FuncICA for Time Series Pattern Discovery

Nishant Mehta* Alexander Gray†

Abstract

We introduce FuncICA, a new independent component analysis method for pattern discovery in inherently functional data, such as time series data. FuncICA can be considered an analog to functional principal component analysis, where instead of extracting components to minimize L_2 reconstruction error, we maximize independence of the components over the functional observations. We develop an algorithm for extracting independent component curves and offer a method for optimizing a smoothing parameter. Results for synthetic, gene expression, and event-related potential data indicate that FuncICA can recover well-known phenomena and improve classification accuracy, highlighting the utility of FuncICA for unsupervised learning in temporal data.

1 Introduction

Algorithmic treatments of independent component analysis (ICA) abound, but our explorations of this space have led to at least one conclusion: all currently existing ICA algorithms treat the problem of ICA in the primal. While the primal form can solve highly interesting problems like blind source separation and compression, a dual form of ICA opens to the door to a world of different, equally important problems. By ICA in the primal we mean that for a linear model $\mathbf{X} = \mathbf{AS}$, where the i^{th} row of $\mathbf{X}$ specifies the observations of a random variable X_i, and the j^{th} column of $\mathbf{X}$ represents the j^{th} joint observation of the joint variable X, ICA seeks a linear unmixing of the random variables $X_1, \ldots X_n$ such that the resulting variables $S_1, \ldots, S_n$ are statistically independent over their observations.

This formulation of ICA naturally lends itself to solving the blind source separation (BSS) problem. In this problem, we observe n signals that are independently and identically distributed (IID)[1], and the observed signals have been mixed (each observation at a time index is the same linear combination of the values of source signals at that time index). In BSS, our goal is to recover the statistically independent source signals.

In this work, we consider ICA in a dual representation. We still have precisely the same observed data $\mathbf{X}$, but we frame the problem as the mixing model $\mathbf{X}^T = \mathbf{AS}^T$ which implies that $\mathbf{A}^{-1}\mathbf{X}^T = S^T$.[2] In the primal, the rows of $\mathbf{A}^{-1}$ are the independent components, which by themselves are of little value except for knowing the inverse mixing coefficients; however, when they are left multiplied with the data, they yield source signals which likely are of value by themselves, such as a particular speaker's speech over some time period. In the dual, the rows of $\mathbf{A}^{-1}$ are time series, which in themselves, are of particular value, while the sources S^T are not of value without the data X^T. The implications of the independent components being time series in the dual is that the independent components represent time series whose inner product over the observations (i.e., their activation, or coefficient) is statistically independent. For instance, an independent component in the dual may be some time series pattern of stock value change that occurs in financial market data that is synchronized to a merger event, with a linear activation coefficient that varies depending on the market capitalization of the new company produced. Further, this temporal progression is statistically independent of the other independent components over the time series observations, and hence it appears to, in isolation, capture some underlying phenomenon of some system (in this case, a financial market).

The ICA in the dual just described is actually temporal ICA in the dual. Researchers in various communities, including neuroscience, psychology, and computer science, have explored another form of ICA, known as spatial ICA, which enables them to analyze functional magnetic resonance imaging (fMRI) data to understand the brain. In spatial ICA, each fMRI image is a random variable, and the voxel indices index the observations. This already appears to be quite similar to our dual representation of temporal ICA, and in fact a naïve implementation of the dual of spatial ICA for this data is equivalent to temporal ICA. For the dual of both temporal and spatial ICA, we will show

*Georgia Institute of Technology

†Georgia Institute of Technology

[1] By IID, we mean that each signal's observations at a given time are assumed to be independent of observations of that signal at other times, and each observation of a given signal is assumed to be drawn from an identical distribution.

[2] For now, assume that A is positive definite and hence invertible.

how, using functional[3] representations of data, we can exploit the smoothness properties of temporally and spatially sampled data respectively to perform noise reduction on our data that improves the utility of our recovered temporal and spatial patterns respectively. Our algorithm, FuncICA, finds precisely these smooth patterns with the level of smoothness optimized by a novel, information-theory-inspired objective that is empirically validated. After introducing FuncICA, our main result, and analyzing its performance, we will argue that both spatial, temporal, and something called spatiotemporal ICA are all not quite the right ICA methods to use for fMRI data. We will then show how, with small changes, FuncICA is the right ICA methodo to use for fMRI data analysis and it can be a powerful tool for spatiotemporal pattern discovery in this exciting domain.

In the rest of this work, we lay the groundwork for and derive an algorithm to compute the general dual ICA method described above, not just for a finite number of random variables, but for an infinite number of random variables, using a functional representation. By exploiting the continuity of time and space, we make this infinite representation tractable; further, by using an optimal regularization parameter, we smooth the observed data before extracting statistically independent patterns. We validate the regularization scheme on synthetic and real electroencephalography (EEG) data. Superior classification results are reported for FuncICA on microarray gene expression data and event-related potential (ERP) data from the brain, as compared to functional principal components analysis (FPCA) and the naïve dual of temporal ICA. Our results indicate that the algorithm produces components that can recover sources, improve classification accuracy by sparsely representing information, and effectively estimate the P300 event-related potential in the brain. Finally, we conclude with a forward-looking, novel framework for fMRI data analysis.

2 Background

Prior to introducing FuncICA, we formally introduce ICA and some functional analysis notation, and we describe functional PCA (FPCA) as an analog to FuncICA that finds uncorrelated but not necessarily statistically independent components.

2.1 ICA ICA seeks a linear transformation from statistically dependent to statistically independent components. The model can be expressed as an instantaneous linear mixing model

$$X^{(t)} = AS^{(t)}, \tag{2.1}$$

where[4] $t \in [T]$, $X^{(t)} \in \mathbf{R}^n$, $A \in \mathbf{R}^{n \times n}$, and $S^{(t)} \in \mathbf{R}^n$. The source signal distributions of $S_1, \ldots, S_n$ realized at indices $[T]$ are statistically independent, but the mixing matrix A clearly erases the independence of the distributions of the observed signals $X_1, \ldots, X_n$. The task is then to estimate the unmixing matrix $W = A^{-1}$ such that

$$Y = WX = WAS. \tag{2.2}$$

It is known that we can recover S only up to a scaling and permutation.

