

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. We show how applying the dual of temporal ICA to temporal data, and likewise applying the dual of spatiotemporal ICA to spatiotemporal data, enables independent component regularization not afforded by the primal forms applied to their original domains. We call this family of regularized dual ICA algorithms FuncICA. 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. In this work, we develop an algorithm for extracting independent component curves, derive a method for optimally smoothing the curves, and validate this method on both synthetic and real datasets. Results for synthetic, gene expression, and electroencephalographic event-related potential data indicate that FuncICA can recover well-known scientific phenomena and improve classification accuracy, highlighting its utility for unsupervised learning in continuous data. We conclude this work with a forward-looking, novel framework for fMRI data analysis by making use of the functional dual of spatiotemporal ICA.

1 Introduction

We propose a novel solution for unsupervised pattern discovery in time series and other structured data. In particular, we are searching for linearly varying patterns of activation. A simple example of such a pattern is an uptrend sustained for a fixed period of time, such that linear variation affects the slope of the trend. Two powerful applications are automatic identification of event-related potentials in electroencephalography (EEG) and discovery of time-varying signaling mechanisms for gene expression. Existing temporal methods are unable to find such time series patterns: functional principal components analysis (FPCA) only finds patterns that have Gaussian variation over the observations, while existing temporal independent component analysis (ICA) methods are concerned with recovering statistically independent signals sampled over time rather than statistically

independent patterns sampled over entire time series observations (e.g. each observation is a 2-second long univariate time series). The former type of source recovery, the primal form of ICA, is not sufficient for discovery of many types of patterns.

1.1 ICA in the primal Algorithmic treatments of 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 with respect to their domain. By ICA in the primal we mean that for a linear model

$$(1.1) \quad X = AS,$$

where the i^{th} row of X specifies the observations of a random variable X_i , and the j^{th} column of X represents the j^{th} joint observation of the joint variable X , ICA seeks a linear unmixing of the random variables X_1, \dots, X_n such that the resulting variables S_1, \dots, 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). The goal in BSS is to recover the statistically independent source signals.

1.2 ICA in the dual While the primal form can solve highly interesting problems like blind source separation [1] and compression, a dual form of ICA opens to the door to a world of different, equally important problems. In this work, we consider ICA in a dual representation. We still have precisely the same observed data X , but we frame the problem as the mixing model $X^T = \tilde{A}S^T$ which implies that $\tilde{A}^{-1}X^T = S^T$.² In the primal, the rows of A^{-1} are the *independent components*, which by themselves are of little value except for

¹By 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.

²For now, assume that A is positive definite ($A \succ 0$) and hence invertible.

*Georgia Institute of Technology

†Georgia Institute of Technology

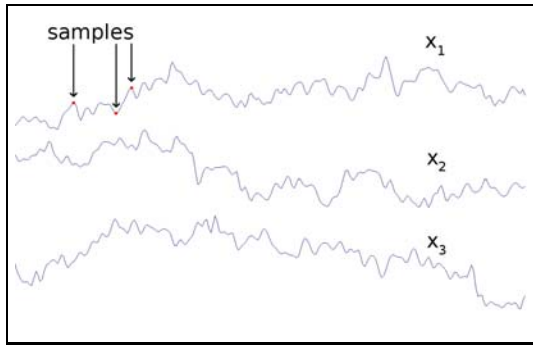


Figure 1: The data for temporal ICA in the primal. Random variables are signals sampled over time indices.

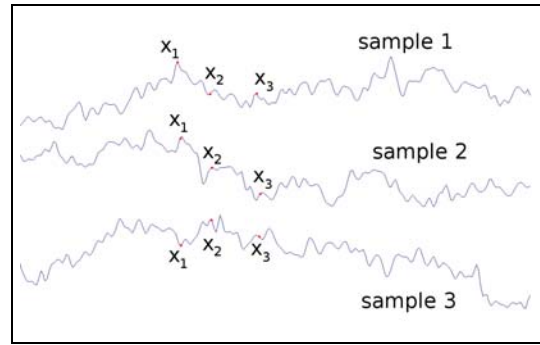


Figure 2: The data for temporal ICA in the dual. Random variables are time indices sampled over time series.

