Probabilistic Independent Component Analysis for Functional Magnetic Resonance Imaging

FMRIB Technical Report TR02CB1

Christian F. Beckmann and Stephen M. Smith

Oxford Centre for Functional Magnetic Resonance Imaging of the Brain (FMRIB),
Department of Clinical Neurology, University of Oxford, John Radcliffe Hospital,
Headley Way, Headington, Oxford, UK

Abstract

We present an integrated approach to Probabilistic ICA for FMRI data that allows for non-square mixing in the presence of Gaussian noise. In order to avoid overfitting, we employ objective estimation of the amount of Gaussian noise through Bayesian analysis of the true dimensionality of the data, i.e. the number of activation and non-Gaussian noise sources. This enables us to carry out probabilistic modelling and achieves an asymptotically unique decomposition of the data. It reduces problems of interpretation, as each final independent component is now much more likely to be due to only one physical or physiological process. We also describe other improvements to standard ICA, such as temporal pre-whitening and variance normalisation of time-series, the latter being particularly useful in the context of dimensionality reduction when weak activation is present. We discuss the use of prior information about the spatiotemporal nature of the source processes, and an alternative-hypothesis testing approach for inference, using Gaussian mixture models. The performance of our approach is illustrated and evaluated on real and artificial FMRI data, and compared to the spatio-temporal accuracy of results obtained from classical ICA and GLM analyses.

1 Introduction

Data analysis often has to proceed on measurements which are intricate mixtures of the initial signal sources of interest. Unless accurate information is available to allow for informed estimation, this is a challenging problem. A possible solution is to employ what is known in the area of signal processing as *blind source separation* (BSS) techniques [27].

The signal of functional magnetic resonance imaging studies is a prime example, comprising different sources of variability, possibly including machine artefacts, physiological pulsation, head motion and haemodynamic changes induced by different experimental conditions. This mixture of signals presents a huge challenge for analytical methods attempting to identify stimulus- or task-related changes.

The vast majority of analytical techniques currently applied to FMRI data test specific hypotheses about the expected BOLD response at the individual voxel locations using simple regression or more sophisticated models like the General Linear Model[1] (GLM) [51]. There, the expected signal changes are specified as regressors of interest in a multiple linear regression framework and the estimated regression coefficients are tested against a null hypothesis that these values are distributed

[1] more correctly referred to simply as 'linear model'

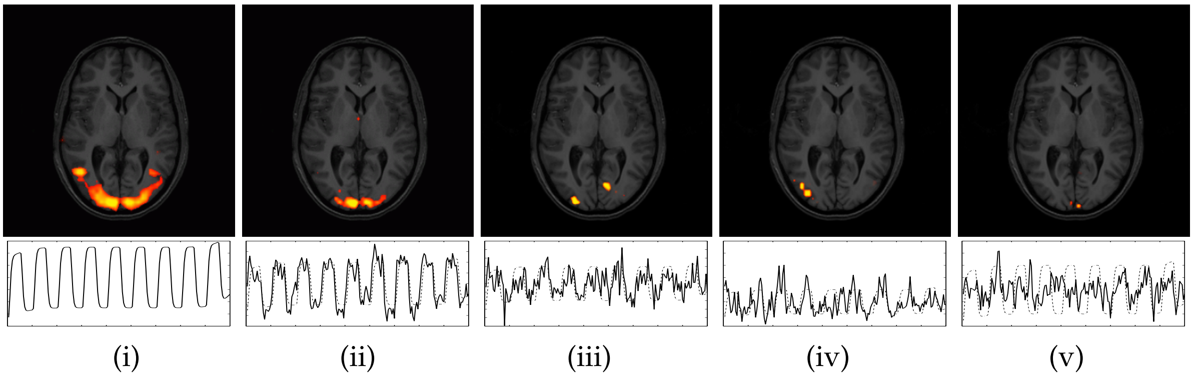


Figure 1: GLM and classical ICA analysis of visual stimulus FMRI data: (i) GLM results using GRF-based inference ($Z > 2.3, p < 0.01$); (ii)–(v) IC maps where the correlation of the extracted time course with the expected BOLD response exceeds 0.3. Both for GLM analysis and ICA, the data was high-pass filtered (Gaussian-weighted local straight line fitting with cutoff of 90s [34]) and spatially smoothed with a Gaussian kernel of 3mm (FWHM). ICA maps were thresholded at $Z > 2.3$ after transformation into "Z-scores" across the spatial domain as described in [36]. Note that this should not be confused with GLM or PICA Z-score thresholding as described in section 4, where Z-scores are formed by dividing by the *voxel-wise* estimated standard deviation of the noise.

according to some known null distribution. These voxel-wise test statistics form summary images known as *statistical parametric maps* which are commonly assessed for statistical significance using voxel-wise null-hypothesis testing or testing for the size or mass of suprathresholded clusters [39]. These approaches are confirmatory in nature and make strong prior assumptions about the spatio-temporal characteristics of signals contained in the data. Naturally, the inferred spatial patterns of activation depend on the validity and accuracy of these assumptions. The most obvious problem with hypothesis-based techniques is the possible presence of unmodelled artefactual signal in the data. Structured noise which is temporally non-orthogonal to an assumed regression model will bias the parameter estimates, and noise orthogonal to the design will inflate the residual error, thus reducing statistical significance. Either way, any discrepancy between the assumed and the 'true' signal space will render the analysis sub–optimal. Furthermore, while there is a growing number of models that explicitly include prior spatial information [22], the standard GLM approach is univariate and essentially discards information about the spatial properties of the data, only inducing spatial smoothness by convolving the individual volumes with a Gaussian smoothing kernel, and returning to spatial considerations only *after* modelling has completed (e.g. *Gaussian Random Field Theory*-based inference [39]).

As one alternative to hypothesis–driven analytical techniques, *Independent Component Analysis* (*ICA*, [15]) has been applied to FMRI data as an exploratory data analysis technique in order to find independently distributed spatial patterns that depict source processes in the data [36, 8]. The basic goal of ICA is to solve the BSS problem by expressing a set of random variables (*observations*) as linear combinations of statistically independent latent component variables (*source signals*).

There have been primarily two different research communities involved in the development of ICA. Firstly, the study of mixed sources is a classical signal processing problem. The seminal work into BSS [27] looked at extensions to standard principal component analysis (PCA). Theoretical work on high order moments provided one of the first solutions to a BSS problem [12]. [27] published a concise presentation of their adaptive algorithm and outlined the transition from PCA to ICA very clearly. Their approach has been further developed by [28] and [14]. Exact conditions for the identifiability of the model can be found in [15] together with an algorithm that approximates source signal distributions using their first few moments, a technique that was also employed by other authors [16].

In parallel to blind source separation studies, unsupervised learning rules based on information theoretic principles were proposed by [32]. These learning rules are based on the principle of redundancy reduction as a coding strategy for neurons of the perceptual system [4].

More recently, [5] and [9] introduced a surprisingly simple blind source separation algorithm for a non-linear feed-forward network from an information maximization viewpoint. This algorithm was subsequently improved, extended and modified [2] and its relation to maximum likelihood estimation and redundancy reduction was investigated [33]. There now exists a variety of alternative algorithms and principled extensions that include work on non-linear mixing, non-instantaneous mixing, incorporation of source structure and observational noise.

Classical Independent Component Analysis has been popularised in the field of FMRI analysis by [36], where the data from the FMRI experiment with n voxels measured at p different time points is written as a $p \times n$ matrix $\boldsymbol{X}$ for which a decomposition is sought such that

$$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{S}. \tag{1}$$

The matrix $\boldsymbol{S}$ is optimized to contain statistically independent spatial maps in its rows, i.e. spatial areas in the brain, each with an internally consistent temporal dynamic, which is characterised by a time-course contained in the associated column of the square *mixing matrix* $\boldsymbol{A}$.

In [36], the sources are estimated by iteratively optimising an *unmixing matrix* $\boldsymbol{W} = \boldsymbol{A}^{-1}$ so that $\boldsymbol{S} = \boldsymbol{W}\boldsymbol{X}$ contains mutually independent rows, using the *infomax* algorithm [9].

The ICA model above, though being a simple linear regression model, differs from the standard GLM as used in neuroimaging in two essential aspects: firstly, the mixing is assumed to be square, i.e. the signal is not constrained to be contained within a lower dimensional signal sub–space. Secondly, the model of equation 1 does not include a noise model. Instead, the data are assumed to be completely characterized by the estimated sources and the mixing matrix. This, in turn, precludes the assessment of statistical significance of the source estimates within the framework of null-hypotheses testing. Both problems are strongly linked in that if we relax the assumption on square mixing by requesting a smaller number of source processes to represent the dynamics in the data, we automatically introduce a mismatch between the best linear model fit and the original data. In analogy to the GLM case, the residual error will be the difference between what the model can explain and what we actually observe.

In the absence of a suitable noise model, slightest differences in the measured hæmodynamic response at two different voxel locations are necessarily treated as 'real effects'. These differences might represent valid spatial variations, e.g. slightly different temporal responses between left and right hemisphere or simply differences in the background noise level (e.g. spatial variations due to the image acquisition, sampling etc.), and may cause clusters of voxels that 'activate' to the same external stimulus to be fragmented into different spatial maps - a split into what in [36] has been termed *consistently* and *transiently task related components* occurs. The noise free generative model precludes any test for significance and threshold techniques like converting the component map values into Z-scores [36] are devoid of statistical meaning and can only be understood as ad-hoc recipes (see section 4).

As an example, figure 1 shows the results of a GLM analysis of a simple FMRI data set from a visual stimulation experiment (30s on/off block design with a black/white checkerboard reversing at 8Hz during the on condition).

Though visual experiments of this kind are generally expected to generate consistent activation maps over a wide range of analysis techniques, the spatial maps of GLM analysis and ICA differ substantially. While one of the ICA maps clearly depicts activation in the primary visual areas (ii), the spatial extent differs from areas found using a GLM analysis (i). This is only partly due to the arbitrary IC map thresholding; the lateral anterior activation foci in the GLM map cannot be obtained within the IC map without dramatically inflating the number of voxels classified as active. In addition to the one highly correlated IC map, three additional component maps have an associated time course which correlates with the experimental paradigm[2] at $r > 0.3$. Their spatial activation patterns, though being well clustered and localised inside visual cortical areas, do not lend themselves to easy interpretation.

This is the classical problem of over-fitting a noise-free generative model to noisy observations [10] and needs to be resolved by setting up a suitable probabilistic model that controls the balance between what is attributable to 'real effects' of interest and what simply is due to observational noise.

In order to address these issues we examine the probabilistic Independent Component Analysis (PICA) model [38, 7] for FMRI data that allows for a non-square mixing process and assumes that the data are confounded by additive Gaussian noise.

In the case of isotropic noise covariance the task of blind source separation can be divided into three stages: (i) estimation of a signal + noise sub–space that contains the source processes and a noise sub–space orthogonal to the first, (ii) estimation of independent components in the signal + noise sub–space and (iii) assessing the statistical significance of estimated sources.

At the first stage we employ probabilistic Principal Component Analysis (PPCA, [49]) in order to find an appropriate linear sub-space which contains the sources. The choice of the number of components to extract is a problem of model order selection. Underestimation of the dimensionality will discard valuable information and result in suboptimal signal extraction. Overestimation, however, results in a large number of spurious components due to underconstrained estimation and a factorization that will overfit the data, harming later inference and dramatically increasing computational costs.

Within the probabilistic PCA framework we will demonstrate that the number of source processes can be inferred from the covariance matrix of the observations using a Bayesian framework that approximates the posterior distribution of the model order [37] and extending this approach to take account of the limited amount of data and the particular structure of FMRI noise [7].

At the second stage the source signals are estimated within the lower- dimensional signal + noise sub–space using a fixed-point iteration scheme [23] that maximises the non-Gaussianity of the source estimates. Finally, at the third level, the extracted spatial maps are converted into 'Z statistic' maps based on the estimated standard error of the residual noise. These maps are assessed for significantly modulated voxels using a Gaussian Mixture Model for the distribution of intensity values.

The paper is organised as follows. Section 2 defines the probabilistic ICA model and discusses the uniqueness of the solution. Estimation of the model order, the mixing process and the sources is outlined in section 3. Section 4 discusses the Gaussian Mixture Model approach to IC map thresholding. Finally, sections 5 and 6 demonstrate the technique on artificial and real FMRI data and some discussion and concluding comments are given in sections 7 and 8.