It can be shown that maximizing independence of the source distributions is equivalent to minimizing the difference between the joint density of the sources and the product of the marginal densities. A natural objective function to measure this difference is the Kullback-Leibler (KL) divergence:

$$\begin{aligned} \mathcal{H}(Y) &= D_{\mathrm{KL}}(p(Y) \| \prod_{i=1}^{n} p(Y_i)) \\ &= \sum_{i=1}^{n} H(Y_i) - H(Y), \end{aligned} \tag{2.3}$$

where $H(Y)$ is the entropy of Y. Notably, the KL divergence of two distributions P and Q is 0 iff all of the Y_i are statistically independent, so minimizing this objective function is firmly grounded. Before describe functional PCA, some notation and basis functional analysis is required.

2.2 Functional representation Let $t \in [0, T]$ and X be a set of n functionals $\{X_1(t), \ldots, X_n(t)\}$, where $X_i(t) \in L^2$. L^2 is a Hilbert space with inner product defined as the integral of the product of two functionals:

$$\langle f, g \rangle = \int f(t)g(t)dt.$$

If we let $\boldsymbol{\beta}(t)$ be a set of m basis functions $\{\beta_1(t), \ldots, \beta_m(t)\}$, then we can represent each functional $X_i(t)$ as a linear combination of the basis functions

$$X_i(t) = \sum_{j=1}^{m} \phi_{i,j}\beta_j(t).$$

The choice of basis is a non-trivial task. For our analysis we have used a flexible cubic b-spline basis,

[3] This use of the word functional refers to the function spaces notion, not to be confused with the functional in fMRI

[4] Here, we adopt the notation $[k] = \{1, 2, \ldots, k\}$.

so that each $\beta_j(t) \in \mathcal{C}_0^2(\mathbf{R})$.[5] Using a nonparametric basis may offer a representation that captures functional variability in the right places, but the simplicity of our choice of basis makes the analysis cleaner and also is very computationally efficient. We will return to this point in the conclusion.

Having established the necessary functional notation, we describe a fuctional PCA, which we will then show to be a close analog and the first step of FuncICA.

2.3 Functional PCA FPCA, introduced by Ramsay and Silverman [1, 2], is in its unregularized form equivalent to the Karhunen-Loève expansion. Ramsay and Silverman have developed an optimally smoothed FPCA to minimize reconstruction error; we discuss one method in Section 4. Via the derivation of FPCA that follows, we hope to convey that FPCA is conceptually equivalent to PCA in a basis function space. This intepretation will ease understanding of FuncICA as well.

For a functional random variable $X(t)$, consider the covariance function

$$\Gamma(s,t) = E[X(s)X(t)]. \tag{2.4}$$

We decompose $\Gamma(s,t)$ in terms of eigenfunctions γ_j:

$$\Gamma(s,t) = \sum_{i=1}^{\infty} \lambda_j \gamma_j(s)\gamma_j(t),$$

where λ_j is the j^{th} eigenvalue of the covariance function. Estimating the sample eigenfunctions $\hat{\gamma}_j(s)$ is equivalent to the following optimization problem:

$$\begin{aligned} \text{maximize} \quad & \int\int \gamma_j(s)\hat{\Gamma}(s,t)\gamma_j(t)\, ds\, dt, \\ \text{subject to} \quad & \int \gamma_j(t)^2 dt = 1 \\ & \int \gamma_i(t)\gamma_j(t)dt = 0 \qquad \text{for } i < j. \end{aligned}$$

If we let f be a vector of coefficients for the basis expansion of $\xi(t)$ in terms of the basis functions $\beta_j(t)$, then

$$\xi(t) = \sum_{j=1}^{m} f_j \beta_j(t).$$

Further, for A being a coefficients matrix for the basis functions over the data, we have

$$X_i(t) = \sum_{j=1}^{m} A_{i,j}\beta_j(t).$$

[5] $\mathcal{C}_0^2(\mathbf{R})$ is the class of functions of $\mathbf{R}$ that have compact support and are continuous up to the second derivative.

Let L be the basis function inner product matrix such that

$$L_{i,j} = \int \beta_i(t)\beta_j(t)dt.$$

Then the principal component score for $X_i(t)$ is

$$\int \xi(t)X_i(t)dt = A_{(i)}L^T f,$$

where $A_{(i)}$ denotes the i^{th} row of A. The PC scores for all of the functional data is then $AL^T f$.

Let $V_{j,k} = \frac{1}{n-1}\sum_{i=1}^{n}(A_{i,j} - \bar{A}_j)(A_{i,k} - \bar{A}_k)$ induce the sample variance matrix of the basis coefficients. Our objective is to maximize with respect to f the objective function $f^T LVL^T f$.

In this form, f is the leading eigenvector. We can estimate the coefficients for the i^{th} eigenfunction by using the Gram-Schmidt process and formulating the objective function using the projected data.

3 FuncICA

FuncICA derivation. A functional version of ICA is equivalent to the original ICA formulation with A now being a Hilbert transform:

$$A : L^2 \mapsto L^2, \tag{3.5}$$

because both the functional observations $X_i(t)$ and the independent functionals S_j are in L^2. This Hilbert operator is somewhat like the analog to the A matrix in ICA, if A were permitted to take on an infinite number of rows and columns. As in the case of FPCA, to gain tractability we consider each functional observation $X_i(t)$ by its expansion in terms of basis functions $\beta_j(t)$, such that

$$X_i(t) = \sum_{j=1}^{m} \psi_{i,j}\beta_j(t). \tag{3.6}$$

To compactly represent this information, let β be a linear map

$$\beta : \mathbf{R}^m \mapsto L^2.$$

To ease the understanding of this operator, we treat it as an analog to a matrix with an infinite number of columns. The j^{th} row is then the functional $\beta_j(t)$. For clarity, if $A \in \mathbf{R}^n \times L^2$ and $B \in \mathbf{R}^m \times L^2$, then each row of the matrices corresponds to a functional, and hence $AB^T = C$ such that

$$C_{i,j} = \int A_i(t)B_j(t)dt,$$

where $A_i(t)$ is the i^{th} row of A and $B_j(t)$ is the j^{th} row of B.