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. The concept of the dual of ICA has been visited before, by Stone et al. [2] and more recently by Choi et al. [3].

In the dual, the rows of \tilde{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 dual of ICA just described is actually the dual of temporal ICA. For clarity, see the primal and dual forms in Figures 1 and 2 respectively. Researchers in various communities, including neuroscience, psychology, and computer science, have explored another form of ICA, known as spatial ICA [4], 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 how, using functional³ 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. At this point in time, the brain is rapidly being imaged via fMRI under a number of stimuli, and it is critical that there exist a method suitable for detecting underlying patterns of activation that can help us better understand neural processes. After introducing FuncICA, our main result, and analyzing its performance, we will argue that both spatial, temporal, and a method known as 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 a more natural method to use for fMRI data analysis and can be a powerful tool for spatiotemporal pattern discovery in this exciting domain.

We begin by providing an overview of ICA in its primal form and some necessary functional notation, as well as a brief summary of FPCA, the closest method to FuncICA. Our contribution follows: we introduce FuncICA and a regularization parameter for smoothing, develop an objective function for optimization of this parameter, validate this objective function for synthetic and real EEG data, and demonstrate superior classification results using FuncICA features for microarray gene expression data and ERP data from the brain, as compared to FPCA and the standard (non-functional) dual of temporal ICA. Our results indicate that the algorithm produces components that can recover sources,

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

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 ICA and FPCA

Prior to introducing FuncICA, we formally introduce ICA and some functional analysis notation. We also describe functional PCA (FPCA) because it is the first step in performing FuncICA, and it is also 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

$$(2.2) \quad X^{(t)} = AS^{(t)},$$

where⁴ $t \in [T]$, $X^{(t)} \in \mathbf{R}^n$, $A \in \mathbf{R}^{n \times n}$, and $S^{(t)} \in \mathbf{R}^n$. The source signal distributions of S_1, \dots, 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, \dots, X_n . The task is then to estimate the unmixing matrix $W = A^{-1}$ such that

$$(2.3) \quad Y \triangleq WX = WAS.$$

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 Kullback-Leibler (KL) divergence between the difference between the joint density of the sources and the product of the marginal densities:

$$(2.4) \quad \mathcal{H}(Y) = D_{\text{KL}} \left(p(Y) \parallel \prod_{i=1}^n p(Y_i) \right) \\ = \sum_{i=1}^n H(Y_i) - H(Y),$$

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 describing 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 $\{X_1(t), \dots, X_n(t)\}$, 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:

$$(2.5) \quad \langle f, g \rangle = \int f(t)g(t)dt.$$

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

$$(2.6) \quad X_i(t) = \sum_{j=1}^m \phi_{i,j} \beta_j(t).$$

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

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

2.3 Functional PCA FPCA, introduced by Ramsay and Silverman [5, 6], is in its unregularized form equivalent to the Karhunen-Loève expansion. Ramsay and Silverman 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 interpretation will ease understanding of FuncICA as well.

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

$$(2.7) \quad \Gamma(s, t) = E[X(s)X(t)].$$

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

$$(2.8) \quad \Gamma(s, t) = \sum_{i=1}^{\infty} \lambda_j \gamma_j(s) \gamma_j(t),$$

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

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

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.

to the following optimization problem:

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

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

$$(2.9) \quad \xi(t) = \sum_{j=1}^m f_j \beta_j(t).$$

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

$$(2.10) \quad X_i(t) = \sum_{j=1}^m A_{i,j} \beta_j(t).$$

Let L be the basis function inner product matrix such that

$$(2.11) \quad L_{i,j} = \int \beta_i(t) \beta_j(t) dt.$$

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

$$(2.12) \quad \int \xi(t) X_i(t) dt = A_{(i)} L^T f,$$

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

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

3.1 FuncICA derivation. A functional version of ICA is equivalent to the original ICA formulation with A now being the linear operator:

$$(3.13) \quad A : L^2 \mapsto L^2,$$

because both the functional observations $X_i(t)$ and the independent functionals S_j are in L^2 . This linear 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

$$(3.14) \quad X_i(t) = \sum_{j=1}^m \psi_{i,j} \beta_j(t).$$

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

$$(3.15) \quad \beta : \mathbf{R}^m \mapsto L^2.$$

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 $\beta_j(t)$. 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

$$(3.16) \quad C_{i,j} = \int A_i(t) B_j(t) dt,$$

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 $\psi \in \mathbf{R}^{n \times m}$. The principal component matrix is

$$(3.17) \quad E = U^T X = U^T \psi\beta = \rho\beta, \quad \text{for } \rho \in \mathbf{R}^{m \times m}.$$

