PROBABILISTIC ICA FOR FMRI - NOISE AND INFERENCE

Christian F. Beckmann[†‡] *and Stephen M. Smith*[†]

[†]Oxford Centre for Functional Magnetic Resonance Imaging of the Brain (FMRIB)
[‡]Medical Vision Laboratory, Department of Engineering, University of Oxford
{beckmann,steve}@fmrib.ox.ac.uk

ABSTRACT

Independent Component Analysis is becoming a popular exploratory method for analysing complex data such as that from FMRI experiments. The application of such 'model-free' methods, however, has been somewhat restricted both by the view that results can be uninterpretable and by the lack of ability to quantify statistical significance. We present an integrated approach to Probabilistic ICA for FMRI data that allows for non-square mixing in the presence of Gaussian noise. We employ an 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. Reduction of the data to this 'true' subspace before the ICA decomposition automatically results in an estimate of the noise, leading to the ability to assign significance to voxels in ICA spatial maps. By this we not only are able to carry out probabilistic modelling, but also reduce problems of interpretation and overfitting. We use an alternative-hypothesis testing approach for inference based on Gaussian + Gamma mixture models. The performance of our approach is illustrated and evaluated on real and complex artificial FMRI data, and compared to the spatio-temporal accuracy of results obtained from standard ICA and standard analyses in the 'General Linear Model' (GLM) framework.

1. PROBABILISTIC ICA MODEL

Similar to the square noise-free case, the probabilistic ICA model is formulated as a generative linear latent variables model: 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 $\boldsymbol{\eta}(t)$:

$$\boldsymbol{x}_i = \boldsymbol{A}\boldsymbol{s}_i + \boldsymbol{\eta}_i \qquad \forall i \in \mathcal{V}. \tag{1}$$

Support from the UK Medical Research Council is gratefully acknowledged.

Here, $\boldsymbol{x}_i$ denotes the individual measurements[1] at voxel location i, $\boldsymbol{s}_i$ denotes the non-Gaussian source signals contained in the data and $\boldsymbol{\eta}_i$ denotes Gaussian noise[2] $\boldsymbol{\eta}_i \sim \mathcal{G}(0, \sigma^2\boldsymbol{\Sigma}_i)$. The 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. The spatial source signals correspond to parameter estimates in the GLM with the additional constraint of being statistically independent of each other.

2. MAXIMUM LIKELIHOOD ESTIMATION

Without loss of generality we will assume that the sources have unit variance. If $\boldsymbol{\Sigma}_i$ is known, we can use its Cholesky decomposition $\boldsymbol{\Sigma}_i = \boldsymbol{K}_i\boldsymbol{K}_i^t$ so that

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

and we obtain a new representation $\bar{\boldsymbol{x}}_i = \bar{\boldsymbol{A}}\boldsymbol{s}_i + \bar{\boldsymbol{\eta}}_i$, where $\bar{\boldsymbol{\eta}}_i = \boldsymbol{K}_i^{-1}\boldsymbol{\eta} \sim \mathcal{G}(0, \sigma^2\boldsymbol{I})$, i.e. where the noise covariance is isotropic at every voxel location [3] To simplify notation, we will henceforth assume isotropic noise and drop the additonal bar.

Noise and signal are uncorrelated, so the data covariance matrix is $\boldsymbol{R}_{\boldsymbol{x}} - \sigma^2\boldsymbol{I} = \boldsymbol{A}\boldsymbol{A}^t$. Let $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}$ (SVD) and let the number of source processes q, be known. Then

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

[1] for simplicity we assume de-meaned data.

[2] The covariance of the noise is allowed to be voxel dependent in order to encode the vastly different noise covariance observed within different tissue types [12].

[3] voxel-wise pre-whitening [12].