Using this notation, $X = \psi\beta$, where each row of X is a functional observation and $\psi \in \mathbf{R}^{n \text{ x } m}$. The principal component matrix is

$$E = U^T X = U^T \psi\beta = \rho\beta, \qquad \text{for } \rho \in \mathbf{R}^{m \text{ x } m}. \tag{3.7}$$

For some Hilbert transform $Z : L^2 \mapsto L^2$, let σ_Z denote the score matrix of Z over the data X. The principal component score matrix is then

$$\sigma_E = EX^T = \rho\beta(\psi\beta)^T = \rho\beta\beta^T\psi^T. \tag{3.8}$$

We can then apply a left linear transformation W to the principal component matrix to obtain some

$$Y = WE = \phi\beta, \tag{3.9}$$

where $\phi = W\rho$. The corresponding score matrix for the functionals in Y is then

$$\sigma_Y = YX^T = WEX^T = W\sigma_E = \phi\beta\beta^T\psi^T. \tag{3.10}$$

If we let σ_Y denote the target score matrix where independence of the marginal distributions of Y_i over the data X is maximized, then the problem of ICA reduces to the familiar setting of optimizing an independence objective function with respect to a finite matrix of parameters W (applied to σ_E).

3.1 Algorithm To optimize W, recall our objective function 2.3. After performing FPCA, we are left with

$$\mathcal{H}(E) = \sum_{i=1}^{n} H(E_i) - H(E). \tag{3.11}$$

Considering rotations of the functional data using the dual orthogonal decomposition provided by E, we consider left rotations W (such that $Y = WE$) that maintain uncorrelatedness of the distributions of the E_i while minimizing $\mathcal{H}(Y)$. Reformulating $\mathcal{H}(Y)$ in terms of Y,W, and E, we have

$$\mathcal{H}(Y) = \textstyle\sum_{i=1}^{n} H(Y_i) - H(Y) = \sum_{i=1}^{n} H(Y_i) - H(WE),$$

which by Theorem 9.6.4 of Cover and Thomas [3] is equal to

$$\sum_{i=1}^{n} H(Y_i) - H(E) - \log(|W|). \tag{3.12}$$

To maintain the uncorrelatedness of the functionals, we restrict W to the set of rotation matrices. Because rotation matrices have unity determinant, it is true that $\log(|W|) = 0$. Discarding the fixed $H(E)$ term, we are left with the new objective function

$$\mathcal{H}^*(Y) = \sum_{i=1}^{n} H(Y_i). \tag{3.13}$$

Employing this objective is equivalent to direct entropy minimization, for which many existing ICA algorithms can be applied. Because of our unique domain of finding an optimal rotation of basis function coefficients, and also due to convincing source separation results for a variety of mixtures, we use the RADICAL ICA algorithm for optimizing W. It is beyond the scope of the paper to give a thorough treatment of RADICAL. The interested reader is encouraged to read [4] for details[6], but we nevertheless provide a brief overview. The core idea of RADICAL is consider pairwise rotations of (Y_i, Y_j) that minimize $H_i + H_j$; hence, the optimization directly minimizes the sum of marginal entropies of pairwise source estimates. In order to estimate the entropy of some Y_i, one could use a nonparametric method that exploits the order statistics of univariate data to estimate entropy consistently [6]. RADICAL employs a modified, more robust version of this estimator.

This work would be far from complete without special considerations for the functional data. A caveat of source separation here is that the distributions estimated are heavily dependent upon the choice of functionals obtained by applying FPCA. Varying the level of smoothing appropriately can provide independent functionals that both have lower marginal entropies and better approximate the source distributions of interest.

4 Optimal Smoothing

Individual functional observations often exhibit a large amount of spiky behavior due to observation error and small variations in the underlying stochastic processes, but the underlying processes of interest often evolve smoothly. This nuisance roughness motivates a method for incorporating a smoothing method into the function extraction. Numerous methods exist for controlling roughness in extracted curves; the method employed here follows a quite general framework used by Ramsay and Silverman [2]. Instead of choosing eigenfunctions with the constraint

$$\int \xi(t)^2 dt = 1,$$

we impose the alternate, roughness-penalizing constraint

$$\int \xi(t)^2 dt + \alpha \int (L\xi(t))^2 dt = 1, \tag{4.14}$$

where $\alpha \geq 0$ and L is a linear combination of differential operators such that

$$L\xi = \sum_{i=0}^{m} a_i D^i \xi. \tag{4.15}$$

[6] RADICAL is competitive with other recent algorithms [5], so our analysis is restricted to this ICA algorithm.

Note that the goodness of a particular penalty operator L is likely domain-dependent. One simple choice of L that we adopt for our experiments is $L = D^2$, a penalty on the norm of the second derivative. This choice yields the new constraint

$$\int \xi(t)^2 dt + \alpha \int (D^2 \xi(t))^2 dt = 1, \tag{4.16}$$

for $\alpha \geq 0$. Optimal selection of L is also a worthy problem to tackle; however, our results suggest that D^2 is sufficient for the data tested here.