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

$$(3.18) \quad \sigma_E = EX^T = \rho\beta(\psi\beta)^T = \rho\beta\beta^T\psi^T.$$

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

$$(3.19) \quad Y = WE = \phi\beta,$$

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

$$(3.20) \quad \sigma_Y = YX^T = WEX^T = W\sigma_E = \phi\beta\beta^T\psi^T.$$

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).

3.2 Algorithm To optimize W , recall our objective function 2.4. After performing FPCA, we are left with

$$(3.21) \quad \mathcal{H}(E) = \sum_{i=1}^n H(E_i) - H(E).$$

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

$$(3.22) \quad \begin{aligned} \mathcal{H}(Y) &= \sum_{i=1}^n H(Y_i) - H(Y) \\ &= \sum_{i=1}^n H(Y_i) - H(WE), \end{aligned}$$

which by Theorem 9.6.4 of Cover and Thomas [7] is equal to

$$(3.23) \quad \sum_{i=1}^n H(Y_i) - H(E) - \log(|W|).$$

To maintain the uncorrelatedness of the functionals, we restrict W to the set of orthogonal (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

$$(3.24) \quad \mathcal{H}^*(Y) = \sum_{i=1}^n H(Y_i).$$

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 .

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 [8]. RADICAL employs a modified, more robust version of this estimator. It is beyond the scope of the paper to give a thorough treatment of RADICAL. The interested reader is encouraged to read [9] for details⁶.

⁶RADICAL is competitive with other recent algorithms [10], so our analysis is restricted to RADICAL.

This work would be far from complete without special considerations for the functional data. A caveat of source separation in the functional case 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 [6]. Instead of choosing eigenfunctions with the constraint

$$(4.25) \quad \int \xi(t)^2 dt = 1,$$

we impose the alternate, roughness-penalizing constraint

$$(4.26) \quad \int \xi(t)^2 dt + \alpha \int (L\xi(t))^2 dt = 1,$$

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

$$(4.27) \quad L\xi = \sum_{i=0}^m a_i D^i \xi.$$

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

$$(4.28) \quad \int \xi(t)^2 dt + \alpha \int (D^2 \xi(t))^2 dt = 1,$$

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 [5]. 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 a 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 is defined as

$$(4.29) \quad J(Y_i) = H(\mathcal{N}(0, 1)) - H(Y_i).$$

Note that $J(Y_i)$ is non-negative because for a fixed variance the Gaussian distribution has maximal entropy [11].

Our objective function is then

$$(4.30) \quad Q(Y) = \sum_{i=1}^p \frac{1}{J(Y_i)}.$$

This objective heavily penalizes nearly-Gaussian sources and mostly ignores sources with high negentropy.

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.

Algorithm 1

- 1: Minimize FPCA LOOCV error to find α_P^* .
 - 2: $Q^{(0)} = \infty$
 - 3: $\alpha^{(1)} = \alpha_P^*$
 - 4: $\tau = 0$
 - 5: **repeat**
 - 6: $\tau = \tau + 1$
 - 7: $Y = \text{FuncICA}(X, \alpha^{(\tau)})$
 - 8: $Q^{(\tau)} = \sum_{i=1}^p \frac{1}{J(Y_i)}$
 - 9: $\alpha^{(\tau+1)} = \gamma \cdot \alpha^{(\tau)}$
 - 10: **until** $Q^{(\tau)} > Q^{(\tau-1)}$
 - 11: **return** $\alpha^{(\tau-1)}$
-

5 Previous Temporal ICA Methods

In this section, we frame FuncICA in the context of other temporal ICA methods that also exploit the temporal structure of the data. Two important points to note 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.

Convolutional 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 [12, 13]; a simple example would be if a source signal had significant auto-correlation and the cross-correlation among the source signals, by virtue of their independence, is zero. 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.

