Signals as Parametric Curves: Application to Independent Component Analysis and Blind Source Separation

Birmingham Hang Guan and Anand Rangarajan

Abstract

Images Stacks as Parametric Surfaces (ISPS) is a powerful model that was originally proposed for image registration. Being closely related to mutual information (MI) – the most classic similarity measure for image registration, ISPS works well across different categories of registration problems. The Signals as Parametric Curves (SPC) model is derived from ISPS extended to 1-dimensional signals. Blind Source Separation (BSS) is a classic problem in signal processing, where Independent Component Analysis (ICA) based approaches are popular and effective. Since MI plays an important role in ICA, based on the close relationship with MI, we apply SPC model to BSS in this paper, and propose a group of geometrical objective functions that are simple yet powerful, and serve as replacements of original MI-based objective functions. Motivated by the geometrical objective functions, we also propose a second-order-statistics approach, FT-PCA. Both geometrical objective functions and FT-PCA consider signals as functions instead of stochastic processes, make use of derivative information of signals, and do not rely on the independence assumption. In this paper, we discuss the reasonability of the assumptions of geometrical objective functions and FT-PCA, and show their effectiveness by synthetic experiments, comparing with other previous classic approaches.

1 Introduction

Independent component analysis (ICA) ([11, 17]) is a well-known topic in machine learning, statistics, and signal processing. The original ICA problem ([10, 3]) and its various extensions ([16, 2]) have been researched through the past 25 years. Being a theoretical topic in statistics, it was originally proposed and applied for signal processing problems, especially blind source separation (BSS) ([10]). ICA-based BSS has various practical applications, like electroencephalographic data analysis (EEG) ([13, 24]). ICA was also applied to image problems, like the work in [23, 26], where the latent independent variable linear mixture model were used for image fusion, etc.

Formally, and ICA problem is described in the following manner: Let \( \mathbf{x} \) be a random vector of size \( N \). The generative model for \( \mathbf{x} \) is via a matrix \( \mathbf{A} \) of size \( N \times M \) and a random vector \( \mathbf{s} \) of size \( M \) whose components are independent of each other, such that

\[
\mathbf{x} = \mathbf{A}\mathbf{s}.
\]

Mutual information (MI), as a natural independence measure of random variables, is considered as a standard approach to solve for the ICA problem:

\[
MI(\mathbf{s}) = \int p_s(\mathbf{s}) \log \frac{p_s(\mathbf{s})}{\prod_i p_{s_i}(s_i)} d\mathbf{s}.
\]

*The authors are with the Department of Computer and Information Science and Engineering, University of Florida, Gainesville, FL, USA. E-mail: \texttt{bkwan.anandr@ufl.edu}
If each $s_i$ are independent, $MI(s) = 0$. Therefore, by minimizing MI, one can get separated random variables as independent as possible.

The definition of a BSS problem has a very close structure: Let $\mathbf{x}(t) = (x_1(t), \ldots, x_N(t))$ be a set of observed signals. Supposing that they are a linear mixture of a set of unknown source signals $\mathbf{s}(t) = (s_1(t), \ldots, s_M(t))$, we can express it as

$$\mathbf{x}(t) = A\mathbf{s}(t)$$

where $A$ is the mixing matrix. There are two main differences between an ICA problem and a BSS problem: First, each component of the $\mathbf{x}$ and $\mathbf{s}$ in the ICA problem is a random variable, whose density is estimated by observations of the sample data; while the components of $\mathbf{x}(t)$ and $\mathbf{s}(t)$ are signals, which can be considered as stochastic processes. The core difference is about the data being discrete or continuous. Second, being random variables, the components of $\mathbf{s}$ in the ICA problem are independent. In BSS, independence is not a necessary assumption.

Apparently, BSS is a highly open and nondeterministic problem, where both $A$ and $\mathbf{s}(t)$ need to be determined. For any approach to solve BSS, additional assumption on sources should be made to decrease the infinitely many solutions down to a small subset. Similar to ICA, the signals $\mathbf{x}(t)$ and $\mathbf{s}(t)$ are commonly assumed to be zero-mean, and signals in $\mathbf{s}(t)$ are uncorrelated with each other. Under this assumption, most approaches applied principal component analysis (PCA) to standardize signals as the first step. As early as 1981 ([25]), it was pointed out that the information in the spectral matrix is not sufficient for separation, and additional assumptions are needed. The work in [22, 9] started to consider higher order statistics, and the assumption of source signals being independence was put forth ([9, 12]). From then on, the assumption of independence was taken for granted by most approaches (and these are not necessarily restricted to just ICA) ([4, 6]) for BSS.

Though the independence assumption in ICA is natural, few papers pointed out the difference in models by applying ICA to BSS. The independence of a set of signals is defined by the independence of the distributions underlying each signal, considering each signal as a sample function from a stationary stochastic process whose distribution at each time point is identical. By the ergodicity of the stationary stochastic processes, the sample values of the signals can be used to estimate the statistical properties of the distributions, and hence can be used to estimate the MI and independence. This is different from ICA, where data are observations of samples of each underlying distribution. Unfortunately, most ICA work, like the work in [18, 10] just mention that consideration of time $t$ should be neglected, and the signal values should be considered as a collection of unordered observations. The drawback of this perspective is that information contained in the “changing over time” of signals are ignored.

Another perspective is to consider time $t$ as a uniform distributed random variable, like the work in [27]. In this way, being continuous functions of time, signals can still be considered as random variables and the “changing over time” information of signals are taken into consideration. However, under this model, independence does not exist since all signals are functions of the same random variable. Therefore, the definition of independence of signals is not well-defined under this perspective.

For BSS, the independence assumption is not necessary. Similar to the work in [4, 5], we are dedicated to consider additional information provided by signals as functions, and solve the BSS problems without the assumption of independence. Signals can be either considered as deterministic functions of time, or stochastic processes, based on its application. Considering signals as deterministic functions, we extend the Images Stacks as Parametric Surfaces model (ISPS), a powerful model originally designed for image registration, to 1D case (we call it Signals as Parametric Curves
(SPC), accordingly), and apply SPC to BSS. Based on the close relationship between SPC and MI, we propose geometrical objective functions that can approximate the MI-based objective functions. We are also able to analyze signals in the frequency domain by the Fourier transform of the signals, and propose the FT-PCA algorithm, which does not rely on the independence of signals, and focuses on the local orthogonality in the frequency domain. For simplicity, in this paper, we only focus on the two-dimensional case (where there are only two observed signals and two source signals), and higher dimensional cases can be extended naturally.

The main content of this paper is as follows: Section 2 briefly introduces well-known previous approaches for BSS, including ICA and second-order-statistics approaches; Section 3 briefly summarizes and analyzes the ICA framework and its assumption; Section 4 applies SPC to BSS and propose geometrical objective functions that are competitive with the traditional MI approach; Section 5 introduces FT-PCA algorithm based on the assumption of kernel-orthogonality in the frequency domain; Section 6 shows synthetic simulation experiments and compare our approaches with several well-known approaches, and shows the effectiveness of our algorithms; and the paper is concluded in Section 7 to highlight our simple yet effective approaches for BSS.

2 Previous Work

Most approaches of BSS can roughly be categorized into two classes: high-order statistics based approaches, or second-order statistics based approaches. ICA approaches stick to the assumption of independence, and try to minimize the entropies of signals to recover sources which are as independent as possible; while joint diagonalization approaches try to make use of information and properties of second-order statistics of signals to solve for the unmixing matrix, bypassing the direct usage of independence to avoid higher-order statistics.

