Appendices

A RKHS-based Independence Measures

A.1 RKHS

A Hilbert space $\mathcal{H}$ is a complete inner product space. An RKHS is a Hilbert space where point evaluation is a continuous linear functional. Thus, by the Riesz representation theorem [1] which states that any functional mapping $\mathcal{H}$ into $\mathbb{R}$ can be represented by an inner product, an RKHS has the Reproducing Property that $f(x) = \langle f, \phi(x)\rangle_{\mathcal{H}}$, where $f : \mathcal{X} \to \mathbb{R}$ is a functional in the RKHS, x belongs to $\mathcal{X}$ and $\phi(x)$ is a feature map from $\mathcal{X}$ to $\mathcal{H}$. The feature map takes the canonical form $\phi(x) = k(x, \cdot)$, where $k(x_1, x_2) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive definite kernel. In other words, the inner product of two feature map can be evaluated by a kernel: $\langle \phi(x), \phi(y)\rangle_{\mathcal{H}} = k(x, y)$.

A.2 MMD

Let x and y be random variables defined on a topological space $\mathcal{X}$, with respective Borel probability measures p and q. Let $\mathcal{T}$ be a class of functions $f : \mathcal{X} \to \mathbb{R}$. The Maximum Mean Discrepancy (MMD) is defined as [2]:

$$\mathbf{MMD}(\mathcal{T}, p, q) := \sup_{f \in \mathcal{T}} (\mathbf{E}_x[f(x)] - \mathbf{E}_y[f(y)]), \tag{21}$$

where $\mathbf{E}_x[f(x)]$ and $\mathbf{E}_y[f(y)]$ denote expectations with respect to p and q.

The MMD can be considered as an integral probability metric [3]. Let $(\mathcal{X}, d)$ be a metric space. We have $p = q$ if and only if $\mathbf{E}_x[f(x)] = \mathbf{E}_y[f(y)]$ for all $f \in C(\mathcal{X})$, where $C(\mathcal{X})$ is the space of bounded continuous functions on $\mathcal{X}$.

In real life, it is not practical to work with a rich function class like $C(\mathcal{X})$. The MMD requires a function class that is rich enough to identify $p = q$ uniquely, yet restrictive enough so that the MMD can be estimated by finite samples. It turns out that the unit ball in an RKHS $\mathcal{H}$ can satisfy both conditions.

Define the mean embedding of p in an RKHS as $\mu_p \in \mathcal{H}$ such that $\mathbf{E}_x f = \langle f, \mu_p\rangle_{\mathcal{H}}$. The formula for the MMD in Equation 21 can be rewritten as:

$$\begin{aligned}
&\mathbf{MMD}^2(\mathcal{T}, p, q) \\
&= \left[\sup_{\|f\|_{\mathcal{H}} \le 1} (\mathbf{E}_x[f(x)] - \mathbf{E}_y[f(y)])\right]^2 && (22)\\
&= \left[\sup_{\|f\|_{\mathcal{H}} \le 1} \langle \mu_p - \mu_q, f\rangle_{\mathcal{H}}\right]^2 && (23)\\
&= \left\|\mu_p - \mu_q\right\|_{\mathcal{H}}^2 && (24)\\
&= \langle \mu_p, \mu_p\rangle_{\mathcal{H}} + \langle \mu_q, \mu_q\rangle_{\mathcal{H}} - 2\langle \mu_p, \mu_q\rangle_{\mathcal{H}} && (25)\\
&= \mathbf{E}_{x,x'}\langle \phi(x), \phi(x')\rangle_{\mathcal{H}} + \mathbf{E}_{y,y'}\langle \phi(y), \phi(y')\rangle_{\mathcal{H}} \\
&- 2\mathbf{E}_{x,y}\langle \phi(x), \phi(y)\rangle_{\mathcal{H}}. && (26)
\end{aligned}$$