Several ICA methods attempt to recover dominant independent components [14, 15], 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 flexibly can facilitate computability by reducing the number of

basis functions used in our representation or being more aggressive in the FPCA dimensionality reduction step⁷. We now present results of FuncICA on synthetic and real-world brain data.

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 contrast the performance of FuncICA with FPCA and the non-functional dual of temporal ICA. We first discuss results on synthetic data as validation of the independent curve extraction and smoothing methods. We then present scientifically intriguing results for feature extraction and classification in EEG and gene expression data. Note that in this section, when we refer to ICA, we mean the standard (non-functional) dual of temporal ICA.

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

$$(6.31) \quad S_1(t) = \frac{1}{\sqrt{2}} \sin(10\pi t) \quad \text{and} \quad S_2(t) = \frac{1}{\sqrt{2}} \cos(10\pi t),$$

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 ($Z_i \sim \text{Laplace}(0, \frac{1}{\sqrt{2}})$). The sources were mixed to form 10^4 observations X_i for $i \in [10^4]$:

$$(6.32) \quad X_i = Z_{1,i} \cdot S_1 + Z_{2,i} \cdot S_2,$$

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

Source recovery accuracy 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 to be the minimum inner product among the source functionals’ inner products with their best fit independent functionals; more formally, accuracy is

$$(6.33) \quad \min \left\{ \langle S_1, (\arg \max_{Y_i} \langle S_1, Y_i \rangle) \rangle, \langle S_2, (\arg \max_{Y_j} \langle S_2, Y_j \rangle) \rangle \right\}.$$

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

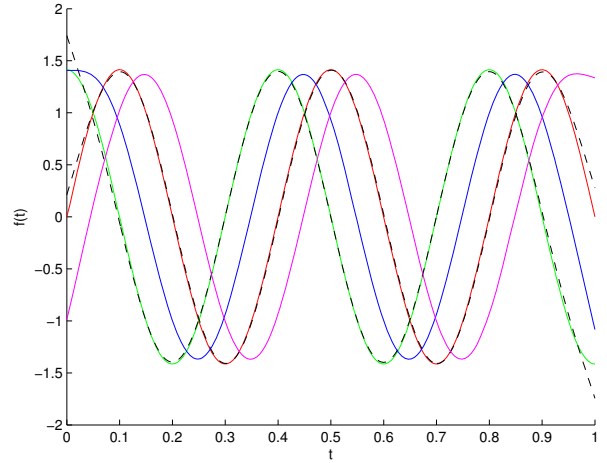


Figure 3: $S_1(t)$ and $S_2(t)$ are shown in red and green respectively. $IC_1(t)$ and $IC_2(t)$ are the black dotted lines. $PC_1(t)$ and $PC_2(t)$ are the blue and magenta lines respectively. The independent functionals are near-perfect replicas of the sources, while the principal component functions are phase-shifted from the sources.

As shown by a typical recovery result in Figure 3, FuncICA achieved near-perfect source recovery, while FPCA nearly always recovered a phase-shifted version of the sources. The mean accuracy of FuncICA 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 entropy estimator [8]) 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.

Robustness with respect to α 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 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 3rd principal component (PC) curve, with negligible change to the extracted IC curves. FPCA therefore seems to be more sensitive to