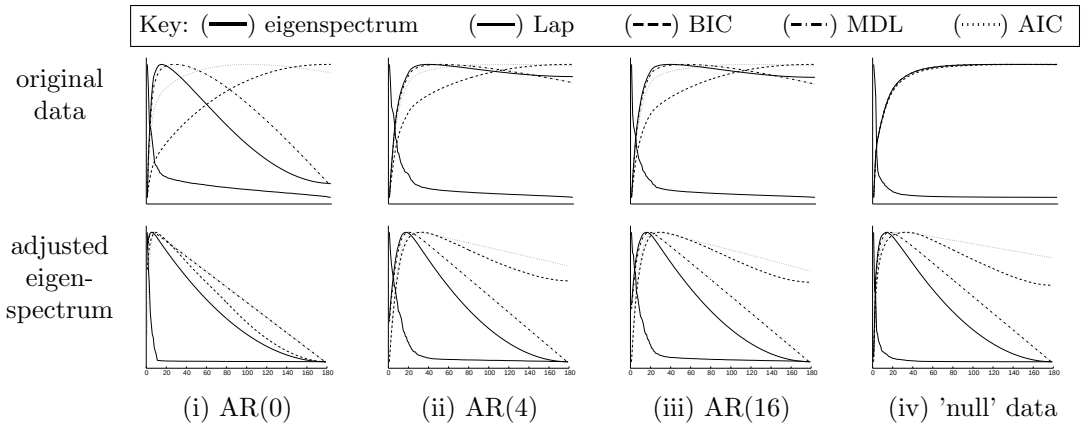

Figure 1: Estimation of the intrinsic dimensionality for 10 sources with non-Gaussian distribution embedded in a 180-dimensional space with different noise characteristics: (i) Gaussian white noise, (ii) AR(4) noise, (iii) AR(16) noise, (iv) resting-state FMRI noise; estimates from the original data (top) and after variance-normalisation and 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.

where $\boldsymbol{U}_q$ and $\boldsymbol{\Lambda}_q$ contain the first q eigenvectors and eigenvalues and where $\boldsymbol{Q}$ denotes a $q \times q$ orthogonal rotation matrix. The maximum likelihood estimates of sources and σ are obtained using generalised least squares:

$$\widehat{\boldsymbol{s}}_{\mathsf{ML}} = (\widehat{\boldsymbol{A}}^t \widehat{\boldsymbol{A}})^{-1} \widehat{\boldsymbol{A}}^t \boldsymbol{x} \quad \text{and} \quad \widehat{\sigma}_{\mathsf{ML}} = \frac{1}{p-q} \sum_{l=q+1}^{p} \lambda_l. \quad (3)$$

Solving the model in the case of an *unknown* noise covariance can then be achieved by iterating estimates $\widehat{\boldsymbol{A}}$ and $\widehat{\boldsymbol{s}}$ and re-estimating the noise covariances from the residuals $\widehat{\boldsymbol{\eta}}$. The form of $\boldsymbol{\Sigma}_i$ needs to be constrained by a suitable parameterisation; here we restrict ourselves to autoregressive noise. Estimation of $\boldsymbol{\Sigma}_i$ from residuals is discussed elsewhere [12].

A consequence of the isotropic noise model is that as an initial pre-processing step we normalise each voxel's time series to unit variance. This pre-conditions the data under the null hypotheses of no signal: the data matrix $\boldsymbol{X}$ is identical (up to second order statistics) to a simple set of realisations from a $\mathcal{G}(0, \boldsymbol{I})$ noise process. Any signal will have to reveal itself via its deviation from Gaussianity.

