10

Application of Polynomial Spline Independent Component Analysis to fMRI Data

Atsushi Kawaguchi[1], Young K. Truong[2] and Xuemei Huang[3]
[1]*Biostatistics Center, Kurume University, Kureme, Fukuoka*
[2]*Department of Biostatistics, University of North Carolina at Chapel Hill, NC*
[3]*Department of Neurology, Penn State University, PA*
[1]*Japan*
[2,3]*USA*

1. Introduction

In independent component analysis (ICA), it is assumed that the components of the observed k-dimensional random vector $\mathbf{x} = (x_1, \ldots, x_k)$ are linear combinations of the components of a latent k-vector $\mathbf{s} = (s_1, \ldots, s_k)$ such that $s_1, \ldots, s_k$ are mutually independent. This is denoted by

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \tag{1}$$

where $\mathbf{A}$ is a $k \times k$ full-rank non-random mixing matrix. The main objective then is to extract the mixing matrix through a set of observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. For a detailed description of this method, including its motivation, existence and relationship with other well known statistical methods such as principal component analysis, factor analysis, see Hyvärinen et al. (2001).

In signal processing applications, it will be convenient to view the observations as values recorded at k locations over time periods $1, \ldots, n$. The number of locations k varies depending on the application area. For instance, $k = 2$ in a blind source seperation problem to $k \approx 10^5$ in a typical human brain imaging data set. The number of time points n also varies and it ranges from $n \approx 10^2$ to $n \approx 10^6$.

The spatial (location) and temporal (time) description of the data has generated a huge number of biomedical applications such as cognitive or genomic research. In this paper, we will focus on human brain data acquired from functional magnetic resonance imaging (fMRI) technique where $k \approx 10^5$ and $n \approx 10^2$. This imaging technique has been used to effectively study brain activities in a non-invasive manner by detecting the associated changes in blood flow. Typically, fMRI data consists of a 3D grid of voxels; each voxel's response signal over time reflects brain activity. However, response signals are often contaminated by other signals and noise, the magnitude of which may be as large as that of the response signal. Therefore, independent component analysis has been applied to extract the spatial and temporal features of fMRI data (Calhoun & Adali, 2006; McKeown et al., 1998).

For fMRI datasets, we remark that it is theoretically possible to search for signals that are independent over space (spatial ICA) or time (temporal ICA). In fact, the above ICA description involving the spatial k and temporal n scales should be called more precisely as

the temporal ICA, while in spatial ICA, k will be treated as time, and n as location. Thus one can see that temporal ICA is just the *transpose* of spatial ICA. However, in practice, it is very difficult to obtain accurate and meaningful results from the temporal ICA of fMRI data because of the correlation among the temporal physiological components. Therefore, the use of spatial ICA is preferred for fMRI analysis (McKeown et al., 1998).

Our ICA on fMRI data is carried out by first reducing the number of independent components (IC) using tools such as principal component analysis (PCA) or singular value decomposition (SVD), followed with an algorithm for determining the ICs. The most commonly used ICA algorithms for analyzing fMRI data are Infomax (Bell & Sejnowski, 1995), FastICA (Hyvärinen & Oja, 1997), and joint approximate diagonalization of eigenmatrices (JADE) (Cardoso & Souloumiac, 1993). Calhoun & Adali (2006) reported that Infomax consistently yielded the most reliable results, followed closely by JADE and FastICA. In this study, we propose a novel ICA algorithm that is a modification of the logspline ICA algorithm (LICA) (Kawaguchi & Truong, 2011) and apply it to fMRI data. In ICA, we employ a likelihood approach to search for ICs by estimating their probability distributions or density functions (pdf). This is equivalent to maximizing the independence among ICs, and it is realized by using polynomial splines to approximate the logarithmic pdf; we call this the logspline model. To account for the sparsity of spatial fMRI maps, we further treat the pdf as a mixture of a logspline and a logistic density function; this approach has proven to be very effective for treating sparse features in data. Using simulated and real data, we compared our method with several well-known methods and demonstrated the relative advantage of our method in extracting ICs.

The remainder of this paper is organized as follows. Section 2 describes the proposed method. Section 3 presents the simulation studies. Section 4 describes the application of the proposed method to real data. Finally, Section 5 presents discussions and concluding remarks of our method.

2. Method

Let $\mathbf{Y}$ denote a $T \times V$ data matrix: each column of this matrix corresponds to a voxel time series, and there are V voxels and T time points. We invoke singular value decomposition (SVD) to yield the approximation $\mathbf{Y} \approx \mathbf{UDX}$, where $\mathbf{U}$ is a $T \times M$ orthogonal matrix, $\mathbf{D} = diag(d_1, d_2, \ldots, d_M)$ with $d_1 \geq d_2 \geq \cdots \geq d_M$, and $\mathbf{X}$ is an $M \times V$ orthogonal matrix. Here, we selected (orthogonal) columns of $\mathbf{U}$ to represent some experimental task functions as well as the physiological components. In addition, the dimension of $\mathbf{D}$ has been reduced by discarding values below a certain threshold; in other words, these values are essentially treated as noise.