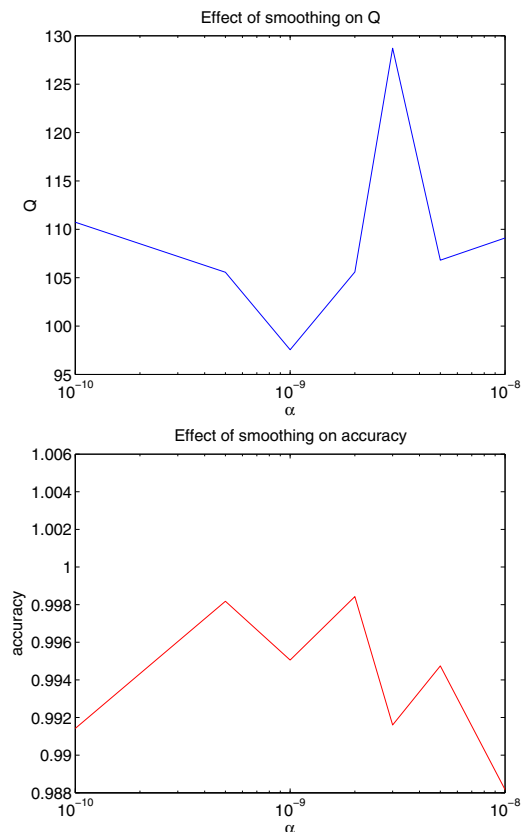


Figure 4: All plots are for the synthetic dataset. Top, the regularization parameter α plotted versus Q . Note the minimum at $\alpha = 1 \cdot 10^{-9}$ and the spurious maximum at $\alpha = 3 \cdot 10^{-9}$. Bottom, accuracy is high in close proximity to $\alpha = 1 \cdot 10^{-9}$, with sharp drops as α increases above $2 \cdot 10^{-9}$ or decreases below $5 \cdot 10^{-9}$.

regularization in cases where the components exhibit variable amounts of smoothness that can inform the separation process, while FuncICA requires no regularization here.

Effect of α on Q and accuracy 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 1 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 4 shows the effect of smoothing on Q and accuracy.

The left plot in Figure 4 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 1 for optimization of α_I^* on synthetic data, we progress to the real-world results.

6.2 Microarray gene expression data. The microarray data obtained by [16] is of α -factor⁸ synchronized temporal gene expression for the yeast cell cycle, with observations of 6178 genes at 18 times in 7 minute 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 [17], who present classification results using FPCA.

Notably, Liebermeister [18] explored gene expression with this data set using the primal form of temporal 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 [17]) 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 eliminate most of the roughness Algorithm 1 would otherwise remove.

Results For the discrimination task, we applied

⁸An 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 to 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 [16]) 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) [19] 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 its non-functional version (RADICAL in this case) for this problem. Our conjectures for the results that we obtained are that FuncICA extracts more meaningful features that are more easily exploited by the SVM classifier.

6.3 P300 event-related potential data. The EEG data set obtained from [20] 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 6, the empirical P300 waveform for this data set is shown; it was calculated by taking the mean of the 2550 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

	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 5: P300 classification accuracy for FuncICA, FPCA, ICA, and empirical P300 waveform. The best accuracy attained in each column is in bold.

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 flash 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.

Results 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 from FuncICA, FPCA, and ICA (for ICA, we use 256 dimensional data reduced by PCA to 12 dimensions) are summarized in Figure 5. 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 6 illustrates the IC curve most similar to the P300 for 2 values of α .

In Figure 8 we demonstrate the success of Algorithm 1 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. To contrast FuncICA with FPCA, the best results for a single PC curve were substantially lower than the results for a single IC curve. As shown by Figure 7, this result is hardly surprising as the curves extracted by FPCA are uninteresting harmonics. The extracted curves being harmonics also makes sense due