2 Probabilistic ICA model

Similar to the square noise-free case, the probabilistic ICA model is formulated as a generative linear latent variables model. It is characterised by assuming that the p-variate vector of observations is generated from a set of q statistically independent non-Gaussian sources via a linear instantaneous mixing process corrupted by additive Gaussian noise $\boldsymbol{\eta}(t)$:

$$\boldsymbol{x}_i = \boldsymbol{A}\boldsymbol{s}_i + \boldsymbol{\mu} + \boldsymbol{\eta}_i \qquad \forall i \in \mathcal{V}. \tag{2}$$

Here, $\boldsymbol{x}_i$ denotes the p-dimensional column vector of individual measurements at voxel location i, $\boldsymbol{s}_i$ denotes the q- dimensional column vector of non-Gaussian source signals contained in the data and $\boldsymbol{\eta}_i$ denotes Gaussian noise $\boldsymbol{\eta}_i \sim \mathcal{N}(0, \sigma^2\boldsymbol{\Sigma}_i)$. We assume that $q < p$, i.e. that there are fewer source

[2] choosing components based on correlation or shared peak frequency response with an assumed evoked haemodynamic response function often appears to work well for simple block paradigms like the one here; for more complicated paradigms choosing activation maps becomes a much harder challenge

processes than observations in time. The covariance of the noise is allowed to be voxel dependent in order to allow for the vastly different noise covariances observed in different tissue types [50].

The vector $\boldsymbol{\mu}$ defines the mean of the observations $\boldsymbol{x}_i$ where the index i is over the set of all voxel locations $\mathcal{V}$ and the $p \times q$ matrix $\boldsymbol{A}$ is assumed to be non-degenerate, i.e. of rank q. Solving the blind separation problem requires finding a linear transformation matrix $\boldsymbol{W}$ such that

$$\widehat{\boldsymbol{s}} = \boldsymbol{W}\boldsymbol{x}$$

is a good approximation to the true source signals $\boldsymbol{s}$.

The PICA model is similar to the standard GLM with the difference that, unlike the design matrix in the GLM, the mixing matrix $\boldsymbol{A}$ is no longer pre-specified prior to model fitting but will be estimated from the data as part of the model fitting. The spatial source signals correspond to parameter estimates in the GLM with the additional constraint of being statistically independent.

The model of equation 2 is closely related to Factor Analysis (FA) [6]. There, the sources are assumed to have a Gaussian distribution and the noise is assumed to have a diagonal covariance matrix. In Factor Analysis, the sources are known as common factors and $\boldsymbol{\eta}$ is a vector of random variables called specific factors. In FA the assumption of independence between the individual source processes reduces to assuming that sources are mutually uncorrelated.

2.1 Uniqueness

[42] extends the standard factor analysis model such that the common and specific variables are independent non-degenerate random variables and examines the implication for the minimum rank of the mixing matrix $\boldsymbol{A}$ in equation 2. Earlier work [41] characterised the multivariate normal distribution through the non-uniqueness of its linear structure, a result which within the ICA literature has been restated as the limitation that only one Gaussian source process, at most, may contribute to the observations for the ICA model to be estimable [15, 23]. Here, a vector variable $\boldsymbol{x}$ is said to have a linear structure if it can be decomposed as

$$\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{s}, \tag{3}$$

where $\boldsymbol{s}$ is a vector of non-degenerate statistically independent one dimensional random variables and $\boldsymbol{A}$ is a matrix of full column rank. The linear structure is said to be *essentially unique* if all the linear decompositions are equivalent in the sense that if the vector variable $\boldsymbol{x}$ allows for two structural representations

$$\boldsymbol{x} = \boldsymbol{\mu}_1 + \boldsymbol{A}_1\boldsymbol{s}_1 \qquad \text{and} \qquad \boldsymbol{x} = \boldsymbol{\mu}_2 + \boldsymbol{A}_2\boldsymbol{s}_2, \tag{4}$$

then every column of $\boldsymbol{A}_1$ is a multiple of some column of $\boldsymbol{A}_2$ and vice versa, i.e. the two matrices are identical modulo scaling and permutation. This again has been noted as a standard restriction within the ICA framework [15].

The main result in [42] is a decomposition theorem that states that if $\boldsymbol{x}$ is a p-variate random variable with a linear structure $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s}$ where all the elements of $\boldsymbol{s}$ are non-Gaussian variables, then there does not exist a non–equivalent linear structure involving the same number or a smaller number of structural variables than that of $\boldsymbol{s}$.

Furthermore, if $\boldsymbol{x}$ is a p-vector random variable with a linear structure $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s}$ then $\boldsymbol{x}$ can be decomposed

$$\boldsymbol{x} = \boldsymbol{x}_1 + \boldsymbol{x}_2$$

where $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are independent, $\boldsymbol{x}_1$ is non-Gaussian and has an essentially unique linear structure and $\boldsymbol{x}_2$ is p-variate normal with a non-unique linear structure.

The proofs involve the characteristic functions of the vector random variables $\boldsymbol{x}$ and $\boldsymbol{s}$ and as such these results are applicable only if the number of observations (i.e. voxels) is sufficiently large to accurately reflect the distribution of these quantities.

The results show, however, that conditioned on knowing the number of source signals contained in the data and under the assumption that the data are generated according to equation 2, i.e. a linear mixture of independent non-Gaussian source signals confounded by Gaussian noise, there is no non-equivalent decomposition into this number of independent non-Gaussian random variables and an associated mixing matrix; the decomposition into independent components is unique, provided we do not attempt to extract more than q source signals from the data.

3 Maximum Likelihood estimation

Throughout the remainder of this paper, we are going to keep the parameter $\boldsymbol{\mu}$ fixed at its ML estimate:

$$\boldsymbol{\mu}_{\mathsf{ML}} = \langle \boldsymbol{x}_i \rangle$$

(assuming zero-mean sources) and will assume that it has been removed from the data. The mean can always be reintroduced after model estimation using $\boldsymbol{x}_i = \boldsymbol{A}(\boldsymbol{s}_i + \boldsymbol{W}\boldsymbol{\mu}_{\mathsf{ML}}) + \widetilde{\boldsymbol{\eta}}_i$, where

$$\widetilde{\boldsymbol{\eta}}_i = \boldsymbol{\eta}_i + (\boldsymbol{I} - \boldsymbol{A}\boldsymbol{W})\boldsymbol{\mu}_{ML}.$$

Without loss of generality we will also assume that the sources have unit variance for we can freely exchange arbitrary scaling factors between the source signals and the associated columns of the mixing matrix $\boldsymbol{A}$.

If the noise covariance $\boldsymbol{\Sigma}_i$ was known we can use its Cholesky decomposition $\boldsymbol{\Sigma}_i = \boldsymbol{K}_i\boldsymbol{K}_i^t$ to rewrite equation 2:

$$\boldsymbol{K}_i^{-1}\boldsymbol{x}_i = \boldsymbol{K}_i^{-1}\boldsymbol{A}\boldsymbol{s}_i + \boldsymbol{K}_i^{-1}\boldsymbol{\eta}_i,$$

and obtain a new representation

$$\bar{\boldsymbol{x}}_i = \bar{\boldsymbol{A}}\boldsymbol{s}_i + \bar{\boldsymbol{\eta}}_i$$

where $\bar{\boldsymbol{\eta}}_i = \boldsymbol{K}_i^{-1}\boldsymbol{\eta} \sim \mathcal{N}(0, \sigma^2\boldsymbol{I})$, i.e. where the noise covariance is isotropic at every voxel location. During this step, the data are pre-whitened with respect to the noise covariance, which is not to be confused with (spatial) pre-whitening of the data matrix using singular-value decomposition. In order to simplify further notation, we are therefore going to assume that the data contain isotropic noise and drop the bars.

Noise and signal are assumed to be uncorrelated and therefore

$$\boldsymbol{R}_{\boldsymbol{x}} - \sigma^2\boldsymbol{I} = \boldsymbol{A}\boldsymbol{A}^t,$$

where $\boldsymbol{R}_{\boldsymbol{x}} = \langle \boldsymbol{x}_i\boldsymbol{x}_i^t \rangle$ denotes the covariance matrix of the observations. Let $\boldsymbol{X}$ be the $p \times N$ matrix of voxel-wise pre-whitened data vectors and let $\boldsymbol{X} = \boldsymbol{U}(N\boldsymbol{\Lambda})^{\frac{1}{2}}\boldsymbol{V}$ be its singular value decomposition. Furthermore let the rank of the mixing matrix $\boldsymbol{A}$, i.e. the number of source processes q, be known. Then

$$\widehat{\boldsymbol{A}}_{\mathsf{ML}} = \boldsymbol{U}_q(\boldsymbol{\Lambda}_q - \sigma^2\boldsymbol{I}_q)^{\frac{1}{2}}\boldsymbol{Q}^t, \tag{5}$$

where $\boldsymbol{U}_q$ and $\boldsymbol{\Lambda}_q$ contain the first q eigenvectors and eigenvalues of $\boldsymbol{U}$ and $\boldsymbol{\Lambda}$ and where $\boldsymbol{Q}$ denotes a $q \times q$ orthogonal rotation matrix.

Estimating the mixing matrix $\boldsymbol{A}$ thus reduces to identifying the square matrix $\boldsymbol{Q}$ after whitening the data with respect to the noise covariance $\boldsymbol{\Sigma}_i$ and projecting the temporally whitened observations onto the space spanned by the q eigenvectors of $\boldsymbol{R}_{\boldsymbol{x}}$ with largest eigenvalues. From $\widehat{\boldsymbol{A}}$, the maximum likelihood source estimates are obtained using generalised least squares:

$$\widehat{\boldsymbol{s}}_{\mathsf{ML}} = (\widehat{\boldsymbol{A}}^t\widehat{\boldsymbol{A}})^{-1}\widehat{\boldsymbol{A}}^t\boldsymbol{x}, \tag{6}$$

and the ML estimate of σ^2 becomes

$$\widehat{\sigma}^2_{\mathsf{ML}} = \frac{1}{p-q}\sum_{l=q+1}^{p}\lambda_l, \tag{7}$$

i.e. is the average of the eigenvalues in the minor subspace spanned by the $p-q$ smallest eigenvectors. Solving the model in the case of a full noise covariance where the noise covariance is unknown can be achieved by iterating estimates for $\widehat{\boldsymbol{A}}$ and $\widehat{\boldsymbol{s}}$ and re-estimating the noise covariances from the residuals $\widehat{\boldsymbol{\eta}}$. The estimation process in the presence of non-isotropic noise is computationally much more involved than estimation in the standard noise free setting. The form of $\boldsymbol{\Sigma}_i$ needs to be constrained, e.g. we can use the common approaches to FMRI noise modelling [11, 50], and restrict ourselves to autoregressive noise. However, since the exploratory approach allows modelling of various sources of variability, e.g. temporally consistent physiological noise, as part of the signal in equation 2, the noise model itself can actually be quite simplistic. Estimation of $\boldsymbol{\Sigma}_i$ from residuals in the case of autocorrelated noise is discussed in detail in [50] which is the approach used in this case.

The maximum likelihood solutions given in equations 5–7 give important insight into the methodology. Firstly, in the case where q and $\boldsymbol{\Sigma}_i$ are known, the maximum likelihood solution for $\widehat{\boldsymbol{A}}$ is contained in the principal eigenspace of $\boldsymbol{R_x}$ of dimension q, i.e. the span of the first q eigenvectors equals the span of the unknown mixing matrix $\boldsymbol{A}$. Projecting the data onto the principal eigenvectors is not just a convenient technique to deal with the high dimensionality in FMRI data but is part of the maximum likelihood solution under the sum of square loss. Even if estimation techniques are employed that do not use an initial PCA step as part of the ICA estimation, the final solution under this model is necessarily contained in the principal subspace. Secondly, combining these results with the uniqueness results stated earlier we see that *only* in the case where the analysis is performed in the appropriate lower-dimensional subspace of dimension q are the source processes uniquely identifiable. Finally, equations 5–7 imply that the standard noise-free ICA approach with dimensionality reduction using PCA implicitly operates under an isotropic noise model.

The remainder of this paper illustrates that by making this specific noise model explicit in the *modelling* and *estimation* stages, we can address important questions of model order selection, estimation and inference in a consistent way.

An immediate consequence of the fact that we are operating under an isotropic noise model is that as an initial pre-processing step we will modify the original data time courses to be normalised to zero mean and unit variance. This appears to be a sensible step in that on the one hand we know that the voxel-wise standard deviation of resting state data varies significantly over the brain but on the other hand, all voxels' time courses are assumed to be generated from the same noise process. This variance-normalisation pre-conditions the data under the 'null hypotheses' of purely Gaussian noise, i.e. in the absence of any signal: the data matrix $\boldsymbol{X}$ is identical up to second order statistics to a simple set of realisations from a $\mathcal{N}(0,1)$ noise process. Any signal component contained in $\boldsymbol{X}$ will have to reveal itself via its deviation from Gaussianity. This will turn out to be of prime importance both for the estimation of the number of sources and the final inferential steps.

After a voxel-wise normalisation of variance, two voxels with comparable noise level that are modulated by the same signal time course, $\boldsymbol{a}_j$ say, but by different amounts will have the same regression coefficient upon regression against $\boldsymbol{a}_j$. The difference in the original amount of modulation is therefore contained in the standard deviation of the residual noise. Forming voxel-wise Z statistics, i.e. dividing the PICA maps by the estimated standard deviation of $\boldsymbol{\eta}$, thus is invariant under the initial variance-normalisation.