2.1. Model order selection

The maximum likelihood estimators depend on knowledge of the latent dimensionality q. In the presence of isotropic noise, the covariance matrix $\boldsymbol{R}_{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{A}^t + \sigma^2 \boldsymbol{I}_p$ will be of full rank where the additional noise has the effect of raising the eigenvalues of the covariance matrix by σ^2 [9]. Inferring the latent dimensionality amounts to test for sphericity of eigenspaces beyond a given threshold level [1]. Simplistic criteria like the reconstruction error or predictive likelihood will naturally predict that the accuracy steadily increases with increased dimensionality. Thus, criteria like retaining 99.9% of the variability result in arbitrary threshold levels [2]. This problem is intensified by the fact that $\boldsymbol{R}_{\boldsymbol{x}}$ is being estimated by the sample covariance $\widetilde{\boldsymbol{R}}_{\boldsymbol{x}}$. In the absence of any source signals, the eigenspectrum of the sample covariance matrix is skewed around the true noise covariance: the eigenspectrum will depict an apparent difference in the significance of individual directions within the noise [3]. In the case of Gaussian noise, the eigenvalues have a Wishart distribution and we can adjust the observed eigenspectrum by the quantiles of the predicted cumulative distribution $G^{-1}(\nu)$ of eigenvalues from Gaussian noise [6] , prior to estimating the model order. If we assume that the source distributions $p(\boldsymbol{s})$ are Gaussian, the model reduces to probabilistic PCA [11] and we can use Bayesian model selection criteria. Here, we are using the Laplace approximation to the posterior distribution of the model evidence that can be calculated efficiently from the adjusted eigenspectrum [8, 1]. While the estimation of the model order is based on the assumption of Gaussian source distribution, [8] provides some empirical evidence that the Laplace approximation still works reasonably well in the case where the source distributions are non-Gaussian. As an example, figure 1 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 noise[4] and real FMRI resting-state noise. 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 in all cases improves the estimation and in most cases, the different estimators give similar results once the data was variance normalised and the eigenspectrum was adjusted. Overall, the Laplace approximation and the Bayesian Information Criterion appear to give consistent and accurate estimates of the latent dimensionality even though the distribution of the embedded sources are non-Gaussian.

3. INFERENCE

After estimating the mixing-matrix $\widehat{\boldsymbol{A}}$, the source estimates are calculated by projecting each voxel's time course onto the time courses contained in the columns of the unmixing matrix $\widehat{\boldsymbol{W}} = (\widehat{\boldsymbol{A}}^t\widehat{\boldsymbol{A}})^{-1}\widehat{\boldsymbol{A}}^t$. In the case where the model order q was estimated correctly, the estimated noise is a linear projection of the true noise and is unconfounded by residual signal. At every voxel location we have pre-conditioned the data such that $\boldsymbol{x}_i$ has unit standard deviation and the estimate of the noise variance σ_i^2 at each voxel location is $\widehat{\sigma}_i^2 = \frac{1}{p-(q+1)}\sum_{l=1}^{p}(\widehat{\boldsymbol{\eta}}_{li} - \mathcal{E}\{\widehat{\boldsymbol{\eta}}_{\cdot i}\})^2$, which, if $p-q$ is reasonably large, will approximately equal σ_i^2, i.e. equal the true variance of the noise. We can then 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 $\widehat{\sigma}_i$.

In order to assess the Z-maps for significantly activated voxels, we employ mixture modelling of the probability density of the Z-statistic spatial maps.

From equation 3 it follows that $\widehat{\boldsymbol{s}}_i = \widehat{\boldsymbol{W}}\boldsymbol{A}\boldsymbol{s}_i + \widehat{\boldsymbol{W}}\boldsymbol{\eta}_i$, i.e. the noise term in equation 1 manifests itself as additive Gaussian noise in the estimated sources. We therefore model the distribution of the spatial intensity values of each Z-map by a mixture of one Gaussian and one or more Gamma distributions, to model background noise and positive and negative BOLD effects [4]. The mixture is fitted using EM. In order to infer the appropriate number of components in this 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. In cases where the number of 'active' voxels is very small,

[4]i.e. auto-regressive noise where the AR parameters were estimated from real resting-state FMRI data.

