A Linear-time Independence Criterion Based on a Finite Basis Approximation

Longfei Yan
Victoria University of Wellington
Australian National University

W. Bastiaan Kleijn
Victoria University of Wellington

Thushara D. Abhayapala
Australian National University

]

Abstract

Detection of statistical dependence between random variables is an essential component in many machine learning algorithms. We propose a novel independence criterion for two random variables with linear-time complexity. We establish that our independence criterion is an upper bound of the Hirschfeld-Gebelein-Rényi maximum correlation coefficient between tested variables. A finite set of basis functions is employed to approximate the mapping functions that can achieve the maximal correlation. Using classic benchmark experiments based on independent component analysis, we demonstrate that our independence criterion performs comparably with the state-of-the-art quadratic-time kernel dependence measures like the Hilbert-Schmidt Independence Criterion, while being more efficient in computation. The experimental results also show that our independence criterion outperforms another contemporary linear-time kernel dependence measure, the Finite Set Independence Criterion. The potential application of our criterion in deep neural networks is validated experimentally.

1 Introduction

Measures of dependence between random variables have been extensively studied in statistics and in science. Statistical quantities such as non-Gaussianity, cross-cumulants and mutual information have been used for dependence detection in Independent Component Analysis (ICA) (Hyvarinen, 1999; Bell and Sejnowski, 1995; Cardoso and Souloumiac, 1993). Among them, mutual information is the most popular dependence measure. It measures the Kullback-Leibler divergence between the joint probability distribution of two random variables and the product of their marginal distributions.

Mutual information can be difficult to measure and optimize with a finite sample (Bach and Jordan, 2002). Alternative measurements in a Reproducing Kernel Hilbert Space (RKHS) have shown superior performance in detecting statistical dependence (Bach and Jordan, 2002; Gretton et al., 2005a,b). Bach and Jordan (2002) developed a kernel method to search for mapping functions in an RKHS that can achieve the maximal correlation, also known as the Hirschfeld-Gebelein-Rényi (HGR) maximum correlation coefficient (Hirschfeld, 1935; Gebelein, 1941; Rényi, 1959), between observed signals. This Kernel ICA approach is robust to near-Gaussianity and outliers. The Hilbert Schmidt Independence Criterion (HSIC) proposed by Gretton et al. (2005a) exhibits the best performance to date. It searches for RKHS functions that can maximize the norm of the cross-covariance operator. It is robust under challenging environments and alert to small deviations from independence.

The advantage of HSIC comes at a price. Its quadratic computational time slows down the calculation. Additionally, for input signals with a large sample size, it is infeasible to use HSIC to compute the statistical independence (Jitkrittum et al., 2017). This disadvantage cannot be overlooked in an era of big data. It is also observed that the stability of HSIC is not guaranteed as it is sensitive to the initial conditions. In ICA experiments, when the initial demixing matrix is not properly guessed, HSIC can suffer from local optima and perform poorly (Shen et al., 2009).

To address the high computational expense of HSIC, the Finite Set Independence Criterion (FSIC) was developed (Jitkrittum et al., 2017). Instead of construct-

Proceedings of the 23rdInternational Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

ing a Gram matrix that requires a computation between every single pair of signal observations, FSIC selects a finite set of test locations to be compared with input signals. Empirically it has been shown that a small set of test locations is sufficient to provide test power (Jitkrittum et al., 2017). This makes FSIC a linear-time algorithm. By focusing on the key regions captured by test locations that signify differences between two input signals, the estimation of the covariance matrix can be made more efficient and accurate.

In order to search for key regions and tune parameters for kernels, FSIC splits the data into training and test sets. Only the training set is exploited for the parameter tuning to avoid overfitting. The test set alone is used for independence measurement. This approach has two problems. Firstly, the training stage prevents FSIC from being an online algorithm. Secondly, the data splitting can be problematic. For example, if the data distribution is not stationary, it is quite likely that the training set will not reflect the true properties of the test set. These disadvantages make it cumbersome to integrate FSIC effectively into machine learning algorithms like ICA.

Comparing different dependence measures in a fair manner is not a trivial task. Given a finite sample, no test of independence is reliable (Gretton et al., 2005b). The universal limit on dependence tests affirms that the lower Type I error a dependence test has, the higher Type II error it can suffer from (Gretton et al., 2005b). In practical applications, one good benchmark to evaluate dependence measures is linear instantaneous ICA (Gretton et al., 2005a). In linear ICA, the aim is to extract original independent source signals when the only available information is a set of linearly mixed signal mixtures (Makino et al., 2007). Performing the signal separation in ICA algorithms depends on good measurement of dependence. The linear instantaneous mixing creates a signal mixture that can be demixed up to a indeterminacy of scale and permutation (Comon, 1994). Thus, evaluating the estimated demixing matrix from the ICA algorithm against the ground truth mixing matrix can indicate the optimization efficacy of the dependence measure employed. The ICA benchmark experiments put forward by Bach and Jordan (2002) have become a standard practice for evaluating many dependence measures (Hyvarinen, 1999; Cardoso and Souloumiac, 1993; Bell and Sejnowski, 1995; Learned-Miller and John III, 2003; Chen and Bickel, 2005; Bach and Jordan, 2002; Gretton et al., 2003; Shen et al., 2009; Pham, 2004).

We propose a new dependence measure called the Finite Basis Independence Criterion (FBIC). It is a linear-time independence measure derived from the fact that the mapping functions attaining the maximal correlation can be approximated in a subspace built by finite basis functions. It follows the spirit of the HGR maximum correlation coefficient. Our experimental results have shown that FBIC can perform similarly to HSIC and FSIC on various data distributions. FBIC is a direct application of Property (F) in Rényi (1959): the maximal correlation between two random variables is invariant with respect to all one-to-one Borel measurable transformations.