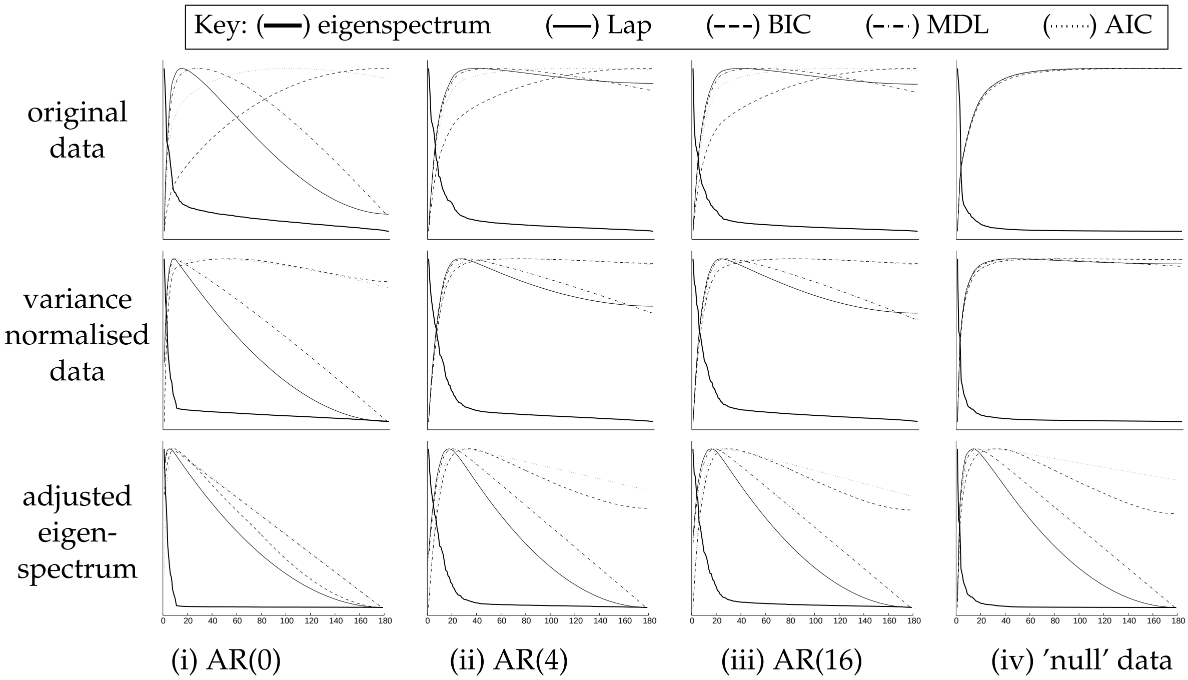
3.1 Model order selection

The maximum likelihood solutions given in equations 5–7 depend on knowledge of the latent dimensionality q. In the noise free case this quantity can easily be deduced from the rank of the covariance of the observations

$$\boldsymbol{R_x} = \langle \boldsymbol{x}_i \boldsymbol{x}_i^t \rangle = \boldsymbol{A}\boldsymbol{A}^t,$$

which is of rank q. In the presence of isotropic noise, however, the covariance matrix of the observations will be the sum of $\boldsymbol{A}\boldsymbol{A}^t$ and the noise covariance [19]

$$\boldsymbol{R_x} = \boldsymbol{A}\boldsymbol{A}^t + \sigma^2 \boldsymbol{I}_p, \tag{8}$$



(i) AR(0) (ii) AR(4) (iii) AR(16) (iv) 'null' data

Figure 2: Estimation of the intrinsic dimensionality for 10 sources with non-Gaussian distribution embedded in a 180 dimensional space with different noise characteristics (see section 5) at different stages of the estimation process: (i) Gaussian white noise, (ii) AR(4) noise, (iii) AR(16) noise, (iv) resting state FMRI noise; estimates from the original data (top), after voxel-wise variance normalisation (middle) and after additionaly adjusting the eigenspectrum using the predictive cumulative distribution $G^{-1}(\nu)$ (bottom). Every graph shows the eigenspectrum of the data covariance matrix and 4 different estimates of the intrinsic dimensionality: Laplace approximation to the model evidence, BIC, MDL and AIC.

i.e. $\boldsymbol{R_x}$ will be of full rank where the additional noise has the effect of raising the eigenvalues of the covariance matrix by σ^2. Inferring the latent dimensionality amounts to identifying the number of identical eigenvalues. In practice, however, the mismatch between the model assumption and the structure of real FMRI data renders the observed eigenspectrum to be less well behaved and we need a statistical test for the equality of eigenvalues beyond a certain threshold [21]. Determining a cutoff value for the eigenvalues using simplistic criteria like the reconstruction error or predictive likelihood will naturally predict that the accuracy steadily increases with increased dimensionality and as such cannot be used to infer latent dimensionality. Thus, criteria like retaining 99.9% of the variability result in arbitrary threshold levels that bear no relevance to the problem of estimating the latent dimensionality correctly.

Many other informal methods have been proposed, the most popular choice being the "scree plot" where one looks for a "knee" in the plot of ordered eigenvalues that signifies a split between significant and presumably unimportant directions of the data. With real FMRI data, however, the decision as to where to choose the cutoff value is not obvious and a choice based on simple visual inspection will be ambiguous (see figure 9(ii) for an example). This problem is intensified by the fact that the data set $\boldsymbol{X}$ is finite and thus $\boldsymbol{R_x}$ is being estimated by the sample covariance of the set of observations $\widetilde{\boldsymbol{R}}_{\boldsymbol{X}}$. Even in the absence of any source signals, i.e. when $\boldsymbol{X}$ contains a finite number of samples from purely Gaussian isotropic noise only, the eigenspectrum of the sample covariance matrix is not identical to σ^2 but instead distributed around the true noise covariance: the eigenspectrum will depict an apparent difference in the significance of individual directions within the noise [19].

In the case of purely Gaussian noise, however, the sample covariance matrix $\widetilde{\boldsymbol{R}}_{\boldsymbol{x}}$ has a Wishart distribution and we can utilise results from random matrix theory on the empirical distribution function $G_n(\nu)$ for the eigenvalues of the covariance matrix of a single random $p \times n$-dimensional matrix $\tilde{\boldsymbol{X}}$ [26]. Suppose that $p/n \rightarrow \gamma$ as $n \rightarrow \infty$ and $0 < \gamma \leq 1$, then $G_n(\nu) \rightarrow G_\gamma(\nu)$ almost surely, where the limiting distribution has a density

$$g(\nu) = \frac{1}{2\pi\gamma\nu}\sqrt{(\nu - b_-)(b_+ - \nu)}, \quad b_- \leq \nu \leq b_+, \tag{9}$$

and where $b_\pm = (1 \pm \sqrt{\gamma})^2$. This can be used to obtain a modification to the standard scree-plot where one compares the eigenspectrum of the observations against the quantiles of the predicted cumulative distribution $G^{-1}(\nu)$, i.e. against the expected eigenspectrum of a random Gaussian matrix. The predicted eigenspectrum of the noise becomes a function of p/n: the larger p/n, the more spread the eigenspectrum. Note that equation 9 is only satisfied for $0 < \gamma \leq 1$, i.e. when the number of samples is equal or larger than the dimensionality of the problem at hand. This approach is similar to [45], where an inverse Wishart prior is placed on the noise covariance matrix in a fully Bayesian source separation model.

If we assume that the source distributions $p(\boldsymbol{s})$ are Gaussian, the probabilistic ICA model (equation 2) reduces to the probabilistic PCA model [49]. In this case, we can use more sophisticated statistical criteria for model order selection. [37] placed PPCA in the Bayesian framework and presented a Laplace approximation to the posterior distribution of the model evidence that can be calculated efficiently from the eigenspectrum of the covariance matrix of observations. When $q < \min(N, p)$, then

$$p(\boldsymbol{X}|q) \approx p(U)\left(\prod_{j=1}^{q}\lambda_j\right)^{-N/2} \widehat{\sigma}_{\mathsf{ML}}^{-N(p-q)}(2\pi)^{(m+q)/2}|A_z|^{-1/2}N^{-q/2}, \tag{10}$$

where $m = pq - q(q+1)/2, \quad N = |\mathcal{V}|$ and $p(U)$ denotes a uniform prior over all eigenvector matrices

$$p(U) = 2^{-q}\prod_{j=1}^{q}\Gamma((p-j+1)/2)\pi^{-(p-i+1)/2},$$

and

$$|A_z| = \prod_{i=1}^{q}\prod_{j=i+1}^{p}N(\hat{\lambda}_j^{-1} - \hat{\lambda}_i^{-1})(\lambda_i - \lambda_j),$$

where the λ_l denote the entries in $\boldsymbol{\Lambda}$ and $\hat{\lambda}_l = \lambda_l$ for $l \leq q$ and $\hat{\lambda}_l = \sigma_{\mathsf{ML}}^2$ otherwise. As our estimate for the latent dimensionality of the data, we choose the value of q that maximises the approximation to the model evidence $p(\boldsymbol{X}|q)$.

In order to account for the limited amount of data, we combine this estimate with the predicted cumulative distribution and replace $\boldsymbol{\Lambda}$ by its adjusted eigenspectrum $\boldsymbol{\Lambda}/G^{-1}(\nu)$ prior to evaluating the model evidence. Other possible choices for model order selection for PPCA include the *Bayesian Information Criterion* (BIC, [29]) the *Akaike Information Criterion* (AIC, [1]) or *Minimum Description Length* (MDL, [43]).

Note that the estimation of the model order in the case of the probabilistic PCA model is based on the assumption of Gaussian source distribution. [37], however, provides some empirical evidence that the Laplace approximation works reasonably well in the case where the source distributions are non-Gaussian. As an example, figure 2 shows the eigenspectrum and different estimators of the intrinsic dimensionality for different artificial data sets, where 10 latent sources with non-Gaussian distribution were introduced into simulated AR data (i.e. auto-regressive noise where the AR parameters were estimated from real resting state FMRI data) and real FMRI resting state noise at peak levels of between 0.3% and 1.6% of the mean signal intensity. Note how the increase in AR order will increase the estimates of the latent dimensionality, simply because there are more eigenvalues that fail the sphericity assumption. Performing variance-normalisation and adjusting the eigenspectrum using $G^{-1}(\nu)$ in all cases improves the estimation. In the case of Gaussian white noise the model assumptions are correct and the adjusted eigenspectrum exactly matches equation 8. In most cases, the different estimators give similar results once the data were variance normalised and the eigenspectrum was adjusted using $G^{-1}(\nu)$. Overall, the Laplace approximation and the Bayesian Information Criterion appear to give consistent estimates of the latent dimensionality even though the distribution of the embedded sources are non-Gaussian.

3.2 Estimation of the unmixing matrix

Recall from equation 5 that in order to estimate the mixing matrix and the sources, we need to optimise an orthogonal rotation matrix in the space of whitened observations:

$$\widehat{\boldsymbol{s}} = \boldsymbol{W}\boldsymbol{x} = \boldsymbol{Q}\tilde{\boldsymbol{x}}, \tag{11}$$

where

$$\tilde{\boldsymbol{x}} = (\boldsymbol{\Lambda}_q - \sigma^2 \boldsymbol{I}_q)^{-1/2} \boldsymbol{U}_q^t \boldsymbol{x}$$

denotes the spatially whitened data.

In order to choose a technique for the unmixing step note that all previous results have highlighted the importance of non-Gaussianity of the source distributions: the split into a non-Gaussian part plus additive Gaussian noise is at the heart of the uniqueness results. Also, the estimation of the intrinsic dimensionality is based on the identification of eigenvectors of the data covariance matrix that violate the sphericity assumption of the isotropic Gaussian noise model. Consistent with this, we will estimate the unmixing matrix $\boldsymbol{W}$ based on the principle of non-Gaussianity. [23] have presented an elegant fixed point algorithm that uses approximations to neg-entropy in order to optimise for non-Gaussian source distributions and give a clear account of the relation between this approach to statistical independence. In brief, the individual sources are obtained by projecting the data $\boldsymbol{x}$ onto the individual rows of $\boldsymbol{Q}$, i.e. the rth source is estimated as

$$\widehat{\boldsymbol{s}}_r = \boldsymbol{v}_r^t \tilde{\boldsymbol{x}},$$

where $\boldsymbol{v}_r^t$ denotes the rth row of $\boldsymbol{Q}$. In order to optimise for non-Gaussian source estimates, [23] propose the following contrast function:

$$J(\boldsymbol{s}_r) \propto \langle F(\widehat{\boldsymbol{s}}_r)\rangle - \langle F(\nu)\rangle, \tag{12}$$

where ν denotes a standardised Gaussian variable and F is a general non-quadratic function that combines the high-order cumulants of $\boldsymbol{s}_r$ in order to approximate the 'true' neg-entropy $J_{\mathcal{N}}(\widehat{\boldsymbol{s}}_r)$. From equation 12, the vector $\boldsymbol{v}_r^t$ is optimised to maximise $J(\widehat{\boldsymbol{s}}_r)$ using an approximative Newton method. This finally leads to the following fixed point iteration scheme:

$$\widehat{\boldsymbol{v}}_r^t \leftarrow \left\langle \boldsymbol{x}F'(\widehat{\boldsymbol{s}}_r) - \langle F''(\widehat{\boldsymbol{s}}_r)\rangle \widehat{\boldsymbol{v}}_r \right\rangle, \tag{13}$$

where F' denotes the derivative of F. This is followed by a re- normalisation step such that $\widehat{\boldsymbol{v}}_r^t$ is of unit length. A proof of convergence and discussion about the choice of the non- linear function can be found in [23]. In order to estimate q sources, this estimation is simply performed q times under the constraint that the vectors $\boldsymbol{v}_l$ are mutually orthogonal. The constraint on the norm and the mutual orthogonality assure that these vectors actually form an orthogonal rotation matrix $\boldsymbol{Q}$. Thus, estimation of the sources is carried out under the assumption that all marginal distributions of $\widehat{\boldsymbol{s}}$ have maximally non-Gaussian distribution.

