Optimal Kernel Combination for Test of Independence against Local Alternatives

Wen-Yu Hua  
Departments of Child and Adolescent Psychiatry  
NYU Langone Medical Center  
littlehanag@gmail.com

Philip Reiss  
Departments of Child and Adolescent Psychiatry  
NYU Langone Medical Center  
Phil.Reiss@nyumc.org

Debashis Ghosh  
Colorado School of Public Health  
University of Colorado  
debashisorama@gmail.com

Abstract

Testing the independence between two random variables $x$ and $y$ is an important problem in statistics and machine learning, where the kernel-based tests of independence is focused to address the study of dependence recently. The advantage of the kernel framework rests on its flexibility in choice of kernel. The Hilbert-Schmidt Independence Criterion (HSIC) was shown to be equivalent to a class of tests, where the tests are based on different distance-induced kernel pairs. In this work, we propose to select the optimal kernel pair by considering local alternatives, and evaluate the efficiency using the quadratic time estimator of HSIC. The local alternative offers the advantage that the measure of efficiency do not depend on a particular alternative, and only requires the knowledge of the asymptotic null distribution of the test. We show in our experiments that the proposed strategy results in higher power than other existing kernel selection approaches.

1 Introduction

Tests of independence have been widely studied in the field of statistics (Bakirov and Szekelyo, 2008; Szekely et al. (2007), and machine learning (Gretton et al. 2005, 2008, 2005; Smola et al. 2007). In particular, the Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al. (2005, 2008); Smola et al. (2007) is a kernel-based independence measure, and it is defined as the maximum mean discrepancy (MMD) between the joint mean embedding and the product of marginal embeddings of $x$ and $y$ (Smola et al., 2007). The advantages of HSIC include consistency against all alternatives (linear or non-linear effects), allowing for variable dimensions of the two input variables, and having the option of using different kernels to represent different assumptions of the underlying variable structures. Recently, HSIC has been shown to be a class of kernel tests (Sejdinovic et al. 2012, 2013), where the tests are based on different combinations of distance-induced kernel
pairs. Based on this finding, a natural extension is to design a strategy that selects the optimal kernel pair from a class of distance-induced kernels.

One intuitive method is to select the kernel combination that maximizes the test statistic (Sriperumbudur et al., 2009), which is equivalent to minimizing the $p$-value; while this approach achieves the highest power in the case of linear correlation among bivariate samples, the maximum statistic is inflated by the polynomial kernel when the dimensions of the variables are large (Section 5.2). Another method for optimal kernel selection is to compute the test statistic that incorporates all the candidate kernels by a weighted average (Wu et al., 2013), but it is difficult to interpret the underlying effects of the variables from the presence of all the kernels. In contrast to these methods that are based on empirical or heuristics results, Gretton et al. (2012) proposed to study the optimality in the scenario of hypothesis testing, where the quadratic time estimator of the HSIC test is reformatted into a linear time estimator that was proved to be Normal distributed asymptotically. Furthermore, Gretton et al. (2012) formulated the optimality in terms of Type II error prob. under the alternative, given the prob. of Type I error (size) $\alpha$ under the null. This approach is elegant but requires a large sample size to achieve the proper power performance.

In this work, we propose to use the efficiency as the criteria for the optimal kernel pair selection. For example, if test A requires 200 samples to achieve a certain power while test B only needs 100 samples for the same power, then test B is twice as efficient. According to this, the classical Neyman-Pearson lemma (the uniformly most powerful tests from Ch. 8 in Casella and Berger (1990)) is considered optimal in the field of parametric tests. For nonparametric approaches such as the test based on HSIC, Hodge-Lehmann (1956) (Ch. 10 in Serfling (1980)) studied the asymptotic efficiency that at a fixed alternative, the optimal one has the faster rate of Type II error prob. approaching to zero than any other tests when size is held at a certain level. However, the asymptotic argument is not trivial, since the Type II error prob. of any test approaches to zero as the sample approaches to infinity.

Therefore, we propose to discuss the efficiency against local alternatives (Pitman efficiency (1949) Ch. 10 in Serfling (1980)), such that as the sample size grows, the alternatives ever closer to the null can be detected by the test. In other words, the Type II error prob. tends to a positive constant given an size $\alpha$. The advantage of local alternatives is that the measure of efficiency obtained does not depend on a particular alternative, it only requires the knowledge of the asymptotic null distribution of the test. Based on the Pitman approach, the optimal choice is the kernel pair that maximizes the power under the local alternative at a given size. We emphasize from the practical point of view that higher powers under “closer” alternatives are more important for the optimality. Our proposed method makes three contributions. First, the test statistic is computed based on the quadratic time estimator of HSIC, which requires fewer samples than the linear time estimator of HSIC to achieve the same power. Second, the optimal kernel pair can be used for the interpretation of input variables. Third, our proposed method results in higher power than other existing methods in our experiments.
2 Preliminaries

In this work, we mainly focus on kernel pair selection in the test of independence using kernels in Hilbert spaces, and distance measures in Euclidean spaces. Therefore, we briefly review some background and definitions of kernels and distance measures.

**Definition 1. Inner product**

Let $\mathcal{H}$ be a vector space over $\mathbb{R}$. A function $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \rightarrow \mathbb{R}$ is said to be an inner product $\mathcal{H}$ if

1. $\langle \alpha f_1 + \beta f_2, g \rangle = \alpha \langle f_1, g \rangle + \beta \langle f_2, g \rangle$
2. $\langle f, g \rangle = \langle g, f \rangle$
3. $\langle f, f \rangle \geq 0$, and $\langle f, f \rangle = 0$ if and only if $f = 0$

A complete inner product space is called a Hilbert space.

**Definition 2. Kernels**

Let $\mathcal{Z}$ be a non-empty set. A function $k : \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R}$ is a kernel if there exists an $\mathbb{R}$-Hilbert space and a mapping $\phi : \mathcal{Z} \rightarrow \mathcal{H}$ such that

$$\forall z, z' \in \mathcal{Z}, k(z, z') := \langle \phi(z), \phi(z') \rangle.$$ (1)

If we have a function with two arguments, how can we determine if it is a valid kernel? We can find a feature mapping and check the conditions defined in (1); however, the feature mapping is not unique. Therefore, the direct property of kernel function is to check the positive definiteness.

**Definition 3. Positive definite**

A symmetric function $k : \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R}$ is positive definite if $\forall n \geq 1$, $\forall (a_1, ..., a_n) \in \mathbb{R}^n$, for all $z_i \in \mathcal{Z}^n$, $i = 1, .., n$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_ia_jk(z_i, z_j) \geq 0.$$

If $\sum_{i=1}^{n} \sum_{j=1}^{n} a_ia_jk(z_i, z_j) > 0$, then the function $k$ is strictly positive definite.

From definition 3, we know that every kernel is a positive definite function. However if a function of two arguments is positive definite, can we conclude that it is a valid kernel? We need to have the following property.