Recently, Móri and Székely (2019) suggested that the strong one-to-one invariance assumption of dependence measures should be replaced by similarity invariance and weak continuity. This relaxation of invariance is claimed to be instrumental for distribution robustness (Móri and Székely, 2019). They have shown that certain pathological random variables artificially constructed may confuse the maximal correlation measurement. However, detailed examination of pros and cons between different invariance assumptions of dependence measures has not been tested experimentally, and is outside of the scope of this paper.

Our contributions include:

- We develop a new framework for an upper bound of independence criterion on the basis of the HGR maximum correlation coefficient.
- We establish the effectiveness of four sets of basis functions, including three sets of RBFs and one set of polynomial functions, both theoretically and experimentally.
- We demonstrate the superior performance of FBIC on dependence detection in a comprehensive ICA benchmark, which shows the validity of applying FBIC in machine learning loss functions through gradient descent.

The remainder of the paper is organized as follows. Section II introduces an evolution from HSIC to FSIC. Section III describes the details of the proposed FBIC algorithm. Section IV explains the experiment design. Section V provides the conclusion and outlines future work.

2 From HSIC to FSIC

Due to their high performance, kernel-based independence measurements like HSIC and FSIC have been studied extensively in recent years (Bach and Jordan, 2002; Gretton et al., 2005a,b; Jitkrittum et al., 2017; Shen et al., 2009). They satisfy the framework introduced by Rényi (1959), stating that the maximal correlation (or cross-covariance) results from sufficiently

rich function classes is zero if and only if the random variables tested are independent. These methods are related to a test statistic comparing distributions called Maximum Mean Discrepancy (MMD) (Gretton et al., 2012). By using the unit ball of an RKHS, the MMD between distributions can be detected.

HSIC is a robust kernel-based independence measurement as it utilizes the sum of the squared singular values in the cross-covariance operator instead of only the largest singular value like the Constrained Covariance (COCO) approach (Gretton et al., 2005b). The summed up quantity is called the squared Hilbert-Schmidt norm, which is the RKHS counterpart of the squared Frobenius norm.

Let $\mathcal{F}$ and $\mathcal{G}$ be two RKHSs with corresponding positive definite kernels $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, respectively. For all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, the Reproducing Property gives us $k(x, x') = \langle \phi(x), \phi(x') \rangle$ and $l(y, y') = \langle \psi(y), \psi(y') \rangle$, where $\phi : \mathcal{X} \to \mathcal{F}$ and $\psi : \mathcal{Y} \to \mathcal{G}$ are feature maps. The mean embeddings of the probability distributions of x and y in $\mathcal{F}$ and $\mathcal{G}$ are defined as μ_x and μ_y:

$$\begin{aligned} \langle \mu_x, f \rangle_{\mathcal{F}} &:= \mathbf{E}_x[\langle \phi(x), f \rangle_{\mathcal{F}}] = \mathbf{E}_x[f(x)], \\ \langle \mu_y, g \rangle_{\mathcal{G}} &:= \mathbf{E}_y[\langle \psi(y), g \rangle_{\mathcal{G}}] = \mathbf{E}_y[g(y)], \end{aligned} \tag{1}$$

where $f \in \mathcal{F}$ and $g \in \mathcal{G}$. Given $a \in \mathcal{F}$ and $b \in \mathcal{G}$, the tensor product operator $a \otimes b : \mathcal{G} \to \mathcal{F}$ is defined as

$$(a \otimes b)g := a\langle b, g \rangle_{\mathcal{G}} \text{ for all } g \in \mathcal{G}. \tag{2}$$

Note that this reduces to $a \otimes b = ab^T$ for finite-dimensional vector spaces. It follows that the cross-covariance operator $C_{xy} : \mathcal{G} \to \mathcal{F}$ is denoted as

$$\begin{aligned} C_{xy} &:= \mathbf{E}_{x,y}[(\phi(x) - \mu_x) \otimes (\psi(y) - \mu_y)] \\ &= \mathbf{E}_{x,y}[\phi(x) \otimes \psi(y)] - \mu_x \otimes \mu_y. \end{aligned} \tag{3}$$

Denoting $\mathbf{E}_{x,y}[\phi(x) \otimes \psi(y)]$ by $\tilde{C}_{xy}$ and the joint distribution over $\mathcal{X} \times \mathcal{Y}$ as P_{xy}. Let $C : \mathcal{G} \to \mathcal{F}$ be a linear operator. The Hilbert-Schmidt norm of C is defined as:

$$\|C\|_{\text{HS}} := \sqrt{\sum_{i,j} \langle C v_j, u_i \rangle_{\mathcal{H}}^2}, \tag{4}$$

where u_i and v_j are orthonormal bases of $\mathcal{F}$ and $\mathcal{G}$ respectively.

A linear operator $C : \mathcal{G} \to \mathcal{F}$ is called a Hilbert-Schmidt operator if its Hilbert-Schmidt norm exists. The set of Hilbert-Schmidt operators is a separable Hilbert space with inner product

$$\langle C, D \rangle_{\text{HS}} = \sum_{i,j} \langle C v_j, u_i \rangle_{\mathcal{H}} \langle D v_j, u_i \rangle_{\mathcal{H}}. \tag{5}$$