The work in [9, 12, 10] firstly introduced the concept of ICA, and created the independence assumption as a foundation of BSS. [3] is another well-known paper that highlighted mutual information based approaches for BSS. In the original ICA framework, the objective function was directly based on the assumption of independence: the Kullback divergence of the joint density and the product of marginal densities, i.e. the mutual information. Nevertheless, its estimation is difficult, and high-order cumulants were introduced to estimate entropies. The work also suggested a standardization step using PCA to standardize the deviation, and pointed out that after PCA, the minimization of MI is equivalent to minimization of negentropies with respect to a sequence of pairwise rotations of signals. The work in [16] put forward the FastICA algorithm. Based on their previous work in estimating entropies ([15]), they suggested a set of contrast functions that are much simpler to compute than high order cumulants in the work in [10]. They also adopted Newton method to decrease the time complexity, so that each signal can be optimized one by one. The FastICA algorithm is very efficient and widely used until now. Another most well-known approach in the ICA category is Kernel ICA, proposed in the work in [2]. The goal of Kernel ICA is to maximize the kernel correlation of whitened signals. It constructs an eigen-decomposition structure, and computes the minimum eigenvalue of a matrix constructed by certain Gram matrices of signal data points. Though the idea is somehow close to our kernel orthogonality, the approach is totally different. It is still within the ICA optimization framework.

Comparing to the ICA series where signals values are used to estimate the independence of the underlying distributions, the second-order-statistics class (we call it SOBI series) tried to take use of other stochastic process properties to bypass the approximation of entropies. AMUSE ([29]) algorithm is an early work of these approaches. Its assumption on source signals is that given some
time shift $\tau$, the auto-correlation matrix is diagonal but not identity, i.e. for $i \neq j$, $E(s_i(t)s_i(t-\tau)) \neq E(s_j(t)s_j(t-\tau))$ and $E(s_i(t)s_j(t-\tau)) = 0$. This assumption grants another eigen-decomposition structure than the PCA step, and makes AMUSE an approach where no optimization is required. However, not all $\tau$ grants diagonal matrices. Once the selected $\tau$ makes the auto-correlation matrix isomorphic to identity, the eigen-decomposition gives trivial results, and AMUSE fails. The work in [6] put forward a joint diagonalization scheme, and an extended algorithm, named SOBI. Instead of a certain $\tau$, SOBI is based on the assumption that $E_{\tau}(s(t+\tau)s^H(t))$ is diagonal, assuming that $s$ is a multivariate stationary process of both $t$ and $\tau$. (It also has an equivalent assumption where the expectation of $\tau$ is defined as arithmetic average of a set of different $\tau$'s.) To select a bunch of different $\tau$ and use the joint diagonalization scheme, SOBI avoided the occurrence of a single trivial $\tau$, and is able to solve the problem by $K$ times matrix diagonalization, where $K$ is the number of $\tau$ selected. A following work in [5] extended this idea to non-stationary signals, where time-frequency distribution (TFD) ([8]) was introduced. Based on similar fact that the spatial TFD matrices (STFD) of signals being diagonal but not identity, eigen-decomposition scheme is also able to be applied to STFD matrices. Since STFD are dependent with time and frequency indices $(t,f)$, and for some special $(t,f)$, the STFD matrix can be rank deficient, they again applied joint diagonalization scheme to solve the problem by a set of different selected $(t,f)$. STFD is close to our approach, except that designed for non-stationary signals, the time-frequency domain analysis was introduced. And similar to SOBI, it adopted selection of parameters and joint diagonalization. Though this approach can handle non-stationary signals and Gaussian signals, it was criticized by complexity and performance ([11, 20]). After this work, many following work came to based on time-frequency analysis and joint diagonalization ([2]). However, most of them, like the work in [14], did not improve the fact that STFD needs local parameter selection and joint diagonalization, and focus on non-stationary signals, which is out of the scope of this paper. The work in [21] is another one close to ours. It also put forward the assumption of disjoint orthogonality. However, it and its following work, like the work in [31], are based on a different problem from $x = As$ where other special conditions are applied, and therefore, are able to solve for more sources than observed signals. This is also not the focus of this paper. Other work on BSS with frequency domain analysis, like the work in [30, 19, 28], though consider the mixing relation between sources and observed signals in frequency domain, are different from our work by assumption, model, and algorithms.

3 ICA Revisited

3.1 A Two-Step Framework

Typically, ICA consists of two steps: the ICA optimization following a prewhitening step, where a PCA is performed. Though in most work ([10, 18]), the prewhitening of the input data $x$ was taken for granted, it is also well known that the purpose of the PCA is not merely to “standardize” $x$ so as to make its covariance identity. The key is whether to accept an additional assumption that $ss^T = I$. This assumption was accepted in the paper of [10] but not in the paper of [16]. Since by the assumption of independence, the source random variable $s$ are uncorrelated. Hence, the assumption of $ss^T = I$ only adds an additional condition that each source random variable has unit variance. In ICA, the scaling of the source random variables is nondeterministic, thus the assumption is reasonable. With this assumption, ICA becomes a two-step algorithm, as analyzed in the following:
Expressing $A$ as its singular value decomposition (SVD)

$$ A = U_A \Sigma_A V_A^T, $$

and given that $ss^T = I$, we have

$$ xx^T = Ass^T A^T = AA^T. $$
i.e.

$$ C_X = U_A \Sigma_A^{-2} U_A^T, $$

where $C_X = xx^T$. Note that $A$ is not orthogonal, otherwise $x$ are uncorrelated and no PCA is needed. Hence, $\Sigma_A^{-2}$ is a diagonal matrix whose main diagonal elements are not equal. And $C_X = U_A \Sigma_A^{-2} U_A^T$ is a unique eigen decomposition. This implies that, applying PCA to $x$, we can solve for both $U_A$ and $\Sigma_A$.

Considering the SVD of the linear mixing matrix $A$, we can call the equation

$$ x = As = U_A \Sigma_A V_A^T s $$
a “rotation-scaling-rotation” procedure (up to some permutation and reflection): $V_A^T$ is the first rotation applied to $s$, $\Sigma_A$ applies scalings to $s$, and $U_A$ is the second rotation. From above we saw that from the mathematical point of view, the PCA step in fact solves for the second rotation $U_A$ and the scaling $\Sigma_A$.

Therefore, a whole ICA procedure should be considered as a two-step framework, which is also very well-known in signal processing literature ([5]): solving for the second rotation $U_A^T$ and the scaling $\Sigma_A$ by PCA; and then solving for the first rotation $V_A^T$ based on other assumptions, like “independence” in the work in [10], or auto-correlation matrices being diagonal in the work in [4].

Let’s call the signals after PCA as $z$, i.e.

$$ z = \Sigma_A U_A^T x, $$

and we have

$$ s = V_A z. $$

In the two-signal cases, the orthonormal matrix $V_A$ is just a rotation matrix, up to some reflection and permutation. And in higher dimensional cases, it is a composition of a series of rotations (and possible reflections) within two-dimensional subspaces. This implies an important fact, which can also be noticed from the MI-based ICA objective functions, that:

The joint entropy of $V_A z$ is invariant to rotation $V_A$.

Therefore, after the first step of an MI-based ICA, the joint entropy is already maximized, and the second step is just a searching for rotations that minimize the summation of each marginal entropy. This agrees with the fact that for any MI-based ICA approach, the true objective function is the summation of negentropies

$$ \sum_i J(p_{z_i}) = \sum_i (H(\phi_{z_i}) - H(p_{z_i})) $$

where $p_{z_i}$ is the density of a random variable $z_i$, and $\phi_{z_i}$ is the Gaussian density with the same mean and variance as $p_{z_i}$. This was interpreted as “Faraway from Gaussian distribution implies independence” ([15]). Note that, during the searching of the rotation angle, the mean (standardized as zero) and variance does not change for each $z_i$, so that $H(\phi_{z_i})$ does not change, and minimizing the negentropy is equivalent to minimizing the sum of marginal entropies.
This also implies that ICA only valid for the case where at most one Gaussian component exists, since if all components are Gaussian, after the first step, the resulted distribution is rotational symmetric, given the assumption that $ss^T = I$.

This fact can be understood as: under the linear mixing model, uncorrelatedness implies maximization of joint entropy, and that independence and uncorrelatedness only differ by a series of rotations. Fig. 1 shows an example where we can observe that the seeking of independence is a seeking of an angle, so that each marginal distribution has as less marginal entropies as possible.