We determine the ICs based on the matrix $\mathbf{X}$ so that $\mathbf{X} = \mathbf{AS}$, where $\mathbf{A}$ is an $M \times M$ mixing matrix and S is an $M \times V$ source matrix. That is, the v-th column of $\mathbf{X}$ is equal to $\mathbf{A}$ multiplied by the v-th column of $\mathbf{S}$, where $v = 1, 2, \ldots, V$. Equivalently, each column of $\mathbf{X}$ is a mixture of M independent sources. Let $\mathbf{S}_v$ denote the source vector at voxel v so that $\mathbf{S}_v = (S_1, S_2, \ldots, S_M)$, $v = 1, 2, \ldots, V$. Suppose that each S_j has a density function f_j for $j = 1, 2, \ldots, M$. Then, the density function of $\mathbf{X}$ can be expressed as $f_{\mathbf{X}}(\mathbf{x}) = \mathsf{det}(\mathbf{W}) \prod_{j=1}^{M} f_j(\mathbf{w}_j \mathbf{x})$, where $\mathbf{W} = \mathbf{A}^{-1}$ and $\mathbf{w}_j$ is the j-th row of $\mathbf{W}$.

We now model each source density according to the mixture with unknown probability a:

$$f_j(x) = a f_{1j}(x) + (1 - a) f_{2j}(x), \tag{2}$$

where the logarithm of $f_{1j}(x)$ is modeled by using polynomial splines

$$\log(f_{1j}(x)) = C(\boldsymbol{\beta}_j) + \beta_{01j}x + \sum_{i=1}^{m_j} \beta_{1ij}(x - r_{ij})_+^3,$$

with $\boldsymbol{\beta}_j = (\beta_{01j}, \beta_{11j}, \ldots, \beta_{1m_jj})$ being a vector of coefficients, $C(\boldsymbol{\beta}_j)$ a normalized constant, r_{ij} the knots; and $f_{2j}(x) = sech^2(x)/2$ is a logistic density function. Here $(y)_+ = \max(y, 0)$.

We denote the vector of parameters in the density function by $\boldsymbol{\theta}$ = (a, $\boldsymbol{\beta}$). The maximum likelihood estimate (MLE) of ($\mathbf{W}$, $\boldsymbol{\theta}$) is obtained by maximizing the likelihood of $\mathbf{X}$ with respect to ($\mathbf{W}$, $\boldsymbol{\theta}$):

$$\ell(\mathbf{W}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{j=1}^{k} \log(f_j(\mathbf{w}_j^T \mathbf{x}_i)).$$

We use a profile likelihood procedure to compute the MLE because a direct computation of the estimates is generally not feasible. The iterative algorithm is shown in Table 1. Note that

1. Initialize $\mathbf{W} = \mathbf{I}$.
2. Repeat until the convergence of $\mathbf{W}$, using the Amari metric.
   (a) Given $\mathbf{W}$, estimate the log density $g_j = \log f_j$ for the jth element X_j of $\mathbf{X}$ (separately for each j) by using the stochastic EM algorithm shown in Appendix 7.
   (b) Given g_j $(j = 1, 2, \ldots, p)$,

$$\mathbf{w}_j \leftarrow \mathsf{ave}[\mathbf{X} g_j'(\mathbf{w}_j^T \mathbf{X})] - \mathsf{ave}[g_j''(\mathbf{w}_j^T \mathbf{X})]\mathbf{w}_j$$

   where $\mathbf{w}_j$ is the jth column of $\mathbf{W}$ and ave is a sample average over $\mathbf{X}$.
   (c) Orthogonalize $\mathbf{W}$

Table 1. Algorithm

the Amari metric (Amari et al., 1996) used in the algorithms is defined as

$$d(\mathbf{P}, \mathbf{Q}) = \frac{1}{p(p-1)} \left\{ \sum_{i=1}^{p} \left(\frac{\sum_{j=1}^{p} |a_{ij}|}{\max_j |a_{ij}|} - 1 \right) + \sum_{j=1}^{p} \left(\frac{\sum_{i=1}^{p} |a_{ij}|}{\max_i |a_{ij}|} - 1 \right) \right\},$$

where $a_{ij} = (\mathbf{P}^{-1}\mathbf{Q})_{ij}$, $\mathbf{P}$, and $\mathbf{Q}$ are $p \times p$ matrices. This metric is normalized, and is between 0 and 1.

Several authors have discussed initial guesses for ICA algorithms. Instead of setting several initial guesses, as discussed in Kawaguchi & Truong (2011), $\mathbf{X}$ is multiplied by $\widehat{\mathbf{W}}$, which is the output of the algorithm when the log density function $g(x)$ is replaced with $g(x) = 1/\{2b^{1/b}\Gamma(1 + 1/b)\} \exp\{-|x|^b/b\}$ with $b = 3$. The final output is obtained in the form $\widehat{\mathbf{W}} = \widehat{\mathbf{W}}\widehat{\mathbf{W}}_0$, where $\widehat{\mathbf{W}}_0$ is the output of the algorithm shown in Table 1.