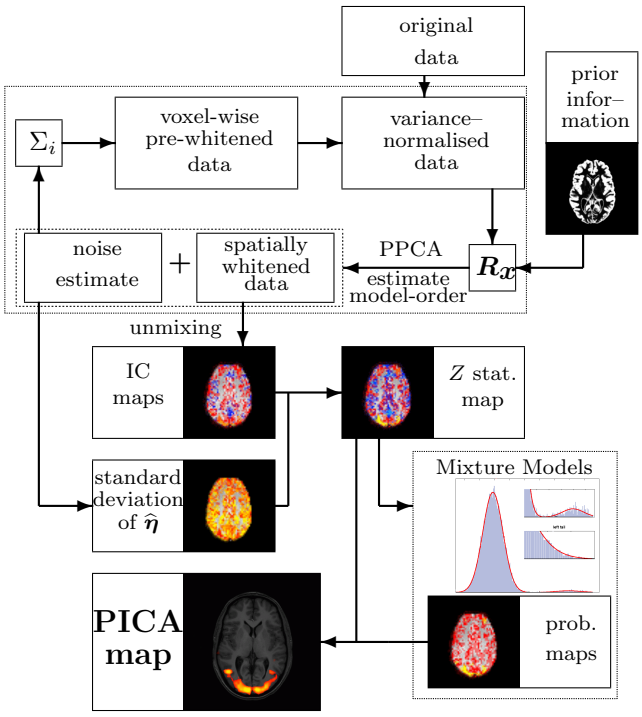

Figure 2: Schematic illustration of PICA.

a single Gaussian mixture may actually have the highest model evidence. In this case, a simple transformation to spatial Z-scores and subsequent thresholding is appropriate[5]. Otherwise we can evaluating the fitted mixture model to calculate the posterior probability of 'activation' as the ratio of the probability of intensity value under the 'noise' Gaussian relative to the sum of probabilities of the value under the 'activation' Gamma densities[6]. Any 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 [4].

4. PICA ALGORITHM OVERVIEW

The individual steps that constitute the Probabilistic Independent Component Analysis are illustrated in figure 2. The de-meaned original data is first normalised to unit variance at each voxel location. If appropriate spatial information is available, this is encoded in the estimation of the sample covariance matrix $\boldsymbol{R_x}$. Individual voxel weights, e.g. gray-matter segmentation,

[5]i.e. reverting to null-hypothesis testing instead of the otherwise preferable alternative-hypothesis testing.

[6]where in this case 'activation' is to be understood as 'cannot be explained as random correlation coefficient'.

can be used to calculate a weighted covariance matrix while voxel-pair weightings can be used to calculate the within-group covariance [2]. Probabilistic PCA is used to infer upon the unknown number of sources and will give an estimate of the noise and a set of spatially whitened observations [1]. We estimate $\mathbf{\Sigma}_i$ from the residuals in order to voxel-wise pre-whiten the data and iterate the entire cycle. From the spatially whitened observations, the individual component maps are estimated using a modified fixed point iteration scheme (FastICA [5]) to optimise for non- Gaussian source estimates via maximising the neg-entropy. These maps are separately transformed to Z scores. Finally, Gaussian/Gamma Mixture Models are fitted to the individual Z maps in order to infer voxel locations that are significantly modulated by the associated time course.

5. EVALUATION DATA

We acquired 180 whole brain volumes (64× 64× 21; 4× 4× 6 mm) of FMRI data on a Varian 3T system under a resting condition and under 30s on/off visual stimulus (black and white checkerboard reversing at 8Hz). The activation data set was analysed using standard GLM techniques as implemented in FEAT [10]; final Z-statistic maps were used to define activation 'masks' by thresholding at $Z > 3.0$ and clustering with $p < 0.01$.

Next, audio- and visual activation patterns were added into (motion-corrected) resting data 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 described above) and convolving with a standard gamma-based HRF kernel (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 being 0.5%, 1%, 3% and 5% [7].

Table 1 and figure 3 summarises the spatio-temporal accuracy of estimation of PICA as compared to standard GLM (mean correlation between the estimated and true time courses over 150 runs and Receiver-Operator Characteristics (ROC) for both GLM and PICA).

In almost all cases, the PICA estimates show an improved ROC curves compared to the GLM results despite the fact that GLM analysis was carried out with perfect knowledge of the regressors of interest.

[7] In the real activation data, the peak activation was 3% – here a % means peak-to-peak activation as a % of mean signal intensity.

	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 levels: correlation between the extracted time courses and the true signal time courses over 150 PICA runs.