### 3.2 The Reasonability of the Independence Assumption

In this section, we discuss the independence assumption formally. The statistical independence can be defined from two different perspectives: the signals being deterministic functions, or stochastic processes.

From the stochastic process point of view, we consider the source signals $s(t) = (s_1(t), \ldots, s_M(t))$ being sample functions of continuous stationary stochastic processes $\tilde{s}(t) = (\tilde{s}_1(t), \ldots, \tilde{s}_M(t))$. For any positive integer $n$, pick time points $t_1, t_2, \ldots, t_n \in D$ and any time interval $\Delta t \in D$, where $D$ is the time domain, for $i = 1, 2, \ldots, M$, the random vector

$$(\tilde{s}_i(t_1), \tilde{s}_i(t_2), \ldots, \tilde{s}_i(t_n))$$

and

$$(\tilde{s}_i(t_1 + \Delta t), \tilde{s}_i(t_2 + \Delta t), \ldots, \tilde{s}_i(t_n + \Delta t))$$

has identical distribution $p_{\tilde{s}_i}$. The independence of the signals are defined as the independence of $p_{\tilde{s}_i}$ for $i = 1, 2, \ldots, M$. By the ergodicity theorem of the stationary stochastic processes, the values of the sample functions – the source signals – can be used to estimate the entropy of underlying distribution, and compute their mutual information. Therefore, assuming independence of the distributions underlying a set of signals is reasonable, and hence ICA can be directly applied to BSS with the independence assumption.

However, “independence” is not the truth, but just an assumption to admit so that ICA can be applied to BSS. It is not perfect, and has the following disadvantages: Firstly, the stochastic process model of signals disregards the derivative information contained in the signals. That is, if we reorder the signal sample values, there are no difference from the stochastic process perspective. We assert that an approach may work for more cases if it takes the derivative information into consideration. Secondly, there exists pairs of source signals that are generated and sampled “independently”, but by ICA, i.e. by the minimization of sum of marginal entropies, the original signals may not be recovered. See Fig. 2. This indicates the fact that the independence assumption may not be the
most reasonable assumption for these source signals. Finally, ICA do not work for the case where at least two source signals are Gaussian, as we mentioned above.

4 Applying SPC to BSS

4.1 The SPC Model

Based on the effectiveness of the MI-based ICA for most cases of BSS, as well as the disadvantages of the stochastic process model of signals underlying ICA for BSS, we need a different approach that is based on the deterministic function model of signals where derivatives of the signals are available, and is closely related to MI. SPC is one of the best choices.

The SPC model is expressed as follows:

Suppose that we have a set of 1D signals \( \{s_1(t), s_2(t), \ldots, s_N(t)\} \) defined on the domain of time \( D \subset \mathbb{R} \). Consider the mapping

\[
S : D \rightarrow \mathbb{R}^{N+1}
\]

by

\[
t \mapsto (t, s_1(t), \ldots, s_N(t)),
\]

and we have a 1D parametric curve \( S(t) \) embedded in an \( N + 1 \) dimensional Euclidean space. And its curve arc length is

\[
\int_D \sqrt{1 + \sum_{i=1}^{N} (s'_i(t))^2} \, dt.
\]

Analogizing the ISSRA objective function in 2D case, we have the following objective function:

\[
\mathcal{O}_1 = \int \sqrt{\frac{N}{\prod_{i=1}^{N} \frac{1}{N} + (s'_i(t))^2}} \, dt
\]

where \( N \) is the number of signals. We can call it the “Signal Parametric Curve Relative Arc Length”, comparing to the name of ISSRA in the ISPS model.

Note that in Eq. 2, comparing to ISSRA, the denominator and numerator of the integrand are flipped. In ISSRA, the joint area is to be minimize for getting images similar, so it is in the numerator. And here we want to minimize \( \mathcal{O}_1 \) to get signals as separated as possible, so we flip the
integrand in order to fit this “opposite” problem, by putting the product of each arc length in the numerator and the joint arc length in the denominator.

4.2 The Relationship with MI

In order to discuss the relationship between SPC and MI, we need to consider the pseudo-SPC model and understand the signals being random variables as functions of time. Consider time $t$ as a uniformly distributed random variable, each signals $s_i(t)$ being a differentiable function of $t$ is also a random variable. To estimate the joint entropy of the “stack of signals”, the pseudo-SPC

$$\tilde{S} : D \rightarrow \mathbb{R}^N$$

by

$$t \mapsto (s_1(t), \ldots, s_N(t))$$

is considered. The difference between SPC $S$ and pseudo-SPC $\tilde{S}$ is that the first dimension $t$ does not appear in $\tilde{S}$, and $\tilde{S}$ is not injective, similar to the relationship between ISPS and pseudo-ISPS.

Unfortunately, similar to the fact that MI-based registration approach is not applied to groupwise case, because of the disagreement of dimensionality, the Lebesgue measure of $\tilde{S}(t)$ embedded in $\mathbb{R}^N$ is zero, and the joint density does not exist. And from the statistics point of view, it is also clear that, since each signal is a function of $t$, there is no independence defined for the set of all signals. This implies that under the pseudo-SPC point of view, MI is not able to be computed to solve the BSS problem.

However, in Section 3.1 we discuss the two-step framework of the ICA problem. We pointed out that in ICA the joint density and joint entropy is never considered. In the second step, no joint entropy is computed, but just the sum of marginal entropies. Fortunately, in the pseudo-SPC perspective, the 1D marginal entropy of each signal is well-defined. And by the close relationship between it and the SPC model, we are still able to apply $O_1$ to BSS to approximate the “MI”, i.e. the sum of marginal entropies, to solve for unmixed signals.

In fact, looking at $O_1$ carefully, we notice that the joint arc length (the denominator) is also invariant to rotations, which means that in the second step of ICA where different rotation matrices are applied, the denominator does not change either. This also meets the fact that $O_1$ is closely related to MI, where the joint entropy part does not change with respect to rotations. And we can simplify $O_1$ to get

$$O_2 = \int \prod_{i=1}^{N} \sqrt{1 + (s'_i(t))^2} dt$$

where only the marginal arc lengths are computed. Clearly, $O_2$ is related to the true objective function, the sum of marginal entropies, in the traditional MI approaches for ICA, and can be considered as the objective function derived from the SPC model.

4.3 Geometrical Objective Functions for BSS

Applying SPC to BSS, and considering the two-step framework of ICA, we proposed the objective function $O_2$, the product of marginal arc lengths, by its close relationship with the objective function of traditional MI approach. Hence, in the second step of ICA, given $z(t)$ as the inputs, we can apply a rotation matrix $R$ to get

$$y(t) = Rz(t)$$
and computes the objective functions of \( y(t) \) to solve for best \( \hat{y}(t) \) that approximates \( s(t) \) best. The optimization can be done either by brute-force search, or gradient descent algorithm since the objective function \( O_2 \) is smooth and convex (see Section 6). In this paper, for simplicity we only do brute-force search for each objective function for comparison.

We also propose some other objective functions which have similar structures as \( O_2 \):

\[
O_3 = \int \prod_{i=1}^{N} |y'_i(t)|dt
\]

\[
O_4 = \int \log \prod_{i=1}^{N} |y'_i(t)|dt
\]

\[
O_5 = \sum_{i=1}^{N} \int \sqrt{1 + (y'_i(t))^2}dt
\]

All these above objective functions come from the arc lengths of each signals, and are named geometrical objective functions for BSS.

Comparing with the objective function of sum of marginal entropies, the advantages of these functions are: they computes easier and faster than estimation of densities; they consider the derivative information of signals; they do not assume the independence, and work for the case where sources are not independent (for example, the counterexample shown in Fig. 2).

Other than this dissertation, there do exist previous work that proposed other functions approximating the traditional MI objective functions. The most famous ones are the following, proposed in the work in [16]:

\[
G_1(y) = \frac{1}{a_1} \log \cosh(a_1 y)
\]

\[
G_2(y) = -\frac{1}{a_2} \exp(-a_2 y^2 / 2)
\]

\[
G_3(y) = \frac{1}{4} y^4
\]

where \( a_1 \) and \( a_2 \) are hyperparameters.