4.1 Objective for α optimization A key inquiry in this work is to identify cases where it is beneficial to choose a smoothing parameter α other than the L_2 optimal reconstruction error parameter α_P^*. α_P^* can be selected by minimizing reconstruction error via leave-one-out cross-validation (LOOCV), where in each epoch one curve is the test set and the remaining curves are the training set [1]. Because our goal is to extract independent functions that maximally depart from Gaussianity (minimal marginal entropy), we conjecture that smoothing is beneficial only when Gaussian sources exist in the data and those sources are sufficiently rougher than the non-Gaussian sources. An example of this would be a high frequency harmonic source with amplitude normally distributed over the observed functionals. The essence of our approach is to assign an FuncICA result (a set of independent functionals) a score that penalizes the result increasingly as the sources become more Gaussian.

The *negentropy* of a unit-variance random variable Y_i be defined as

$$J(Y_i) = H(\mathcal{N}(0,1)) - H(Y_i). \tag{4.17}$$

Note that $J(Y_i)$ is non-negative because for a fixed variance the Gaussian distribution has maximal entropy [7].

Our objective function is then

$$Q(Y) = \sum_{i=1}^{p} \frac{1}{J(Y_i)}. \tag{4.18}$$

The algorithm for optimizing α is shown in Algorithm 1. The idea is to optimize Q starting at α_P^*, the optimal α chosen by FPCA, and to slowly increase α as long as Q decreases. Choice of the parameter γ (for $\gamma > 1$) can be decided by an optimizer.

We will present results of this algorithm on synthetic and real-world brain data.

Algorithm 1

Minimize FPCA LOOCV error to find α_P^*.
$Q^{(0)} = \infty$
$\alpha^{(1)} = \alpha_P^*$
$\tau = 0$
repeat
  $\tau = \tau + 1$
  $Y = \text{FuncICA}(X, \alpha^{(\tau)})$
  $Q^{(\tau)} = \sum_{i=1}^{p} \frac{1}{J(Y_i)}$
  $\alpha^{(\tau+1)} = \gamma \cdot \alpha^{(\tau)}$
until $Q^{(\tau)} > Q^{(\tau-1)}$
return $\alpha^{(\tau-1)}$

5 Previous Temporal ICA Methods

We briefly overview some of the previous work with primal temporal ICA methods that exploit temporal dependence. Two important points in this section are:

1. These methods solve a problem different than the one solved by FuncICA.
2. The methods make use of temporal dependence in a way that is inconsistent with the problem solved by FuncICA.

Convolutive mixing is a variation of the instantaneous linear mixing model, where the goal is to isolate an audio source signal from echoey microphone recordings. A closely related problem is source separation where the sources exhibit temporal structure; a simple example would be if a source signal had significant autocorrelation and the cross-correlation among the source signals, by virtue of their independence, is zero[8, 9]. The family of methods for solving these problems attempt to recover independent components whose activations vary independently over time, and hence the temporal structure of the problem can be used by minimizing the cross-correlation between the observed signals. In the dual, the independent components are time series and the goal is to attain statistical independence of their activation over the functional observations. In this setting, the cross-correlation between signals is not of interest, because the independent components themselves are unconstrained; it is only their activation over the functional data that is optimized to be statistically independent.

A few ICA methods attempt to recover dominant independent components, with a goal similar to our goal of discovering time series patterns. A common strategy here is to first do ICA on whitened data that has *not* been dimensionality-reduced, and then to use some properties of the components to select the most dominant components. For n signals, this becomes intractable in the primal formulation, due to $O(n^3)$ scaling properties for ICA algorithms. Additionally, by using the primal interpretation of statistical independence, the recovered time series patterns are not guaranteed to have statistically independent activation over the time

series observations.

In contrast, FuncICA's computational complexity is limited by the number of basis functions it uses to represent each observed time series, rather than the number of observed time series. Further, by using FPCA prior to optimizing the independence objective, the dimensionality of the problem is further reduced. As a result, the algorithm's computational complexity is linear in n, and we are able to show results of FuncICA for datasets containing a very high number of signals; our synthetic dataset consists of 10^4 time series in which we mine for independent component curves. Hence, from a practical computability standpoint, we can flexibly increase computability by reducing the number of basis functions used in our representation or being more aggressive in the FPCA dimensionality reduction step[7].

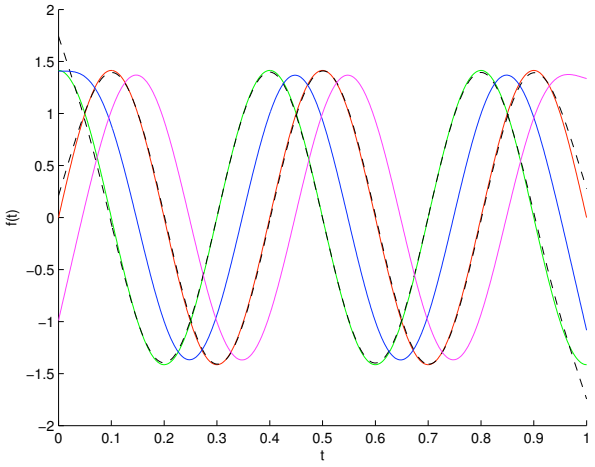

Figure 1: $S_1(t)$ and $S_2(t)$ are shown in red and green respectively. $IC_1(t)$ and $IC_2(t)$ are the black dotted lines. $PC_1(t)$ and $PC_2(t)$ are the blue and magenta lines respectively. The independent functionals are near-perfect replicas of the sources, while the principal component functions are phase-shifted from the sources.

6 Experiments and Results

We tested FuncICA's performance on synthetic data and two real-world data sets. Our goal is to demonstrate that FuncICA is capable of extracting highly interesting components that sparsely encode the most relevant information for functional data sets, and further to constrast the performance of FuncICA with FPCA. We first discuss results on synthetic data as validation of the independent curve extraction and smoothing methods. We then present scientifically encouraging results for feature extraction and classification in EEG and gene expression data.