We can now define HSIC as the squared Hilbert-Schmidt norm of the associated cross-covariance operator C_{xy}:

$$\begin{aligned} &\mathbf{HSIC}(P_{xy}, \mathcal{F}, \mathcal{G}) := \|C_{xy}\|_{\text{HS}}^2 \\ &= \left\|\tilde{C}_{xy} - \mu_x \otimes \mu_y\right\|_{\text{HS}}^2 \\ &= \langle \tilde{C}_{xy}, \tilde{C}_{xy} \rangle_{\text{HS}} + \langle \mu_x \otimes \mu_y, \mu_x \otimes \mu_y \rangle_{\text{HS}} - 2\langle \tilde{C}_{xy}, \mu_x \otimes \mu_y \rangle_{\text{HS}} \\ &= \mathbf{E}_{x,y}\mathbf{E}_{x',y'}[k(x, x')l(y, y')] + \mathbf{E}_x\mathbf{E}_y\mathbf{E}_{x'}\mathbf{E}_{y'}[k(x, x')l(y, y')] \\ &\quad - 2\mathbf{E}_{x,y}\mathbf{E}_{x'}\mathbf{E}_{y'}[k(x, x')l(y, y')]. \end{aligned} \tag{6}$$

Denote the marginal distributions over $\mathcal{X}$ and $\mathcal{Y}$ as P_x and P_y respectively. Let $\mathcal{T}$ be the unit ball in an RKHS associated with a kernel $v : \mathcal{F} \times \mathcal{G} \to \mathbb{R}$. HSIC associated with C_{xy} is equivalent to the squared MMD between the joint distribution P_{xy} and the product of its marginal distributions $P_x P_y$ (Bueno Larraz, 2015):

$$\mathbf{MMD}^2(\mathcal{T}, P_{xy}, P_x P_y) = \mathbf{HSIC}(P_{xy}, \mathcal{F}, \mathcal{G}). \tag{7}$$

FSIC is related to MMD in the sense that it also measures a difference between mean embeddings in RKHSs. However, the maximum distance aspect is not required in FSIC. Define the empirical measure $\nu := \frac{1}{J}\sum_{i=1}^{J} \delta_{(v_i, w_i)}$ over J test locations $V_J := \{(v_i, w_i)\}_{i=1}^{J} \subset \mathcal{X} \times \mathcal{Y}$, where δ_t denotes the Dirac measure centered on t. Let μ_{xy}, μ_x and μ_y represent mean embeddings of P_{xy}, P_x and P_y. Define random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$. Using the $L^2(\mathcal{X} \times \mathcal{Y}, \nu)$ to measure the distance between μ_{xy} and $\mu_x \mu_y$, FSIC is defined by:

$$\mathbf{FSIC}^2[X, Y] := \|\mu_{xy} - \mu_x \mu_y\|_{L^2(\mathcal{X} \times \mathcal{Y}, \nu)}^2 \tag{8}$$

$$= \int_{\mathcal{X}} \int_{\mathcal{Y}} (\mu_{xy}(x, y) - \mu_x(x)\mu_y(y))^2 d\nu(x, y) \tag{9}$$

$$= \frac{1}{J}\sum_{i=1}^{J} u(v_i, w_i)^2, \text{where} \tag{10}$$

$$u(v, w) := \mu_{xy}(v, w) - \mu_x(v)\mu_y(w) \tag{11}$$

$$= \mathbf{E}_{x,y}[k(x, v)l(y, w)] - \mathbf{E}_x[k(x, v)]\mathbf{E}_y[l(y, w)] \tag{12}$$

$$= \mathbf{cov}_{xy}[k(x, v), l(y, w)]. \tag{13}$$

FSIC is a good measure for independence testing purpose, but it is not easy to employ it in machine learning algorithms like ICA due to the training data requirement for the test location optimization. To design a linear-time independence criterion that can be easily integrated into a machine learning algorithm, we propose FBIC that requires no training data.

3 FBIC

Define nondegenerate univariate random variables X and Y whose observations $x = (x_1, ..., x_m)^T$ and $y = (y_1, ..., y_m)^T$ can represent signals of interest. Without loss of generality, we can scale observed signals x and y such that they are within the closed domain interval $[0, 1]$. Let $L^2[0, 1]$ be the square-integrable function space such that for $f(t) \in L^2[0, 1]$

$$\|f(t)\|_2 = \left(\int_0^1 |f(t)|^2 dt\right)^{1/2} < \infty, \tag{14}$$

where t is a real value defined on the closed domain interval $[0, 1]$. Let $C[0, 1]$ be the function space of real-valued continuous functions defined on the closed interval $[0, 1]$. We now define a function space $T = L^2[0, 1] \cap C[0, 1]$, which is an equivalence class of functions that are both square-integrable and continuous almost everywhere on the closed domain interval $[0, 1]$. A finite dimensional subspace $V \subseteq T$ can be built upon a set of basis functions $\mathcal{B} = \{p_i : [0, 1] \to \mathbb{R}\}$, where $i \in \{1, 2, ..., N\}$ and N is a finite positive integer.

It has been shown in (Bach and Jordan, 2002) that the maximal correlation can be used to measure independence between tested variables. The maximal correlation between x and y is:

$$\rho_{\max}[x, y] = \sup_{f_1, f_2 \in T} \Big(\rho[f_1(x), f_2(y)]\Big) = \sup_{f_1, f_2 \in T} \frac{\mathbf{cov}(f_1(x), f_2(y))}{\sigma(f_1(x))\sigma(f_2(y))}, \tag{15}$$

where $\rho[\cdot, \cdot]$ denotes the Pearson Correlation, and $\sigma(\cdot)$ denotes the standard deviation.