**Definition 4. Reproducing kernel Hilbert space (RKHS)**

Let $\mathcal{H}$ be a Hilbert space of real-valued functions defined on $\mathcal{Z}$. A function $k : \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R}$ is called a reproducing kernel of $\mathcal{H}$ if:

1. $\forall z \in \mathcal{Z}$, $k(\cdot, z) \in \mathcal{H}$, and
2. $\forall z \in \mathcal{Z}$, $\forall f \in \mathcal{H}$, $\langle f, k(\cdot, z) \rangle = f(z)$. 
If $\mathcal{H}$ has a reproducing kernel, it is called a reproducing kernel Hilbert space (RKHS).

Therefore, if a function $k(z, z')$ of two arguments is symmetric and positive definite, then it is a valid unique reproducing kernel defined (Moore-Aronszajn theorem).

We then provide some reviews of distance measure defined on semi-metric spaces of negative type.

**Definition 5. Semi-metric space**

Let $\mathcal{Z}$ be a non-empty set and let $\rho : \mathcal{Z} \times \mathcal{Z} \to [0, \infty)$ be a function such that $\forall z, z' \in \mathcal{Z}$

1. $\rho(z, z') = 0$ iff $z = z'$, and
2. $\rho(z, z') = \rho(z', z)$.

Then $(\mathcal{Z}, \rho)$ is a semimetric space and $\rho$ is a semimetric on $\mathcal{Z}$ (not enforced triangle inequality).

If triangle inequality is held, then $(\mathcal{Z}, \rho)$ is a metric space.

**Definition 6. Negative type**

The semimetric space $(\mathcal{Z}, \rho)$ is said to have negative type if $\forall n \geq 2, z_1, ..., z_n \in \mathcal{Z}$, and $\alpha_1, ..., \alpha_n \in \mathbb{R}$, with $\sum_{i=1}^{n} \alpha_i = 0$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \rho(z_i, z_j) \leq 0.$$

Notice that all Euclidean spaces are of negative type.

### 3 Hilbert Schmidt Independence Criterion

#### 3.1 HSIC for a class of distance-induced kernels

The goal of our work is to provide a kernel combination selection strategy on an unifying class for test of independence, where the class members are and HSIC Gretton et al. (2005, 2008); Smola et al. (2007) and distance covariance Bakirov and Szekely (2008); Szekely et al. (2007). It has been shown that distance covariance is actually a member of the family of HSIC methods when it is discussed under the semimetric space of negative type Sejdinovic et al. (2012, 2013). We now introduce HSIC on a family of distance-induced kernels $\mathcal{K}_\rho$ based on Sejdinovic et al. (2012, 2013):

$$\mathcal{K}_\rho = \left\{ k \in \mathcal{K}_\rho, \text{s.t. } k(z, z') = \frac{1}{2} [\rho(z, z_0) + \rho(z', z_0) - \rho(z, z')] \right\},$$

where $\mathcal{Z}$ is a nonempty set with $z_0 \in \mathcal{Z}$ and $\rho$ is a measure of negative type on a semi-metric space $\mathcal{Z}$, such that the distance-induced kernel $k \in \mathcal{K}_\rho$ is generated by a semi-metric $\rho$ and centered at $z_0$. Let $(\mathcal{X} \subseteq \mathbb{R}^p, \rho_x)$ and $(\mathcal{Y} \subseteq \mathbb{R}^q, \rho_y)$ be semi-metric
spaces of negative type, and \( X \sim P_x \in \mathcal{M}_p(X) \) and \( Y \sim P_y \in \mathcal{M}_p(Y) \) with a joint distribution \( P_{xy} \in \mathcal{M}_{p_x p_y}(X \times Y) \). Denote \( k(x, x') = \langle \phi(x), \phi(x') \rangle \in \mathcal{K}_{p_x} \) on \( X \) with respect to a reproducing kernel Hilbert space (RKHS) \( \mathcal{F} \) and an associated feature mapping \( \phi(x) \in \mathcal{F} \). Likewise, \( l(y, y') = \langle \psi(y), \psi(y') \rangle \) is another distance-induced kernel that generates \( \rho_y \) on \( Y \) with an associated RKHS \( \mathcal{G} \) and a feature mapping \( \psi(y) \in \mathcal{G} \). The HSIC can be measured as:

\[
\text{HSIC}(P_{xy}, \mathcal{F}, \mathcal{G}) = \| \mu_{P_{xy}} - \mu_{P_x P_y} \|^2
\]

\[
= E_{xx'yy'}[k(x, x')l(y, y')] + E_{xx'}[k(x, x')E_{yy'}[l(y, y')]]
\]

\[
- 2E_{xy}[k(x, x')E_{yy'}[l(y, y')]],
\]

where \( x' \) and \( y' \) are the independent copies of \( x \) and \( y \) respectively. The first equation is the MMD between the joint mean embedding \( \mu_{P_{xy}} \) and the product of marginal embeddings \( \mu_{P_x} \mu_{P_y} \) that determines whether two random variables \( x \) and \( y \) are independent \cite{Smola:2007}.

### 3.2 Empirical estimate of HSIC and asymptotic distributions

For the test of independence, the empirical estimate of HSIC(\( P_{xy}, \mathcal{F}, \mathcal{G} \)) for the random samples \((x, y) = (x_1, y_1), \ldots, (x_n, y_n)\) is as follows. Define

\[
S_1 = \frac{1}{n^2} \sum_{ij} k_{ij} l_{ij}, \quad S_2 = \frac{1}{n^2} \sum_{ij} k_{ij} \frac{1}{n} \sum_{ij} l_{ij}, \quad \text{and} \quad S_3 = \frac{1}{n^2} \sum_{ijr} k_{ir} l_{jr}. \quad (4)
\]

Then,

\[
\text{HSIC}_n(x, y) = S_1 + S_2 - 2S_3 = \frac{\text{trace}(KHLH)}{n^2},
\]

where \( K \) is an \( n \times n \) matrix with entries \( k_{ij} = k(x_i, x_j) \), \( L \) is another \( n \times n \) matrix with entries \( l_{ij} = l(y_i, y_j) \), and \( H = I - \frac{1}{n} 11^T \).

Notice that \cite{Gretton:2008}, \cite{Szekely:2007} showed the asymptotic distribution of HSIC test under \( H_0 : P_{xy} = P_x P_y \):

\[
\frac{n \times T_n}{n \times \text{HSIC}_n(x, y)} 
\rightarrow_D Q, \quad (5)
\]

where \( S_2 \) is defined in (4), and \( Q \) follows a weighted sum of independent Chi-square variables with \( E(Q) = 1 \) \cite{Szekely:2007}. \cite{Gretton:2008} also discussed that the empirical estimator of HSIC follows an asymptotic Gaussian distribution under the alternative hypothesis \( (P_{xy} \neq P_x P_y) \). As a result, the asymptotic distribution between the null and the alternative are different, it is hard to evaluate the power under these difference. Therefore, we propose to look at the local alternative surrounding the null hypothesis.