The purpose of spatial ICA is to obtain independent spatial maps and the corresponding temporal activation profiles (time courses). By multiplying $\mathbf{X}$ with $\widehat{\mathbf{W}}$, we can obtain the estimates of the spatial map $\widehat{\mathbf{S}}$ as $\widehat{\mathbf{S}} = \widehat{\mathbf{W}}\mathbf{X}$. On the other hand, the corresponding time courses are obtained in the form $\widehat{\mathbf{A}} = \widehat{\mathbf{W}}(\mathbf{UD})^{-1}$.

3. Simulation study

In this section, we conducted a simulation study to compare the proposed method with existing methods such as Infomax (Bell & Sejnowski, 1995), fastICA (Hyvärinen & Oja, 1997), and KDICA (Chen & Bickel, 2006). We designed our comparative study by using data that emulated the properties of fMRI data. The spatial sources **S** consisted of a set of 250 $\times$ 250 pixels. These spatial sources were modulated with four corresponding time courses **A** of length 128 to form a 62,500 $\times$ 128 dataset. The spatial source images **S** shown in the left-hand side of Figure 1 are created by generating random numbers from normal density functions with mean 0 and standard deviation 0.15 for a non-activation region, and mean 1 and standard deviation 0.15 for an activation region. The activated regions consist of squares of d_i pixels on a side, for i = 1, 2, 3, 4, that are located at different corners. We consider two situations: d_i's are the same among the four components ($d_1 = d_2 = d_3 = d_4 = d$) and d_i's are different. For the former, we used d = 20, 30, 40, and 50. For the latter, we generated uniform random numbers between 20 and 50 for each d_i. The temporal source signals in the right-hand side of Figure 1 are the stimulus sequences convolved with an ideal hemodynamic response function as a task-related component, and sin curves with frequencies of 2, 17, and 32 as other sources. We generated the task-related component by using the R package `fmri` with onset times (11,75) and a duration of 11. We repeated the above procedure 10 times for the case in which d_i's were the same and 50 times for the case in which d_i's were different.

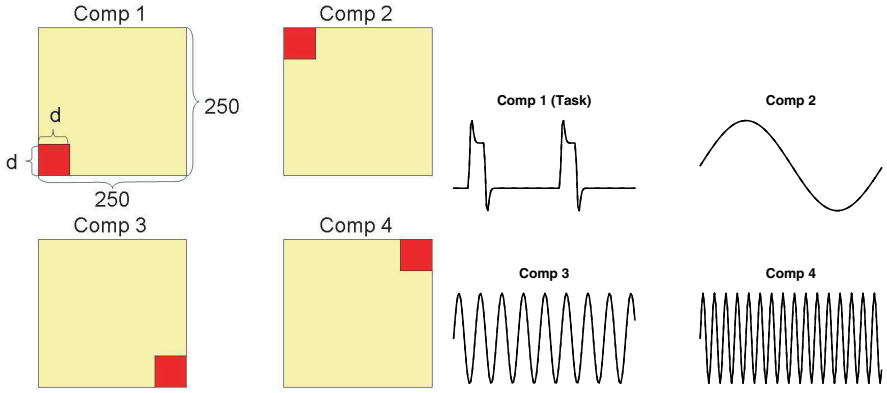

Fig. 1. Spatial and Temporal Simulation Data

Both the spatial and the temporal accuracies of ICA were assessed by R-square fitting of a linear regression model. The evaluation was carried out as follows. For every estimated time course, the R-square is computed from the linear regression model with the response being each of the estimates and the predictor being true, that is, the stimulus sequence (`Comp 1` on the right-hand side in Figure 1). The component that has the maximum R-square is considered to be task-related. We used the R-square value of this component for the comparison with the existing methods with respect to temporal accuracy and to determine the corresponding spatial map. The intensities of the spatial map are vectorized and used in the linear regression model as the response with the vectorized true (`Comp 1` on the left-hand side in Figure 1) as the predictor to compute R-square for the spatial accuracy.

The averaged R-squares over simulations are summarized in Tables 2 and 3 for the temporal and spatial data, respectively. When the sizes of the activation region were the same among all

	Infomax	fastICA	KDICA	PSICA
d=50	0.627	0.852	0.679	0.843
d=40	0.456	0.460	0.472	0.735
d=30	0.408	0.463	0.424	0.586
d=20	0.358	0.270	0.709	0.518
average	0.462	0.511	0.571	0.670
rand	0.623	0.651	0.529	0.699

Table 2. Temporal R-square for simulation data. The mean over d = 20, 30, 40, and 50 is calculated in the row labeled as average. The rand row shows the average over 50 replications when d_i's were chosen randomly from the range 20 to 50.