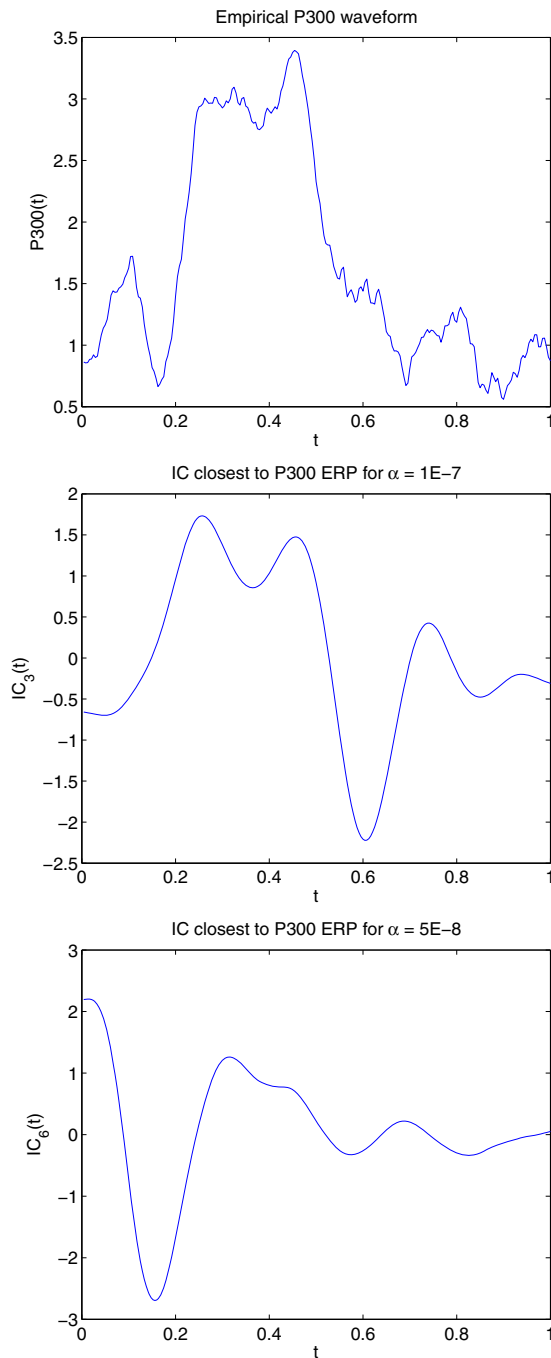


Figure 6: Top to bottom. (1) Empirical P300 waveform calculated from 2550 trials. (2) Closest IC to P300 for $\alpha = 5 \cdot 10^{-5}$. (3) Closest IC to P300 for $\alpha = 1 \cdot 10^{-7}$.

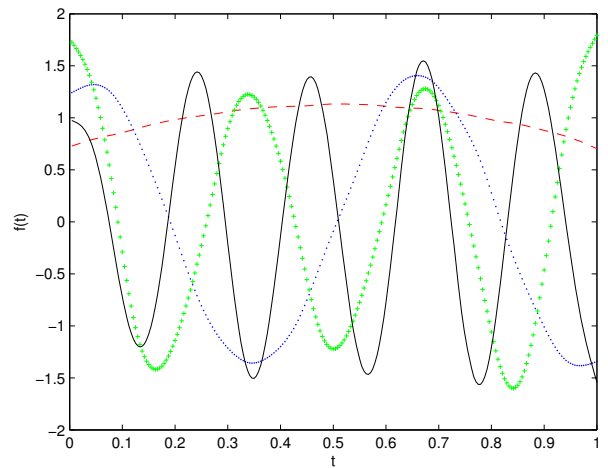


Figure 7: PC curves 1, 4, 7, and 10 plotted as a dashed red line, blue dotted line, green dashes line, and solid black line respectively - Note that the components are almost entirely described by harmonic behavior.

to the dominance of harmonics within EEG, and hence 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 5, 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 FuncICA better captures the underlying event-related potential of interest when humans are presented with unexpected stimuli.

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

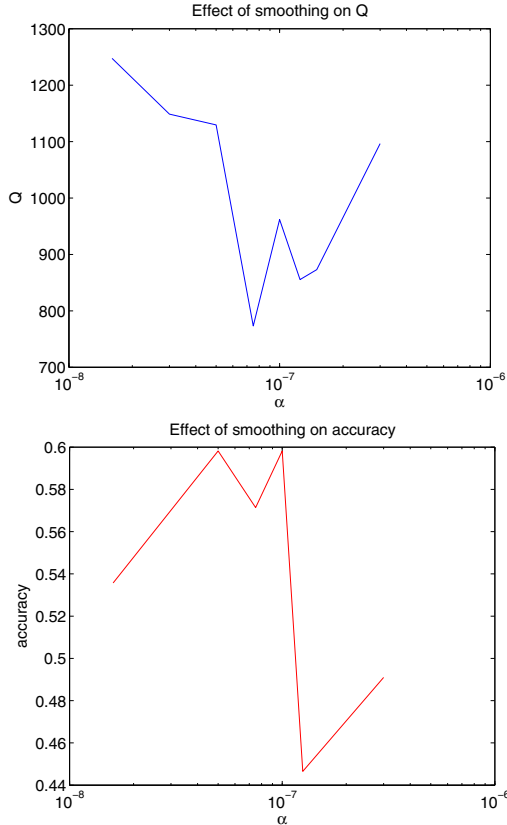


Figure 8: All plots are for the P300 dataset. Top, the regularization parameter α plotted versus Q . The local minimum encountered when moving leftward from $\alpha_P^* \approx 1.6 \cdot 10^{-8}$ (as per Algorithm 1) occurs at $\alpha_I^* = 7.5 \cdot 10^{-8}$. Bottom, α_I^* is very close to the optimal neighborhood for accuracy.