Let $X := \{x_1, \cdots, x_m\}$ and $Y := \{y_1, \cdots, y_n\}$ be observations independently and identically drawn from p and q. By using the empirical estimates of the feature space based on X and Y, we obtain

$$\mathbf{MMD}^2(\mathcal{T}, X, Y) = \left[\frac{1}{m^2}\sum_{i,j=1}^{m} k(x_i, x_j) - \frac{2}{mn}\sum_{i,j=1}^{m,n} k(x_i, y_j) + \frac{1}{n^2}\sum_{i,j=1}^{n} k(y_i, y_j)\right]. \tag{27}$$

A.3 More Details on HSIC

A.3.1 Proof for the Cross-Covariance Operator

$$C_{xy} := \mathbf{E}_{x,y}[\phi(x) \otimes \psi(y)] - \mu_x \otimes \mu_y. \tag{28}$$

Proof.

$$\begin{aligned}
\langle f, C_{xy} g\rangle_{\mathcal{F}} &= \langle C_{xy}, f \otimes g\rangle_{\text{HS}} \\
&= \mathbf{E}_{x,y}\langle(\phi(x) - \mu_x) \otimes (\psi(y) - \mu_y), f \otimes g\rangle_{\text{HS}} \\
&= \mathbf{E}_{x,y}[\langle f, \phi(x) - \mu_x\rangle\langle g, \psi(y) - \mu_y\rangle] \\
&= \mathbf{E}_{x,y}[(\langle f, \phi(x)\rangle - \langle f, \mu_x\rangle)(\langle g, \psi(y)\rangle - \langle g, \mu_y\rangle)] \\
&= \mathbf{E}_{x,y}[(f(x) - \mathbf{E}_x[f(x)])(g(y) - \mathbf{E}_y[g(y)])] \\
&= \mathbf{E}_{x,y}[f(x)g(y)] - \mathbf{E}_x[f(x)]\mathbf{E}_y[g(y)] \\
&= \mathbf{E}_{x,y}\langle\phi(x) \otimes \psi(y), f \otimes g\rangle - \langle\mu_x \otimes \mu_y, f \otimes g\rangle \\
&= \mathbf{E}_{x,y}\langle\phi(x) \otimes \psi(y), f \otimes g\rangle - \mathbf{E}_{x,y}\langle\mu_x \otimes \mu_y, f \otimes g\rangle \\
&= \mathbf{E}_{x,y}\langle\phi(x) \otimes \psi(y) - \mu_x \otimes \mu_y, f \otimes g\rangle.
\end{aligned}$$

□

A.3.2 Proof for Equivalence between MMD and HSIC

Proof. Define a kernel v in the tensor product space $\mathcal{F} \times \mathcal{G}$ as $v((x, y)(x', y')) = k(x, x')l(y, y')$. The RKHS associated with v is $\mathcal{H}_v$. We have:

$$\begin{aligned}
\mathbf{MMD}^2(\mathcal{T}, P_{xy}, P_xP_y) &= \left\|\mu P_{xy} - \mu P_x P_y\right\|^2_{\mathcal{H}_v} \\
&= \left\|\mathbf{E}_{x,y} v((x,y), \cdot) - \mathbf{E}_x\mathbf{E}_y v((x,y), \cdot)\right\|^2_{\mathcal{H}_v} \\
&= \left\|\mathbf{E}_{x,y} k(x, \cdot)l(y, \cdot) - \mathbf{E}_x\mathbf{E}_y k(x, \cdot)l(y, \cdot)\right\|^2_{\mathcal{H}_v} \\
&= \left\|\mathbf{E}_{x,y} k(x, \cdot)l(y, \cdot) - \mathbf{E}_x k(x, \cdot)\mathbf{E}_y l(y, \cdot)\right\|^2_{\mathcal{H}_v} \\
&= \langle\mathbf{E}_{x,y} k(x, \cdot)l(y, \cdot), \mathbf{E}_{x',y'} k(x', \cdot)l(y', \cdot)\rangle_{\mathcal{H}_v} \\
&\quad + \langle\mathbf{E}_x k(x, \cdot)\mathbf{E}_y l(y, \cdot), \mathbf{E}_{x'} k(x', \cdot)\mathbf{E}_{y'} l(y', \cdot)\rangle_{\mathcal{H}_v} \\
&\quad - 2\langle\mathbf{E}_{x,y} k(x, \cdot)l(y, \cdot), \mathbf{E}_{x'} k(x', \cdot)\mathbf{E}_{y'} l(y', \cdot)\rangle_{\mathcal{H}_v} \\
&= \mathbf{E}_{x,y}\mathbf{E}_{x',y'}[k(x, x')l(y, y')] + \mathbf{E}_x\mathbf{E}_y\mathbf{E}_{x'}\mathbf{E}_{y'}[k(x, x')l(y, y')] \\
&\quad - 2\mathbf{E}_{x,y}\mathbf{E}_{x'}\mathbf{E}_{y'}[k(x, x')l(y, y')].
\end{aligned}$$