We use the function space T for possible mapping functions employed in the maximal correlation. This is because the Pearson Correlation requires the chosen functions to be square-integrable. The continuous property implies that the chosen functions are Borel-measurable. Define f_1^* and f_2^* such that $\rho_{\max}[x, y] = \rho[f_1^*(x), f_2^*(y)]$. We want to use the best approximations q_1 and q_2 of f_1^* and f_2^* in V that minimize the difference between $\rho[f_1^*(x), f_2^*(y)]$ and $\rho[q_1(x), q_2(y)]$. From the definition of V, we can write $q_1(x) = \sum_{i=1}^N a_i p_i(x)$ and $q_2(y) = \sum_{j=1}^N b_j p_j(y)$. As the standard deviation can be scaled by a multiplication factor and it will be divided for normalization in Pearson Correlation in the end, we can assume that all mapping functions in T have unit standard deviation when applied to test variables.

3.1 Definition of FBIC

We propose that the Finite Basis Independence Criterion (FBIC) can imply independence by measuring the pair-wise correlations between the basis function mappings from tested variables. This can be expressed with a set of basis functions $\{p_i\}$ from $\mathcal{B}$:

$$\begin{aligned}\mathbf{FBIC}(\mathcal{B}, X, Y) &= \sum_{i=1}^N \sum_{j=1}^N \Big|\rho_{ij}[x, y]\Big| \\ &= \sum_{i=1}^N \sum_{j=1}^N \Big|\mathbf{E}[(p_i(x) - \mathbf{E}[p_i(x)])(p_j(y) - \mathbf{E}[p_j(y)])]\Big|.\end{aligned} \tag{16}$$

Let Q_x and Q_y be the probability density functions of x and y respectively. Let $X := \{x_1, ..., x_m\}$ and $Y := \{y_1, ..., y_m\}$ be observations independently and identically drawn from Q_x and Q_y. The empirical FBIC is:

$$\begin{aligned}\widehat{\mathbf{FBIC}}(\mathcal{B}, X, Y) = \sum_{i=1}^N \sum_{j=1}^N \Big| \frac{1}{m} \sum_{k=1}^m [(p_i(x_k) \\ - \frac{1}{m} \sum_{s=1}^m p_i(x_s))(p_j(y_k) - \frac{1}{m} \sum_{t=1}^m p_j(y_t))]\Big|.\end{aligned} \tag{17}$$

3.2 Basis Functions in FBIC

The choice of basis function sets is crucial for the good performance of the FBIC. Our first category of basis functions are inspired by the RBF neural networks, which have the universal approximation capacity (Wu et al., 2012). A RBF $K(\cdot)$ is a radially symmetric function such that $\|x\|_p = \|y\|_p$ implies $K(x) = K(y)$ (Park and Sandberg, 1991). Here we use a general norm notation $\|\cdot\|_p$ to include more than the Euclidean norm for the construction of RBFs.

A finite set of RBFs can approximate any function in T arbitrarily well. This can be corroborated by the Theorem 1 of (Park and Sandberg, 1991), which states that for functions that are integrable, bounded, continuous almost everywhere and has non-zero integral, a finite sum of these functions with coefficients is dense in p-integrable function spaces for every $p \in [1, \infty)$. RBFs we choose should satisfy these properties and are, hence dense in T, a subset of the square-integrable function space.

Three sets of RBFs are tested in this paper. They are Gaussian RBFs, Laplacian RBFs and Inverse Multiquadratic RBFs. They are listed below:

$$\begin{cases} \text{Gaussian:} & K(x) = e^{-\epsilon(x-c)^2} \\ \text{Laplacian:} & K(x) = e^{-\epsilon|x-c|} \\ \text{Inverse Multiquadratic:} & K(x) = \frac{1}{\sqrt{1+\epsilon(x-c)^2}}. \end{cases} \tag{18}$$

We define $\epsilon \in (0, \infty)$ as the shape parameter and $c \in [0, 1]$ as the displacement parameter for the RBFs in Equation 18. The smaller the ϵ, the flatter the shape of RBFs is (Mongillo, 2011). RBFs of the same category share the same shape parameter. The displacement parameters are equally distanced grid points. For example, when there are 10 RBFs in a set, $c = \{0, 0.1, 0.2, \cdots, 0.9\}$.

Another category of basis functions is the Shifted Legendre Polynomials. They are orthogonal polynomials that can play the role of orthogonal basis functions. This is a natural choice for a finite set of basis functions as orthogonality is numerically effective for function approximation (Powell, 1981).

Let $P_n(x)$ be a polynomial of degree n. Legendre Polynomials satisfy the condition that

$$\int_{-1}^{1} P_m(x) P_n(x) dx = 0, \text{if } n \neq m. \tag{19}$$

$P_0(x)$ is generally set to 1 for standardization. Shifted Legendre Polynomials are defined as $\widetilde{P}_n(x) = P_n(2x-1)$. They satisfy the condition that

$$\int_0^1 \widetilde{P}_m(x)\widetilde{P}_n(x)dx = 0, \text{if } n \neq m. \tag{20}$$

Shifted Legendre Polynomials are preferred as they are orthogonal on the interval $[0,1]$, which corresponds to our function space T. For one set of Shifted Legendre Polynomial basis functions, they are distinguished by different polynomial degrees n.

The Weierstrass's theorem states that a continuous function on a closed interval can be approximated arbitrarily well by polynomials (Jackson, 1934). Therefore, Shifted Legendre Polynomials can be regarded as the orthogonal basis for the construction of any continuous function on a closed interval (Kolmogorov and Fomin, 1961). Since any polynomial can be decomposed into Shifted Legendre Polynomials and there exists some good polynomial approximation for any function in T, this suggests that Shifted Legendre Polynomials can approximate any function in T arbitrarily well given a sufficient number of degrees.

3.3 Justifications for FBIC

Theorem 1. *Given a sufficient set of basis functions, FBIC gives an upper bound to the maximal correlation between tested variables with an arbitrarily small positive error bound* Δ.