---

1As defined in \cite{Sejdinovic:2013}, let \( \nu \in \mathcal{M}(Z) \) be a finite \( \theta \)-moment w.r.t. a semimetric \( \rho \) on \( Z \) of negative type:

\[
\mathcal{M}_\rho(Z) = \left\{ \nu \in \mathcal{M}(Z) : \exists z_0 \in Z \\text{s.t.} \int \rho^\theta(z, z_0) d|\nu|(z) < \infty \right\}, \quad (3)
\]
4 Hypothesis testing, local alternatives and optimal kernel pair selection

With the representation of HSIC in MMD format for test of independence, one can convert the hypothesis from $H_0 : P_{xy} = P_x P_y$ versus $H_1 : P_{xy} \neq P_x P_y$ into $H_0 : \mu_{P_{xy}} = \mu_{P_x P_y}$ versus $H_1 : \mu_{P_{xy}} = \mu_{P_x P_y} + \Delta$, where $\Delta \neq 0$ is defined in Hilbert space such that $\mu_{P_x P_y} + \Delta$ is still a valid mean embedding measure. Given the observation data $(x, y)$, hypothesis testing assesses the compatibility of the data with the null hypothesis. Since there are exactly two possible outcomes (reject, or fail to reject $H_0$), the false inferences can therefore only be made in two way. Type I errors happen when the null hypothesis is incorrectly rejected, and Type II errors occur when it is incorrectly accepted. For any test, we call $\alpha$ the size of the test which is the prob. of Type I error, and denote $\beta$ as the prob. of Type II error. The power of a test is the prob. of rejecting $H_0$ when it is false, which is $1 - \beta$. In practice, the exact size and power are not computable through the tests because the test distributions are often unknown. In the case of HSIC test, we consider the asymptotic performances instead; let $r_n$ be a sequence of critical values associated to the HSIC test (5) with sample size $n$, , and the limiting size and power of the HSIC test are:

$$\alpha = \lim_{n \to \infty} P(n \times T_n \geq r_n|H_0), \quad \beta = \lim_{n \to \infty} P(n \times T_n \geq r_n|H_1).$$

A kernel pair of HSIC test can be treated as an individual test of independence, while the different kernel pairs are the independence tests for the same $H_0$. A general way to compare the tests is to select the most efficient test, where the test with the highest power given an upper bound of $\alpha$ is desired. In this work, we consider the local alternatives that surround the null, i.e.,

$$H_{local} : \mu_{P_{xy}} = \mu_{P_x P_y} + \Delta_n,$$

for measuring efficiency. The claim of the distribution of the HSIC test under local alternatives is provided by the following statement:

**Theorem 1.** If $\|\Delta_n\|$ can be expressed as $cn^{-1/2}$, where $c$ is an arbitrary constant, then the power function of the HSIC test (5) is independent of sample size $n$, but proportional to $c^2$.

**Proof:**
We first discuss the case that under the null hypothesis. Let $r_n$ are sequence critical values and given an asymptotic size $\alpha$, the limiting size:

$$P(n \times T_n > r_n|H_0) \to P(Q > t_\alpha) = \alpha.$$

Here, $\lim_n r_n = t_\alpha$, where $t_\alpha$ is a threshold associated to distribution $Q$ at level $\alpha$, where $Q$ is defined in (5).

The power function of HSIC test under local alternatives is computed based on B2 in
Gretton et al. (2012) that the empirical HSIC is as follows:

\[
\text{HSIC}_n(x, y) = \frac{1}{n^2} \sum_{ij} k(x_i, x_j)l(y_i, y_j) + \frac{1}{n^4} \sum_{ijqr} k(x_i, x_j)l(y_q, y_r) - \frac{2}{n^3} \sum_{ijq} k(x_i, x_q)l(y_j, y_q)
\]

\[
= \frac{1}{n^2} \sum_{ij} \langle \phi(x_i)\psi(y_i), \phi(x_j)\psi(y_j) \rangle + \frac{1}{n^4} \sum_{ijqr} \langle \phi(x_i)\psi(y_j), \phi(x_q)\psi(y_r) \rangle
\]

\[
- \frac{2}{n^3} \sum_{ijr} \langle \phi(x_i)\psi(y_j), \phi(x_r)\psi(y_r) \rangle
\]

(7)

The first, second, and third terms of (7) are centered by their mean embeddings, which are:

\[
\frac{1}{n^2} \sum_{ij} \langle \phi(x_i)\psi(y_i) - \mu_{Pxy}, \phi(x_j)\psi(y_j) - \mu_{Pxy} \rangle,
\]

\[
\frac{1}{n^4} \sum_{ijqr} \langle \phi(x_i)\psi(y_j) - \mu_{Pxy}, \phi(x_q)\psi(y_r) - \mu_{Pxy} \rangle,
\]

\[
\frac{2}{n^3} \sum_{ijr} \langle \phi(x_i)\psi(y_j) - \mu_{Pxy}, \phi(x_r)\psi(y_r) - \mu_{Pxy} \rangle.
\]

Therefore, the empirical centered HSIC, denoted as \(\text{cHSIC}_n\), is then:

\[
\text{cHSIC}_n(x, y) = \frac{1}{n^2} \sum_{ij} \langle \phi(x_i)\psi(y_i) - \mu_{Pxy}, \phi(x_j)\psi(y_j) - \mu_{Pxy} \rangle
\]

\[
+ \frac{1}{n^4} \sum_{ijqr} \langle \phi(x_i)\psi(y_j) - \mu_{Pxy}, \phi(x_q)\psi(y_r) - \mu_{Pxy} \rangle
\]

\[
- \frac{2}{n^3} \sum_{ijr} \langle \phi(x_i)\psi(y_j) - \mu_{Pxy}, \phi(x_r)\psi(y_r) - \mu_{Pxy} \rangle
\]

With the statement of \(H_{\text{Local}}\), \(\mu_{Pxy} = \mu_{Pxy} + \Delta_n\); the difference between \(\text{HSIC}_n\) and \(\text{cHSIC}_n\) is
\[
\text{HSIC}_n(x, y) - \text{cHSIC}_n(x, y) = -2 \sum_{i}^{n} \langle \phi(x_i)\psi(y_i), \mu_{\text{P}_{xy}} \rangle + \langle \mu_{\text{P}_{xy}}, \mu_{\text{P}_{xy}} \rangle \\
-2 \sum_{i,j}^{n} \langle \phi(x_i)\psi(y_j), \mu_{\text{P}_{xy}} \rangle + \langle \mu_{\text{P}_{xy}}, \mu_{\text{P}_{xy}} \rangle \\
+2 \sum_{i}^{n} \langle \phi(x_i)\psi(y_i), \mu_{\text{P}_{xy}} \rangle + \frac{2}{n^2} \sum_{i,j}^{n} \langle \phi(x_i)\psi(y_j), \mu_{\text{P}_{xy}} \rangle \\
\]
\[
-2 \langle \mu_{\text{P}_{xy}}, \mu_{\text{P}_{xy}} \rangle \\
= \frac{2}{n^2} \sum_{i,j}^{n} \langle \phi(x_i)\psi(y_j), \Delta_n \rangle - \frac{2}{n} \sum_{i}^{n} \langle \phi(x_i)\psi(y_i), \Delta_n \rangle + n \langle \Delta_n, \Delta_n \rangle.
\]