	Infomax	fastICA	KDICA	PSICA
d=50	0.801	0.765	0.641	0.761
d=40	0.462	0.502	0.545	0.726
d=30	0.409	0.528	0.552	0.680
d=20	0.323	0.478	0.624	0.587
average	0.499	0.568	0.591	0.688
rand	0.537	0.607	0.579	0.643

Table 3. Spatial R-square for simulation data. The mean over d = 20, 30, 40, and 50 is calculated in the row labeled as average. The rand row shows the average over 50 replications when d_i's were chosen randomly from the range 20 to 50.

components, R-squares of the proposed method were significantly larger than those of others for moderate sizes (d = 40 and 30) for both temporal and spatial data. For d = 50, fastICA had the largest R-square for both temporal and spatial data, with the difference from the result of the proposed method being small. For d = 20, KDICA had the largest R-square for both temporal and spatial data, with the difference from the result of the proposed method being significant for temporal data but not for spatial data. With respect to the average for d = 50, 40, 30, and 20, the proposed method had the largest R-square value than the others did. When d_i was determined randomly, which might be more practical, we observed that the largest R-square value in the rand row of the table was achieved by the proposed method.

4. Application

To demonstrate the applicability of the proposed method to real data, we separate fMRI data into independent spatial components that can be used to determine three-dimensional brain maps. To study brain regions that are related to different finger tapping movements, fMRI data were obtained from a twin pair (Twin 1 and Twin 2) performing different tasks alternately.The paradigm shown in Figure 2 consisted of externally guided (EG) or internally guided (IG) movements based on three different finger sequencing movements performed alternately by either the right or the left hand.

The fMRI dataset has 128 scans that were acquired using a modified 3T Siemens MAGNETOM Vision system. Each acquisition consists of 49 contiguous slices. Each slice contains 64×64 voxels. Hence, each scan produces $64\times64\times49$ voxels. The size of each voxel is 3 mm$\times$3 mm$\times$3 mm. Each acquisition took 2.9388 s, with the scan-to-scan repetition time (TR) set to 3 s. The dataset was pre-processed using SPM5 (Friston et al., 1995). The preprocessing

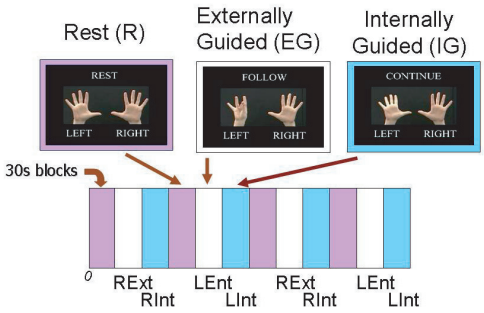

Fig. 2. Experimental Paradigm

included slice timing, realignment, and smoothing. We masked the image outside the human head using the GIFT software package (Group ICA of fMRI Toolbox, Calhoun et al., 2001). We used 21 components for Twin 1 and 30 for Twin 2; these were estimated using the minimum description length (MDL) criteria.

We applied four ICA algorithms—Infomax (Bell & Sejnowski, 1995), fastICA (Hyvärinen & Oja, 1997), KDICA (Chen & Bickel, 2006), and the proposed method (PSICA)—to the twins' data. The R-square statistic was calculated from the fitted multiple linear regression model with the estimated time course as the response. The predictors were the right EG, right IG, left EG, and left IG, which consists of the expected BOLD response for the task indicator function given by the argument as a convolution with the hemodynamic response function modeled by the difference between two gamma functions. Table 4 shows the corresponding R-square statistics. From this table, we can see that the proposed method extracted more correlated components for a task than did the other methods for both twins.

	Infomax	fastICA	KDICA	PSICA
Twin 1	0.640	0.666	0.655	0.680
Twin 2	0.847	0.661	0.805	0.862

Table 4. Temporal R-square statistics for the twin data

Figure 3 shows one of the resulting spatial maps of SPICA for Twins 1 and 2 respectively, in which the right motor area is highly activated and the corresponding time course shows a fit to the left-hand task paradigm.

We mention a few important observations in this real human brain analysis:

1. After the analysis, it was revealed to us that Twin 1 had shown signs and symptoms (tremors and slowed movements) of the Parkinson's disease (PD), while Twin 2 was considered normal at the time the data were collected. This may help to explain why Twin 2 — the normal subject has higher R-squares in three of the four methods (Table 4). In these methods, fastICA shows practically no difference of the twins.
2. In interpreting results from ICA, one should note that ICA is ambiguous about the sign: $\mathbf{x} = \mathbf{A}\mathbf{s} = (-\mathbf{A})(-\mathbf{s})$. This fact has produced different colour scales in the spatial maps (located in the lower right corner). With this in mind, one can say that Twin 2 or the normal subject has a higher intensity or activation level in the right motor area (because of the left-hand task paradigm).