Proof. Define $\gamma_{ij} = \hat{a_i}\hat{b_j}$. Let $\hat{a_i}$ and $\hat{b_j}$ be scaled coefficients a_i and b_j from the definitions of q_1 and q_2. $\hat{a_i}$ and $\hat{b_j}$ are within $[-1,1]$. The coefficients can be scaled as the Pearson Correlation is scale invariant. From the definition of FBIC in Equation 16, we can derive an inequality as follows:

$$\begin{aligned}
&\mathbf{FBIC}[\mathcal{B},X,Y] = \sum_{i=1}^{N}\sum_{j=1}^{N}\left|\mathbf{E}[(p_i(x)-\mathbf{E}[p_i(x)])(p_j(y)-\mathbf{E}[p_j(y)])]\right| \\
&= \sum_{i=1}^{N}\sum_{j=1}^{N}\left|\mathbf{E}[p_i(x)p_j(y)] - \mathbf{E}[p_i(x)]\mathbf{E}[p_j(y)]\right| \\
&\geq \sum_{i=1}^{N}\sum_{j=1}^{N}|\gamma_{ij}|\left|\mathbf{E}[p_i(x)p_j(y)] - \mathbf{E}[p_i(x)]\mathbf{E}[p_j(y)]\right| \qquad (21)\\
&\geq \left|\sum_{i=1}^{N}\sum_{j=1}^{N}\gamma_{ij}(\mathbf{E}[p_i(x)p_j(y)] - \mathbf{E}[p_i(x)]\mathbf{E}[p_j(y)])\right| \qquad (22)\\
&= \left|\rho[q_1(x),q_2(y)]\right|. \qquad (23)
\end{aligned}$$

Equation 21 is obtained by multiplying each term with $|\gamma_{ij}|$. Equation 23 follows from definitions of q_1, q_2 and the Pearson Correlation. By the universal approximation property of RBFs and polynomials, given a sufficient set of basis functions, we have $||\rho[q_1(x),q_2(y)]| - \rho[f_1^*(x),f_2^*(y)]| = \Delta$, where Δ is arbitrarily close to zero. This shows $\mathbf{FBIC}[\mathcal{B},X,Y] + \Delta \geq \rho_{\max}[x,y]$. □

Theorem 2. *Given a sufficient set of basis functions, FBIC between tested variables is arbitrarily close to zero if and only if* $X \perp\!\!\!\perp Y$.

Proof. The sufficient condition of this theorem is trivial as independent variables always have 0 correlation no matter what basis functions are used for mapping. The positive error bound Δ reaches zero for independent variables. The necessary condition can be proved by using the upper bound property in Theorem 1. Since the maximal correlation is non-negative, it is arbitrarily close to zero if its upper bound is arbitrarily close to 0. □

Additionally, we can use properties of the Laplace transform to prove the validity of FBIC with Gaussian RBFs. Define a Borel probability measure $\mu : X \times Y \to \mathbb{R}^+$, where $\iint \mu(x,y)dxdy = 1$. We use $K(x) = e^{-\epsilon(x-c)^2}$ for the Gaussian RBFs. Define $K_s(x) = e^{-\epsilon(x-s)^2} = e^{-\epsilon(x^2-2xs+s^2)}$ and $K_t(y) = e^{-\epsilon(y-t)^2} = e^{-\epsilon(y^2-2yt+t^2)}$, where s and t are displacement parameters and $s,t \in [0,1]$. Now we can establish a theorem that FBIC with Gaussian RBFs will equate the Borel probability measure of the joint probability distribution of x and y to the product of the marginal probability distributions of x and y if and only if x and y and independent.

Theorem 3. $\forall s,t \iint K_s(x)K_t(y)\mu(x,y)dxdy = \iint K_s(x)\mu(x,y)dxdy \iint K_t(y)\mu(x,y)dxdy$, *if and only if* $\mu(x,y) = \int \mu(x,s)ds \int \mu(y,t)dt$.

Proof. Let $\eta(x,y) = \mu(x,y) - \int \mu(x,s)ds \int \mu(y,t)dt$. We need to show $\forall s,t \iint K_s(x)K_t(y)\eta(x,y)dxdy = 0 \iff x \perp\!\!\!\perp y$. The necessary condition is trivial to prove as independent variables have the property that $\mu(x,y) = \int \mu(x,s)ds \int \mu(y,t)dt$. The sufficient condition is proved as follows. Assume $\forall s,t \iint K_s(x)K_t(y)\eta(x,y)dxdy = 0$. We have $\forall s,t \iint e^{\epsilon(2xs+2yt))}e^{-\epsilon(x^2+y^2)}\eta(x,y)dxdy = 0$. Let $\eta_1(x,y) = e^{-\epsilon(x^2+y^2)}\eta(x,y)$. This gives us $\forall s,t \iint e^{\epsilon(2xs+2yt}\eta_1(x,y)dxdy = 0$. We can observe that this is exactly the Laplace transform of $\eta_1(x,y)$ so that it satisfies the condition that $\forall s,t\ [\mathcal{L}(\eta_1)](-2\epsilon s,-2\epsilon t) = 0 \iff \eta_1(x,y) = 0$. □

Complexity: The computational complexity of FBIC is linear with respect to the number of data points. The computation of the Pearson Correlation is $O(n)$, where n is the sample size. The dimensionality of tested variables is 2. Let the number of basis functions be k. We can express the computational complexity of FBIC as:

$$O(\mathbf{FBIC}) = O(k^2 n) \approx O(n). \tag{24}$$

3.4 FBIC vs FSIC

FBIC and FSIC differ in some significant aspects:

- FBIC is inspired by the HGR maximum correlation coefficient, whereas FSIC takes advantage of constrained covariance on the basis of MMD.
- FBIC can use many different RBFs and even orthogonal polynomials, whereas FSIC depends on characteristic kernels (Sriperumbudur et al., 2010) like Gaussian kernels.
- FBIC has evenly spaced RBFs, whereas FSIC employs randomly spaced kernels.