In Section 6.1 we show the function graph of each of the above objective functions. The results showed that all these geometrical objectives and the contrast functions agree at similar global minimum, up to some approximation error, which indicates that all these objective functions have similar behaviors in the BSS problems, and are effective approaches. However, among them, the geometrical objectives have significant better precision, especially \( O_3 \) and \( O_5 \), which indicates that the geometrical objective functions not only share good properties with the contrast functions, but also have better performance. Therefore, they are competitive replacements of contrast functions, and have both theoretical and practical potentials.

5 Frequency Domain Approaches and the New Assumption

5.1 Motivation

Among the new objective functions proposed above, \( O_3(R) = \int |y'_1(t) y'_2(t)|dt \) has the simplest formula. An immediate question then comes up: does it work if we simplify it further by taking
assumption that can construct a PCA-like structure for solving BSS, bypassing the optimization.

However, this approach is inspiring, which gives the motivation of this section: finding a reasonable enough as an assumption.

derivatives. Empirically, we may assert that the orthogonality on derivative signals is not good enough as an assumption.

is easier to find counterexamples for the assumption of orthogonal derivatives than the assumption of independence. Fig. 3 shows one of the examples where the original signals do not have orthogonal derivatives. As what we discuss in Section 1, a BSS problem is highly open, and it can only be solved with restrictions based on assumptions. However, assumptions are not true away the absolute value sign, i.e. \( \hat{O}_3(R) = \int y_1'(t)y_2'(t)dt \)? From the experiment results in Section 6.1, we can observe that it has worse performance than \( O_3 \), but its error was acceptable for a practical BSS task.

For different \( s_1(t) \) and \( s_2(t) \), most likely \( \int (s_1'(t))^2dt \neq \int (s_2'(t))^2dt \), then suppose that the minimization of \( \hat{O}_3(R) \) leads to

\[
\min \int y_1'(t)y_2'(t)dt = \int s_1'(t)s_2'(t)dt = 0,
\]

This induces the actual assumption of \( \hat{O}_3 \), other than approximating MI. Accepting this assumption, we can solve BSS by solving a PCA problem on the derivatives of given signals \( z_1(t) \) and \( z_2(t) \):

Suppose that \( \hat{R} \) is the correct rotation matrix to be solved, i.e. \( s(t) = \hat{R}z(t) \). Taking derivatives on both sides, we have

\[
s'(t) = \hat{R}z'(t).
\]

Hence,

\[
\int z'(t)(z'(t))^Tdt = \hat{R}^T \int s'(t)(s'(t))^Tdt \hat{R}.
\]

Since \( \int (s_1'(t))^2dt \neq \int (s_2'(t))^2dt \) and \( \int s_1'(t)s_2'(t)dt = 0, \int s'(t)(s'(t))^Tdt \) is a nontrivial diagonal matrix. Therefore, similar to the first PCA step, by eigen decomposition of \( \int z'(t)(z'(t))^Tdt \), we are able to get \( \hat{R} \). We call this approach Derivative-PCA.

The Derivative-PCA approach is based on the assumption that \( \int (s_1'(t))^2dt \neq \int (s_2'(t))^2dt \) and \( \int s_1'(t)s_2'(t)dt = 0, \int s'(t)(s'(t))^Tdt \). In other words, for a BSS problem, suppose that we do not have any assumption on independence, but an assumption on orthogonality of derivatives of the source signals, and then we are able to solve for the second rotation by another PCA step on the derivative signals. As what we discuss in Section 1 a BSS problem is highly open, and it can only be solved with restrictions based on assumptions. However, assumptions are not true or false. An assumption works if it meets the real cases. Just like what we discussed above: in most cases the independence assumption works, but there are also counterexamples. Similarly, if most source signals that are sampled “independently” have orthogonal derivatives, then the above approach would give correct approximations to the sources. Unfortunately, practically speaking, it is easier to find counterexamples for the assumption of orthogonal derivatives than the assumption of independence. Fig. 3 shows one of the examples where the original signals do not have orthogonal derivatives. Empirically, we may assert that the orthogonality on derivative signals is not good enough as an assumption.

However, this approach is inspiring, which gives the motivation of this section: finding a reasonable assumption that can construct a PCA-like structure for solving BSS, bypassing the optimization

![Figure 3:](image)

Figure 3: Countereexamples of the assumption of derivative orthogonality. Each figure shows a scatter plot of the derivatives of a pair of source signals. From the figures we can observe that the derivatives are not orthogonal. The covariance of each pair of derivatives are: (in order) 0.0195, 0.0881, 0.0580, which are relatively large.
procedure. This approach falls in the second-order-statistics category for BSS. And as we discussed in Section 2, AMUSE, SOBI, and STFD are well-known approaches in this category. In the following, we propose a new approach FT-PCA following this idea, and discuss the reasonability of its assumption comparing with AMUSE and SOBI.

### 5.2 The Fourier Transform Approach

The goal of the second step is to solve for \( s(t) \) and \( \hat{R} \) from \( s(t) = \hat{R}z(t) \). Applying Fourier transform (FT) on both sides, we have

\[
S(\omega) = \hat{R}Z(\omega),
\]

where \( \omega \in \mathbb{R} \) is the frequency, \( S(\omega) \) is the FT of \( s(t) \), and \( Z(\omega) \) is the FT of \( z(t) \).

By Parseval’s Theorem, we know that

\[
\int |Z_i(\omega)|^2 d\omega = \int (z_i(t))^2 dt = 1
\]

for \( t = 1, 2 \), and

\[
\int S_i(\omega)|^2 d\omega = \int (s_i(t))^2 dt = 1
\]

for \( t = 1, 2 \), and

\[
\int Z_1(\omega)\overline{Z_2(\omega)}d\omega = \int z_1(t)z_2(t)dt = 0
\]

\[
\int S_1(\omega)\overline{S_2(\omega)}d\omega = \int s_1(t)s_2(t)dt = 0.
\]

Therefore, \( \int Z(\omega)(Z(\omega))^H d\omega = \int S(\omega)(S(\omega))^H d\omega = I \) are both the identity matrix.

The above transformation gives trivial results since \( \hat{R} \) is not able to be solved from

\[
\int Z(\omega)(Z(\omega))^H d\omega = \hat{R}^T \int S(\omega)(S(\omega))^H d\omega \hat{R}
\]

which is equivalent to \( I = \hat{R}^T \hat{R} = I \). In fact, any change of basis applied to the function space of the signals have similar results, due to the Parseval’s Theorem. However, inspired by the above derivative orthogonality assumption, we can apply kernel tricks as follows:

Multiplying both sides of \( S(\omega) = \hat{R}Z(\omega) \) by a certain complex function \( \phi(\omega) \) that is nonzero on a set with positive Lebesgue measure, we have

\[
\phi(\omega)S(\omega) = \phi(\omega)\hat{R}Z(\omega).
\]

Let \( K(\omega) = \phi(\omega)\overline{\phi(\omega)} \) be the kernel function, then the elements of the covariance matrices \( \int K(\omega)Z(\omega)(Z(\omega))^H d\omega \) and \( \int K(\omega)S(\omega)(S(\omega))^H d\omega \) become inner products of FT of signals in the kernel space defined by \( K(\omega) \), and we have

\[
\int K(\omega)Z(\omega)(Z(\omega))^H d\omega = \hat{R}^T \int K(\omega)S(\omega)(S(\omega))^H d\omega \hat{R}.
\]

For convenience, let us name each of the elements in the above matrices as follows:

\[
S_{ij} = \int K(\omega)S_i(\omega)\overline{S_j(\omega)}d\omega
\]
and
\[ Z_{ij} = \int K(\omega)Z_i(\omega)\overline{Z_j(\omega)}d\omega \]
under the case where \( K(\omega) \) has no ambiguity, then we can write that
\[ \int K(\omega)Z(\omega)(\overline{Z(\omega)})^Hd\omega = \begin{pmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{pmatrix} \]
and
\[ \int K(\omega)S(\omega)(\overline{S(\omega)})^Hd\omega = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}. \]

Suppose there exists a kernel space defined by \( K(\omega) \), so that \( S_{11} \neq S_{22} \) and \( S_{12} = 0 \), then \( \int K(\omega)S(\omega)(\overline{S(\omega)})^Hd\omega \) is a nontrivial diagonal matrix, and \( \hat{R} \) can be solved by eigen decomposition of \( \int K(\omega)Z(\omega)(\overline{Z(\omega)})^Hd\omega \) by the uniqueness property of eigen decompositions. Formally, suppose that \( \int K(\omega)S(\omega)(\overline{S(\omega)})^Hd\omega \) is a diagonal but not the identity matrix, the eigen decomposition of \( \int K(\omega)Z(\omega)(\overline{Z(\omega)})^Hd\omega \) can be written as
\[ \int K(\omega)Z(\omega)(\overline{Z(\omega)})^Hd\omega = E^T\Lambda E. \]

Then \( \int K(\omega)S(\omega)(\overline{S(\omega)})^Hd\omega \) and \( \Lambda \) only differ by row switching, and \( E \) and \( \hat{R} \) only differ by row switching and signs. This approach of solving for the second rotation in BSS is called FT-PCA.