6.1 Synthetic data For our experiments with synthetic data, we created 2 independent functions by using orthogonal functions with unit norm. These consisted of

$$S_1(t) = \frac{1}{\sqrt{2}}\sin(10\pi t) \qquad \text{and} \qquad S_2(t) = \frac{1}{\sqrt{2}}\cos(10\pi t),$$

sampled at 1000 points with uniform spacing. We generated 10^4 realizations of 2 IID Laplace random variables Z_1 and Z_2 with zero mean and unit variance ($Z_i \sim \text{Laplace}(0, \frac{1}{\sqrt{2}})$). We mixed the sources to form 10^4 observations X_i for $i \in [10^4]$:

$$X_i = Z_{1,i} \cdot S_1 + Z_{2,i} \cdot S_2, \tag{6.19}$$

where $Z_{i,j}$ represents the j^{th} realization of Z_i.

We ran FuncICA on multiple data sets generated as described above; 120 cubic-bspline basis functions were used, and α was set to zero since no observation noise or other roughness was added to the data.

In order to rigorously evaluate source recovery performance, we define a measure of accuracy as the minimum inner product among the source functionals' inner products with their best fit independent functionals; more formally, accuracy is

$$\min\{\langle S_1, (\arg\max_{Y_i}\langle S_1, Y_i\rangle)\rangle, \langle S_2, (\arg\max_{Y_j}\langle S_2, Y_j\rangle)\rangle\}.$$

As shown by a typical recovery result in Figure 1, FuncICA achieved near-perfect source recovery , while FPCA nearly always recovered a phase-shifted version of the sources. The mean accuracy of ICAFunc over 5 realizations of the synthetic data was 0.9999, while FPCA's mean accuracy over the same 5 realizations was only 0.9051. The reason for the discrepancy is that FPCA tends to select more Gaussian source distributions as a result of minimizing reconstruction error. This is evidenced from the entropy estimates (using Vasicek's estimator) for the 2 PC distributions being 1.3963 and 1.3610 respectively, while entropy estimates for the two IC distributions were lower at 1.3647 and 1.3296 respectively, yielding a difference of the sums of the marginal entropies of 0.0629.

Under two Laplace distributions mixed with an additive high frequency Gaussian component - that is, a function form of $\sin(40\pi t)$ with linear activation

[7] Using FPCA for dimensionality reduction is preferred. Choice of the number of basis functions appears subjective, while FPCA minimizes a clear objective

through the data varying as $\mathcal{N}(0, \sigma)$ - FuncICA and FPCA perform very differently when $\alpha = 0$. While FuncICA separates the Laplace-distributed sources well, FPCA extracts 4 functionals that all have significant high frequency harmonics. As α was increased, the high frequency harmonic component was driven into the 3$^{\text{rd}}$ principal component (PC) curve, with negligible change to the extracted IC curves. FPCA therefore seems to be more sensitive to regularization in cases where the components exhibit variable amounts of smoothness that can inform the separation process, while FuncICA requires no regularization here.

Additionally, FuncICA was able to extract the Laplacian-distributed components from mixtures of two high-frequency Gaussian sources and two Laplacian sources. The two Laplacian sources are as described before, and the two Gaussian sources are $\sin(40\pi t)$ and $\cos(40\pi t)$, again distributed according to $\mathcal{N}(0, \sigma)$. We applied Algorithm S to automatically find an optimal value for the regularization parameter by increasing α, starting from $\alpha_P^* = 0$ in this case, until Q reaches a minima. Figure 2 shows the effect of smoothing on Q and accuracy.

The left plot in Figure 2 shows that Q reaches a minimum at $\alpha_I^* = 1 \cdot 10^{-9}$. Looking at the right plot, the accuracy measure is quite close to the optimal value for this choice of α, and for a small window around α_I^*, accuracy remains high. Significant departure from α_I^* results in large drops in accuracy. The intuition behind this result is that higher values of Q indicate independent functionals that are more Gaussian distributed, and the rest follows from the well known result that Gaussian sources impede source recovery of non-Gaussian sources. Minimizing Q serves to dampen the high frequency Gaussian components and impede their recovery, but we are careful to stop the search for the optimal α once Q hits a local minimum, because we otherwise risk finding another local minimum at a higher α where the non-Gaussian sources are also dampened. Although not shown in the plots, another minima of Q occurs at $\alpha = 1 \cdot 10^{-5}$, but the accuracy there is substantially lower (close to 0.92) because the non-Gaussian sources $S_1(t)$ and $S_2(t)$ have been dampened. Having validated Algorithm S for optimization of α_I^* on synthetic data, we progress to real-world results.

Microarray gene expression data.

Previous work in the domain of temporal gene expression indicates that such patterns of activation exist [10], but thus far, FPCA is the core functional method that has been applied to this data [11]. Although this work focuses on temporal observations, it is applicable to any problems with discrete samples of a continuous process (e.g. spatial data).

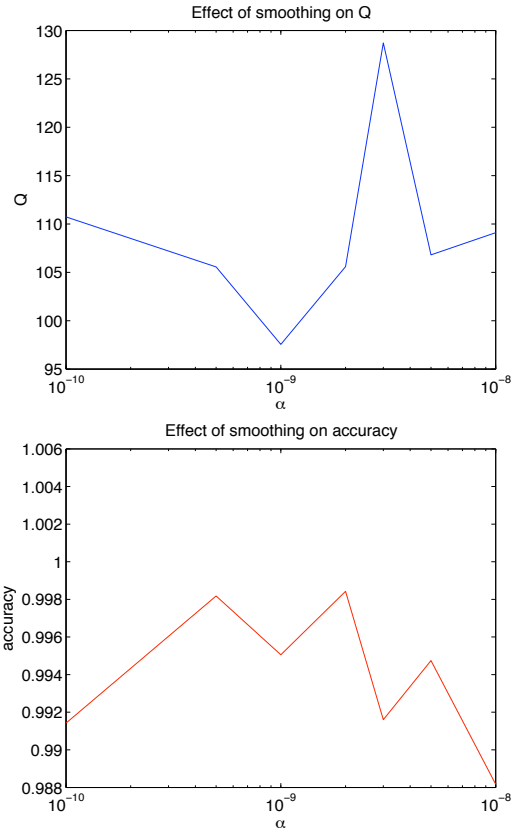

Figure 2: At left, the regularization parameter α plotted versus Q. Note the minimum at $\alpha = 1 \cdot 10^{-9}$ and the spurious maximum at $\alpha = 3 \cdot 10^{-9}$. The right plot indicates that accuracy is high in close proximity to $\alpha = 1 \cdot 10^{-9}$, with sharp drops as α increases above $2 \cdot 10^{-9}$ or decreases below $5 \cdot 10^{-9}$.