□

A.3.3 Empirical HSIC

Let $Z \subseteq \mathcal{X} \times \mathcal{Y}$ be samples independently drawn from P_{xy}. Let us use shorthand notations $k_{ij} = k(x_i, x_j)$ and $l_{ij} = l(y_i, y_j)$. Gram matrices K and L can be defined by $K_{ij} = k_{ij}$ and $L_{ij} = l_{ij}$. The empirical HSIC is given by:

$$\begin{aligned}
&\widehat{\mathbf{HSIC}}(Z, \mathcal{F}, \mathcal{G}) \\
&:= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} k_{ij}l_{ij} + \frac{1}{n^4}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} k_{ij}\right)\left(\sum_{i=1}^{n}\sum_{j=1}^{n} l_{ij}\right) \\
&- \frac{2}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{q=1}^{n} k_{ij}l_{iq} \\
&= \frac{1}{n^2}\text{tr}(KL) + \frac{1}{n^4}(1_n^T K 1_n)(1_n^T L 1_n) - \frac{2}{n^3} 1_n^T K L 1_n \\
&= \frac{1}{n^2}\text{tr}(KHLH),
\end{aligned} \tag{29}$$

where 1_n is an $n \times n$ matrix of ones and $H := I - \frac{1}{n}1_n 1_n^T$.

A.4 More Details on FSIC

A.4.1 Assumptions of FSIC

A positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be analytic on its domain $\mathcal{X} \times \mathcal{X}$ if for all $\mathbf{v} \in \mathcal{X}$, $f(\mathbf{x}) := k(\mathbf{x}, \mathbf{v})$ is an analytic function on $\mathcal{X}$ [4]. Denote $\mathscr{P}$ the set of all Borel probability measures on a topological space (M, $\mathcal{A}$). For a set $\mathscr{L} \subset \mathscr{P}$, $\gamma(\mathbb{P}, \mathbb{Q})$ is a metric for any $\mathbb{P}, \mathbb{Q} \in \mathscr{L}$. A bounded measurable positive definite kernel k is characteristic if $\gamma(\mathbb{P}, \mathbb{Q}) = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}$ [5].

FSIC is a novel approach to measure independence in a linear time manner. It evaluates a random metric between two probability distributions at a finite number of points [6, 4]. This is possible when FSIC satisfies the assumption that the kernels $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ are bounded by B_k and B_l respectively, and the product kernel $g((\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}')) := k(\mathbf{x}, \mathbf{x}')l(\mathbf{y}, \mathbf{y}')$ is characteristic and analytic on $(\mathcal{X} \times \mathcal{Y}) \times (\mathcal{X} \times \mathcal{Y})$ [5].

A.4.2 Empirical FSIC

Let $X := \{x_1, ..., x_n\}$ and $Y := \{y_1, ..., y_n\}$ be observations independently and identically drawn from P_x and P_y. The empirical estimate of FSIC can be written as:

$$\widehat{\mathbf{FSIC}}^2(X, Y) = \frac{1}{J}\sum_{i=1}^{J} \hat{u}(v_i, w_i)^2, \text{where} \tag{30}$$

$$\hat{u}(v, w) := \hat{\mu}_{xy}(v, w) - \widehat{\mu_x\mu_y}(v, w) \tag{31}$$

$$= \frac{1}{n}\sum_{i=1}^{n} k(x_i, v)l(y_i, w) - \frac{1}{n^2}\sum_{i=1}^{n} k(x_i, v)\sum_{j=1}^{n} l(y_i, w). \tag{32}$$

B FBIC Experiments

B.1 Probability Density Functions for Signal Generation

Table 1: This table lists the probability density functions used for signal generation. The kurtosis of each probability density function is also presented. Degree of Freedom is denoted DOF.