Hence
\[
n \times \text{HSIC}_n(x, y) = \quad (8)
\]
\[
n \times \text{cHSIC}_n(x, y) + \frac{2}{n^2} \sum_{i,j}^{n} \langle \phi(x_i)\psi(y_j), \Delta_n \rangle - \frac{2}{n} \sum_{i}^{n} \langle \phi(x_i)\psi(y_i), \Delta_n \rangle + n \langle \Delta_n, \Delta_n \rangle.
\]

Hence,
\[
n \times \frac{\text{HSIC}_n(x, y)}{S_2} = \quad (9)
\]
\[
n \times \frac{\text{cHSIC}_n(x, y)}{S_2} + \frac{2}{n} \sum_{i,j}^{n} \langle \phi(x_i)\psi(y_j), \Delta_n \rangle - \frac{2}{n} \sum_{i}^{n} \langle \phi(x_i)\psi(y_i), \Delta_n \rangle + \frac{n}{S_2} \langle \Delta_n, \Delta_n \rangle.
\]

Therefore, as \( n \) goes to infinity, \( n \times \text{cHSIC}_n(x, y) S_2^{-1} \to P Q \) under the null hypothesis. Also, by SLLN, \( S_2 \to E(S_2) \) almost surely, and given \( \| \Delta_n \| = cn^{-1/2} \), we obtain \( nS_2^{-1} \times \langle \Delta_n, \Delta_n \rangle = c^2 S_2^{-1} \simeq c^2 \); therefore, \( \| \Delta_n \|^2 \) dominates the second and third terms of (9) in terms of \( c^2 \). So, the local limiting power function given the threshold \( t_{\alpha_0} \) at a size \( \alpha \):
\[
P(n \times T_n > t_{\alpha}) \to P(Q + o(c^2) + c^2 > t_{\alpha}) \\
\simeq P(Q + c^2 > t_{\alpha}) \\
= \pi(c),
\]
where (9) is independent of \( n \), and increases as \( c \) increased.

Notice that the local limiting power function \( \pi(c) = P(Q + c^2 > t_{\alpha}) \) of HSIC test is independent of the sample size \( n \) but proportional to \( c \). Therefore, we are able to select the optimal kernel combination according to the maximum local limiting power \( \pi(c) \), by maximizing the offset \( c^2 \). Furthermore, \( c^2 \) can be decomposed into sample size \( n \) and the difference \( \delta^2 \):
\[
c^2 \simeq n \times \left( \frac{\| \mu_{\text{P}_{xy}} - \mu_{\text{P}_{xy}} - cn^{-1/2} \|^2}{\| \mu_{\text{P}_{xy}} \|^2} - \frac{\| \mu_{\text{P}_{xy}} - \mu_{\text{P}_{xy}} \|^2}{\| \mu_{\text{P}_{xy}} \|^2} \right) \equiv n \times \delta^2.
\]

\[8\]
Since $n$ is fixed when we evaluate the power among the tests, the optimal kernel combination can be determined by maximizing $\delta^2$:

$$
(k^*, l^*) = \arg \sup_{k,l \in \mathcal{K}} \delta^2_{k,l},
$$}

where $\delta^2_{k,l}$ is defined in (11) given a kernel pair $(k, l)$. In practice, we do not know the true difference $\delta^2$, we therefore use the empirical estimates for $\delta^2_{k,l}$, such that

$$
(\hat{k}^*, \hat{l}^*) = \arg \sup_{k,l \in \mathcal{K}} \hat{\delta}^2_{k,l},
$$}

and the empirical $\hat{\delta}^2_{k,l}$ is defined as

$$
\hat{\delta}^2_{k,l} = \text{HSIC}_n(x, y|\mathcal{H}_{\text{local}})S_2^{-1} - \text{HSIC}_n(x, y|\mathcal{H}_0)S_2^{-1},
$$}

where $\text{HSIC}_n(x, y|\mathcal{H}_{\text{local}})S_2^{-1}$ and $\text{HSIC}_n(x, y|\mathcal{H}_0)S_2^{-1}$ are the empirical estimates of $\|\mu_{P_{xy}} - \mu_{P_{x}P_{y}} - cn^{-1/2}\|_2^2 \times \|\mu_{P_{xy}}\|_2^2$ and $\|\mu_{P_{xy}} - \mu_{P_{x}P_{y}}\|_2^2$, respectively.

The following theorem shows the convergence of $(\hat{k}^*, \hat{l}^*)$ to $(k^*, l^*)$.

**Theorem 2.** Consider the finite kernel combinations $(k, l)$'s that are in a class of distance-induced kernels $\mathcal{K}$ in (12), both $k$ and $l$ are therefore bounded. If $E(S_2) = \|\mu_{P_{xy}}\|_2^2$ is bounded away from zero, then $\left|\sup_{k,l \in \mathcal{K}} \hat{\delta}^2_{k,l} - \sup_{k,l \in \mathcal{K}} \delta^2_{k,l}\right| = O_p(n^{-1})$. Therefore, $(\hat{k}^*, \hat{l}^*)$ converges to $(k^*, l^*)$ in probability.

**Proof:**

We start with the bounded difference:

$$
\left|\sup_{k,l \in \mathcal{K}} \hat{\delta}^2_{k,l} - \sup_{k,l \in \mathcal{K}} \delta^2_{k,l}\right| \leq \sup_{k,l \in \mathcal{K}} \left|\hat{\delta}^2_{k,l} - \delta^2_{k,l}\right|
$$

$$
= \sup_{k,l \in \mathcal{K}} \left|\text{HSIC}_n(x, y|\mathcal{H}_{\text{local}})S_2^{-1} - \text{HSIC}_n(x, y|\mathcal{H}_0)S_2^{-1}
- E(\text{HSIC}_n(x, y|\mathcal{H}_{\text{local}})E^{-1}(S_2) + E(\text{HSIC}_n(x, y|\mathcal{H}_0))E^{-1}(S_2)\right|
$$

$$
\leq \sup_{k,l \in \mathcal{K}} \frac{E(S_2)}{S_2E(S_2)} \left|\text{HSIC}_n(x, y|\mathcal{H}_{\text{local}}) - \text{HSIC}_n(x, y|\mathcal{H}_0)
- E(\text{HSIC}_n(x, y|\mathcal{H}_{\text{local}}) - \text{HSIC}_n(x, y|\mathcal{H}_0))\right|
$$

$$
+ \sup_{k,l \in \mathcal{K}} \frac{E(\text{HSIC}_n(x, y|\mathcal{H}_{\text{local}}) - \text{HSIC}_n(x, y|\mathcal{H}_0))}{S_2E(S_2)} \left|S_2 - E(S_2)\right|
$$

$$
\leq \xi_1 \left(\sup_{k,l \in \mathcal{K}} \left|\text{HSIC}_n(x, y|\mathcal{H}_{\text{local}}) - E(\text{HSIC}_n(x, y|\mathcal{H}_{\text{local}}))\right|ight)
+ \xi_2 \sup_{k,l \in \mathcal{K}} \left|S_2 - E(S_2)\right|,
$$

where

(i) $\xi_1 \leq \sup_{k,l \in \mathcal{K}} \left|\text{HSIC}_n(x, y|\mathcal{H}_{\text{local}}) - E(\text{HSIC}_n(x, y|\mathcal{H}_{\text{local}}))\right|$

(ii) and

(iii) $\xi_2 \leq \sup_{k,l \in \mathcal{K}} \left|S_2 - E(S_2)\right|$. 