The choice of the nonlinear function is domain specific and in our case will be strongly linked to the inferential steps that are being performed after IC estimation (see section 4 below).

3.3 Incorporation of prior knowledge

Within the framework of the standard GLM, spatial and temporal information like the assumed spatial smoothness of the areas of activation or temporal autocorrelation is incorporated into the modelling process by temporal and/or spatial filtering of the data prior to model fitting, e.g. the temporal characteristic of the hæmodynamic response is commonly encoded via the assumed and normally fixed convolution kernel.

The spatial and temporal filtering steps can also be used for data pre-processing for ICA. In the case of spatial smoothing note that since the inferential steps (see section 4 below) are not based on Gaussian Random Field theory [52], we have the additional freedom of choosing more sophisticated smoothing techniques that do not simply convolve the data using a Gaussian kernel. Non-linear smoothing like the SUSAN filter [47] allow for the reduction of noise whilst preserving the underlying spatial structure and as a consequence reduce the commonly observed effect of estimated spatial pattern of activation 'bleeding' into non-plausible anatomical structure like CSF or white matter.

In the temporal domain, temporal highpass filtering is of importance since in FMRI low frequency drifts are commonly observed which can significantly contribute to the overall variance of an individual voxels' time course. If these temporal drifts are not removed, they will be reflected in the low-frequency part of the eigenvectors of the covariance matrix of the observations $\boldsymbol{R_x}$ and increase the estimate for the rank of $\boldsymbol{A}$. If the spatial variation between voxels' time courses is low, these areas of variability can be estimated as a separate source, e.g. B_0 signal field inhomogeneities. If, however, the low frequency variations are substantially different between voxels, these effects ought to be removed prior to the analysis. For the experiments presented in this paper, we used linear highpass temporal filtering via Gaussian-weighted least squares straight line fitting [34].

In addition to these data pre-processing steps note that the estimates for the mixing matrix and the sources (equation 5) involve the estimate of the eigenvectors U and the eigenspectrum $\boldsymbol{\Lambda}$ of the data covariance matrix $\boldsymbol{R_x} = \sum_i w_i(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^t$, where w_i is the contribution of voxel i's time course to the covariance matrix. Typically, $w_i = \frac{1}{N} \quad \forall i$. In the case where prior information on the importance of individual voxels is available, we can simple encode this by choosing w_i appropriately. As an example consider the case where we have results from an image segmentation into tissue types available: if $\boldsymbol{p}$ is a vector where the individual entries p_i denote the estimated probability of voxel i being within gray-matter we can choose $w_i = p_i$ and the covariance is weighted by the probability of gray-matter membership. Simple approaches to performing ICA on the cortical surface (e.g. [20]) are special cases of this, binarising $\boldsymbol{p}$ and therefore losing valuable partial volume information. In this more general setting, however, the uncertainty in the segmentation will also be incorporated.

In order to incorporate more complex spatial information note that we can rewrite $\boldsymbol{R_x}$ in the following form:

$$\begin{aligned} \boldsymbol{R_x} &= \underbrace{\frac{1}{2}\sum_{ij} w_i w_j m_{ij}(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^t}_{R_w} \\ &\quad + \underbrace{\frac{1}{2}\sum_{ij} w_i w_j (1 - m_{ij})(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^t}_{R_b}, \end{aligned} \tag{14}$$

i.e. the canonical covariance matrix can be split into within-group and between-group covariance terms. The matrix $\boldsymbol{M} = (m_{ij}); m_{ij} \in [0, 1]$ defines a weighted graph of N nodes and can encode any possible association between any pair of voxels that we want to introduce into the estimation. We can restrict calculation to the first term in equation 14 and perform the eigenvalue analysis on only the part of the covariance matrix generated by voxel pairs we believe to be associated with each other. In its general form the matrix $\boldsymbol{M}$ has N^2 entries which for typical FMRI data sets requires vast amounts of memory. Often, however, the matrix $\boldsymbol{M}$ can be sparse only having $\mathcal{O}(N)$ non-zero entries whilst

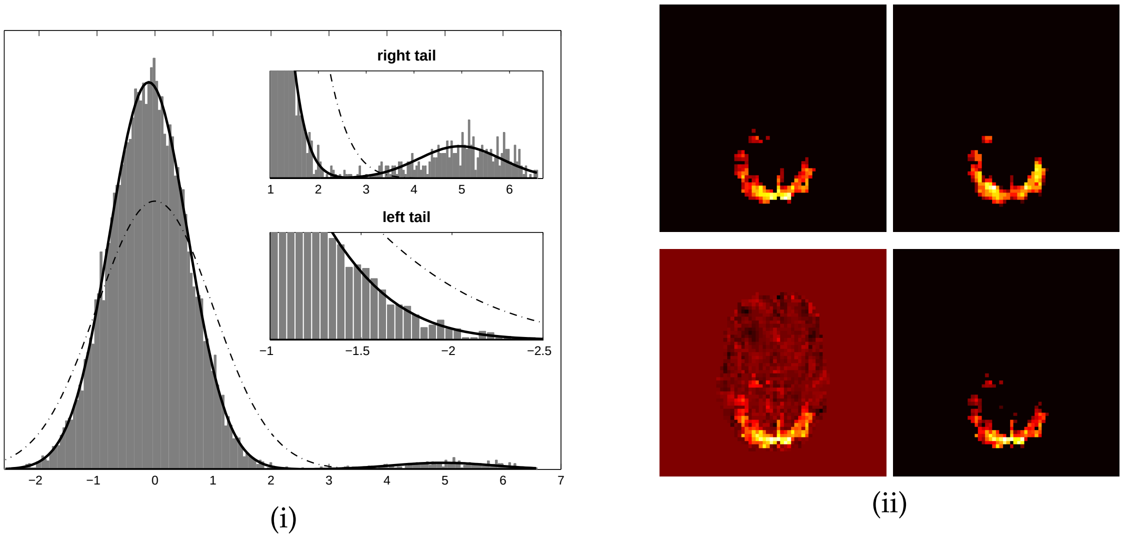


Figure 3: Gaussian Mixture Model for detecting activation: (i) histogram of the intensity values within a Z map together with the fit of a GMM with two Gaussians and the ML fit of a single Gaussian (dash-dotted). The single Gaussian fits poorly to the histogram of intensity values so transformation into spatial Z-scores and subsequent thresholding leads to meaningless threshold levels; (ii) top row: 'true' activation mask (left) and GLM results (right). The spatial smoothing required by Gaussian Random Field Theory causes the peak activation focus to be displaced; bottom row: IC map before and after thresholding using the Gaussian mixture model.

still encoding a variety of spatial models, e.g. we can constrain the calculation to voxel pairs within a certain neighbourhood of fixed size.

In addition to spatial information, assumptions on the nature of the time courses can be incorporated using regularized principal component analysis techniques [40]. Instead of filtering the data, constraints can be imposed on the eigenvectors, e.g. constraints on the smoothness can be included by penalizing the roughness using the integrated square of the second derivative. Alternatively it is possible to penalize the diffusion in frequency space, i.e. impose the constraint that the eigenvectors have a sparse frequency representation.

4 Inference

After estimating the mixing-matrix $\widehat{\boldsymbol{A}}$, the source estimates are calculated according to equation 6 by projecting each voxel's time course onto the time courses contained in the columns of the unmixing matrix $\widehat{\boldsymbol{W}} = (\widehat{\boldsymbol{A}}^t\widehat{\boldsymbol{A}})^{-1}\widehat{\boldsymbol{A}}^t$.

[36] suggest transforming the spatial maps to Z-scores (transform the spatial maps to have zero mean and unit variance) and thresholding at some level (e.g, $|Z| > 2.0$). The spatial maps, however, are the result of an ICA decomposition where the estimation optimises for non-Gaussianity of the distribution of spatial intensities. This is explicit in the case of the fixed-point iteration algorithm employed here, but also true for the Infomax or similar algorithms where the optimisation for non-Gaussian sources is implicit in the choice of nonlinearity. As a consequence, the spatial intensity histogram of an individual IC map is not Gaussian and a simple transformation to voxel-wise Z-scores and subsequent thresholding will necessarily result in an arbitrary and uncontrolled false-positive rate: the estimated mean and variance will not relate to an underlying null-distribution. Figure 1 shows an example where the estimated Gaussian (dash-dotted line) neither represents the 'background noise' Gaussian nor the entire image histogram, and any threshold value based on the expected number of false-positives becomes meaningless with respect to the spatial map.

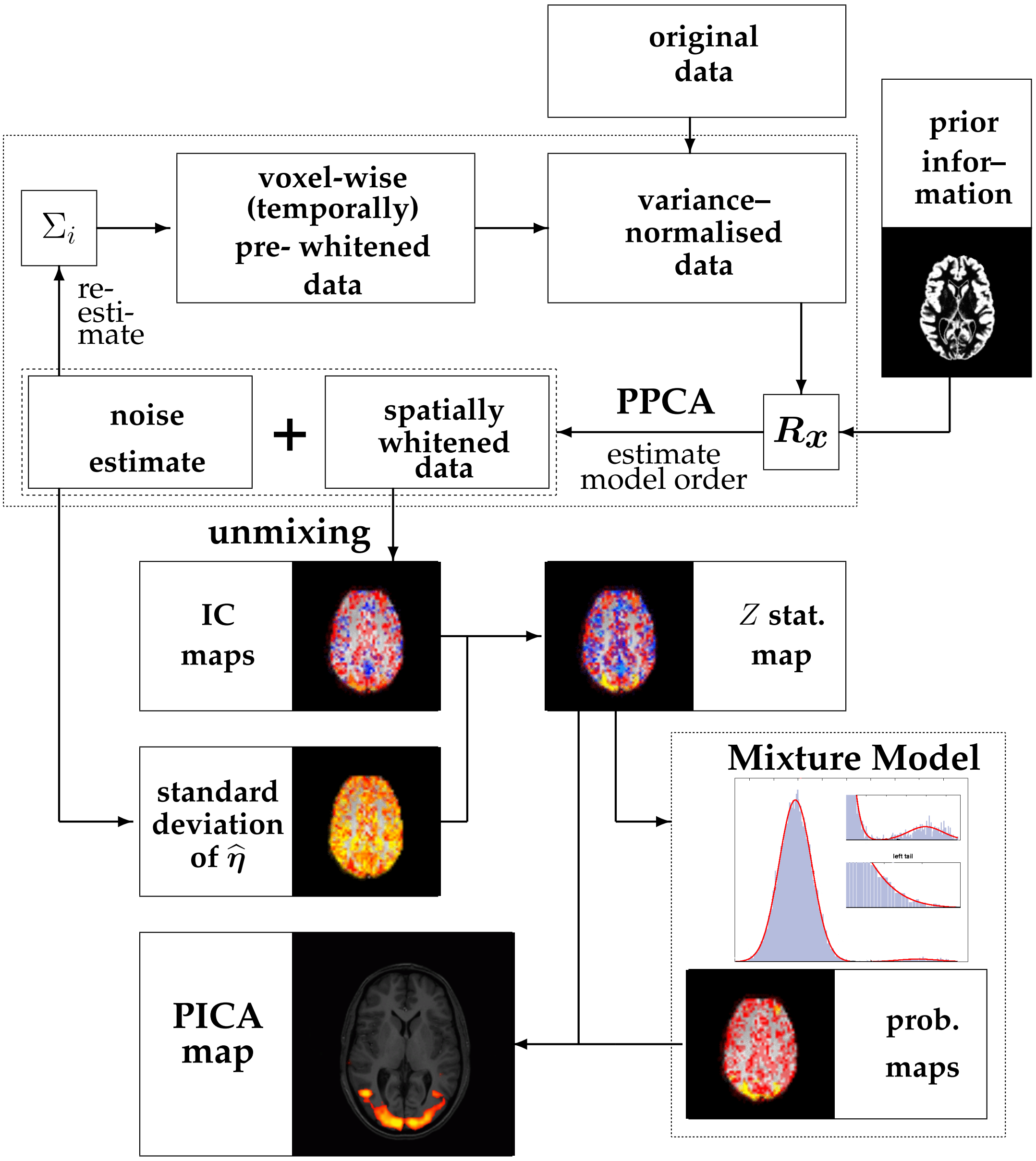


Figure 4: Schematic illustration of the analysis steps involved in estimating the PICA model.

