hut.fi/projects/ica/fastica

Kiers, H.A.L., & Harshman, R.A. (1997). Relating two proposed methods for speedup of algorithms for fitting two- and three-way principal component and related multilinear models. *Chemometrics and Intelligent Laboratory Systems*, *36*, 31–40.

Kroonenberg, P.M. (1983). *Three-mode principal component analysis*, Leiden: DSWO Press.

Kruskal, J.B. (1977). Three-way arrays: rank and uniqueness of trilinear decompositions, with applications to arithmetic complexity and statistics. *Linear Algebra and its Applications*, *18*, 95–138.

McKeown, M.J., Jung, T-P., Makeig, S., Brown, G., Kindermann, S.S., Lee, T-W., & Sejnowski, T.J. (1998). Spatially independent activity patterns in functional MRI data during the Stroop color-naming task. *Proceedings of the National Academy of Sciences of the U.S.A.*, *95*, 803–810.

McKeown, M.J., Makeig, S., Brown, G.G., Jung, T-P., Kindermann, S.S., Bell, A.J., & Sejnowski, T.J. (1998). Analysis of fMRI data by blind separation into independent spatial components. *Human Brain Mapping*, *6*, 160–188.

McKeown, M.J., Hansen, L.K., & Sejnowski, T.J. (2003). Independent component analysis of functional MRI: what is signal and what is noise? *Current Opinion in Neurobiology*, *13*, 620–629.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, *2*, 559–572.

Penny, W.D., Roberts S.J., & Everson, R.M. (2001). ICA: model order selection and dynamic source models. In Robert, S., & Everson, R. (Eds.), *Independent Component Analysis, Principles and Practice* (pp. 299–314) Cambridge: Cambridge University Press.

Pham, D.-T., & Cardoso, J.F. (2001). Blind separation of instantaneous mixtures of non stationary sources. *IEEE Transactions on Signal Processing*, *49*, 1837–1848.

Rajih, H., & Comon, P. (2005). Enhanced line search: a novel method to accelerate Parafac. In *Proceedings of the 13th European Signal Processing Conference*, Antalya, Turkey.

Sidiropoulos, N.D., & Bro, R. (2000). On the uniqueness of multilinear decomposition of N-way arrays. *Journal of Chemometrics*, *14*, 229–239.

Sidiropoulos, N.D. (2004). Low-rank decomposition of multi-way arrays: A signal processing perspective. In *Proceedings of IEEE Sensor Array and Multichannel (SAM) signal processing Workshop*, Barcelona, Spain.

Smilde, A., Bro, R., & Geladi, P. (2004). *Multi-way Analysis: Applications in the Chemical Sciences*. Chichester: Wiley.

Stegeman, A. (2006a). Degeneracy in Candecomp/Parafac explained for $p \times p \times 2$ arrays of rank $p + 1$ or higher. *Psychometrika*, *71*, 483–501.

Stegeman, A. (2006b). Low-rank approximation of generic $p \times q \times 2$ arrays and diverging components in the Candecomp/Parafac model. Technical Report. Available online at http://www.gmw.rug.nl/∼stegeman

Stegeman, A. (2006c). Degeneracy in Candecomp/Parafac explained for several three-sliced arrays with a two-valued typical rank. Technical Report. Available online at http://www.gmw.rug.nl/∼stegeman

Thomas, C.G., Harshman, R.A., & Menon, R.S. (2002). Noise reduction in BOLD-based fMRI using component analysis. *NeuroImage*, *17*, 1521–1537.

Tipping, M.E., & Bishop, C.M. (1999). Mixtures of probabilistic principal component analyzers. *Neural Computation*, *11*, 443–482.

Tomasi, G., & Bro, R. (2006). A Comparison of algorithms for fitting the Parafac model. *Computational Statistics & Data Analysis*, *50*, 1700–1734.

Tucker, L.R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, *31*, 279–311.

Vega-Montoto, L., & Wentzell, P.D. (2003). Maximum likelihood parallel factor analysis (MLPARAFAC). *Journal of Chemometrics 17*, 237–253.

Wink, A.M., & Roerdink, J.B.T.M. (2006). BOLD noise assumptions in fMRI. *International Journal of Biomedical Imaging*, in press.