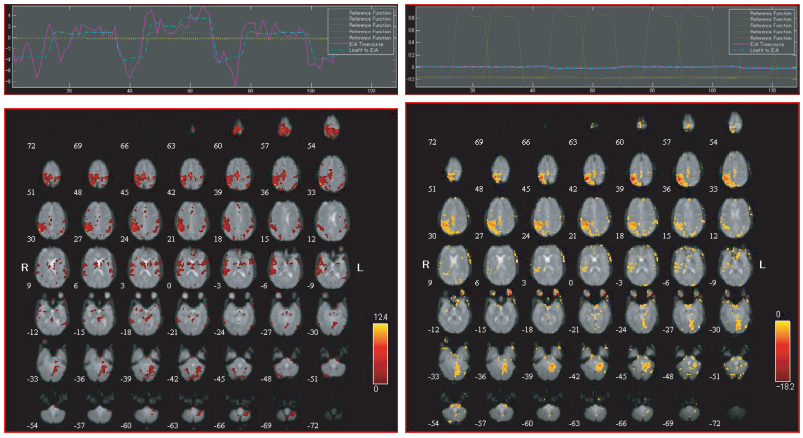

Fig. 3. Spatial Images for Twin 1 (left) and Twin 2 (right)

3. Further examination of the spatial maps indicates that the normal subject (on the right panel) has a more focused location of the motor area, see particularly the red region in slices 51, 48, 45, 42, 39, 36 and 33. The activated motor area of the PD twin (the left panel) is not as sharply defined.

5. Discussion and conclusion

In this study, we developed an ICA algorithm based on a maximum likelihood approach using a mixture of logspline and logistic density models with adaptive knot locations. The first concern about this approach is that its model dimension seems to be much higher than those of its peers. Here model dimensionality is defined as the number of model parameters including possibly the spline knot locations. Depending on how noisy the data are, the built-in model selection procedure (which is based on AIC or BIC) works in a sensible adaptive way: there is constantly a trade-off in balancing the bias and variance of the estimate of the parameter since the optimal strategy is to minimize the mean square error loss at the expense of the model dimension. Moreover, the logistic component is included to reduce the model dimension from the spline part in handling the sparsity of the spatial map. The main issue then is the time required to extract the ICs this way. It is considerably more time consuming, but the accuracy is very rewarding. The improvement over its peers performance was demonstrated numerically in Tables 2 and 3 using the R-square as a criterion.

It is important to point out that we should also provide a sensitivity and specificity analysis of the activated spatial locations as described in Lee et al. (2011), where popular methods such as Infomax and fastICA were shown to have a higher false-positive/nagative rate. This implies that brain activation should be studied more carefully, and one should avoid using methods that tend to yield false activation.

As in our previous approahces to ICA, the key feature has always been the flexibility modeling the source. In Kawaguchi & Truong (2011), the marginal distribution of the temporal source component was modelled by the logspline methodology and we noted the improvement over its peers. The comparatrive study was based on a wide variety of density functions, some are known to be very challenging to estimate. Further details of this approach can be found in

Kawaguchi & Truong (2011). In pursuing spatial ICA for fMRI based on human brain data, we observed that simply taking the transpose of the temporal ICA approach mentioned in the introduction did not always work. This is due to the fact that the spatial activation maps are very sparse: density estimation using the logspline approach in the presence of sparsity has never been investigated before. One of the our findings is that the logspline estimate of the spatial distribution is too noisy, perhaps the model dimension is too high. Thus the logistic component is added to our previous temporal ICA procedure in order to address this issue. The advantage over the simple tranposition of the temporal approach has been clearly shown in this paper.

The mixture modeling has been used previously for the detection of brain activation in fMRI data (Everitt & Bullmore, 1999; Hartvig & Jensen, 2000; Neumann et al., 2008). In fMRI data, the density functions of spatial sources are known to be supergaussion with heavy tails due to the fact that brain activation is sparse and highly localized (McKeown et al., 1998), and often skewed due to larger signal amplitudes in activated regions (Stone et al., 2002). Cordes & Nandy (2007) modeled source densities as improved exponential power family. Our modeling would be more flexible than these approaches.

In addition, the method may have some important extensions. Namely, it has been an important problem as how to assess the variability of ICA, especially how the variance of the spatial map can be best displayed. One way to examine the variation of the mixing coefficient estimates is to use bootstrap method while preserving information about the spatial structure. For example, in spatial ICA, one can generate bootstrap random samples from the logspline density estimates of the source over space. Mix these samples using the estimate mixing coefficients to yield the observed fMRI (BOLD) signals, which will then pass through ICA to produce the so called bootstrapped spatial maps and mixing coefficients. We outline this as an algorithm:

1. $\mathbf{x} \approx \hat{\mathbf{A}}\hat{\mathbf{s}}$ via our ICA algorithm.
2. $\hat{\mathbf{s}} \rightarrow \mathbf{s}^*$ which is a bootstrapped source sample drawnn from the distribution of $\hat{\mathbf{s}}$.
3. $\mathbf{x}^* := \hat{\mathbf{A}}\mathbf{s}^*$ to yield bootstrapped observed samples.
4. $\mathbf{x}^* = \hat{\mathbf{A}}^*\hat{\mathbf{s}}^*$ using our ICA algorithm.
5. Repeat until a desirable number of bootstrap samples is achieved.