Instead, consider the estimated residual noise at a single voxel location i:

$$\widehat{\boldsymbol{\eta}}_i = \boldsymbol{P}\boldsymbol{x}_i,$$

where $\boldsymbol{P} = \boldsymbol{I} - \widehat{\boldsymbol{W}}^t\widehat{\boldsymbol{W}}$ is the residual generating projection matrix. In the case where the model order q was estimated correctly, the columns of the estimated mixing matrix $\widehat{\boldsymbol{A}}$ will span the entire signal space, i.e. $\mathrm{span}(\widehat{\boldsymbol{A}}) \supset \mathrm{span}(\boldsymbol{A})$ so that $\boldsymbol{PA} = 0$. Therefore

$$\widehat{\boldsymbol{\eta}}_i = \boldsymbol{P}\boldsymbol{x} = \boldsymbol{PAs} + \boldsymbol{P}\boldsymbol{\eta} = \boldsymbol{P}\boldsymbol{\eta},$$

i.e. the estimated noise is a linear projection of the true noise and is unconfounded by residual signal. The estimate of the noise variance σ_i^2 at each voxel location is

$$\widehat{\sigma}_i^2 = \widehat{\boldsymbol{\eta}}_i^t\widehat{\boldsymbol{\eta}}_i/\mathrm{trace}(\boldsymbol{P}),$$

which, if $p - q$ is reasonably large, will approximately equal σ_i^2, i.e. equal the true variance of the noise [25]. We can thus convert the individual spatial IC maps $\boldsymbol{s}_{r\cdot}$ into Z-statistic maps $\boldsymbol{z}_{r\cdot}$ by dividing the raw IC estimate by the estimate of the voxel-wise noise standard deviation.

Under the null-hypothesis of no signal and after variance-normalisation, the estimated sources are just random regression coefficients which, after this transformation, will have a clearly defined and spatially stationary voxel-wise false-positive rate at any given threshold level.[3] While, for reasons outlined above, the null-hypothesis test is generally not appropriate, the voxel-wise normalisation also has important implication under the alternative hypothesis; it normalises what has been estimated as effect (the raw IC maps) relative to what has been estimated as noise and thus makes different voxel locations comparable in terms of their signal-to noise characteristics for a now given basis (the estimated mixing matrix). This is important since the mixing matrix itself is data- driven. As such, the estimated mixing matrix will give a better temporal representation at different voxel locations than at others and this change in 'specificity' is reflected in the relative value of residual noise.

In order to assess the Z-maps for significantly activated voxels, we follow [18] and [22] and employ mixture modelling of the probability density for spatial map of Z-scores.

Equation 6 implies that

$$\widehat{\boldsymbol{s}}_i = \widehat{\boldsymbol{W}}\boldsymbol{A}\boldsymbol{s}_i + \widehat{\boldsymbol{W}}\boldsymbol{\eta}_i,$$

i.e. in the signal space defined by the mixing matrix $\boldsymbol{A}$, the additional noise term in equation 2 manifests itself as an additive Gaussian noise term. The same is true after transformation of the intensity values to Z-scores. We therefore model the distribution of the spatial intensity values of the rth Z-map $\boldsymbol{z}_{r\cdot}$ by K mixtures of 1-dimensional Gaussian distributions [10]

$$p(\boldsymbol{z}_{r\cdot}|\boldsymbol{\theta}_K) = \sum_{l=1}^{K} \pi_{r,l}\mathcal{N}_{z_r}[\mu_{r,l}, \sigma_{r,l}^2], \tag{15}$$

where $\boldsymbol{\theta}_K$ denotes the vector of all parameters $\boldsymbol{\theta}_K = \{\boldsymbol{\pi}_K, \boldsymbol{\mu}_K, \boldsymbol{\sigma}_K\}$ and $\boldsymbol{\pi}_K, \boldsymbol{\mu}_K$ and $\boldsymbol{\sigma}_K$ are the vectors of the K mixture coefficients, means and variances. Voxels that are not influenced by a specific time course in $\widehat{\boldsymbol{A}}$ will simply have a random regression coefficient and will be Gaussian distributed. The distribution of intensity values for areas that are influenced by the associated time course, however, can be arbitrary and we will use the fact that the Gaussian mixture model of equation 15 is universal in that any source probability density can be approximated by a sufficient number of mixtures [10]. As an alternative to this approach, [22] fits a mixture of one Gaussian and two Gamma distributions to model the probability density of background noise, positive and negative BOLD effects. The model of equation 15 is fitted using the expectation-maximization (EM) algorithm [17]. In order to infer the appropriate number of components in the mixture model we successively fit models with an increasing number of mixtures and use an approximation to the Bayesian model evidence to define a stopping rule (see [44] for details). Our experiments suggest that this typically results in a model with 2-3 mixtures.

In cases where the number of 'active' voxels is small, however, a single Gaussian mixture may actually have the highest model evidence, simply due to the fact that the model evidence is only approximated in the current approach. In this case, however, a transformation to spatial Z-scores and subsequent thresholding is appropriate, i.e. reverting to null hypothesis testing instead of the otherwise preferable alternative hypothesis testing.

If the mixture model contains more than a single Gaussian, we can calculate the probability of any intensity value being background noise by evaluating the probability density function of the single Gaussian that models the density of background noise. Conversely, we can evaluate the set of additional Gaussians and calculate the probability under the alternative hypothesis of 'activation'[4]

[3] Strictly speaking these will be T- distributed, not normal.

[4] where in this case 'activation' is to be understood as 'cannot be explained as random correlation coefficient to the associated time course'

with respect to the associated time course, i.e. we obtain the estimate of the posterior probability for activation of voxel i in the Z-score map r as [18]:

$$Pr(\text{activation}|\boldsymbol{z}_{r,i}) = \frac{\sum_{l=2}^{K} \widehat{\pi}_{r,l}\mathcal{N}_{z_r}[\widehat{\mu}_{r,l}, \widehat{\sigma}^2_{r,l}]}{p(\boldsymbol{z}_{r\cdot}|\widehat{\boldsymbol{\theta}}_K)},$$

where without loss of generality we assume that the first term in the mixture models the background noise. Identification of the Gaussian that models the background is straightforward since it typically coincides with the dominant mode of the intensity histogram.

Figure 3 illustrates the process for a spatial map extracted from a data set with artificial activation introduced into FMRI resting data (see section 5 for details). Voxels with an estimated posterior probability of activation exceeding a certain threshold value are labeled active. The threshold level, though arbitrary, directly relates to the loss function we like to associate with the estimation process, e.g. a threshold level of 0.5 places an equal loss on false positives and false negatives [22]. Alternatively, because we have explicitly modelled the probabilities under the null and alternative hypothesis, we can choose a threshold level based on the desired false positive rate over the entire brain or at the cluster level simply by evaluating the probabilities under the null and alternative hypotheses.

4.1 Illustration

The individual steps that constitute the Probabilistic Independent Component Analysis are illustrated in figure 4. The de-meaned original data are first temporally pre-whitened using knowledge about the noise covariance $\boldsymbol{\Sigma}_i$ at each voxel location. The covariance of the data is calculated from the data after normalization of the voxel-wise standard deviation. In the case where spatial information is available, this is encoded in the estimation of the sample covariance matrix $\boldsymbol{R_x}$. This is used as part of the probabilistic PCA steps to infer upon the unknown number of sources contained in the data, which will provide us with an estimate of the noise and a set of spatially whitened observations. We can re-estimate $\boldsymbol{\Sigma}_i$ from the residuals and iterate the entire cycle. In practice, the output results do not suggest a strong dependency on the form of $\boldsymbol{\Sigma}$ and preliminary results suggest that it is sufficient to iterate these steps only once. From the spatially whitened observations, the individual component maps are estimated using the fixed point iteration scheme (equation 13). These maps are separately transformed to Z scores using the estimated standard deviation of the noise. In contrast to raw IC estimates, the Z score maps depend on the amount of variability explained by the entire decomposition at each voxel location. Finally, Gaussian Mixture Models are fitted to the individual Z maps in order to infer voxel locations that are significantly modulated by the associated time course in order to allow for meaningful thresholding of the Z images.

5 Evaluation data

We illustrate the method above on a set of artificial and real FMRI data under resting condition and visual / audio-visual stimulation. The different artificial data sets were generated so as to cover the spectrum from data that conforms to the modelling assumptions of equation 2 to real FMRI data.

5.1 FMRI data

We acquired whole brain volumes (64× 64× 21; 4× 4× 6 mm, $N = 18470$ inter-cranial voxels) of FMRI data on a Varian 3T system (TR=3sec; TE=30ms) under (i) resting condition and under different experimental stimuli (180 volumes each): (ii) 30s on/off visual stimulus (black and white checkerboard reversing at 8Hz), (iii) 30s on/off visual stimulus (coloured checkerboard reversing at 8Hz) and 45s on/off auditory stimulus (radio recording). The data were corrected for subject motion using MCFLIRT [24] to perform 6 parameter rigid-body motion correction. The corrected data was temporally high pass filtered (Gaussian-weighted LSF straight line subtraction, with $\sigma = 75.0s$ [34]) and masked of non- brain voxels using BET [46].

5.2 Artificial activation in real FMRI resting state data

The activation data set (iii) was analysed using standard GLM techniques as implemented in FEAT [46]. Final Z statistic maps were used to define activation 'masks' by thresholding at $Z > 3.0$ and clustering with $p < 0.01$, subtracting this threshold and re-scaling so that the final mask range was within the range $[0, 1]$, where any value $0 < value \leq 1$ signifies 'level' of activation. These masks were transformed into the space of the resting data set using FLIRT [24].

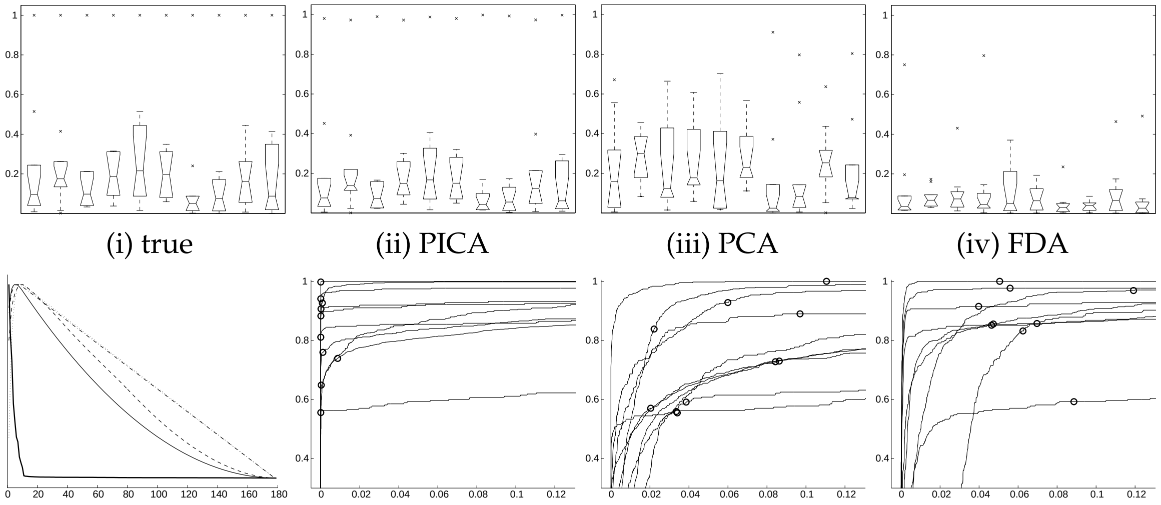
Next, activation was linearly added into the resting data (i) using artificial timecourses, modulated spatially by the activation masks described above. The timecourses were created by taking simple box-car designs (matching the paradigm of the activation data (iii) described above) and convolving with a standard gamma-based HRF kernel function (std.dev.=3s, mean lag=6s). Various overall levels of activation were added to create various test data sets, with the maximum resulting activation signal of 0.5%,1%, 3% and 5% times the mean baseline signal intensity. The average activation level within the clusters was ~ 0.25 of the peak activation level. In the real activation data, the highest activation was $\sim$ 3% peak to peak. Note that this is more realistic than the artificial data presented in [31] where all activated voxels have identical Z scores. The above procedure was carried out for auditory and visual 'activation' using a separate spatial activation mask and activation timecourses.

5.3 Artificial signal in synthetic noise

In a similar way to that outlined above, we added various source signals into Gaussian and autoregressive noise. The background noise parameters (i.e. voxel wise mean and std. deviation in the case of Gaussian noise and AR parameters in the case of autoregressive noise) were estimated from the resting state FMRI data. We then added 10 spatial maps and associated time courses taken from ICA decompositions of various other true FMRI data sets. The sources were chosen to represent different source processes that commonly are identified in real FMRI data, e.g. high frequency noise within the ventricular system, fluctuations in the B_0 field homogeneity spatially located near tissue-air boundaries, activation maps etc. This should not bias the results of the comparison in favor of PICA given that these spatial maps originated from different data sets and as such are not mutually spatially independent. Similarly, the associated time courses are not uncorrelated. This, we belive, is a more faithful representation of FMRI data. In FMRI it is often possible (and almost always advisable) to create experiments under an orthogonal experimental design which hopefully renders the temporal responses to external stimulation to be mutually orthogonal. Any additional source processes, however, can have arbitrary correlation with any column of the design. We therefore did not impose any constraints on the associated time courses.