Probability Density Function	Kurtosis
Student, 3 DOF	∞
Double exponential	3.00
Uniform	-1.20
Student, 5 DOF	6.00
Exponential	6.00
2 double exponential	1.11
Symmetric 2 Gaussians, multimodal	-1.68
Symmetric 2 Gaussians, transmodal	-0.74
Symmetric 2 Gaussians, unimodal	-0.50
Asymmetric 2 Gaussians, multimodal	-0.53
Asymmetric 2 Gaussians, transmodal	-0.67
Asymmetric 2 Gaussians, unimodal	-0.47
Symmetric 4 Gaussians, multimodal	-0.82
Symmetric 4 Gaussians, transmodal	-0.62
Symmetric 4 Gaussians, unimodal	-0.80
Asymmetric 4 Gaussians, multimodal	-0.77
Asymmetric 4 Gaussians, transmodal	-0.29
Asymmetric 4 Gaussians, unimodal	-0.67

B.2 Number of Basis Functions

To illustrate how the number of basis functions can affect the quality of approximation, 100 experiments were performed on two-channel mixtures of randomly selected distributions. Mixtures had length 250. We fixed Gaussian RBFs with shape parameter 200 and varied the number of basis functions only. We set the number of basis functions to be 5, 10, 15 and 20. The Amari distances resulting from the selected number of basis functions were 8.77, 6.20, 5.92 and 8.34 respectively. When the number of basis functions increased from 5 to 10 then to 15, the quality of approximation indicated by the Amari distance improved alongside. However, when the number of basis functions further increased to 20, the quality of approximation started to degrade.

B.3 RBFs vs Shifted Legendre Polynomial Basis Functions

Legendre polynomials have a fixed region of support whereas RBFs utilized in FBIC have infinite support. Initially we thought that the approximation quality of Legendre polynomials would be better than those of RBFs when applied to variables with finite support. However, the experimental results for data generated from uniform distributions favored RBFs. Therefore, pinpointing the scenario where Legendre polynomials in FBIC can work best remains an open problem.

C FSIC-based ICA

A novel FSIC-based ICA algorithm was implemented to compare with its FBIC counterpart. The neural network used for demixing matrix estimation shared the same architecture with that of the FBIC-based algorithm. The difference was that the neural network cost function was replaced with the $\widehat{\text{NFSIC}}^2$ statistic proposed in [6]. As there was no training stage in our ICA experiments, test locations of NFSIC were randomly picked from Gaussian distributions with corresponding mean and variance of signal samples. The number of test locations were set to 10, 50 or 100. For 2-channel signal mixtures with sample length 250, the best results were achieved when there were 50 test locations. For 2-channel signal mixtures with sample length 1000 and 4-channel signal mixtures with sample length 1000, the best results were achieved when there were 100 test locations. For 4-channel signal mixtures with sample length 4000, the best results were achieved when there were 10 test locations.

References

[1] R. Michael and S. Barry, “Methods of modern mathematical physics,” 1975.

[2] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, “A kernel two-sample test,” *Journal of Machine Learning Research*, vol. 13, no. Mar, pp. 723–773, 2012.

[3] A. Müller, “Integral probability metrics and their generating classes of functions,” *Advances in Applied Probability*, vol. 29, no. 2, pp. 429–443, 1997.

[4] K. P. Chwialkowski, A. Ramdas, D. Sejdinovic, and A. Gretton, “Fast two-sample testing with analytic representations of probability measures,” in *Advances in Neural Information Processing Systems*, 2015, pp. 1981–1989.

[5] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. Lanckriet, “Hilbert space embeddings and metrics on probability measures,” *Journal of Machine Learning Research*, vol. 11, no. Apr, pp. 1517–1561, 2010.

[6] W. Jitkrittum, Z. Szabó, and A. Gretton, “An adaptive test of independence with analytic kernel embeddings,” *arXiv preprint arXiv:1610.04782*, 2016.