The microarray data obtained by [12] is of α-factor[8] synchronized temporal gene expression for the yeast cell cycle, with observations of 6178 genes at 18 times in 7 min. increments. Of interest in this domain is identification of co-regulated genes related to specific cell cycle phases. Among the phases G_1, S, S/G_2, G_2/M, and M/G_1, we focus on discriminating between genes related to G_1 versus non-G_1 phase regulation. The motivation for this comparison is in part due to [11], who present classification results using FPCA.

Notably, Liebermeister [10] explored gene expression with this data set using classical ICA. Because Liebermeister did not use a functional representation, the analysis was able to consider each gene's expression (at discrete time indices) over multiple cell cycles, with each cell cycle being synchronized by a different method. By contrast, our functional representation currently prohibits use of multiple curves per observation, and as a result we (in agreement with [11]) restrict our analysis to cell cycle gene expression synchronized by α-factor mating pheromone. Though this outcome may seem less preferable, a functional representation facilitates natural methods for incorporating genes that have missing observations at a few time indices. Use of a nonzero α is not necessary here, because the sparse temporal sampling of the gene expression process causes the spline-fitting stage to eliminates most of the roughness Algorithm S would otherwise remove.

For the discrimination task, we ran FuncICA on data from a 5672-gene subset corresponding to all genes where the α-factor synchronized data is missing at most one value. We tested the relevance of the extracted IC curves to a classification task involving 48 genes related to G1 phase and 50 genes related to non-G1 phases (identified by traditional methods by [12]) that have data available and are missing no more than one time-indexed observation. Missing values are handled automatically by the cubic b-spline fitting process. We measure the quality of a set of features according to performance of a support vector machine (SVM) [13] with Gaussian kernel, and we use leave-one-out cross-validation to select both a suitable SVM-regularization parameter and kernel bandwidth.

SVM experiments were run using all 11 PC curve score features versus all 11 IC curve features. The PC features resulted in 9.2% error, while the IC features yielded 7.1% error. Removing 2 features determined to be detrimental from the IC feature set (IC curves 5 and 10) reduced error to 6.1%, whereas removing the last 3 PC curves yielded error of 8.2%. Unfortunately, ICA does not have a natural solution for missing data imputation and so we cannot provide results comparing FuncICA to standard ICA (RADICAL in this case) for this problem. Our conjectures for these results are that ICA extracts more meaningful features that are more easily exploited by the SVM classifier.

P300 event-related potential data. The EEG data set obtained from [14] includes 15300 1-second observations, where for each time index in a given observation there are 64 voltage values corresponding to scalp electrical activity recorded during a brain-computer interface (BCI) experimental paradigm. During the experiment, a subject is attending to a particular letter on a screen containing a 6 by 6 grid of letters and numbers. Each row and column is flashed once in a random sequence, and the 1 second signal observations correspond to the data recorded after each flash. The BCI community is interested in identifying the row and column (the letter) to which a subject is attending. Much previous work has characterized an event-related potential (ERP) known as the P300, due to its signature appearance as a positivity in the EEG approximately 300 ms concluding the presentation of an attended stimulus. In Figure 4, the empirical P300 waveform for this data set is shown; it was calculated by taking the mean of the trials in which the P300 is known to activate. While the P300 ERP exists in most if not all subjects, the actual timing and waveform exhibited vary to some degree by subject and various physiological factors affecting attention. Therefore, an automated process for estimating the waveform from an unlabeled set of trials will prove highly useful for classification problems where we wish to pick out the trials that flashes an attended letter.

The problem is to decide which letter was attended to, given 15 trials of each row and column flashing. Classification ideally takes place from 1 trial of each row and column flashing, but in practice the low SNR makes a larger number of trials necessary. All of our experiments are done from recordings at the electrode Pz, with a common average reference (CAR) filter which instantaneously adjusts the value of the signal at an electrode by the average of all 64 channels. We did not apply any further spatial nor frequency filtering. The result is a data set that is raw data with the exception of the CAR filter; poor results are obtained from EEG data without considerable filtering. Our classification accuracy rates demonstrate recovery of the P300 waveform under different choices for the parameter α. The classifier was simply to pick the row/column that had the highest mean activation for the used feature. The results for the best component

[8] An unfortunate intersection of notation has occurred. Note that α-factor refers to a mating pheromone within biology, which is completely separate from our α which refers to a regularization parameter.

	FuncICA	ICA	FPCA	Empirical P300
column	**64.3%**	35.7%	37.5%	35.7%
row	**55.4%**	48.2%	37.5%	46.4%
letter	**30.4%**	17.9%	16.1%	17.9%

Figure 3: P300 task accuracy for FuncICA, FPCA, ICA, and empirical P300 waveform

from FuncICA, FPCA, and ICA (using 256 dimensional data reduced by PCA to 12 dimensions) are summarized in Figure 3. The best results for a single IC curve were obtained with $\alpha = 10^{-7}$, with column, row, and letter classification accuracy of 64.3%, 55.4%, and 30.4% respectively. Note that letter classification requires correct row and column classification for 15 trials where the attended letter is constant. Figure 4 illustrates the IC curve most similar to the P300 for 2 values of α.