9
We first focus on the bounded difference under $\mathcal{H}_{\text{local}}$, and the bounded difference can be expanded as:

\[
(i) = \sup_{k,l \in K} \left| \text{HSIC}_n(x,y|\mathcal{H}_{\text{local}}) - E(\text{HSIC}_n(x,y|\mathcal{H}_{\text{local}})) \right|
\]

\[
= \sup_{k,l \in K} \left| \text{cHSIC}_n(x,y) - \frac{2}{n^3} \sum_i \sum_q \langle \phi(x_i)\psi(y_i) - \phi(x_q)\psi(y_r), \Delta_n \rangle + \|\Delta_n\|^2 \right|
\]

\[
- \left\| \mu_{P_{xy}} - \mu_{P_{x}P_{y}} \right\|^2 + 2\langle \mu_{P_{xy}} - \mu_{P_{x}P_{y}}, \Delta_n \rangle - \|\Delta_n\|^2 \right|
\]

\[
\leq \sup_{k,l \in K} \left| \text{cHSIC}_n(x,y) - \left\| \mu_{P_{xy}} - \mu_{P_{x}P_{y}} \right\|^2 \right| + \sup_{k,l \in K} \left| \frac{2}{n^2} \sum_i \langle \phi(x_i)\psi(y_i) - \mu_{P_{xy}}, \Delta_n \rangle \right|
\]

\[
+ \sup_{k,l \in K} \left| \frac{2}{n^2} \sum_q \langle \phi(x_q)\psi(y_r) - \mu_{P_{xy}}, \Delta_n \rangle \right|
\]

\[
\leq E \left( \sup_{k,l \in K} \left| \text{cHSIC}_n(x,y) - \left\| \mu_{P_{xy}} - \mu_{P_{x}P_{y}} \right\|^2 \right| \right) + \frac{\tilde{D}}{n} \sqrt{\frac{2 \ln 2}{\xi}}
\]

The first equation above is based on (8), and the second equation is using triangle inequality. The proof of uniform convergence of means involves three steps: a concentration inequality, symmetrization and simple restrictions.

In (12), the difference of (a) is first concentrated by McDiarmid inequality, where an $(x_i, y_i)$ pair changes the quality of (a) by no more than $\tilde{D}n^{-2}$, therefore, with probability $1 - \xi$:

\[
\sup_{k,l \in K} \left| \text{cHSIC}_n(x,y) - \left\| \mu_{P_{xy}} - \mu_{P_{x}P_{y}} \right\|^2 \right|
\]

\[
\leq E \left( \sup_{k,l \in K} \left| \text{cHSIC}_n(x,y) - \left\| \mu_{P_{xy}} - \mu_{P_{x}P_{y}} \right\|^2 \right| \right) + \frac{\tilde{D}}{n} \sqrt{\frac{2 \ln 2}{\xi}}
\]

Second, we symmetrize by replace $\left\| \mu_{P_{xy}} - \mu_{P_{x}P_{y}} \right\|^2$ by $\text{cHSIC}_n(x',y')$, where $(x',y')$ are the i.i.d. copies of $(x,y)$, and $\text{cHSIC}_n(x,y) = \sum_{ij} \tilde{k}(x_i, x_j)\tilde{l}(y_i, y_j)/n^2$. In particular,
given the observed sample \((x, y) = \{x_i, y_i\}_{i=1}^n\), we have

\[
(a_1) = E \sup_{k, l \in \mathcal{K}} \left| E \left[ \frac{1}{n^2} \sum_{ij} \tilde{k}(x_i, x_j) \tilde{l}(y_i, y_j) - \frac{1}{n^2} \sum_{ij} \tilde{k}(x'_i, x'_j) \tilde{l}(y'_i, y'_j) \right] \right| x, y \leq EE \sup_{k, l \in \mathcal{K}} \left| \frac{1}{n^2} \sum_{ij} \tilde{k}(x_i, x_j) \tilde{l}(y_i, y_j) - \frac{1}{n^2} \sum_{ij} \tilde{k}(x'_i, x'_j) \tilde{l}(y'_i, y'_j) \right| x, y
\]

\[
= E \sup_{k, l \in \mathcal{K}} \left| \frac{1}{n^2} \sum_{ij} \nu_i \nu_j \tilde{k}(x_i, x_j) \tilde{l}(y_i, y_j) \right| x, y \leq 2E \sup_{k, l \in \mathcal{K}} \left| \frac{1}{n^2} \sum_{ij} \nu_i \nu_j \tilde{k}(x_i, x_j) \tilde{l}(y_i, y_j) \right|
= 2\| R_n \|_{\mathcal{K}},
\]

where \( \nu \in \pm 1 \), for all \( i, j \). The term \( R_n = \frac{1}{n^2} \sum_{ij} \nu_i \nu_j \tilde{k}(x_i, x_j) \tilde{l}(y_i, y_j) \) is defined as Rademacher process, and \( E\| R_n \|_{\mathcal{K}} \) is a Rademacher complexity of \( \mathcal{K} \), which leads us to consider the simple restrictions over data. Consider,

\[
\hat{R}_n = E \sup_{k, l \in \mathcal{K}} \left| \frac{1}{n^2} \sum_{ij} \nu_i \nu_j \tilde{k}(x_i, x_j) \tilde{l}(y_i, y_j) \right| x, y \leq E \left| \sum_{ij} \nu_i \nu_j \tilde{k}(x_i, x_j) \tilde{l}(y_i, y_j) \right|
= \frac{1}{n^2} E \left| \sum_{i} \nu_i^2 \tilde{k}(x_i, x_i) \tilde{l}(y_i, y_i) \right| x, y
\]

Therefore, \( E\| R_n \|_{\mathcal{K}} = E(\hat{R}_n) \leq 4\hat{D}n^{-1} \). Hence, it results in our main conclusion of part (a) that with the probability at least \( 1 - \xi \),

\[
\sup_{k, l \in \mathcal{K}} \left| c:\text{HSIC}_n(x, y) - \| \mu_{P_{xy}} - \mu_{P_{x}P_{y}} \| \right| \leq \frac{8\hat{D}}{n} + \frac{\hat{D}}{n} \sqrt{\frac{1}{2} \ln \frac{2}{\xi}}
\]

The McDiarmid inequality is first applied on part (b) in (12), then the supremum of the empirical process of (b) is concentrated to its expectation, and the expectation has the same rate as the expected supremum of of the Rademacher processes with the probability of at least \( 1 - \xi \):

\[
\sup_{k, l \in \mathcal{K}} \left| \frac{2}{n} \sum_{i} \langle \phi(x_i) \psi(y_i) - \mu_{P_{xy}}, \Delta_n \rangle \right| \leq 2E \sup_{k, l \in \mathcal{K}} \left| \frac{2c}{n} \sum_{i} \nu_i \phi(x_i) \psi(y_i) \langle \Delta_n, \| \Delta_n \| \rangle \right| + \frac{D}{n} \sqrt{\frac{1}{2} \ln \frac{2}{\xi}} \leq \frac{4c\sqrt{D}}{n} + \frac{D}{n} \sqrt{\frac{1}{2} \ln \frac{2}{\xi}}
\]

The last equation is based on Lemma 22 [Bartlett and Mendelson (2003)], where \( \| \Delta_n \| = cn^{-1/2} \), \( c \in \mathbb{R} \) is a constant. Similarly, part (c) in (12) is bounded by \( \frac{4c\sqrt{D}}{n\sqrt{n}} + \frac{D}{n} \sqrt{\frac{1}{2} \ln \frac{2}{\xi}} \).
with probability $1 - \xi$. Therefore, (b) and (c) in (12) are both bounded in probability with the rates $n^{-1}$.