Note that in the ideal case where \( S_{12} = S_{21} = 0 \), \( \int K(\omega)S(\omega)(\overline{S(\omega)})^Hd\omega \) is a real matrix. And \( \hat{R} \) is real, \( \int K(\omega)Z(\omega)(\overline{Z(\omega)})^Hd\omega \) is also a real matrix. Therefore, under the ideal kernel \( K(\omega) \), we only need to consider the real part of the matrix \( \int K(\omega)Z(\omega)(\overline{Z(\omega)})^Hd\omega \), and consider the imaginary part as error.

The key points of FT-PCA are the reasonability of assuming the kernel orthogonality, i.e. \( S_{ij} = \int K(\omega)S_i(\omega)\overline{S_j(\omega)}d\omega = 0 \) for \( i \neq j \), and if this is reasonable, how to find the kernel \( K(\omega) \).

From Section 5.1 we know that the approach of Derivative-PCA works for some inputs, but does not work for others. Note that by applying FT to both sides of Eq. (3) we have
\[ \omega S(\omega) = \omega \hat{R}Z(\omega), \]
and the Derivative-PCA approach is just a special case of FT-PCA where the kernel \( K_1(\omega) = \omega^2 \). This candidate kernel works for some inputs, but not perfect since there exist counterexamples.

By noticing the function graph of \( K(\omega) = \omega^2 \), we observe that this kernel is similar to a window function that focuses on the high frequency intervals of the source signals, and hence, an immediate alternative option comes up:
\[ K_2(\omega) = \frac{1}{1 + |\omega|} \]
which grants low frequency parts of the signals more weights. See Fig. 4. Certainly, we can generalize it by
\[ K_3(\omega) = \frac{1}{1 + |\omega - \omega_0|} \]  (5)
where \( \omega_0 \) is the center of this window-like function. By picking different \( \omega_0 \)'s, \( K(\omega) \) focuses on different intervals of the frequency domain by giving that interval higher weights, so as to grant \( S_{12} \) close to zero and \( S_{11} \neq S_{22} \). If there exists an ideal \( \omega_0 \) so that the kernel orthogonality assumption holds, then FT-PCA can theoretically solve the BSS problems. In Section 5.3 we show that the kernel orthogonality assumption is reasonable, and in Section 5.4 we show that the ideal \( \omega_0 \) is not available, but provide a strategy to search for good \( \omega_0 \) to practically solve BSS using FT-PCA.

With \( \omega_0 \) as the hyper-parameter, we have the FT-PCA Algorithm described as [1].
Algorithm 1 The FT-PCA algorithm with $\omega_0$ as a hyper-parameter.

**Input:** signals $x(t)$, the hyper parameter $\omega_0$.

**First Step:**
1. Centering $x$ by $x \leftarrow x - \bar{x}$;
2. Let $x = U\Sigma V^T$ be the SVD of $x$;
3. Compute $z = \Sigma^{-1}U^T x$;

**Second Step:**
4. Compute the FT of $z$ as $Z(\omega)$;
5. $K(\omega) = \frac{1}{1+|\omega-\hat{\omega}_0|}$;
6. Compute eigen decomposition of the matrix $\text{Re}(\int K(\omega)Z(\omega)(Z(\omega))^H \,d\omega) = EE^T$;
7. $S(\omega) = E^T Z(\omega)$;
8. Compute the inverse FT of $S(\omega)$ as $s$.

**Output:** the separated signals $s(t)$.

---

Figure 4: The function graph of $K_1(\omega)$ and $K_2(\omega)$. **Left:** the function graph of $K_1(\omega)$ which gives higher weights to high frequency parts; **Right:** the function graph of $K_2(\omega)$ which gives higher weights to lower frequency parts. By extending $K_2(\omega)$ to $K_3(\omega)$, the center shift $\omega_0$ grants the kernel focusing on frequency parts defined by user.
5.3 The Reasonability of the Assumption

Firstly, let’s discuss the assumptions of the previous approaches. The assumptions of the second-order-statistics approaches SOBI and AMUSE are deficient. The assumption of AMUSE is that

\[ \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} s(t + \tau)s(t)^* = \text{diag}[\rho_1(\tau), \ldots, \rho_n(\tau)] \]

for a certain \( \tau \). And the assumption of SOBI shares the same formula, except that it is true for a set of different \( \tau \)’s. Both algorithms did not give a clear approach to determine which \( \tau \) satisfies the assumption. If we assume that for any \( \tau \),

\[ s_i(t + \tau)s_j(t) = 0 \]

it implies that the correlation of the two functions

\[ s_i(t) * s_j(t) = \int f(\tau)g(t + \tau)d\tau = 0 \]

and hence,

\[ \overline{S_i(\omega)}S_j(\omega) = 0 \]

where \( S_i(\omega) \) is the Fourier transform of \( s_i(t) \). Obviously, except for special cases, the multiplication of the Fourier transform of two signals cannot be a zero function. Therefore, this assumption is too strong. If we cannot assume the autocorrelation being zero for any time shift \( \tau \), the assumption is not complete since we are not able to aware of how to select the correct time shifts. And most importantly, it does not make physical sense why such time shifts should exist so that the autocorrelation of the source signals are zero. The SOBI algorithm introduces the joint diagonalization strategy, to select a collection of \( \tau \)’s, and compute based on the average pattern of the covariance matrix, in order to bypass the deficiency of their assumption. The case of STFD is similar, where the assumption is that special time and frequency shifts \( t \) and \( f \) can be selected so that the covariance matrix has the diagonal structure. This is not guaranteed by theory, since only very strong assumption can guarantee the diagonal structure for any shifts \( t \) and \( f \). And the algorithm has to apply joint diagonalization. Therefore, since the assumption of SOBI and STFD are either too strong or not applicable, their algorithms are heuristic.

Comparing to the incompleteness of the assumptions of previous approaches, we discuss the assumption of FT-PCA in the following.