Table 5. Bootstrap Algorithm

The bootstrapped sample $\mathbf{s}^*$ can be regarded as a by-product of the adequately modelled spatial map density function. The algorithm can be described similarly for temporal ICA. Thus it is feasible to develop the statistical inference framework for assessing the variability of the estimator of the mixing matrix via the bootstrap method while preserving information about the spatial or temporal structure.

In extending our temporal ICA to spatial ICA, we merely added the logistic component to the logspline piece, which is essentially a one-dimensional density estimation, or marginal density estimation procedure. Alternatively, in order to capture the actual spatial feature of the three dimensional brain, or the two dimensional map, one can incorporate the spatial correlation structure of the spatial map by introducing tensor products of spline functions or the interaction terms in the logspline formulation. For temporal ICA, this can be implemented by using time series models to account for the source serial correlations. Indeed, Lee et al. (2011) has reported that there is noticeable improvement over the marginal density based ICA

procedures. It will be important to see if the same will hold for the above spatial ICA approach using tensor products of splines.

Another issue that we have not addressed is how to extend our method to compare groups of subjects. This is known as the group ICA problem. In principle, we can follow the Group ICA of fMRI Toolbox (Calhoun et al., 2001) by simply concatenating the observed data matrices. This will certainly increase the computational complexity and one has to address the efficiency problem as well.

Finally, we recall that prior to applying any of the ICA algorithms, one must carry out a dimension reduction step on the observed data matrix first. In temporal ICA with T and V as time and space scales, V will be reduced by, typically, employing the principal component analysis (PCA), while the time factor T will be reduced in the spatial ICA. We have found that even greater improvement can be achieved by using informative dimension reduction methods such as singular value decomposition (SVD) by choosing the eigen-vectors to relate to the experimental task paradigm closely. This is being referred to as a *supervised SVD dimension reduction* procedure (Bai et al., 2008) and has been used effectively in Lee et al. (2011).

In conclusion, the results presented in this paper can be viewed as a tool for setting up a new framework for addressing some of known issues in applying ICA to fMRI or other brain imaging modalities such as EEG or neural spike sorting problems. We have demonstrated that the key element here is the flexibility in modeling the source distribution and that was achieved by using polynomial splines as an approximation tool. We also used a mixture of distribution approach to account for the spatial distribution in ICA for fMRI data analysis. Although there are still many issues to be addressed, we have illustrated the usefulness of our approach to fMRI brain activation detection in both simulated and real data analysis.

6. Acknowledgment

We are grateful to Dr. Aiyou Chen for providing the KDICA programming code. We are also deeply grateful to Dr. Mechelle Lewis for her insight about the Twin data set, and her fruitful discussion on our analysis. This research was supported in part by the Banyu Fellowship Program sponsored by the Banyu Life Science Foundation International and by Grants-in-Aid from the Ministry of Education, Culture, Sport, Science and Technology of Japan (21700312) to AK, and by NSF DMS-0707090 to YT.

7. Appendix

A. Stochastic EM algorithm for mixture density estimation

In statistics, an expectation-maximization (EM) algorithm is an iterative method for finding maximum likelihood estimates (MLE) of parameters in statistical models. Typically these models involve latent variables in addition to unknown parameters and known data observations (Dempster et al., 1977). In our mixture model (2), parameter a is associated with the latent variable of the number of non-activated voxels in a given sample, and the unknown parameter β is related to the distribution of the fMRI intensity, coming from the logspline component.

The EM algorithm is particularly useful when the score function cannot be solved directly. The algorithm iteration alternates between performing an expectation (E) step, which computes the expectation of the log-likelihood evaluated using the current estimate for the

parameters, and a maximization (M) step, which locates parameters maximizing the expected log-likelihood resulted in the E step. A version of this algorithm called a stochastic EM algorithm was introduced in (Celeux & Diebolt, 1992) to avoid stabilization on saddle points in parametric mixture models by incorporating the stochastic step (S-step) into the EM algorithm. (Bordes et al., 2007) generalized it to semiparametric mixture models by using kernel density estimation.

Suppose we have observations $x_1, x_2, \ldots, x_n$, and the observations are grouped by the k-means clustering method with k being the integer part of $n/10$. Let us denote the number of members in each group by n_g ($g = 1, 2, \ldots, k$) and $\boldsymbol{x}_g = (x_{i_{g1}}, x_{i_{g2}}, \ldots, x_{i_{gn_g}})$. The algorithm used in this paper is given below.