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 [2], which seeks to simultaneously seek statistical independence in both the temporal and spatial sense via

a convex combination of their respective objectives.

Notably, there is substantial temporal and spatial dependence in neural activation patterns, and so we strongly argue that a more appropriate method is to consider the dual of spatiotemporal ICA, where the temporal and spatial dependencies are unconstrained. The dual of spatiotemporal ICA is in fact a generalization of FuncICA to spatiotemporal data; this version of FuncICA results in a powerful alternative method to 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 to push our spline representation⁹ into a four-dimensional space (three spatial dimensions plus time). Such basis functions have been used previously in positron emission tomography (PET) analysis [21]. The second generalization required is to specify another regularization term that penalizes roughness over both space and time. After these two modifications, the FuncICA algorithm outlined in the previous sections directly can be used to mine fMRI data for spatiotemporal patterns of activation. In the near future, we will explore this application with a highly interesting classification task using real fMRI data.

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.

From a theoretical perspective, we have shown how applying the dual of temporal ICA to temporal data, and likewise applying the dual of spatiotemporal ICA to spatiotemporal data, enables independent component

⁹Alternatively, we could consider nonparametric basis functions.

regularization not afforded by the primal forms applied to their original domains. As a consequence of this duality result, we have proposed a new method for mining fMRI data for spatiotemporal patterns whose activation is statistically independent over a large set of trials.

More pragmatically, our 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. In general, claims can be made about reduction of high frequency noise. We hope that researchers working with functional data, such as time series and spatial data, use FuncICA to gain new insights into countless domains.

References

- [1] A.J. Bell and T.J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.
- [2] 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.
- [3] S. Choi, A. Cichocki, H.M. Park, and S.Y. Lee. Blind source separation and independent component analysis: A review. *Neural Information Processing-Letters and Reviews*, 6(1):1–57, 2005.
- [4] 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):160–188, 1998.
- [5] J. Ramsay and B. Silverman. *Functional data analysis*. Springer, 2005.
- [6] J. Ramsay and B. Silverman. *Applied functional data analysis: methods and case studies*. Springer, 2002.
- [7] T.M. Cover and J.A. Thomas. *Elements of information theory*. Wiley New York, 1991.
- [8] O. Vasicek. A test for normality based on sample entropy. *J.R. Stat. Soc. Ser. B.*, 38(1):54–59, 1976.
- [9] E.G. Learned-Miller et al. ICA using spacings estimates of entropy. *JMLR*, 4(1271-1295):1–2, 2003.
- [10] 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.
- [11] A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4-5):411–430, 2000.
- [12] J.V. Stone. Blind source separation using temporal predictability. *Neural Comp.*, 13(7):1559–1574, 2001.
- [13] B.A. Pearlmutter, L.C. Parra, et al. Maximum likelihood blind source separation: a context-sensitive generalization of ICA. *NIPS*, 9:613, 1997.
- [14] Y.M. Cheung and L. Xu. An empirical method to select dominant independent components in ICA for time series analysis. In *Neural Networks, 1999. IJCNN'99. International Joint Conference on*, volume 6, 1999.
- [15] JM Gorriz, CG Puntonet, M. Salmeron, and EW Lang. Time series prediction using ICA algorithms. In *Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, 2003. Proceedings of the Second IEEE International Workshop on*, pages 226–230, 2003.
- [16] 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.
- [17] X. Leng and H.G. Muller. Classification using functional data analysis for temporal gene expression data. *Bioinformatics*, 22(1):68–76, 2006.
- [18] W. Liebermeister. Linear modes of gene expression determined by independent component analysis. *Bioinformatics*, 18(1):51–60, 2002.
- [19] T. Joachims. Making large-scale support vector machine learning practical. *Advances in Kernel Methods: Support Vector Learning*, 1999.
- [20] 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.
- [21] TE Nichols, J. Qi, E. Asma, and RM Leahy. Spatiotemporal reconstruction of list-mode PET data. *Medical Imaging, IEEE Transactions on*, 21(4):396–404, 2002.