Putting them all together,

$$\sup_{k,l \in K} \left| \text{HSIC}_n(x, y|\mathcal{H}_{local}) - E(\text{HSIC}_n(x, y|\mathcal{H}_{local})) \right| = O_p(n^{-1}).$$

The bounded difference between $\text{HSIC}_n(x, y|\mathcal{H}_0)$ and its expectation is similar to (13), because the $\Delta_n$ under the null is zero, therefore, we can focus on part (a) in (12), which is (13). The only difference here is that under the null hypothesis, the expectation of $\text{HSIC}_n(x, y|\mathcal{H}_0)$ equals to zero. Therefore, with the probability of at least $1 - \xi$, we have:

$$(ii) = \sup_{k,l \in K} \left| c\text{HSIC}_n(x, y|\mathcal{H}_0) \right| \leq \frac{2\hat{D}}{n} + \frac{\hat{D}}{n} \sqrt{\frac{1}{2} \ln \frac{2}{\xi}}.$$

Hence,

$$\sup_{k,l \in K} \left| \text{HSIC}_n(x, y|\mathcal{H}_0) - E(\text{HSIC}_n(x, y|\mathcal{H}_0)) \right| = O_p(n^{-1}).$$

In summary, according to SLLN, $S_2$ converges to $E(S_2)$ almost surely, therefore, $S_2$ is bounded. $\xi_1$ and $\xi_2$ are the constants for the boundedness of $E(S_2)/S_2E(S_2)$ and $(E(\text{HSIC}_n(x, y|\mathcal{H}_{local}) - \text{HSIC}_n(x, y|\mathcal{H}_0))/S_2E(S_2)$. In addition, (iii) is one of the component of (i) and (ii), and (i) and (ii) are both bounded in probability with the rate $n^{-1}$ by McDiarmid’s inequality McDiarmid (1989), and Rademacher complexity bound in Bartlett and Mendelson (2003). Therefore, $\sup_{k,l \in K} \hat{d}_{k,l}^2 - \sup_{k,l \in K} \delta_{k,l}^2 = O_p(n^{-1})$, and $(\hat{k}^*, \hat{l}^*)$ converges to $(k^*, l^*)$ in probability, where $(\hat{k}^*, \hat{l}^*)$ and $(k^*, l^*)$ are the corresponding optimizers of $\sup_{k,l \in K} \hat{d}_{k,l}^2$ and $\sup_{k,l \in K} \delta_{k,l}^2$.

□

4.1 Local-alternative Power Maximization (LaPM) Algorithm

We have so far described maximizing power for local alternatives as a strategy for the optimal kernel combination $(k^*, l^*)$, which is to consider the maximum difference $\delta_{k,l}^2$ among different input kernel pairs. Here, we adopt permutation to estimate the difference, by subtracting the average $T_n$’s in (5) over the permuted samples from the $T_n$ calculated on original samples, where we assume that the $T_n$ computed on the original data is from the alternative. We denote our proposed algorithm as Local-alternative Power Maximization (LaPM), and is detailed in the following:

Initialize: Consider a class of distance-induced kernels $\mathcal{K}$ in (2), we evenly split the $2n$ data into training and test points, where the training samples are independent of the test samples.

1. Compute $T_n(x, y|k, l)$ in (5) for each kernel pair $(k, l \in \mathcal{K})$ using the training set.
2. The indices of $y$ are randomly permuted $b$ times, denoted as $\tilde{y}_i; i = 1, \ldots, b$, for a given pair of kernel $(k, l)$, i.e., $\frac{1}{b} \sum_{j=1}^{b} T_n(x, \tilde{y}_j|k, l)$. 

12
3. \( \hat{T}_n(x, y|k, l) \) in (step 1) minus \( \frac{1}{b} \sum_{j=1}^{b} T_n(x, \tilde{y}_j|k, l) \) (in step 2).
4. Select the \((\hat{k}^*, \hat{l}^*) \in \mathcal{K}\) that maximizes \(\hat{\delta}_{k,l}^2\) from training samples.
5. Apply the kernels \(k^*,\) and \(l^*\) on the test samples, compute \(\tilde{T}_n(x, y|\hat{k}^*, \hat{l}^*)\) and then permute \(y\)'s indices \(B\) times over the test samples to evaluate test significance.

### 4.2 Two modified existing methods

For our experiment in Section 4, two existing methods are included for the performance evaluation. Sriperumbudur et al. (2009); Gretton et al. (2012), where their original discussions were focused on two samples test using MMD. For our evaluations, we modify their algorithms into the settings for test of independence.

First, we present the work of Sriperumbudur et al. (2009) for kernel combination approach for test of independence, where their original statistic (minimum \(p\)-value) as the optimal kernel pair, and we denote the modified version as the \(\text{max-HSIC}\) algorithm:

**Initialize:**

Consider a class of distance-induced kernels \(\mathcal{K}\). We evenly split the \(2n\) data into the training samples and test samples.

1. Compute \(T_n(x, y|k, l)\) in (5) for each kernel pair \((k, l \in \mathcal{K}\) using the training samples.
2. Select the \((\hat{k}^*, \hat{l}^*) \in \mathcal{K}\) that maximizes \(T_n(x, y|k, l)\) from training samples.
3. Apply the kernels \(\hat{k}^*,\) and \(\hat{l}^*\) on the test samples, compute \(\tilde{T}_n(x, y|\hat{k}^*, \hat{l}^*)\) and then permute \(y\)'s indices \(B\) times over the test samples for evaluating test significance.

The second method to be evaluated in our experiments is the work of Gretton et al. (2012). In order to modify it for test of independence, we first present the linear time estimate of HSIC, and then discuss its asymptotic distribution for hypothesis testing.

**Lemma 1.** Given distance-induced kernels \((k, l \in \mathcal{K})\),

\[
lHSIC_n(x, y|k, l) := \frac{1}{m} \sum_{i=1}^{m} h((x_{4i-3}, y_{4i-3}, \ldots, (x_{4i}, y_{4i})) \tag{14}
\]

is a linear unbiased estimator for HSIC\((P_{xy}, \mathcal{F}, \mathcal{G})\), where \(m := \lfloor \frac{n}{4} \rfloor\), \(h((x_1, y_1), \ldots, (x_4, y_4)) := f_1(x_1, x_2, x_3, x_4)f_2(y_1, y_2, y_3, y_4), f_1(x_1, x_2, x_3, x_4) := k(x_1, x_2) - k(x_1, x_3) - k(x_2, x_3) + k(x_3, x_4), \) and \(f_2(y_1, y_2, y_3, y_4) := l(y_1, y_2) - l(y_1, y_3) - l(y_2, y_4) + l(y_3, y_4)\).