(1) E-step: Compute $\tau(j|\boldsymbol{x}_g)$ ($g = 1, 2, \ldots, k$, $j = 1, 2$) using

$$\tau(j|\boldsymbol{x}_g) = \frac{1}{n_g} \sum_{h=1}^{n_g} \widetilde{\tau}(j|x_{i_{gh}})$$

where $\widetilde{\tau}(j|x) = af_j(x)/f(x)$.

(2) S-step: Draw $z(\boldsymbol{x}_g)$ randomly from a Bernoulli distribution with probability of $\tau(1|\boldsymbol{x}_g)$ and define $z(\boldsymbol{x}_{i_{gh}}) = 1$ if $z(\boldsymbol{x}_g) = 1$ and $z(\boldsymbol{x}_{i_{gh}}) = 1$ otherwise for $g = 1, 2, \ldots, k$ and $h = 1, 2, \ldots, n_g$.

(3) M-step: The estimator of a is given by

$$\hat{a} = \frac{1}{n} \sum_{i=1}^{n} z(x_i).$$

f_1 is estimated by maximizing the likelihood described in Appendix 7 based on x_i for $i \in \{i;\ z(x_i) = 1\}$.

These steps are repeated until convergence. For the log spline density f_1, the maximum likelihood estimation is applied. The data-driven knot locations in f_1 are optimized as described in Appendix 7. We use the k-means method to initialize a and f_1. From the observations that are separated by the k-means method, those having a larger mean are used to initialize f_1. It is possible that the stochastic EM algorithm may not converge but be stable (Bordes et al., 2007; Celeux & Diebolt, 1992). Therefore, we use a large number of iterations so as to stabilize the estimate of f. We then select $\hat{f}$ as the final estimate, whose likelihood $\sum_{i=1}^{n} \log \hat{f}(x_i)$ is the maximum among the iterations.

B. Logspline density estimation

Let X be a random variable having a continuous and positive density function. The log density of X is modeled by

$$g(x) = \log(f(x)) = C(\boldsymbol{\beta}) + \beta_{01}x + \sum_{i=1}^{m} \beta_{1i}(x - r_i)_+^3,$$

where $\boldsymbol{\beta} = (\beta_{01}, \beta_{11}, \ldots, \beta_{1m})$ is a vector of coefficients, $C(\boldsymbol{\beta})$ is a normalized constant, r_{ji} are the knots, and $(a)_+ = \max(a, 0)$. Let $X_1, \ldots, X_n$ be independent random variables having the same distribution as X. The log-likelihood function corresponding to the logspline family is given by $\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} g(X_i)$. The maximum likelihood estimate $\hat{\boldsymbol{\beta}}$ is obtained by maximizing the log-likelihood function. This methodology was introduced by Stone (1990)

and the software was implemented by Kooperberg & Stone (1991). An ICA algorithm based on the logspline density estimation was initiated by Kawaguchi & Truong (2011).

The knot selection methodology involves initial knot placement, stepwise knot addition, stepwise knot deletion, and final model selection based on the information criterion. We set the initial knot placement to be the minimum, median, and maximum values of the distribution of data. At each addition step, we first find a good location for a new knot in each of the intervals $(L, r_1), (r_1, r_2), \ldots, (r_{K-1}, r_K), (r_K, U)$ determined by the existing knots r_1, $r_2, \ldots, r_K$ and some constants L and U. Let $X_{(1)}, \ldots, X_{(n)}$ be the data written in nondecreasing order. Set $l_1 = 0$ and $u_K = n$. Define l_i and u_i by

$$l_i = d_{\min} + \max\{j : 1 \le j \le n \text{ and } X_{(j)} \le r_i\}, \quad i = 2, \ldots, K$$

and

$$u_i = -d_{\min} + \max\{j : 1 \le j \le n \text{ and } X_{(j)} \ge r_i\}, \quad i = 1, \ldots, K-1,$$

where $d_{\min}$ is the minimum distance between consecutive knots in order statistics.

For $i = 0, \ldots, K$ and for the model with X_{j_i} as a new knot where $j_i = [(l_i + u_i)/2]$ with $[x]$ being the integer part of x, we compute the Rao statistics R_i defined by

$$R_i = \frac{[\boldsymbol{S}(\hat{\boldsymbol{\beta}})]_i}{\sqrt{[\boldsymbol{I}^{-1}(\hat{\boldsymbol{\beta}})]_{ii}}},$$