5.1. Accuracy and dimensionality

Within the estimation steps, the choice of number of components was determined from the estimate of the Bayesian evidence. 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 $\boldsymbol{A}$. In order to assess the dependency of the spatio-temporal accuracy on estimated number of source processes, we tested PICA results obtained after projecting the data into subspaces of increasing dimensionality. Figure 4 shows the results of the temporal correlation and the final false-positive and false-negative rates over the range of possible dimensions for the data set with 3% peak level activation. For both the spatial and temporal accuracy 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 will lead to increasingly poor estimates. For this data set, the Laplace approximation to the evidence for model order (figure 4) appears to work well.

5.2. Real FMRI data

Figure 5 shows the PICA results on the visual stimulation study. Based on the estimate of the model order, the data was projected onto the first 27 eigenvectors prior to the unmixing. Comparing the results with figure 5(bottom) we get a much better correspondence between the areas of activation estimated from the GLM approach and the main PICA estimate. 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. For standard ICA the detected activation appears fragmented into different spatial maps, a consequence of over-fitting the noise- free generative model to noisy observations. In the absence of a suitable noise model, any modulation in the temporal response to the same common signal between two voxels

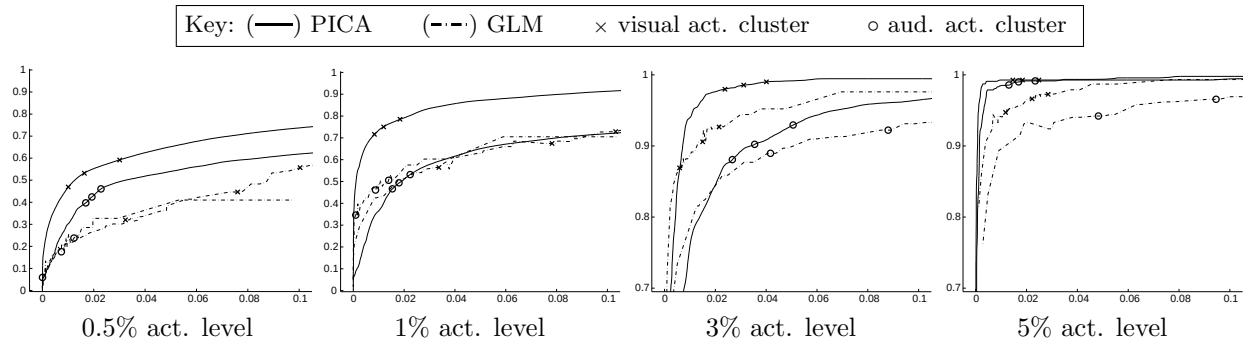

Figure 3: Spatial accuracy at different activation levels: ROC curves for PICA (solid lines - mean over 150 runs) vs. ROC curves for GLM-based Z-statistical maps thresholded using cluster-based thresholding at different Z and p levels . Markers indicate typical threshold levels.

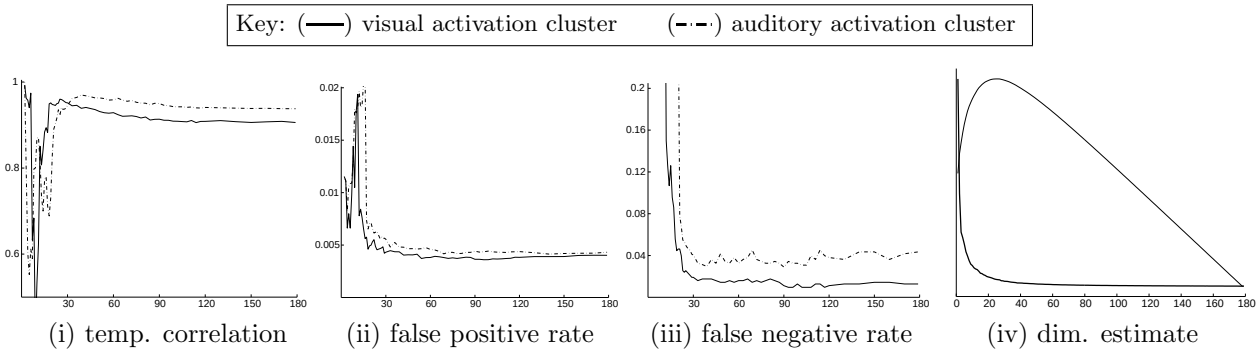

Figure 4: Spatio-temporal accuracy as a function of assumed dimensionality for the simulated audio-visual activation data at the 3% level: (i) correlation between the extracted time course and the true signal time course; (ii) false positive rate; (iii) false negative rate; (iv) eigenspectrum of the data covariance matrix together with Laplace approximation to the model order.