**Proof:**

The proof is according to Lyons et al. (2013). Let \((\mathcal{X}, \rho_x)\) be a semimetric space of negative type, and \((\mathcal{Y}, \rho_y)\) is another semimetric space of negative type, and let \(x \sim P_x \in \mathcal{M}^2_{\rho_x}(\mathcal{X})\) and \(y \sim P_y \in \mathcal{M}^2_{\rho_y}(\mathcal{Y})\) with a joint distribution \(P_{xy} \in \mathcal{M}^2_{\rho_x \rho_y}(\mathcal{X} \times \mathcal{Y})\). \(k\) and \(l\) are two kernels on \(\mathcal{X}\) and \(\mathcal{Y}\) that are induced by \(\rho_x\) and \(\rho_y\). Now, consider
\((x_i, y_i); i = 1, \ldots, 4\) are the random samples of \(P_{x,y}\). By triangle inequality, 
\[
|f_1(x_1, x_2, x_3, x_4)| \leq g_1(x_1, x_3, x_4) := 2 \max\{k(x_1, x_3), k(x_3, x_4)\},
\]
\[
|f_2(y_1, y_2, y_3, y_4)| \leq g_2(y_2, y_4, y_3) := 2 \max\{l(y_2, y_4), l(y_4, y_3)\},
\]
where \(g_1(x_1, x_3, x_4)\) and \(g_2(y_2, y_4, y_3)\) are integrable. By Fubini’s theorem, the expectation of \(h\) equals to \(\text{HSIC}(P_{x,y}, \mathcal{F}, \mathcal{G})\), i.e., \(E(h) = \text{HSIC}(P_{x,y}, \mathcal{F}, \mathcal{G})\). Therefore, \(\text{IHSIC}_n(x, y|k, l)\) is an unbiased estimator and can be computed in linear time.

Similar to [Gretton et al. (2012)](Gretton et al. 2012), the asymptotic distribution of \(\text{IHSIC}_n\) can be derived by Central Limit Theorem, under both null and alternative hypotheses:
\[
\sqrt{n}(\text{IHSIC}_n(x, y|k, l) - \text{HSIC}(P_{x,y}, \mathcal{F}, \mathcal{G})) \to N(0, 4\sigma^2_{k, l}),
\]
where \(\sigma^2 = [E(h^2) - E^2(h)]\), under the assumption of \(0 < E(h^2) < \infty\). Therefore, the test based on the linear estimator \([13]\) of asymptotic level \(\alpha\) has the threshold \(t_{\alpha|k, l} = n^{-1/2}\sigma_{k, l}\Phi^{-1}(1 - \alpha)\), where \(\Phi^{-1}\) is the inverse CDF of standard normal. The limiting power under \(H_1: \text{HSIC}(P_{x,y}, \mathcal{F}, \mathcal{G}) > 0\) is
\[
P(\text{IHSIC}_n > t_{\alpha|k, l}) = 1 - \Phi\left(\Phi^{-1}(1 - \alpha) - \frac{\sqrt{n}\text{HSIC}(P_{x,y}, \mathcal{F}, \mathcal{G})}{2\sigma_{k, l}}\right). \tag{15}
\]
\([15]\) decreases as the ratio \(\text{HSIC}(P_{x,y}, \mathcal{F}, \mathcal{G})\sigma_{k, l}^{-1}\) increases, since \(\Phi\) is a monotone function. Consequently, the optimal kernel combination \((k^*, l^*)\) can be determined by maximizing the ratio \(\text{HSIC}(P_{x,y}, \mathcal{F}, \mathcal{G})\sigma_{k, l}^{-1}\) over the class \(\mathcal{K}\) given \(t_{\alpha|k^*, l^*}\). However, the true parameters are not known with finite samples, therefore, we use the sample empirical estimators given \(t_{\alpha|k^*, l^*}\) for the optimality, such that
\[
(k^*, l^*) = \arg \sup_{k, l \in \mathcal{K}} \text{IHSIC}\hat{\sigma}_{k, l}^{-1}. \tag{16}
\]
Finally, we denote the modified algorithm based on [Gretton et al. (2012)](Gretton et al. 2012) as the max-ratio algorithm:
Initialize:
Consider a class of distance-induced kernels \(\mathcal{K}\). Evenly split the \(2n\) data into the training samples and test samples.
1. Compute the empirical linear estimator \(\text{IHSIC}_n\), and \(\hat{\sigma}\) under the training samples.
2. Plug \(\text{IHSIC}\) and \(\hat{\sigma}\) into \([16]\) to select the \((k^*, l^*)\), and \(t_{\alpha|k^*, l^*}\).
3. Apply the kernel \((k^*, l^*)\) on the test samples and compute \(\text{IHSIC}_n\), and \(\hat{\sigma}\).
4. Evaluate the significance of \((\sqrt{n} \times \text{IHSIC}_n)/(2\hat{\sigma})\) based on standard normal.

5 Experiments and results

We used \(b = 100\) for LaPM, \(B = 10^4\) permutations, and the computation of empirical size and powers for all numerical analyses were repeated 1000 times. Also, the test significance was set at level 0.05.
5.1 Toy example simulation

The first simulation was designed based on the correlations between bivariate samples under the local alternatives Anderson et al. (1994), such that the \((x, y)\) samples were generated from a Bivariate normal (BVN) with mean zero and a \(2 \times 2\) covariance matrix \(\Sigma\), where the diagonals are \(\sigma^2\) and the off-diagonals are \(cn^{-1/2}\). We compared our proposed LaPM method with max-HSIC, and max-ratio with distance-induced kernels of class \(K\) that included \(L_2\) distance, linear, quadratic, and Gaussian RBF kernels with \(\rho = 0.1, 0.5, 1, 5, 10\) for both \(k\) and \(l\), therefore making the cardinality \(|K| = 8\). We also included a baseline method of Pearson’s correlation coefficient test (\(\rho\)-test) in this evaluation.