In Figure 6 we demonstrate the success of Algorithm S for this data set. The value $\alpha_I^* = 7.5 \cdot 10^{-8}$ returned by the algorithm corresponds to a neighborhood of α values that provide high accuracy. The accuracy at α_I^* itself is quite close to better accuracy values in its neighborhood. We interpret these results as validation of our algorithm, though we also believe that providing more formal guarantees is highly desirable. To contrast FuncICA with FPCA, the best results for a single PC curve where substantially lower than the results for a single IC curve. As shown by Figure 4, this result is hardly surprising as the curves extracted by FPCA are uninteresting harmonics. This result makes sense due to the dominance of harmonics within EEG, but a L_2-loss minimizing representation is less useful for identifying specific ERPs. Particularly interesting is the performance of ICA for this problem. As shown in Figure 3, ICA performed quite poorly compared to FuncICA. We infer that the reason for this performance is due to a lack of smoothing used by ICA to extract its components; this inference is partially validated by the fact that FuncICA performs optimally on this problem with nonzero choice of α. The most surprising result is that using the empirical P300 waveform as the component for classification results in substantially lower classification accuracy than using the best IC curve. A possible explanation for this result is the use of the empirical mean without employing any functional smoothing techniques. Nevertheless, the implications for the result is that a technique such as FuncICA better captures the underlying event-related potential of interest when humans are presented with unexpected stimuli.

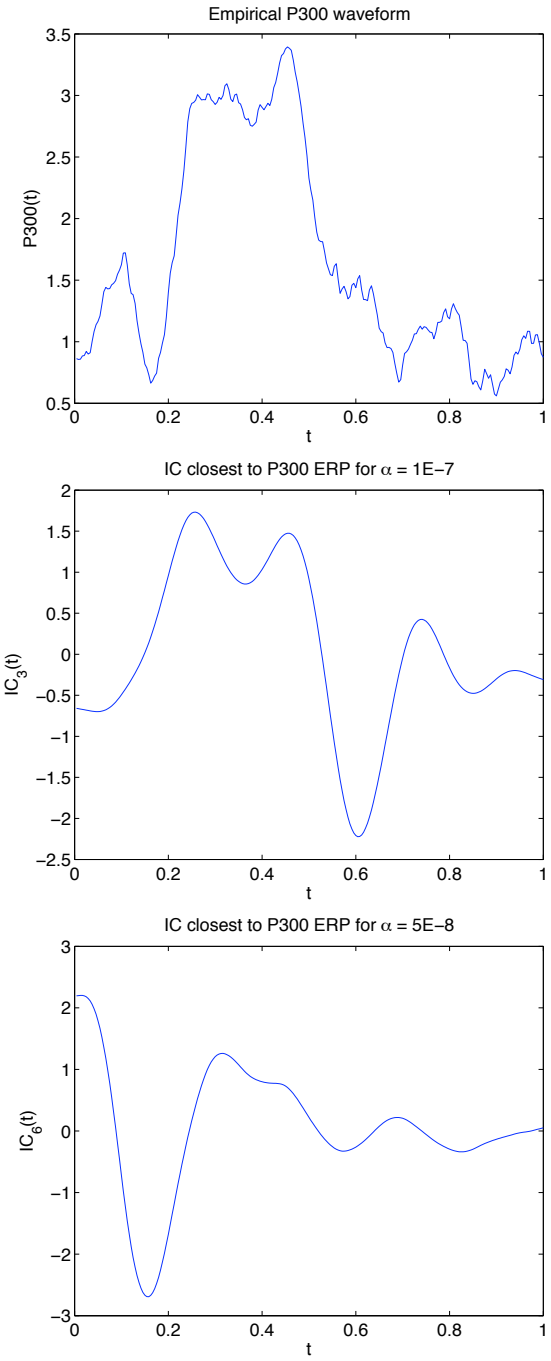

Figure 4: Top to bottom. (1) Empirical P300 waveform calculated from 2550 trials. (2) Closest IC to P300 for $\alpha = 5 \cdot 10^{-5}$. (3) Closest IC to P300 for $\alpha = 1 \cdot 10^{-7}$.

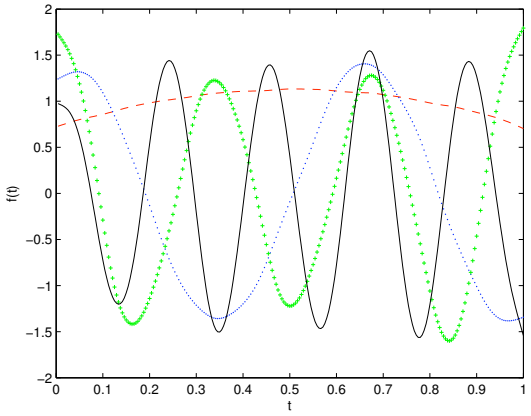

Figure 5: PC curves 1, 4, 7, and 10 plotted as a dashed red line, blue dotted line, green dashes line, and solid black line respectively - Note that the components are almost entirely described by harmonic behavior.

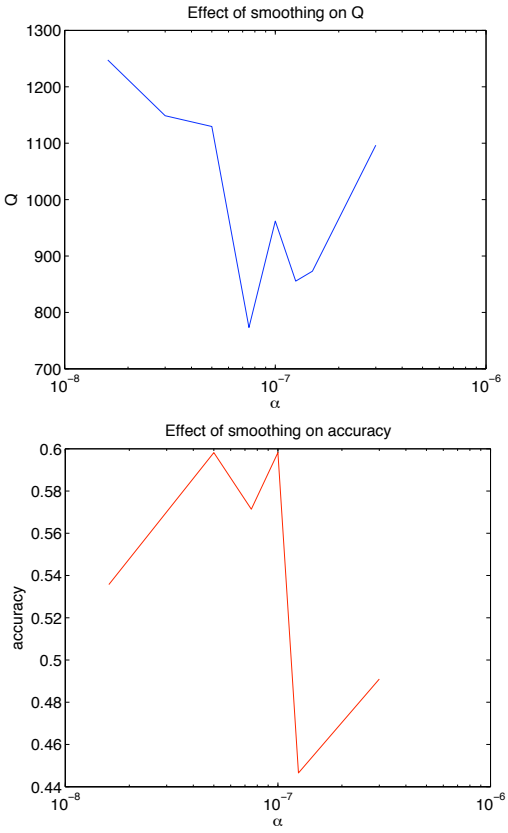

Figure 6: At left, the regularization parameter α plotted versus Q. The local minimum encountered when moving leftward from $\alpha_P^* \approx 1.6 \cdot 10^{-8}$ (as per Algorithm S) occurs at $\alpha_I^* = 7.5 * 10^{-8}$. This value is very close to the optimal neighborhood for accuracy shown in the right plot.