The first issue that we need to discuss is the existence of \( K(\omega) \). Suppose that \( |S_1(\omega)|^2 \neq |S_2(\omega)|^2 \) on a subset \( D \subset \mathbb{R} \) where \( m(D) > 0 \). (Without loss of generality, we can suppose that \( |S_1(\omega)|^2 - |S_2(\omega)|^2 > 0 \) on \( D \), since if there exists a subset with positive Lebesgue measure such that \( |S_1(\omega)|^2 \neq |S_2(\omega)|^2 \), we can always pick a subset of it such that \(|S_1(\omega)|^2 > |S_2(\omega)|^2 \) or \( |S_1(\omega)|^2 < |S_2(\omega)|^2 \).) Then the nonnegative kernel function \( K(\omega) \) with \( \int K(\omega)d\omega > 0 \) exists so that \( S_{11} \neq S_{22} \), since we can always pick

\[ K(\omega) = 1_D(\omega) \]

where

\[ \int K(\omega)d\omega = m(D) > 0. \]

On the other hand, suppose that \( |S_1(\omega)|^2 \neq |S_2(\omega)|^2 \) only on a subset of the frequency domain with zero Lebesgue measure, then \( K(\omega) \) does not exist. Because

\[ \int (|S_1(\omega)|^2 - |S_2(\omega)|^2)^2 d\omega = 0, \]
and hence, for any $K(\omega)$,
\[
|S_{11} - S_{22}| = \left| \int K(\omega)(|S_1(\omega)|^2 - |S_2(\omega)|^2) d\omega \right|
\leq \int |K(\omega)(|S_1(\omega)|^2 - |S_2(\omega)|^2)| d\omega
\leq \left( \int (K(\omega))^2 d\omega \right)^{1/2} \left( \int (|S_1(\omega)|^2 - |S_2(\omega)|^2)^2 d\omega \right)^{1/2}
= 0
\]

This indicates that, if the two source signals have the same energy density almost everywhere, no kernel functions exist so that the two signals can be separated by FT-PCA. Therefore, we have an necessary condition for the source signals: FT-PCA does not work for signals whose power spectral densities are the same. This necessary condition excludes the cases where the source signals are too close, for example $s_1(t) = \sin t$ and $s_2(t) = \cos t$.

Secondly, suppose that there exists an interval $D \subset \mathbb{R}$ such that all the following conditions are satisfied:

1. $S_1(\omega) \neq 0$ on a subset $D_1 \subset D$ with $m(D_1) > 0$
2. $S_2(\omega) \neq 0$ on a subset $D_2 \subset D$ with $m(D_2) > 0$
3. $D_1 \cap D_2 = \emptyset$

Then we immediately have that $\int_D |S_1(\omega)|^2 d\omega \neq \int_D |S_1(\omega)|^2 d\omega$ and $\int_D S_1(\omega)S_2(\omega)d\omega = 0$. The physical meaning of the these conditions can be interpreted directly: if there exists an interval on which the two source signals have exclusive spectral density, then FT-PCA works. This sufficient condition gives us a clear intuition of the reasonability of FT-PCA. The nature of the second-order-statistics approach is to find a subset of the domain (either time domain or frequency domain) with positive Lebesgue measure where the source signals are clearly different. Since the linear combination matrix $A$ is applied to the whole domain, the BSS problem can be solved algebraically by finding a subset where the source signals have the characteristics to be separated. And the reason to pick frequency domain as the approach is clear: in practice, it makes sense that different source signals almost always have different density distributions, and it is almost always possible to find subsets (no matter how small it is) where the spectrums are approximately exclusive. On the other hand, using other possible assumptions is less practical, for example, trying to find an interval in time domain where the signals have exclusive subsets is unlikely, and thus these kinds of approaches do not work.

Practically, since no spectral functions contain subsets where the spectral power is exactly zero, within acceptable error, if there exists an interval on which one density function has large values while the other is close to zero, and vise versa, then the above conditions can be approximately satisfied. See Fig. 5. And in practice we do not use a true window function as the kernel but the Eq. 5 in order to make the approximation more smooth.

On the other hand, from the experimental point of view, we observe that, for each pair of source signals we examined, there always exists a best $\omega_0$ such that $\int K(\omega)S(\omega)(S(\omega))^H d\omega$ is close to a nontrivial diagonal matrix most. And suppose that we know this specific $\omega_0$ for this pair of source signals, we are able to solve the BSS nearly perfectly using FT-PCA, where the error is extremely small. See Fig. 6. This also supports the reasonability of the assumption of FT-PCA.
Figure 5: An example of an interval where two source signals has approximately exclusive spectral density. The figures show the segments of the Fourier transform of each signal in this specific interval (only the real parts). The variance of the left signal in this interval is 0.1626; the variance of the right signal in this interval is 0.0980 and the covariance of them is -0.0036+0.0017i.

Figure 6: Examples of best $\omega_0$’s that diagonalize $\int K(\omega)S(\omega)(S(\omega))^H d\omega$. Each figure shows the Re($S_{12}$) with respect to $\omega_0$ for a different pair of input source signals. In each example, there always exists a best $\omega_0$ so that Re($S_{12}$) is close to zero the most. With these best $\omega_0$’s, the FT-PCA result has errors (from left to right): 0.000130, 0.0133, 0.00028, 0.00003, which are much smaller than any other approaches. Please note that, these $\omega_0$’s are the ideal cases, based on the analysis of Section 5.4, we know that there do not exist approaches to search for these ideal $\omega_0$’s. We can only use heuristic strategies to searching for good $\omega_0$’s which has larger errors than these perfect solutions. This figure is shown to support the reasonability of the assumption of FT-PCA.
5.4 The Heuristic Strategy to Search for $\omega_0$

Unfortunately, though FT-PCA has reasonable assumptions and solid theory, practically it is not easy to search for the ideal interval where the source signals are exclusive only based on the input signals $z(t)$. This means that for the searching of $\omega_0$, there are no theory to guarantee the optimization. See the following analysis:

Our task is to apply different $\omega_0$ as the shifts of the kernel function in

$$K(\omega) = \frac{1}{1 + |\omega - \omega_0|},$$

and search for best $\omega_0$ so that $|S_{11} - S_{22}|$ is not close to zero while $|S_{12}|$ is minimized, based on the values of $Z_{ij}$ for $i, j = 1, 2$ that we computed according to each $\omega_0$ that we apply. Practically, we need $\int K(\omega)S(\omega)(S(\omega))^H d\omega$ more “diagonal” than $\int K(\omega)Z(\omega)(Z(\omega))^H d\omega$.

Without loss of generality, we can write the rotation matrix

$$\tilde{R} = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}$$

for a certain rotation angle $\theta$. Then, by

$$\int K(\omega)Z(\omega)(Z(\omega))^H d\omega = \tilde{R}^T \int K(\omega)S(\omega)(S(\omega))^H d\omega \tilde{R}$$

we have

$$Z_{11} = \cos^2 \theta S_{11} + 2 \cos \theta \sin \theta \Re(S_{12}) + \sin^2 \theta S_{22},$$

$$Z_{12} = \cos \theta \sin \theta (S_{22} - S_{11}) + (\cos^2 \theta - \sin^2 \theta) \Re(S_{12}) + \Im(S_{12}),$$

$$Z_{21} = \cos \theta \sin \theta (S_{22} - S_{11}) + (\cos^2 \theta - \sin^2 \theta) \Re(S_{12}) - \Im(S_{12}),$$

$$Z_{22} = \sin^2 \theta S_{11} - 2 \cos \theta \sin \theta \Re(S_{12}) + \cos^2 \theta S_{22}. \quad (6)$$

Clearly, each $Z_{ij}$ is a mixture of $S_{11}, S_{22},$ and $S_{12}$, and $|S_{11} - S_{22}|$ and $|S_{12}|$ cannot be solved separately by $Z_{ij}$ without knowing $\theta$. And we are not able to understand the changing of $|S_{11} - S_{22}|$ and $|S_{12}|$ by observing the changing of $Z_{ij}$, either. Hence, theoretically there is no way to guarantee that the optimized $\omega_0$ can be searched based on the $Z_{ij}$ values we observed.

However, there exist heuristic strategies to search for good $\omega_0$.