necessarily is treated as a 'real effect'. These modulations might represent valid spatial variations or simply differences in the background noise level. In the PICA approach this is 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.

Within the set of PICA 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 V5/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. In the case of standard ICA, only unilateral secondary activations are identified.This is not attributable to the difference in the thresholding itself; the raw IC maps do not allow for bilateral activation patterns. 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 the identified areas so that after transforming the raw IC estimates $\boldsymbol{s}_i$ into Z-scores, the well localised areas emerge.

Note that in both examples the PICA maps are actual Z-statistic maps and as such are comparable against output from a standard GLM analysis. Standard ICA maps, for reasons outlined above, are simply raw parameter estimates and as such purely descriptive.

6. CONCLUSION

We have presented an adaptation to classical Independent Component Analysis and address a variety of important issues that exist in ICA applied to FMRI data. Most importantly, we can statistically infer areas within

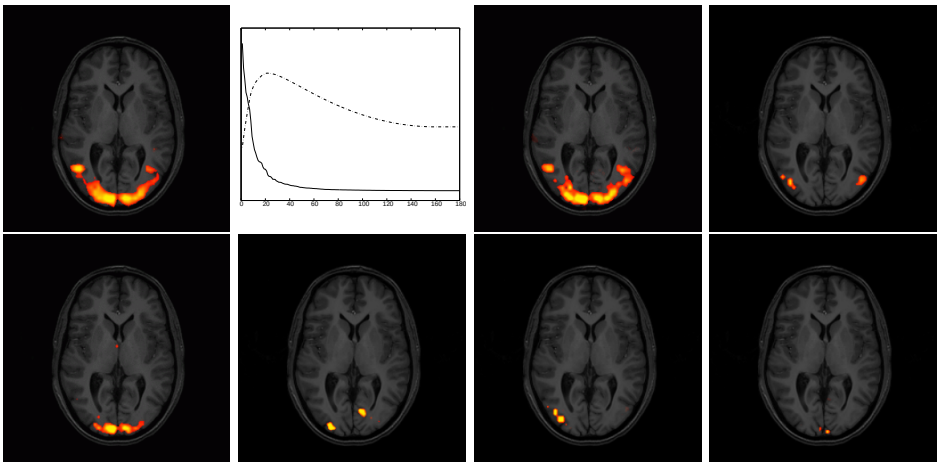

Figure 5: GLM, PICA and classical ICA analysis of visual stimulus FMRI data: (top) GLM results at $Z > 2.3, p < 0.01$, Eigenspectrum of the data covariance matrix with estimate of the latent dimensionality and spatial maps of PICA; (bottom) spatial maps from standard ICA (as presented in [7]).

the spatial maps that are modulated significantly by the associated time courses.

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 an example 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 there has been implemented as MELODIC, a software tool freely available as part of FSL (FMRIB's Software Library – `http://www.fmrib.ox.ac.uk/fsl`).

7. REFERENCES

[1] 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.

[2] C.F. Beckmann and S.M. Smith. Probabilistic independent component analysis in fmri. In *Proc. Int. Soc. of Magnetic Resonance in Medicine*, 2002.

[3] 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.

[4] N.V. Hartvig. Spatial mixture modelling of fMRI data. *Human Brain Mapping*, 2000.

[5] A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–1492, 1997.

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

[7] 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.

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

[9] S.J. Roberts and R. Everson, editors. *Independent Component Analysis: Principles and Practice*. Cambridge University Press, 2001.

[10] 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.

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

[12] 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.