6 Results

For the artificial data sets where ground truth is available, we follow [31] and report the quality of source identification over a range of possible threshold levels in the form of ROC curves, i.e. as a plot of the false positive rate versus the true positive rate at different threshold levels. We report the temporal accuracy as the (normalised) correlation between the estimated time course and the 'true' timecourse after projection into the same signal space as defined by the PPCA decomposition, i.e. correlation between the estimated time course $\widehat{\boldsymbol{a}}_{j'}$ and $\boldsymbol{U}_q\boldsymbol{U}_q^t\boldsymbol{a}_j$, where j' and j index the corresponding columns in the estimated mixing matrix $\widehat{\boldsymbol{A}}$ and the true mixing matrix $\boldsymbol{A}$ and where $\boldsymbol{U}_q$ is the matrix of the q major eigenvectors of the data covariance. By calculating the temporal correlation



(i) true (ii) PICA (iii) PCA (iv) FDA

(v) dim.-estimate (vi) PICA (vii) PCA (viii) FDA

Figure 5: Spatio-temporal accuracy of different exploratory data analysis techniques on artificial FMRI data (10 artificial signals embedded in 180-dimensional Gaussian noise as described in section 5): (i) correlation structure of the time courses of the 10 sources; (ii)-(iv) correlation structure of time courses estimated using PICA, PCA and FDA respectively; (v) eigenspectrum of the data together with different estimators for the number of sources at ; (vi)-(viii) ROC curves: false-positives rate (0–0.12) vs. true- positives rate (0.4–1) for the estimated spatial maps, markers indicate the results for the canonical threshold level 0.5.

with respect to projected rather than the 'true' original time courses we ensure that the measure of temporal accuracy is unconfounded by the dimensionality reduction itself.

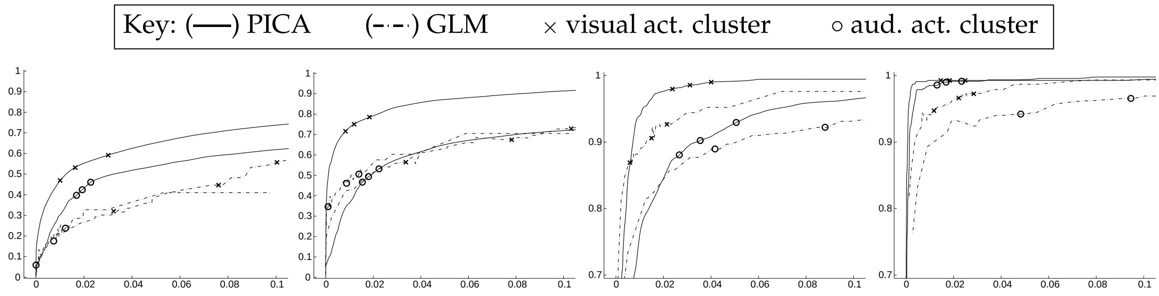
6.1 Artificial signal in synthetic noise

This data set approximately conforms to the model assumptions from equation 2 and is used here to illustrate the difference between a PICA decomposition and results from standard PCA and regularised PCA, or *functional data analysis*. FDA was carried out as described in [40] using a set of 60 B-spline basis functions.

Figure 5 summarises the spatio-temporal accuracy of the decompositions for all three techniques. For both FDA and PCA, the first 10 estimated sources were chosen for the comparison; in the case of PICA the model selection correctly identifies the number true number of hidden sources (figure 5), so only 10 source signals are estimated. The top row shows boxplots of the cross-correlation between the 10 true and the 10 estimated time courses (i.e. the first boxplot summarises the temporal cross-correlation between source nr. 1 and all other time courses) while the bottom row shows ROC curves for each of the 10 associated spatial maps. For both PCA and FDA, the estimated time courses differ substantially from the 'true' time courses. While in almost all cases both techniques estimate at least one time course with a significant correlation, the overall correlation structure is not preserved (figure 5 (i) compared to (iii) and (iv)) . This is a simple consequence of the fact that both PCA and FDA estimate orthogonal sets of time courses. Note that FDA appears to perform worse in terms of the estimation of the time courses but outperforms PCA in the spatial domain. In the case of PICA, the underlying sources are much better identified, both in the temporal and the spatial domain: the correlation structure of the estimated time courses is close to that of the true source signals. At the same time the PICA decomposition results in an improved ROC characteristics with highest true-positive rates at any false- positives level. In almost all cases the canonical threshold level of 0.5 results in 0 false positives. In this case, the difference between PICA and PCA is purely

	0.5%	1%	3%	5%
vis.	0.33 ± 0.03	0.62 ± 0.01	0.9 ± 0	0.95 ± 0
aud.	0.29 ± 0.01	0.5 ± 0.01	0.87 ± 0	0.94 ± 0

Table 1: Temporal accuracy at different activation levels: correlation between the extracted time courses and the true signal time courses over 150 runs.



0.5% act. level 1% act. level 3% act. level 5% act. level

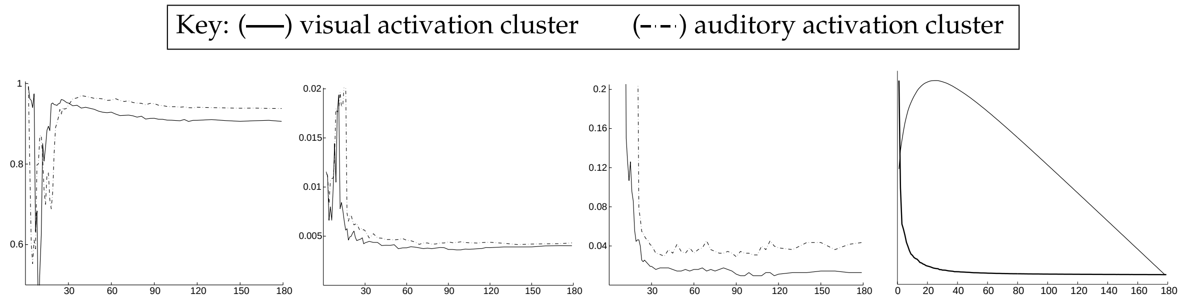
Figure 6: Spatial accuracy at different activation levels: ROC curves for PICA (solid lines - mean over 150 runs) vs. ROC curves for Z statistical maps thresholded using Gaussian Random Field theory at different Z and p levels. Markers indicate typical threshold levels: 0.33,0.5 and 0.66 (PICA alternative hypothesis test); $Z > 1.6, 2.3$ and 3.1 (GLM null-hypothesis test).

due to the additional orthogonal rotation matrix $\boldsymbol{Q}$. In the case of PICA the underlying sources are much better identified in that in all cases exactly one of the estimated time courses has a very high correlation with the true signal time course. Consequently, the covariance structure between different time courses is almost identical to the true covariance structure. It is interesting to note that, in all cases, the estimated time courses have a slightly smaller cross-covariance structure than the true time courses. This is an effect quite different to the assumptions that lead into the investigation of 'Spatiotemporal Independent Component Analysis' [48] . There, the authors speculated that in in the case of an ICA decomposition based on optimising spatial independence between estimated source signals, suboptimal solutions emerge since the decomposition will tend towards unplausible solutions in the 'dual' temporal domain in order to satisfy the independence in the spatial domain. In our experience, however, spatial and temporal accuracy appear to be strongly related. This is, in fact, what is to be expected given the uniqueness results presented earlier.

6.2 Spatio-temporal accuracy of PICA

We analysed the data set with artificial activation introduced into baseline FMRI data using PICA and GLM. The exact activation time courses were used within the GLM as regressors of interest. This will introduce a small bias in favour of the GLM analysis which we prefer compared to the alternative where we would have to artificially encode the more plausible ignorance that normally exists over the exact shape of the signal within the data. In order to estimate the consistency of the probabilistic ICA approach, the analysis was repeated 150 times in order to evaluate the repeatability between runs[5]. Table 1 summarises the mean correlation between the estimated and true time courses over all 150 runs while figure 6 shows the ROC curves for both GLM and PICA Naturally, the estimation accuracy improves with increased signal level. Note that the ROC curves for the GLM are not monotonically increasing. This is a direct consequence of the ambiguity built into the statistical thresholding steps

[5] The ICA decomposition begins with a random unmixing matrix and therefore does not necessarily give the same decomposition every time.



(i) temp. correlation (ii) false positive rate (iii) false negative rate (iv) dim. estimate

Figure 7: Spatio-temporal accuracy as a function of assumed dimensionality (i.e. when retaining different amounts of variance) for the simulated audio-visual activation data at the 3% level: (i) correlation between the extracted time course and the true signal time course; (ii) false positive rate; (iii) false negative rate (visual stim.: solid, aud. stim.: dashed); (iv) eigenspectrum of the data covariance matrix together with Laplace approximation to the model order.

based on Gaussian Random Field Theory[6], where a Z-threshold level is combined with a significance level for cluster heights or size. In the present case, we evaluated different sets with Z ranging from 1.1 to 7.0 and p ranging from 0.0005 to 0.1. For fixed Z or fixed p, a monotonically increasing ROC curve can be plotted but for reasons of simplicity, we ordered all results by increasing false positive rate and in the case of multiple true positive outcomes only used the best one. In almost all cases, the PICA estimates show an improved ROC characteristics compared to the GLM results despite the fact that GLM analysis was carried out under the ideal condition of perfect knowledge of the regressors of interest. This is due to the fact that a standard GLM analysis is adversely affected by the presence of un-modelled structured noise in this data. The PICA decomposition, on the other hand, estimates sufficiently strong structured noise as separate components resulting in increased spatial accuracy for the activation component. Note that in the case of the GLM a difference in the Z threshold even within a commonly-used range of values leads to substantially different quality of estimation.

6.3 Accuracy and dimensionality

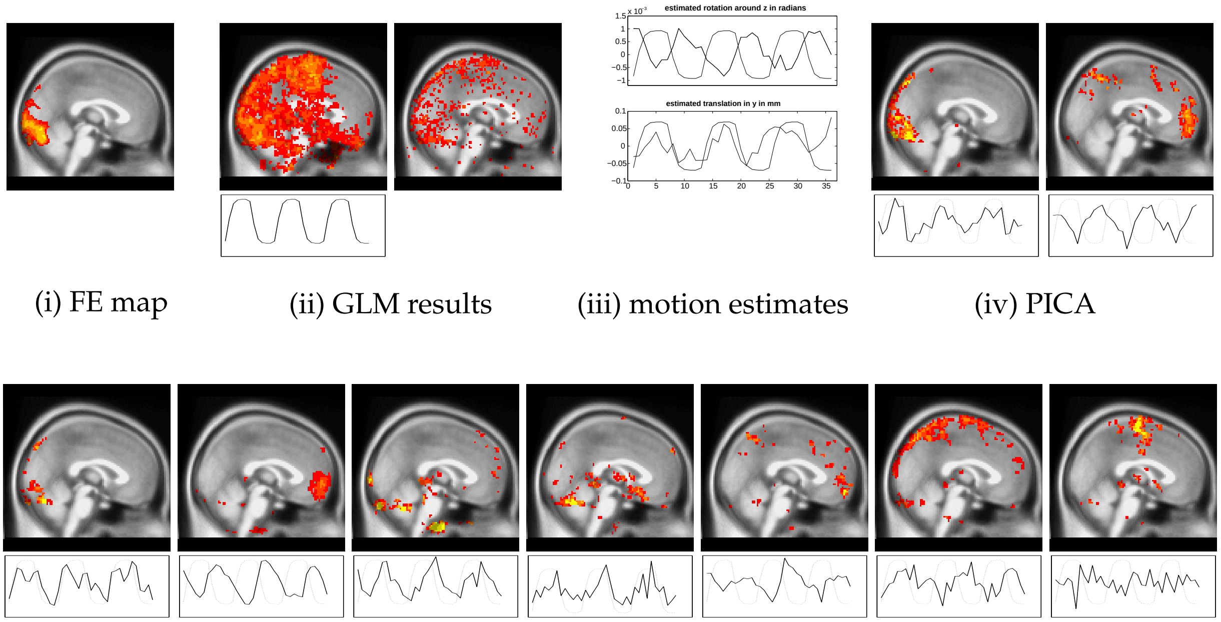
Within the estimation steps, the choice of number of components was determined from the estimate of the Bayesian evidence (equation 10). A different choice of q gives rise to a different model with different quality of estimation. Under the model, the optimal number of components should match the column rank of $\boldsymbol{A}$, where the number of components is restricted to the 'true' number of source processes in order to avoid arbitrary splits of identified sources into separate component maps ('overfitting').