From the eigen decomposition structure of the equation

$$\int K(\omega)Z(\omega)(Z(\omega))^H d\omega = \tilde{R}^T \int K(\omega)S(\omega)(S(\omega))^H d\omega \tilde{R}$$

and the relationship of traces and determinants, we observe that:

$$Z_{11} + Z_{22} = S_{11} + S_{22} \quad (7)$$

and

$$Z_{11}Z_{22} - |Z_{12}|^2 = S_{11}S_{22} - |S_{12}|^2. \quad (8)$$

From Eq. 7 and 8 we have

$$(Z_{11} - Z_{12})^2 + 4|Z_{12}|^2 = (S_{11} - S_{22})^2 + 4|S_{12}|^2$$
where the left hand side is available, while the right hand side contains the sum of squares of the two terms that we care about most. For convenience, let’s name them as follows:

\[
f_1 = (Z_{11} - Z_{12})^2 + 4Z_{12}^2
\]

\[
f_2 = (S_{11} - S_{22})^2
\]

\[
f_3 = 4|S_{12}|^2.
\]

Since \( f_3 \) is close to zero, the values of \( f_1 \) is dominated by \( f_2 \). However, maximizing \( f_1 \) is not a good strategy, since we do not need \( F_2 \) to be maximized but just not close to zero, and practically by maximizing \( f_1 \), the part of \( f_3 \) gets larger which leads to worse solution, since the diagonal property of \( \int K(\omega)S(\omega)(S(\omega))^Hd\omega \) is more sensitive with the changing of \( f_3 \). From Fig. 7 we can have an intuitive idea about the absolute values of \( f_1, f_2, f_3 \). From the figure, as well as observations on other examples, we notice that, minimizing instead of maximizing \( f_1 \) should be a better solution, since the magnitude of \( f_3 \) will be controlled while \( f_1 \) decreases, which can guarantee \( f_3 \) being close to zero. After the minimum of \( f_1 \) is found, we need to move the shift \( \omega_0 \) steps away from the arg min \( f_1 \). This is because that there exists a small interval around arg min \( f_1 \) where \( f_2 \) drops heavily so as to be even less than \( f_3 \), and practically this will lead to \( \int K(\omega)S(\omega)(S(\omega))^Hd\omega \) getting too close to the identity matrix. By moving away a certain distance from the minimum point, we are able to have \( f_2 \) significantly larger than \( f_3 \), and practically this can be considered as \( |S_{11} - S_{22}| \) being far from zero while \( |S_{12}| \) is close to zero. From the experient results shown in Section 6 through this heuristic strategy we were able to get notable results even better than SOBI.

6 Experiments

6.1 The Behaviors of the Geometrical Objective Functions

Firstly, we show the function graphs of different objective functions proposed in Section 4.3 and compare them with MI and contrast functions proposed for FastICA. We randomly picked two natural signals (two segments of true audio files) \( s_1(t) \) and \( s_2(t) \), normalized them by removing their mean and standardizing their covariance matrix, and then applied a random \( 2 \times 2 \) mixing matrix \( A \) to get the mixed signal observations \( x_1(t) \) and \( x_2(t) \). Then the first step – PCA was operated on the observed signals to get the standardized signals \( z_1(t) \) and \( z_2(t) \). Finally, we searched the angle \( \theta \) from \( -\pi \) to \( \pi \), and computed the recovered signals \( y_1(t) \) and \( y_2(t) \) by

Figure 7: The function graphs of \( f_1, f_2, f_3 \) with respect to the kernel shift \( \omega_0 \). **Left:** The overall function graphs. **Mid:** A closer view by showing the function values within range \([0, 1]\). **Right:** The function graphs around the global minimum, where \( f_2 \) falls below \( f_3 \). The good \( \omega_0 \) can be selected at the points outside the crossing of \( f_2 \) and \( f_3 \), where \( f_3 \) is small enough and \( f_2 \) is not close to zero.
Figure 8: Comparison of different objective functions for the second step. **Left**: The overall objective function graphs. **Right**: The objective function graphs in a neighborhood of the global minimum.

Table 1: Average error of each objective function with MI.

|       | $O_1$ | $O_2$ | $O_3$ | $O_4$ | $O_5$ | $\tilde{O}_3$ | $G_1$ | $G_2$ | $G_3$ |
|-------|-------|-------|-------|-------|-------|----------------|-------|-------|-------|
| Error (rad) | 0.0623 | 0.0623 | **0.0621** | 0.4484 | 0.0624 | 0.0680 | 0.3457 | 0.0706 | 0.3463 |

$$y_\theta(t) = R(\theta)z(t)$$

where $R(\theta)$ is the rotation matrix

$$
\begin{pmatrix}
\cos \theta & -\sin \theta \\
\sin \theta & \cos \theta
\end{pmatrix}
$$

of each rotation angle $\theta$, and computed the values of each objective functions, i.e. $\text{Obj}(y_\theta(t))$. Fig. 8 shows the objective function graphs with respect to the rotation angle. The objective values shown in the figure were normalized (removing means and divided by standard deviations) so that they are comparable. In the figures, 'mi' means the original MI objective, i.e. the sum of marginal entropies; 'issra' is the objective function $O_1$ in Section 4.3; 'con1' to 'con3' are the FastICA contrast functions $G_1$ to $G_3$; and 'obj1' is $O_3$, 'obj2' is $O_2$, 'obj3' is $O_4$, and 'obj4' is $O_5$, in Section 4.3.

From Fig. 8 we observe that for all the objective functions, their function graph has similar shapes, and the global minimum are very close. This shows the fact that all the proposed geometrical objective functions are good approximations to the MI objective (which is in fact the summation of marginal entropies).

We also did synthetic experiments to investigate the behaviors of each geometrical objective functions, as well as the MI objective and FastICA contrast functions. The synthetic experiment was done the same way as above: Firstly we randomly picked a pair of source signals, and standardize them. Then we applied random mixing matrix to generate the observed signals. For the BSS process, we did the PCA-ICA steps, and in the ICA step, we optimized each objective function to get its solution, as well as the computation time. The synthetic experiment was repeated for 20 times, and the errors and CPU time were provided as mean/std of the 20 results for each objective function.

Table 1 shows the differences between the MI results and each other objective function, where $\tilde{O}_3 = \int \prod_i y'_i(t) \, dt$ is the objective function of the Derivative-PCA. The errors were computed as the differences of the resulted rotation angles. From the table, we observe that $O_3$ approximates the MI objective function best.
Table 2: Average error of each objective function with the ground truths and average computational time of each objective functions.

|           | MI     | O₂     | O₃     | O₅     | O₃     | G₁     | G₂     | G₃     |
|-----------|--------|--------|--------|--------|--------|--------|--------|--------|
| Error (deg)| 0.7683 | 0.7685 | 0.7735 | 0.7535 | 0.7535 | 4.8035 | 2.2080 | 1.4273 |
| Std       | 1.0697 | 1.0679 | 1.0840 | 1.0668 | 1.0703 | 2.1372 | 4.3855 | 1.1455 |
| Time (sec)| 6.14   | 10.40  | 9.55   | 2.13   | 2.62   | 2.14   | 3.00   | 1.39   |

Table 2 shows the mean errors computed between the solution of each objective function and the ground truth. The errors were the differences between the resulted rotation angles of each objective function and the true rotation angle. Please note that, since we standardized the source signals, the source signals and the first step (PCA) results exactly differ by a rotation, and hence we can compare the true rotation angle with each approach. The optimizations were done by brute-force search. The CPU time was the average of all 20 repeated experiments. In each experiment, each objective function was computed 1801 times (from $-180^\circ$ to $180^\circ$ with step size as 0.1°).

From the table we can observe that all the geometrical objective functions worked well, much better than all contrast functions. And the best ones: $O₃$ and $O₅$ worked even slightly better than MI. Among the geometrical objectives, $O₃$ has the least computational time due to its simple formula. The experiment results supports our assertion that the geometrical objective functions are good candidates for the ICA step of BSS problems, especially $O₃$ and $O₅$ which compute simply and fast, and has promisingly good precision for mixed signal recovering. Additionally, being preliminary convex functions with their derivatives available, $O₃$ and $O₅$ can be applied to a gradient-based optimization algorithm, and serve as a good algorithm for ICA-based BSS, replacing the traditional MI objective functions. The only disadvantage of $O₃$ is that the time complexity gets higher than FastICA as the number of signals grows. And $O₅$ can be adapted to the FastICA algorithm for it has the form of aggregation.