Yet FBIC and FSIC are related when we examine their mathematical foundations. The HGR maximum correlation coefficient and the MMD criterion have possibly different function classes, optimize for different directions (one for maximal similarity, the other for maximal difference) and measure different statistical quantities. Nonetheless, the functions utilized in the MMD criterion are Borel-measurable with constrained variance in the Reproducing Kernel Hilbert Spaces. This implies that the class of functions for the MMD criterion is a subset of Borel-measurable functions that can be used for the HGR maximum correlation coefficient. Different optimization directions are more like a difference of optimization flavours leading to the same result: yielding zero if and only if the tested variables are independent. Both criteria measure covariance, though one is normalized (i.e., correlation) and the other is not normalized in the RKHSs.

Neither the HGR maximum correlation coefficient nor the MMD criterion is superior to the other in theory. Different independence criteria are preferable under different circumstances. In circumstances where training before using is not feasible, FBIC is preferable to FSIC. We will show how FBIC compares with FSIC in Section 4.

4 Experimental Studies

We use the classic ICA benchmark experiments put forward by (Bach and Jordan, 2002). The perfect demixing matrix V is known in advance and the optimized demixing matrix W can be evaluated by the Amari distance (Amari et al., 1996):

$$d(V,W) = \frac{1}{2m(m-1)} \sum_{i=1}^{m} \left(\frac{\sum_{j=1}^{m} |a_{ij}|}{\max_j |a_{ij}|} - 1 \right) + \frac{1}{2m(m-1)} \sum_{j=1}^{m} \left(\frac{\sum_{i=1}^{m} |a_{ij}|}{\max_i |a_{ij}|} - 1 \right), \quad (25)$$

where $a_{ij} = (V^{-1}W)_{ij}$. This distance is ranged between 0 and 1. It is invariant to permutation and scaling. When $V = W$, the Amari distance between them is zero. In our experiments, the smaller the distance, the better the result.

We submitted our experiments to a grid computing computer cluster controlled by the Sun Grid Engine to run them in parallel. The CPU model was Intel(R) Core(TM) i7-8700 with 3.20GHz. There were 6 cores on each computer.

4.1 Simulated Data

To test the applicability of FBIC in different scenarios, source signals generated from a comprehensive set of probability density functions were used for ICA experiments. Both supergaussian and subgaussian distributions were covered. Mode and symmetry variation were also considered. All probability density functions had zero mean and unit variance. The details of each probability density function are demonstrated in Table 1 of Appendix.

In our experiment, two probability density functions were randomly chosen from the set. Signal samples $\mathbf{S} \in \mathbb{R}^{J \times M}$ were generated according to their probability density functions independently, where J is the number of sources and M is the sample length. They were then mixed by a random mixing matrix $\mathbf{A} \in \mathbb{R}^{J \times J}$ with a bounded condition number between 1 and 2, which gave the observed signals $\mathbf{X} = \mathbf{AS}$. The goal of the ICA experiment is to estimate the demixing matrix W while only knowing that the original unmixed signals are statistically independent and they are linearly mixed.

As the source signals were generated randomly from random distributions, it was a difficult task to perform ICA algorithms on these signals. Unlike experiments on speech or image processing, we cannot take advantage of extra pattern information to separate the mixture of random signals. The successful performance of ICA algorithms in this dataset relied heavily on the performance measure used, which made this benchmark dataset desirable for evaluating different dependence measures.

Since the aim of our ICA experiments is to compare different independence criteria, it is unnecessary to test the robustness of different ICA algorithms when outliers are present in the observed mixture. In scenarios where noisy outliers occur, it is more appropriate to apply outlier removal process before applying independence criterion. We evaluate different independence criteria through their performance on dependence detection only.

4.2 Preprocessing

We firstly prewhitened the observed signals to reduce the search space of the optimization process. Prewhitening is a popular BSS technique that eliminates the linear dependency between tested signals. It transformed the observed signals $\mathbf{X}$ to have unit covariance matrix. By multiplying a whitening matrix $\mathbf{Z}$, we got the transformed signals $\tilde{\mathbf{X}} = \mathbf{ZX}$ such that $\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T = \mathbf{I}$. The whitening matrix $\mathbf{Z}$ can be obtained by first performing a singular value decomposition to the mixing matrix and get $\mathbf{A} = \mathbf{UDQ}^T$ (Golub and Reinsch, 1971). $\mathbf{U}$ and $\mathbf{Q}$ are orthonormal matrices and $\mathbf{D}$ is a diagonal matrix containing singular values. The whitening matrix can be thus defined as $\mathbf{Z} = \mathbf{D}^{-1}\mathbf{U}^T$.

To make a fair comparison with the HSIC-based ICA experiments, we used the Jade algorithm (Cardoso and Souloumiac, 1993) to initialize the first guess of the demixing matrix. This was also the initialization set up in the HSIC-based ICA algorithm (Shen et al., 2009).

4.3 Neural Network settings

A simple neural network was utilized to optimize the demixing matrix. It consisted of two layers only: one input layer and one output layer. In practice, each neuron in the input layer was activated by a standard sigmoid function $\mathcal{S}(x) = \frac{1}{1+\exp(-x)}$, where x is the input signal for the neuron. The output layer had the identity activation function.

The number of neurons in the input layer and output layer was equal to the number of original sources. The input signals were observed mixed signals. The output signals were supposed to be demixed signals without considering the scale or order. The weight matrix connecting the two layers thus played the role of the demixing matrix.