Figure 2 has demonstrated that in the case of data that conforms to the model, the model order can be inferred accurately. In the case of real FMRI data, estimation of the number of source processes is a much more difficult task. In order to assess the dependency between the estimated number of source processes and the spatio-temporal accuracy of the estimation, we performed ROC analysis on spatial maps obtained after projecting the data into subspaces of increasing dimensionality[7].

Figure 7 shows the results of the temporal correlation and the final false-positive rates and false-negative rates over the range of possible dimensions for the data set with 3% peak level activation, where the spatial maps were thresholded at the 0.5 level. Both for the spatial and temporal accuracy

[6] We here compare PICA results against results obtained from the GLM with GRF-based inference. Cluster-based thresholding appears to be generally accepted as the method of choice in the case of reasonably sized and well-localised activation patterns like the ones used in this example.

[7] where the ROC analysis is performed on the spatial map with highest temporal correlation between the true and estimated time courses


estimated rotation around z in radians
x 10⁻³
1.5
1
0.5
0
−0.5
−1
estimated translation in y in mm
0.1
0.05
0
−0.05
−0.1
0
5
10
15
20
25
30
35

(i) FE map (ii) GLM results (iii) motion estimates (iv) PICA

(v) maps from classical ICA

Figure 8: Analysis of visual stimulation data: (i) map from a fixed effects analysis of the non-motion confounded 31 data sets for reference, (ii) FEAT Z-statistical maps ($Z > 3.0, p < 0.01$) obtained from GLM fitting of the motion-confounded data (left) and after the inclusion of estimated motion parameters as additional regressors of no interest (right), (iii) estimated motion parameters on this one data set that show a high absolute correlation with the stimulus, (iv) spatial maps from PICA performed in a space spanned by the 7 dominant eigenvectors, (v) set of spatial maps from a standard ICA analysis where the data was projected into a 29 (out of a possible 35) dimensional subspace of the data that retains $> 90\%$ of the overall variability in the data. For ICA and PICA all maps are shown where the associated time course has its peak power at the frequency of the stimulus.

these plots suggest that the quality of estimation does not improve once the source signals are being estimated in a subspace with more than about 30 dimensions. These results appear to be consistent for both artificial activation patterns and time courses. Reducing the number of sources below 30, however, will lead to increasingly poor estimates.

Overfitting would necessarily result in an increase of the false-negative rate, an effect that is shown in figure 7 (iii) for the auditory activation cluster. For this particular data set the effect is very subtle since the data has been generated without any voxel-wise variation of the temporal signal introduced into resting-state FMRI data. The artificial time courses are consistent within the clusters and therefore these clusters are less likely than in 'real' FMRI data to be incorrectly split into different spatial maps. Though the quality of estimation does not degenerate badly with increased dimensionality it is still essential to find a good estimate of the lower dimensional subspace. It not only dramatically decreases the computational load but more importantly provides better estimates for the noise, which is essential as part of the inferential steps. For this data set, the Laplace approximation to the evidence for model order (figure 7(iv)) appears to work well.

6.4 Real FMRI data

For the first example, we used data courtesy of Dr. Dave McGonigle that previously had been used to evaluate the between-session variability in FMRI [35]. In brief, the experiment involved 33 sessions of runs under motor, cognitive and visual stimulation. The data presented here is one of the two visual stimulation sessions of 36 volumes each that proved unacceptably difficult to analyse using a

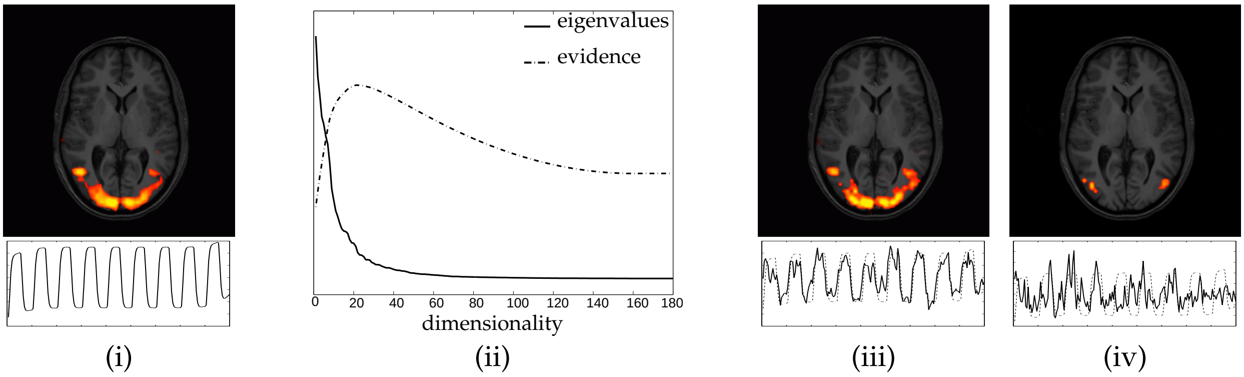


Figure 9: GLM vs. PICA on visual stimulation data: (i) FEAT results and regressor time course, (ii) Eigenspectrum of the data covariance matrix and estimate of the latent dimensionality using equation 10, (iii) & (iv) spatial maps and associated time courses of PICA results, all maps with $r > 0.3$ between the estimated and expected time course are shown.

model based approach and had therefore been excluded from the previous analysis due to obvious motion artifacts. It is used here to illustrate the advantages of model-free data analysis techniques in cases where the data does not conform to simple a priori hypotheses.

Figure 8(i) show the results from a fixed effects analysis over the 31 non-confounded data sets after each set was analysed separately using FEAT. It shows the general visual activation pattern that emerged from the analysis of sessions that were not heavily confounded by subject motion. In contrast, figure 8(ii) shows sagittal maximum intensity projections of Z score maps from a GLM regression of one of the confounded data sets against the expected response. There are large amounts of non-plausible and 'spurious' activation. These results were obtained after initial rigid-body motion correction using MCFLIRT. Visual inspection of the data after correction suggested that the algorithm was able to realign the volumes reasonably well with no 'noticable' misalignment of neighbouring volumes. The estimated motion parameters in figure 8(iii) suggest that the poor localisation of visual cortical areas in the Z maps is not due to high magnitude of motion but instead is a result of a strong correlation between certain motion parameters and the stimulus sequence (stimulus correlated motion). Within the GLM framework, the classical approach is to include the estimated motion parameters as nuisance regressors. In this case, however, the GLM results do not improve and still do not uniquely identify visual cortical areas (figure 8(ii), second map).

In the case of a PICA analysis of the motion confounded data set, only seven component maps remained after dimensionality reduction of which only 2 maps have an associated time course where the highest power is at the frequency of stimulus presentation (figure 8(iv)). The results from a probabilistic independent component analysis clearly improve upon the GLM results in that the first PICA map shows a clean and well localised area of activation within the visual cortex similar to the area identified by the fixed effects analysis while the second map has large values at the intensity boundraries of the original EPI data and has an associated time course with high correlation to the estimated rotation around the Z-axis (iii, top). In comparison, figure 8(v) shows the result of a standard ICA decomposition, where the data was projected onto the dominant 29 eigenvectors in order to retain $> 90\%$ of the variability in the data. Using the same criterion for the selection of maps as before, seven components emerge (here ordered with decreasing absolute correlation from left to right after thresholding by converting each intensity value into a Z score and only retaining voxels with $Z > 2.3$). It is difficult to assess the differences between figure 8(iv) and (v) with respect to estimated motion. For the visual activation, however, the comparison suggests that results from classical ICA do actually overfit the data in that different features that appear both in the PICA map and fixed effects map are distributed across different spatial maps.

As a second example, figure 9 shows the PICA results on the visual stimulation study used within

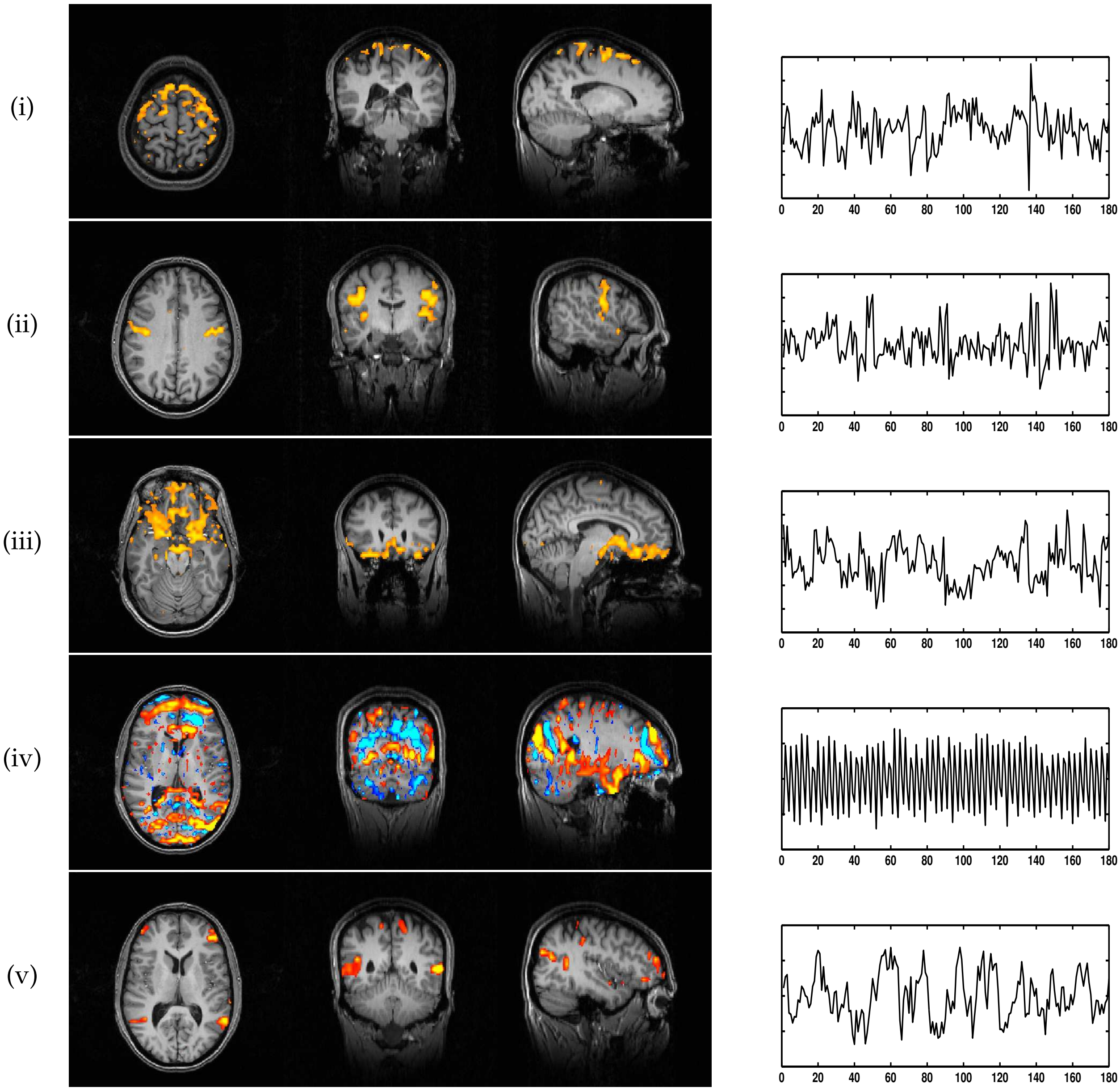


Figure 10: Additional PICA maps from the visual activation data: (i) head motion (translation in Z), (ii) sensory motor activation, (iii) signal fluctuations in areas close to the sinuses (possibly due to interaction of B_0 field inhomogeneity with head motion), (iv) high frequency MR 'ghost' and (v) 'resting- state' fluctuations/ physiological noise.

the introduction to illustrate the problem of overfitting. Based on the estimate of the model order, the data was projected onto the first 27 eigenvectors prior to the unmixing. Comparing figure 9(i) and (ii) we get a much better correspondence between the areas of activation estimated from the GLM approach and the main PICA estimate (compared to figure 1(ii)). This is reassuring, since simple visual experiments of this kind are known to activate large visual cortical areas which should be reliably identifiable over a whole range of analysis techniques. Within the set of IC maps a second source estimate has an associated time course that correlates with the assumed response at $r > 0.3$. This map depicts a bilateral pattern of activation within visual cortical areas, possibly V3/MT, areas known to be involved in the processing of visual motion. This is highly plausible given that under the stimulation condition the volunteer was presented with a checkerboard reversing at 8Hz. The associated time course is very similar to the time course associated with the spatial map (iv) in figure 1, but in the case of standard ICA, only a unilateral activation is identified. This is not attributable to the difference in the thresholding itself; the raw IC map in figure 1 does not allow for a bilateral activation pattern. Instead, it turns out to be direct consequence of the existence of a noise model: the standard deviation of the residual noise in the PICA decomposition is comparably small within these areas. After transforming the raw IC estimates $\boldsymbol{s}_i$ into Z-scores, the well localised areas emerge. In addition, figure 10 shows a selection of maps found during the same PICA decomposition on this data, depicting e.g. physiological 'noise', motion and scanner artefacts. Note that in both examples the PICA maps are actual Z statistical maps and as such are much easier to compare against output from a standard GLM analysis. Standard ICA maps, for reasons outlined above, are simply raw parameter estimates and as such purely descriptive.