6.2 The Comparison of the Geometrical Objective Functions and the FT-ICA approach

In this experiment, we compare the MI objective with the above proposed objective $O₃$, as well as the contrast function $G₂$ (which approximates MI well and has fastest computational time in the above experiments) and the second-order approaches: SOBI and FT-PCA. The inputs are 72 pairs of real source signals (audio segments) that were standardized to have zero mean and identity covariance. Random mixing matrices were applied to each pair of source signals, and the compared approaches were applied to solve for the second rotation. The errors were computed as the differences of the rotation angles solved by each approach with the ground truth. Table 3 shows the average error for all 72 rounds of experiments, without noise or signal-noise-ratio (SNR) being 100, 50, and 20. For the FT-PCA approach, we adopted the heuristic strategy that we described in Section 5.4, where the searching radius after minimization of $f₁$ is fixed as 100 and the step size of $ω₀$ is 0.001.

From the table, we observe that, when there are no noise, or the SNR = 100 and 50, every approach works well. Errors of all approaches except Derivative-PCA are less than 2 degree, which indicates that all these approaches have practically acceptable precision. For the optimization approaches (MI, $G₂$, and $O₃$), MI and $O₃$ worked better, and $G₂$ worked worse. And MI is slightly worse than $O₃$ in average though they are very close. For the second-order-statistics approaches, FT-PCA worked best (its precision is very close to MI), and Derivative-PCA worked worst. When the SNR is 20, none of these approaches worked. The above results supports that, for optimization based
Table 3: The error table of different compared approaches for the synthetic BSS experiment.

| Error (deg) | MI      | $G_2$ | $O_3$ | FT-PCA  | Derivative-PCA | SOBI |
|-------------|---------|-------|-------|---------|----------------|------|
| No Noise    | 0.7668  | 1.0995| 0.7807| 0.8735  | 2.7195         | 1.0195|
| SNR = 100   | 0.7970  | 1.1447| 0.7878| 0.8158  | 2.8338         | 0.9767|
| SNR = 50    | 0.8777  | 1.1796| 0.8724| 1.0363  | 2.9054         | 1.0750|
| SNR = 20    | 8.026   | 5.612 | 5.779 | 23.856  | 24.253         | 17.942|

approaches, $O_3$ is indeed a good objective function, which is significantly better than the contrast functions and competitive with MI. For second-order-statistics approaches, FT-PCA works better than SOBI, hence it is an effective approach for BSS problems that is based on different assumptions than independence and has simple and fast algorithm.

7 Conclusions

In this paper, we highlight two main contributions. First, we point out the model of ICA-based approaches for BSS, and based on the relationship between SPC and MI, we apply SPC to BSS to propose geometrical objective functions based on the property of the joint signals, whose computational time and precision are both excellent. Second, we proposed a new second-order-statistics approach, FT-PCA, that assumes the kernel orthogonality of signals in the frequency domain, and solve the BSS problem by applying Fourier transforms and solve a second eigen decomposition. Comparing with other second-order-statistics approaches, FT-PCA has a more reasonable assumption that bypasses the independence concepts, and has a simple and fast algorithm that does not require any optimization or joint diagonalization, given good hyper-parameters. We also propose heuristic strategies for searching good hyper-parameter, which was proven efficient in the experiment section.

A immediate future work is to extend the idea of FT-PCA to nonstationary signals, and propose a generalized algorithm that works for signals which have different frequency distribution for different time intervals. Another potential future work is to apply FT-PCA to nonlinear ICA problems.

Acknowledgments

We acknowledge helpful conversations with Mingyuan Gao and Yuan Zhou.

References

[1] F. Abrard and Y. Deville. A time–frequency blind signal separation method applicable to underdetermined mixtures of dependent sources. *Signal processing*, 85(7):1389–1403, 2005.

[2] F. R. Bach and M. I. Jordan. Kernel independent component analysis. *Journal of machine learning research*, 3(Jul):1–48, 2002.

[3] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, 1995.
[4] A. Belouchrani, K. Abed-Meraih, J.-F. Cardoso, and E. Moulines. A blind source separation technique using second-order statistics. *IEEE Transactions on signal processing*, 45(2):434–444, 1997.

[5] A. Belouchrani and M. G. Amin. Blind source separation based on time-frequency signal representations. *IEEE Transactions on Signal Processing*, 46(11):2888–2897, 1998.

[6] J.-F. Cardoso. Blind signal separation: statistical principles. *Proceedings of the IEEE*, 86(10):2009–2025, 1998.

[7] G. Chabriel, M. Kleinsteuber, E. Moreau, H. Shen, P. Tichavsky, and A. Yeredor. Joint matrices decompositions and blind source separation: A survey of methods, identification, and applications. *IEEE Signal Processing Magazine*, 31(3):34–43, 2014.

[8] L. Cohen. *Time-frequency analysis*, volume 778. Prentice hall, 1995.

[9] P. Comon. Separation of sources using higher-order cumulants. In *Advanced Algorithms and Architectures for Signal Processing IV*, volume 1152, pages 170–184. International Society for Optics and Photonics, 1989.

[10] P. Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

[11] P. Comon and C. Jutten. *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press, 2010.

[12] P. Comon, C. Jutten, and J. Herault. Blind separation of sources, part ii: Problems statement. *Signal processing*, 24(1):11–20, 1991.

[13] A. Delorme and S. Makeig. Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *Journal of neuroscience methods*, 134(1):9–21, 2004.

[14] C. Févotte and C. Doncarli. Two contributions to blind source separation using time-frequency distributions. *IEEE Signal Processing Letters*, 11(3):386–389, 2004.

[15] A. Hyvärinen. New approximations of differential entropy for independent component analysis and projection pursuit. In *Advances in neural information processing systems*, pages 273–279, 1998.

[16] A. Hyvarinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE transactions on Neural Networks*, 10(3):626–634, 1999.

[17] A. Hyvärinen, J. Karhunen, and E. Oja. *What is independent component analysis?* Wiley Online Library, 2001.

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

[19] S. Ikeda and N. Murata. A method of ica in time-frequency domain. In *in Proc. ICA*. Citeseer, 1999.

[20] C. J. James and C. W. Hesse. Independent component analysis for biomedical signals. *Physiological measurement*, 26(1):R15, 2004.
[21] A. Jourjine, S. Rickard, and O. Yilmaz. Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures. In Acoustics, Speech, and Signal Processing, 2000. ICASSP’00. Proceedings. 2000 IEEE International Conference on, volume 5, pages 2985–2988. IEEE, 2000.

[22] J. Lacoume and P. Ruiz. Sources indentification: a solution based on the cumulants. In Spectrum Estimation and Modeling, 1988., Fourth Annual ASSP Workshop on, pages 199–203. IEEE, 1988.

[23] T.-W. Lee and M. S. Lewicki. Unsupervised image classification, segmentation, and enhancement using ica mixture models. IEEE Transactions on Image Processing, 11(3):270–279, 2002.

[24] S. Makeig, A. J. Bell, T.-P. Jung, and T. J. Sejnowski. Independent component analysis of electroencephalographic data. In Advances in neural information processing systems, pages 145–151, 1996.

[25] H. Mermoz. Spatial processing beyond adaptive beamforming. The Journal of the Acoustical Society of America, 70(1):74–79, 1981.

[26] N. Mitianoudis and T. Stathaki. Pixel-based and region-based image fusion schemes using ica bases. Information Fusion, 8(2):131–142, 2007.

[27] A. Rajwade, A. Banerjee, and A. Rangarajan. Probability density estimation using isocontours and isosurfaces: applications to information-theoretic image registration. IEEE transactions on pattern analysis and machine intelligence, 31(3):475–491, 2009.

[28] H. Saruwatari, S. Kurita, and K. Takeda. Blind source separation combining frequency-domain ica and beamforming. In Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP’01). 2001 IEEE International Conference on, volume 5, pages 2733–2736. IEEE, 2001.

[29] L. Tong, V. Soon, Y. Huang, and R. Liu. Amuse: a new blind identification algorithm. In Circuits and Systems, 1990., IEEE International Symposium on, pages 1784–1787. IEEE, 1990.

[30] E. Weinstein, M. Feder, and A. V. Oppenheim. Multi-channel signal separation by decorrelation. IEEE transactions on Speech and Audio Processing, 1(4):405–413, 1993.

[31] O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. IEEE Transactions on signal processing, 52(7):1830–1847, 2004.