7 Extension to fMRI data

We have shown promising results of the dual of temporal ICA for time series data; also of interest is spatial data, and beyond that, spatiotemporal data. Functional magnetic resonance imaging (fMRI) data consists of three-dimensional spatial recordings of neural activations, with as many as 5 such recordings a second when the number of acquired slices is limited (i.e., one of the spatial dimensions is limited). With the advent of fMRI recording that has not only high spatial resolution but also decent temporal resolution, spatial ICA becomes less preferable when mining for complex, spatiotemporal neural activation patterns.

We earlier mentioned that the dual of temporal ICA involves finding time series patterns whose variation over the observed time series is statistically independent. The dual of spatial ICA similarly involves finding spatial patterns whose variation over the observed spatial maps is statistically independent. For the case of fMRI, it seems natural that, given the spatial and temporal nature of the data, we should seek both statistical independence in the primal temporal ICA sense (signals that vary statistically independently over time) and the primal spatial ICA sense (spatial maps that vary statistically independently over space). Stone explored precisely just such a method by creating spatiotemporal ICA [15], which seeks to simultaneously seek statistical independence in both the temporal and spatial sense via a convex combination of their respective objectives.

A generalization of FuncICA to spatiotemporal data results in a powerful alternative method to spatiotemporal ICA, which we refer to as the dual of spatiotemporal ICA. Rather than minimizing a convex combination of the spatial and temporal dependence measure objectives, we instead consider the video of the brain recorded after presentation of some stimulus to be a single observation. Given n presentations of stimuli, we then have n observations, each of which is a video representing spatiotemporal neural activation in response to a stimulus. The generalization needed from our previous description of FuncICA is a spline, or other basis function, representation that can model a three-dimesional spaptial pattern varying over time. Given such basis functions, and an alternate notion of regularization that penalizes roughness both over space and time, the FuncICA algorithm outlined in the previous sections directly can be used to mine fMRI data for spatiotemporal patterns of activation.

8 Conclusions

We have introduced the first algorithm for ICA on inherently functional data. After proposing an algorithm for optimal smoothing of the extracted IC curves, we validated this method on synthetic data and real-world gene expression and EEG data. FuncICA offers a principled method for discovering IC curves in what is a very undercomplete problem. In comparisons with FPCA, we have shown that FuncICA extracts patterns that are not only quite different from PC curves, but also are capable of describing interesting phenomena. FuncICA is capable of dampening high-frequency curves that vary according to the Gaussian distribution, and the algorithm also is relatively insensitive to instantaneous observation noise.

The algorithm efficiently encodes information via IC curves that provide superior classification performance for identifying covarying sets of genes. For EEG data, FuncICA can extract the P300 event-related potential using unlabeled trials; previous work requires labeled trials to identify the P300. The method of ICA introduced here is applicable to a dual class of variation over sets of time series rather than variation within individual time series.

This work is a novel addition to the set of ICA tools that analyze data with temporal structure. An inherent issue is the unsolved problem of balancing reconstruction error and source separation. Without *a priori* knowledge of the nature of the sources, it is difficult to design a robust smoothing method. Our algorithm succeeds because of the presence of high frequency noise in the synthetic and EEG data. We hope that researchers working with functional data use FuncICA in addition to FPCA, as an alternative that can provide new insights in countless domains.

References

[1] J. Ramsay and B. Silverman. *Functional data analysis*. Springer, 2005.

[2] J. Ramsay and B. Silverman. *Applied functional data analysis: methods and case studies*. Springer, 2002.

[3] T.M. Cover and J.A. Thomas. *Elements of information theory*. Wiley New York, 1991.

[4] E.G. Learned-Miller et al. ICA using spacings estimates of entropy. *JMLR*, 4(1271-1295):1–2, 2003.

[5] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *ALT*, volume 3734, pages 63–77. Springer, 2005.

[6] O. Vasicek. A test for normality based on sample entropy. *J.R. Stat. Soc. Ser. B.*, 38(1):54–59, 1976.

[7] A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4-5):411–430, 2000.

[8] J.V. Stone. Blind source separation using temporal predictability. *Neural Comp.*, 13(7):1559–1574, 2001.

[9] B.A. Pearlmutter, L.C. Parra, et al. Maximum likelihood blind source separation:a context-sensitive generalization of ICA. *NIPS*, 9:613, 1997.

[10] W. Liebermeister. Linear modes of gene expression determined by independent component analysis. *Bioinformatics*, 18(1):51–60, 2002.

[11] X. Leng and H.G. Muller. Classification using functional data analysis for temporal gene expression data. *Bioinformatics*, 22(1):68–76, 2006.

[12] P.T. Spellman et al. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, 9(12):3273–3297, 1998.

[13] T. Joachims. Making large-scale support vector machine learning practical. *Advances in Kernel Methods: Support Vector Learning*, 1999.

[14] B. Blankertz, K.R. Muller, et al. The BCI competition III: validating alternative approaches to actual BCI problems. *IEEE Trans. Neural Sys. Rehab. Eng.*, 14(2):153–159, 2006.

[15] J. V Stone, J. Porrill, C. Buchel, and K. Friston. Spatial, temporal, and spatiotemporal independent component analysis of fMRI data. In *Proceedings of the 18th Leeds Statistical Research Workshop on Spatial-Temporal Modelling and Its Applications*, pages 23–28. Leeds University Press, 1999.