7 Discussion

The Probabilistic Independent Component Analysis model presented in this paper is aimed at solving the problem of overfitting in classical ICA applied to FMRI data. This is approached by including a noise term in the classical ICA decomposition which renders the model identical to the standard GLM with the conceptual difference that the number and the shape of the regressors is estimated rather than pre-specified. As in the standard GLM, the noise is assumed to be additive and Gaussian distributed. Structured noise (e.g. physiological noise) is most likely to appear in the data as structured non-Gaussian noise, and as such is estimated as one (or more) of the underlying sources; this should not be confused with the Gaussian noise eliminated during the PPCA stage. If different source processes (e.g. activation and physiological noise components) are partially (temporally) correlated, they can still be separated from each other in the PICA unmixing as long as they are spatially distinct and not perfectly temporally correlated. If (as e.g. suggested by [30]) a noise component combines non-additively with a signal source, then indeed the linear mixing model used here will be imperfect. In this case, however, the nonlinear interaction should (to first order) appear as a third PICA component, in exactly the same way as modelling nonlinear interactions as a third explanatory variable in GLM modelling attempts to do.

Some of the methodological steps presented in section 2 build on ideas and techniques from standard parametric FMRI modelling; for example, the estimation of the voxel-wise covariance for pre-whitening is an extension of the technique presented in [50]. Also, the use of mixture models for inference has been motivated by work from [18] and [22], where mixture models were used for statistical maps generated from parametric FMRI activation modelling and links to the work on explicit source density modelling for ICA [3, 13, 45].

The proposed methodology can be extended in various ways. In the present implementation, we chose to discard an explicit source model from the estimation stages and use the Gaussian mixture model only after estimation is completed for the inferential steps. In a more integrated approach, the mixture model could be re-estimated after every iteration. This could then be used as an alternative to neg-entropy estimation in order to explicitly quantify the non-Gaussianity of source processes. [13] approximate full posterior distributions for all model parameters of a PICA model using the variational Bayesian framework. The technique is conceptually attractive, but suffers from a substantial increase in computational load and as such does not yet appear to be applicable to FMRI data. Also, our technique only encodes spatial neighbourhood information via the covariance of the observations that feeds into the PPCA step. In order to incorporate spatial information explicitly into the ICA estimation, a spatial Markov model can be used to represent the joint probability density of neighbouring samples.

8 Conclusion

We extended the classical Independent Component Analysis framework to allow for non-square mixing in the presence of Gaussian noise. By including a noise model, we are able to address a variety of important issues that exist in ICA applied to FMRI data. Most importantly, we address the issue of overfitting and can associate statistical significance to individual voxels within spatial maps that are modulated significantly by the associated time courses. Within the method, we take into account the very specific form of FMRI data: the general characteristics of autoregressive noise, possibly varying at different voxel locations, the necessity of voxel-wise variance normalisation and the fact that other image derived spatial information often is available and should be allowed to aid the estimation of signals of interest.

This model improves on the robustness and interpretability of IC maps as currently generated on FMRI data: experiments on artificial data suggest that the proposed methodology can accurately extract various sources of variability, not only from artificial noise that conforms to the model, but from artificial data generated from real FMRI noise.

The technique was illustrated on two examples of real FMRI data where the probabilistic independent component model is able to produce relevant patterns of activation that can neither be generated within the standard GLM nor standard ICA frameworks. We believe that PICA is a powerful technique complementary to existing methods that allows exploration of the complex structure of FMRI data in a statistically meaningful way.

The research described in this paper has been implemented as MELODIC (Multivariate Exploratory Linear Optimized Decomposition into Independent Components - a standalone C++ program). MELODIC is freely available as part of FSL (FMRIB's Software Library - `www.fmrib.ox.ac.uk/fsl`).

9 Acknowledgements

The authors thank Mark Jenkinson, Mark Woolrich, Jonathan Marchini, Didier Leibovici, Alison Noble and other members of the FMRIB Image Analysis Group for their input into this work and Dave McGonigle for providing parts of the data used within this paper. Support from the UK Medical Research Council and GSK is gratefully acknowledged.

References

[1] H. Akaike. Fitting autoregressive models for regression. *Annals of the Institute of Statistical Mathematics*, 21:243–247, 1969.

[2] S.I. Amari. Neural learning in structured parameter spaces – natural Riemannian gradient. In *Advances in Neural Information Processing Systems 9. Proceedings of the 1996 Conference*, pages 127–133, 1997.

[3] H. Attias. Independent Factor Analysis. *Neural Computation*, 11:803–851, 1999.

[4] F. Attneave. Some informational aspects of visual perception. *Psychological Review*, 61:183–193, 1954.

[5] Y. Baram and Z. Roth. Density shaping by neural networks with application to calssification, estimation and forecasting. Technical Report CIS-94-20, Center for Intelligent Syatems, Technion, Israel Institute for Technology, Haifa, Israel, 1994.

[6] D.J. Bartholomew. *Latent Variable Models and Factor Analysis*. Charles Griffin & Co. Ltd, London, 1987.

[7] C.F. Beckmann, J.A. Noble, and S.M. Smith. Investigating the intrinsic dimensionality of FMRI data for ICA. In *Seventh Int. Conf. on Functional Mapping of the Human Brain*, 2001.

[8] C.F. Beckmann, I. Tracey, J.A. Noble, and S.M. Smith. Combining ICA and GLM: A hybrid approach to FMRI analysis. In *Sixth Int. Conf. on Functional Mapping of the Human Brain*, page 643, 2000.

[9] A.J. Bell and T.J. Sejnowski. An information-maximisation approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

[10] C.M. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, 1995.

[11] E. Bullmore, M. Brammer, S.C.R. Williams, S. Rabe-Hesketh, N. Janot, A. David, J. Mellers, R. Howard, and P. Sham. Statistical methods of estimation and inference for functional MR image analysis. *Magnetic Resonance in Medicine*, 35(2):261–277, 1996.

[12] J.F. Cardoso. Source separation using higher order moments. In *Proc. ICASSP'89*, pages 2109–2112, 1989.

[13] R. Choudrey and S. Roberts. Flexible bayesian independent component analysis for blind source separation. In *Proceedings of ICA-2001*, 2001. To appear, see `http://www.robots.ox.ac.uk/~sjrob/pubs.html`.

[14] A. Cichocki, R. Unbehauen, L. Moszczynski, and E. Rummert. A new on-line adaptive algorithm for blind separation of source signals. In *Proc. Int. Symposium on Artificial Neural Networks ISANN'94*, pages 406–411, Tainan, Taiwan, 1994.

[15] P. Comon. Independent component analysis – a new concept? *Signal Processing*, 36:287–314, 1994.

[16] G. Deco and D. Obradovic. Linear redundancy reduction learning. *Neural Networks*, 8(5):751–755, 1995.

[17] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society series B*, 39:1–38, 1977.

[18] B.S Everitt and E.T. Bullmore. Mixture model mapping of brain activation in functional magnetic resonance images. *Human Brain Mapping*, 7:1–14, 1999.

[19] R.M. Everson and S.J. Roberts. Inferring the eigenvalues of covariance matrices from limited, noisy data. *IEEE Transactions on Signal Processing*, 48(7):2083–2091, 2000.

[20] E. Formisano, F. Esposito, F. Di Salle, and R Goebel. Cortes-based Independent Component Analysis of fMRI time-series. In *Seventh Int. Conf. on Functional Mapping of the Human Brain*, 2001.

[21] L.K. Hansen, J. Larsen, F. A. Nielsen, S.C. Strother, E. Rostrup, R. Savoy, N. Lange, J. Sidtis, C. Svarer, and O.B. Paulson. Generalizable Patterns in Neuroimaging: How Many Principal Components. *NeuroImage*, 9:534–544, 1999.

[22] N.V. Hartvig and J. Jensen. Spatial mixture modelling of fmri data. *Human Brain Mapping*, 11(4):233–248, 2000.

[23] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley, 2001.

[24] M. Jenkinson, P. Bannister, and S.M. Smith. Improved optimisation for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, 2002. in publication.

[25] P. Jezzard, P.M. Matthews, and S.M. Smith, editors. *Functional MRI: An Introduction to Methods*. OUP, Oxford, 2001.

[26] I.M. Johnstone. On the distribution of the largest principal component. Technical report, Department of Statistics, Stanford University, 2000.

[27] C. Jutten and J. Herault. Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24:1–10, 1991.

[28] J. Karhunen and J Joutsensalo. Generalizations of principal component analysis, optimization problems, and neural networks. *Neural Networks*, 8(4):549–562, 1995.

[29] R.E. Kass and A.E. Raftery. Bayes factors and model uncertainty. Technical Report 254, University of Washington, 1993.

[30] G. Kruger and G.H. Glover. The physiological noise in oxygen-sensitive magnetic resonance imaging. *Magnetic Resonance in Medicine*, 46:631–637, 2001.

[31] N. Lange, S.C. Strother, J.R. Anderson, F.A. Nielsen, A.P. Holmes, T. Kolenda, R. Savoy, and L.K. Hansen. Plurality and resemblance in fMRI data analysis. *NeuroImage*, 10:282–303, 1999.

[32] R. Linsker. Self-organization in a perceptual network. *Computer*, 21:105–117, 1988.

[33] D.J.C. MacKay. Maximum likelihood and covariant algorithms for independent component analysis. in preparation, 1996.

[34] J.L. Marchini and B.D. Ripley. A new statistical approach to detecting significant activation in functional MRI. *NeuroImage*, 12(4):366–380, 2000.

[35] D.J. McGonigle, A.M. Howseman, B.S. Athwal, K. J. Friston, R.S.J. Frackowiak, and A.P. Holmes. Variability in fMRI: An examination of intersession differences. *NeuroImage*, 11:708–734, 2000.

[36] M. J. McKeown, S. Makeig, G. G. Brown, T. P. Jung, S. S. Kindermann, A. J. Bell, and T. J. Sejnowski. Analysis of fMRI data by blind separation into independent spatial components. *Human Brain Mapping*, 6(3):160–88, 1998.

[37] T.P Minka. Automatic choice of dimensionality for PCA. Technical Report 514, MIT, 2000.

[38] W. Penny, S. Roberts, and R Everson. ICA: Model order selection and dynamic source models. In S. Roberts and R. Everson, editors, *Independent Component Analysis: Principle and Practice*, chapter 12. CUP, 2001.

[39] J.-B. Poline, K.J. Worsley, A.C. Evans, and K. Friston. Combining spatial extent and peak intensity to test for activations in functional imaging. *NeuroImage*, 5:83–96, 1997.

[40] J.O. Ramsay and B.W. Silverman. *Functional Data Analysis*. Springer, 1997.

[41] C. R. Rao. Characterization of the distribution of random variables in linear structural relations. *Synkhya*, Ser. A(28):252–260, 1966.

[42] C. R. Rao. A decomposition theorem for vector variables with a linear structure. *The Annals of Mathematical Statistics*, 40(5):1845–1849, 1969.

[43] J. Rissanen. Modelling by shortest data description. *Automatica*, 14:465–471, 1978.

[44] S.J Roberts, D. Husmeier, I. Rezek, and W. Penny. Bayesian approaches to gaussian mixture modelling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1133–1142, 1998.

[45] D. B. Rowe. Bayesian Source Separation for Reference Function Determination in fMRI. *Magnetic Resonance in Medicine*, (46):374–378, 2001.

[46] S. Smith, P. Bannister, C. Beckmann, M. Brady, S. Clare, D. Flitney, P. Hansen, M. Jenkinson, D. Leibovici, B. Ripley, M Woolrich, and Y. Zhang. FSL: New tools for functional and structural brain image analysis. In *Seventh Int. Conf. on Functional Mapping of the Human Brain*, 2001.

[47] S.M. Smith and J.M. Brady. SUSAN - a new approach to low level image processing. *International Journal of Computer Vision*, 23(1):45–78, May 1997.

[48] J.V. Stone, J. Porrill, N.R. Porter, and I.D. Wilkinson. Spatiotemporal independent component analysis of event-related fMRI data using skewed probability density functions. *NeuroImage*, 15:407–421, 2002.

[49] M.E. Tipping and C.M. Bishop. Mixtures of probabilistic principal component analyzers. *Neural Computation*, 11(2):443–482, 1999.

[50] M.W. Woolrich, B.D. Ripley, J.M. Brady, and S.M. Smith. Temporal autocorrelation in univariate linear modelling of FMRI data. *NeuroImage*, 14(6):1370–1386, 2001.

[51] K.J. Worsley and K.J. Friston. Analysis of fMRI time series revisited - again. *NeuroImage*, 2:173–181, 1995.

[52] K.J. Worsley, S. Marrett, P. Neelin, A.C. Vandal, K.J. Friston, and A.C. Evans. A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping*, 4:58–73, 1996.