The cost function was $\mathcal{L} = \mathcal{F} + \mathcal{O}$, which was the sum of FBIC measurement between output signals and the orthogonality regularizer of the demixing matrix. The orthogonality regularizer was calculated by $\mathcal{O} = \text{sum}(|\mathbf{A}^T\mathbf{A}|) - \text{trace}(\mathbf{A}^T\mathbf{A})$, where $\text{sum}(\cdot)$ returned the total of all the entries in the matrix given. The role of $\mathcal{O}$ was to ensure the demixing matrix would not introduce any linear dependency after prewhitening. An additional Pearson Correlation between tested variables x and y after sigmoid activation, $\rho[\mathcal{S}(x), \mathcal{S}(y)]$, was added to the cost function to aid the optimization.

The Adam optimizer was used to perform the gradient descent. The learning rate was set to 0.0001. The maximum epoch was 100000. If the loss did not decrease after 1000 epochs, the training of the neural network would have an early stopping.

4.4 Parameter Choices for RBFs

The parameters of FBIC RBFs were selected by grid search. They were fixed regardless of the number of channels or the length of samples. For Gaussian RBFs, we used shape parameter 200 and the step parameter was increased by 0.1 stepwise from 0 to 0.9. For Laplace RBFs, we used shape parameter 20 and the step parameter was increased by 0.05 stepwise from 0 to 0.95. For Inverse Multiquadratic RBFs, we used shape parameter 900 and the step parameter was increased by 0.1 stepwise from 0 to 0.9.

The Shifted Legendre Polynomials did not require shape or step parameters. We used polynomial degrees from 2 to 20. Degree 0 and 1 were not chosen because they were either constant or identity mapping, which would not contribute much to reveal the nonlinear correlation between tested variables.

In FBIC, The number of basis functions does not always scale positively with the quality of approximation. Jitkrittum et al. (2017) observed a similar phenomenon in FSIC as well. When there are not enough basis functions, increasing the number of basis functions will naturally improve the quality of approximation. However, too many redundant basis functions do not contribute to a better quality of approximation. It makes the upper bound of the maximal correlation coefficient from FBIC unnecessarily high. The experimental details are provided in Appendix.

4.5 Multi-Channel Extension of FBIC

By the definition of the maximal correlation, it is natural to apply FBIC to two random variables. For dependence detection amongst multiple random variables, a simple extension of FBIC is to calculate all the pairwise FBIC scores between variables and sum them up. Although limiting FBIC to pairwise correlation is suboptimal as multivariate dependence may not be detected, we still found favorable results for multichannel ICA experiments as will be shown in Table 1.

4.6 Results and Discussions

As HSIC exhibited superior performance over other dependence measures in the classic ICA benchmark dataset (Gretton et al., 2005a; Shen et al., 2009), we compared the performance of FBIC mainly with FastKICA (Shen et al., 2009), a fast HSIC-based Kernel ICA based on the incomplete Cholesky decomposition. Its kernel width was set to 0.5 as suggested by the original authors. Its maximum iteration number was set

Table 1: The Amari distance of demixed signals from n sources based on different independence measures. The sample length is m. Rep. represents the number of experiment replications. The Gaussian RBFs in FBIC are denoted g. The Laplace RBFs in FBIC are denoted l. The Inverse Multiquadratic RBFs are denoted imq. The Shifted Legendre Polynomials are denoted lp.

n	m	Rep.	FBIC$_g$	FBIC$_l$	FBIC$_{imq}$	FBIC$_{lp}$	FKICA	NFSIC	FICA	Jade
2	250	1000	5.8	**5.7**	6.4	6.6	6.3	6.5	12.3	9.1
2	1000	1000	2.4	**2.1**	2.7	2.5	2.8	2.5	6.2	4.3
4	1000	100	3.4	**3.1**	3.7	3.8	3.3	3.5	6.7	5.1
4	4000	100	1.7	**1.5**	1.8	1.8	**1.5**	4.6	3.2	2.7

to 20 and the convergence threshold was set to 1e-6. Results from FastICA (Hyvarinen, 1999) and Jade (Cardoso and Souloumiac, 1993) algorithms were also included. In Table 1, we use FKICA to denote FastKICA and FICA to denote FastICA. To compare FBIC and FSIC directly, we implemented a novel ICA algorithm based on Normalized FSIC, which is denoted NFSIC in the table. The implementation details are provided in Appendix.

From Table 1 we can observe that FBIC with Laplacian RBFs performed the best in all the experiments. This is consistent with the finding that the Laplacian kernel in HSIC performed better than the Gaussian kernel (Gretton et al., 2005a). FBIC with other basis functions was also better than FKICA in two-channel experiments. For four-channel experiments, the performance of FBIC with Laplace RBFs was no worse than FKICA, while the performance of FBIC with other basis functions were acceptable. FBIC with Inverse Multiquadratic RBFs and Shifted Legendre Polynomials performed less well than Laplace and Gaussian RBFs. In all experiments, the performance of both FBIC-based and HSIC-based algorithms surpass those of FICA and Jade by a large margin. The FSIC-based algorithm performed well in all but the four-channel experiments with long sequences. This can be explained by the absence of optimization of test locations. The deviation between random and optimal test locations becomes larger when the signals are longer.

A smaller step parameter increment was beneficial for FBIC with Laplacian RBFs, but FBIC with other RBFs did not benefit. One explanation is that the shape parameter we selected for other RBFs may not be well suited for more fine-grained approximation. Another possibility is that the sharp peak of Laplace RBFs makes them more suited for smaller step size. More experiments may shed light on the relationship between the targeted distributions and the best FBIC parameter settings.