where $\boldsymbol{S}(\hat{\boldsymbol{\beta}})$ is the score function, that is, the vector with entries $\partial \ell(\hat{\boldsymbol{\beta}})/\partial \beta_j$, and $\boldsymbol{I}(\hat{\boldsymbol{\beta}})$ is the matrix whose entry in row j and column k is given by $-\partial^2 \ell(\hat{\boldsymbol{\beta}})/\partial \beta_j \partial \beta_k$. We place the potential new knot in the interval $[X_{l_{i^*}}, X_{u_{i^*}}]$ where $i^* = \operatorname{argmax} R_i$. Within this interval, we further optimize the location of the new knot. To do this, we proceed by computing the Rao statistics R_l for the model with $X_{(l)}$ as the knot with $l = [(l_{i^*} + j_{i^*})/2]$ and R_u for the model with $X_{(u)}$ as the knot with $u = [(j_{i^*} + u_{i^*})/2]$. If $R_{i^*} \ge R_l$ and $R_{i^*} \ge R_u$, we place the new knot at $X_{(i^*)}$; if $R_{i^*} < R_l$ and $R_l \ge R_u$, we continue searching for a knot location in the interval $[X_{(l_{i^*})}, X_{(j_{i^*})}]$; and if $R_{i^*} < R_u$ and $R_l < R_u$, we continue searching for a knot location in the interval $[X_{(j_{i^*})}, X_{(u_{i^*})}]$.

After a maximum number of knots $K_{\max} = \min(4n^{1/5}, n/4, N, 30)$, where N is the number of distinct X_i's, we continue with stepwise knot deletion. During knot deletion, we successively remove the knot that has minimum Wald statistics, defined by

$$W_i = \frac{\hat{\beta}_i}{\sqrt{[\boldsymbol{I}^{-1}(\hat{\boldsymbol{\beta}})]_{ii}}}$$

of the existing knots.

Among all the models that are fit during the sequence of knot addition and knot deletion, we choose the model that minimizes the Bayesian information criterion (BIC) defined by $BIC = -2\ell(\hat{\boldsymbol{\beta}}) + m \log(n)$.

8. References

Amari, S., Cichocki, A. & Yang, H. H. (1996). A new learning algorithm for blind signal separation, *Advances in Neural Information Processing Systems* 8: 757–763.

Bai, P., Shen, H., Huang, X. & Truong, Y. (2008). A Supervised Singular Value Decomposition for Independent Component Analysis of fMRI, *Statistica Sinica* 18: 1233–1252.

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

Bordes, L., Chauveau, D. & Vandekerkhove, P. (2007). A stochastic em algorithm for a semiparametric mixture model, *Comput. Stat. Data Anal.* 51: 5429–5443.

Calhoun, V. D. & Adali, T. (2006). Unmixing fmri with independent component analysis, *Engineering in Medicine and Biology Magazine, IEEE* 25: 79–90.

Calhoun, V. D., Adali, T., Pearlson, G. D. & Pekar, J. J. (2001). A method for making group inferences from functional mri data using independent component analysis, *Human Brain Mapping* 14: 140–151.

Cardoso, J. F. & Souloumiac, A. (1993). Blind beamforming for non gaussian signals, *IEE-Proc.-F* 140: 362–370.

Celeux, G. & Diebolt, J. (1992). A stochastic approximation type EM algorithm for the mixture problem, *Stochastics Stochastics Rep.* 41(1-2): 119–134.

Chen, A. & Bickel, P. J. (2006). Efficient independent component analysis, *Annals of Statistics* 34: 2825–2855.

Cordes, D. & Nandy, R. (2007). Independent component analysis in the presence of noise in fmri, *Magnetic Resonance Imaging* 25(9): 1237–1248.

Dempster, A., Laird, N. & Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 1–38.

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

Friston, K., Holmes, A., Worsley, K., Poline, J., Frith, C. & Frackowiak, R. (1995). Statistical parametric maps in functional imaging: A general linear approach, *Human Brain Mapping* 2: 189–210.

Hartvig, N. V. & Jensen, J. L. (2000). Spatial mixture modeling of fmri data, *Human Brain Mapping* 11: 233–248.

Hyvärinen, A., Karhunen, J. & Oja, E. (2001). *Independent Component Analysis*, John Wiley & Sons.

Hyvärinen, A. & Oja, E. (1997). A fast fixed point algorithm for independent component analysis, *Neural Computation* 9: 1483–1492.

Kawaguchi, A. & Truong, K. Y. (2011). Logspline independent component analysis, *Bulletin of Informatics and Cybernetics* 43: 83–94.

Kooperberg, C. & Stone, C. (1991). A study of logspline density estimation, *Computational Statistics & Data Analysis* 12(3): 327–347.

Lee, S., Shen, H., Truong, Y., Lewis, M. & Huang, X. (2011). Independent component analysis involving autocorrelated sources with an application to functional magnetic resonance imaging, *Journal of the American Statistical Association* 106(495): 1009–1024.

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

Neumann, J., von Cramon, D. Y. & Lohmann, G. (2008). Model-based clustering of meta-analytic functional imaging data, *Human Brain Mapping* 29(2): 177–192.

Stone, C. (1990). Large-sample inference for log-spline models, *The Annals of Statistics* 18(2): 717–741.

Stone, J. V., Porrill, J., Porter, N. R. & Wilkinson, I. D. (2002). Spatiotemporal independent component analysis of event-related fmri data using skewed probability density functions, *Neuroimage* 15: 407–421.