We also discovered that all the shape parameters selected for the three sets of RBFs were relatively large. This indicates that sharper RBFs are more suitable for approximating best mapping functions $f_1^*(\cdot)$ and $f_2^*(\cdot)$ of the HGR maximum correlation coefficient. The justification is that $f_1^*(\cdot)$ and $f_2^*(\cdot)$ tend to be functions that magnify areas where two probability density functions of the tested variables differ the most. The magnification can be accomplished by assigning a sharp peak to the mapping function around the area to be magnified. The sharper RBFs therefore are better candidates for $f_1^*(\cdot)$ and $f_2^*(\cdot)$ approximation.

Parameter selection has always been a challenge for using RBFs or kernels. In HSIC, scale parameters for kernels have to be decided. In FSIC, scale parameters as well as test locations require careful tuning. Regarding FBIC, it is also vital to select the right shape parameters for RBFs. Since the experiments we conducted included 18 different distributions that could represent a wide variety of real-life distributions, we recommend using the shape parameters we found as a starting point for more customized parameter tuning. On the other hand, FBIC with Shifted Legendre Polynomials could perform well without tuning parameters. Therefore, it can be used as a FBIC baseline. The FBIC algorithms using RBFs should perform at least as well as the FBIC baseline.

5 Conclusions

We have proposed the Finite Basis Independence Criterion, a linear-time independence measurement that can be optimized through gradient descent. By approximating the best mapping function with a finite set of basis functions, FBIC establishes an upper bound to the Hirschfeld-Gebelein-Rényi maximum correlation coefficient. The ICA algorithm based on FBIC outperforms the analogous ICA algorithm based on the state-of-the-art kernel independence measurements HSIC and FSIC in two-channel experiments. For four-channel experiments, FBIC-based ICA can still perform competitively against FKICA when the Laplace RBFs are utilized, though only pairwise correlations are calculated.

Acknowledgments

We thank Hung Pham and Richard Arnold for help with proofs. We thank Yuan Yao for helpful discussions.

References

Shun-ichi Amari, Andrzej Cichocki, and Howard Hua Yang. A new learning algorithm for blind signal separation. In *Advances in neural information processing systems*, pages 757–763, 1996.

Francis R Bach and Michael I Jordan. Kernel independent component analysis. *Journal of machine learning research*, 3(Jul):1–48, 2002.

Anthony J Bell and Terrence J Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, 1995.

Beatriz Bueno Larraz. Independence measures. Master's thesis, Universidad Autónoma de Madrid, Escuela Politécnica Superior, Spain, 2015.

Jean-François Cardoso and Antoine Souloumiac. Blind beamforming for non-Gaussian signals. In *IEE proceedings F (radar and signal processing)*, volume 140, pages 362–370. IET, 1993.

Aiyou Chen and Peter J Bickel. Consistent independent component analysis and prewhitening. *IEEE Transactions on Signal Processing*, 53(10):3625–3632, 2005.

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

Hans Gebelein. Das statistische problem der korrelation als variations-und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, 21(6):364–379, 1941.

Gene H Golub and Christian Reinsch. Singular value decomposition and least squares solutions. In *Linear Algebra*, pages 134–151. Springer, 1971.

Arthur Gretton, Ralf Herbrich, and Alexander J Smola. The kernel mutual information. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 4, pages IV–880. IEEE, 2003.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005a.

Arthur Gretton, Alexander J Smola, Olivier Bousquet, Ralf Herbrich, Andrei Belitski, Mark Augath, Yusuke Murayama, Jon Pauls, Bernhard Schölkopf, and Nikos K Logothetis. Kernel constrained covariance for dependence measurement. In *AISTATS*, volume 10, pages 112–119, 2005b.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

Hermann O Hirschfeld. A connection between correlation and contingency. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 31, pages 520–524. Cambridge University Press, 1935.

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

Dunham Jackson. A proof of weierstrass's theorem. *The American Mathematical Monthly*, 41(5):309–312, 1934.

Wittawat Jitkrittum, Zoltén Szabó, and Arthur Gretton. An adaptive test of independence with analytic kernel embeddings. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1742–1751. JMLR. org, 2017.

Andrej Nikolaevich Kolmogorov and Sergeĭ Vasil'evich Fomin. *Elements of the theory of functions and functional analysis*. Academic Press, 1961.

Erik G Learned-Miller and W Fisher John III. Ica using spacings estimates of entropy. *Journal of machine learning research*, 4(Dec):1271–1295, 2003.

Shoji Makino, Te-Won Lee, and Hiroshi Sawada. *Blind Speech Separation (Signals and Communication Technology)*. Springer Netherlands, 2007.

Michael Mongillo. Choosing basis functions and shape parameters for radial basis function methods. *SIAM undergraduate research online*, 4(190-209):2–6, 2011.

Tamás F Móri and Gábor J Székely. Four simple axioms of dependence measures. *Metrika*, 82(1):1–16, 2019.

Jooyoung Park and Irwin W Sandberg. Universal approximation using radial-basis-function networks. *Neural computation*, 3(2):246–257, 1991.

Dinh-Tuan Pham. Fast algorithms for mutual information based independent component analysis. *IEEE Transactions on Signal Processing*, 52(10): 2690–2700, 2004.

Michael James David Powell. *Approximation theory and methods*. Cambridge university press, 1981.

Alfréd Rényi. On measures of dependence. *Acta mathematica hungarica*, 10(3-4):441–451, 1959.

Hao Shen, Stefanie Jegelka, and Arthur Gretton. Fast kernel-based independent component analysis. *IEEE Transactions on Signal Processing*, 57(9): 3498–3511, 2009.

Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert RG Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(Apr):1517–1561, 2010.

Yue Wu, Hui Wang, Biaobiao Zhang, and K-L Du. Using radial basis function networks for function approximation and classification. *ISRN Applied Mathematics*, 2012, 2012.