![Simulation Results](image)

Figure 1 displays the simulation results. The left panel shows the scatter plots of four different strength \(c\)’s, and \(\sigma^2 = 1\) for 100 and 500 test samples, respectively. We can observe the positive correlations between \(x\)’s and \(y\)’s when \(n = 100, c = 6,\) and \(\sigma^2 = 1\), but the correlation became weaker when \(n = 500, c = 6,\) and \(\sigma^2 = 1\), this is because of the design of simulation is based on local alternatives. The right part of Figure 1 shows that the empirical size of all approaches were all close to size 0.05 when \(c = 0\), which suggests all methods are able to control Type I errors, and the empirical powers \((c > 0)\) all decreased when the \(\sigma^2\) increased. In addition, the increase in sample size did not affect the empirical powers much due to the local alternative effects of our simulation setting. Of the evaluated approaches, both max-HSIC and LaPM were very close to the baseline approach of \(\rho\)-test; but, the max-ratio that was based on the linear estimator strategy showed less power on the correlation detection. This reflected the fact that max-HSIC and LaPM are computed based on \(n\) samples, while max-ratio is calculated using \(n/4\) samples.
5.2 Simulation based on Alzheimer’s Disease Neuroimaging Initiative study

In this experiment, we consider associations between $x$ and $y$ in a more higher-dimensional setting. The simulation design was based on the Alzheimer’s Disease Neuroimaging Initiative (ADNI) study dataset ADNI (2003). ADNI dataset contains brain magnetic resonance imaging (MRI) scans of the enrolled subjects, where the goal of the ADNI study is to find the correlation between the genetic variants (in terms of single nucleotide polymorphisms (SNPs)) and the changes in brain volume (in terms of brain MRI scans). We utilized the same simulation design as the work in Hua and Ghosh (2014), where a linear model $y_r = h(x) + \epsilon_r$ with $r = 1, \ldots, q$ was used to associate the phenotypes ($y = (y_1, \ldots, y_q)$) and genotypes ($x = (x_1, \ldots, x_p)$). The structure of the responses ($y_1, \ldots, y_q$) was modelled using a covariance matrix $\Sigma$ based on the eight ($q = 8$) positive correlated frontal cortex regions using the 358 mild cognitive impairment (MCI) subjects (Figure 2a), since the MRI scans of the MCI samples are relatively more uniform than both the healthy and disease groups Vounou et al. (2010). Therefore, the multivariate responses were according to all $\epsilon$’s that were generated from multivariate normal MVN(0, $\Sigma$). For generating the predictors (i.e., the genotype effects), 141 SNPs on gene FLJ16124 were used for the genotype elements, i.e. $x = (x_1, \ldots, x_{141})$, from Hua and Ghosh (2014). The effect of $h(x) = k(x, x')$ was defined as $h(x_1, \ldots, x_{141}) = c \times h_1$, while only the first 5 SNPs, $(x_1, \ldots, x_5)$ of 141 $x$’s were the causative SNPs, such that $h_1(x_1, \ldots, x_5) = 2\cos(x_1) - 3x_2^2 + 2\exp(-x_3)x_4 - 1.6\sin(x_5)\cos(x_3) + 4x_1x_5$, and $c$ was the association strength, such that $c = 0, 2, 4, 6$.

200 and 500 (2n) samples were generated, and the empirical size ($c = 0$) and powers ($c = 0.05, 0.1$) were computed based on 1000 runs and the significance level of 0.05. The kernel choice for $k$ included $L_2$ distance, linear, quadratic and identical by state (IBS) Wessel and Schork (2006); the kernel choice for $l$ were $L_2$ distance, linear, quadratic, and Gaussian RBF with $\rho \in 0.1, 0.5, 1, 5, 10$, where $|K| = 22$. In addition, we added another kernel selection method called the perturbation method from Wu et al. (2013), denoted as the avg method in this experiment, which analyzes the association based the single phenotype kernel machine regression (KMR) model Liu et al. (2007), and uses a weighted average among the multiple candidate kernels for the optimal selection. The advantage of the avg method lies in its ability to incorporate all samples for kernel selection, while LaPM, max-HSIC, and max-ratio all require the data to be split into training and test sets. To evaluate the avg method under this simulation, we used the first principal component of the eight phenotypes as the single phenotype for KMR, and included all four kernels for the genotype effects. Furthermore, we evaluated the avg method using both the entire sample size, as well as the second half only setting (corresponding to test samples), denoted as avg$_w$ and avg$_h$.

Figure 2a shows the frequency counts of the ADNI simulation for $K$, where the max-HSIC only chose the kernel pairs $(k, l)$=(quadratic,linear) or (quadratic,RBF with $\rho = 10$) for all strengths of associations $c = 0.0, 0.05, 0.1$, which was due to the quadratic effects of high dimensional variables (141 SNPs) from the quadratic kernel; furthermore, since max-HSIC only selects the largest statistic among the candidate kernel pairs, therefore (quadratic,linear) and (quadratic,RBF with $\rho = 10$) were the two combinations se-
Figure 2: (a) all-pairwise correlations of the 358 MCI subjects using eight \((q = 8)\) prefrontal cortex regions (1-2: left and right \((l/r)\) anterior dorsolateral, 3-4: \(l/r\) posterior dorsolateral, 5-6: \(l/r\) anterior medial, 7-8: \(l/r\) posterior medial) of the 119 region of interests (ROIs) for the corrected structure in \(\varepsilon\). (b) Kernel pair selection frequencies based on 250 training samples for the 1000-run simulation in 5.2. Top to bottom: max-HSIC and the proposed LaPM, where the x-axis indicates the 22 candidate kernel pairs, bar colors represent the association strength \(c = 0, 0.05\) and 0.1 as blue, red and green.

Selected based on such largest-statistic strategy (max-HSIC). In the contrary, our proposed method (LaPM) demonstrated more uniformity in the kernel selection when there was no association, and selecting the appropriate kernel pairs in \((k^*, l^*) = (\text{quadratic, linear})\) as the associations strength increased. Due to the lack of empirical power, we skipped the frequency count of max-ratio, and in fact the frequency distributions of max-ratio are relatively uniform among all kernel pairs in the class \(\mathcal{K}\).

In summary, the (quadratic, linear) kernel combination from our proposed method achieved a higher power than max-HSIC. This suggest that there exists strong pairwise interaction effects among the SNPs in \(FLJ16124\) that associate with eight brain regions. Table 1 shows the empirical size and powers for the max-HSIC, LaPM, max-ratio, and avgs results, and the values of empirical size of max-HSIC, LaPM and max-ratio were all close to \(\alpha = 0.05\), which suggests that all approaches were able to control Type I error \(\alpha\). However, the values of empirical size of avg\(_h\) and avg\(_w\) were less than 0.05, which suggests the test of avg approach is on the conservative side. For the empirical powers on the test samples, LaPM achieved the highest power with avg\(_h\) coming in second, followed by max-HSIC and then max-ratio: the performance of the avg\(_h\) method can be affected by bad kernels due to its averaging process; the max-HSIC’s largest-statistic strategy were over-influenced by the quadratic kernel’s inflation effect; and the max-ratio required more samples in the linear estimator. Finally, avg\(_w\) achieved the highest empirical power but required more samples (the entire sample size).
### Table 1: Empirical size and powers from simulation 5.2.

| method    | n  | size  | power  | n  | size  | power  |
|-----------|----|-------|--------|----|-------|--------|
|           | c = 0 | c = 0.05 | c = 0.1 | c = 0 | c = 0.05 | c = 0.1 |
| max-HSIC  | 100 | 0.042 | 0.281 | 0.598 | 250 | 0.043 | 0.541 | 0.577 |
| LaPM      | 100 | 0.046 | 0.381 | 0.943 | 250 | 0.049 | 0.891 | 1.000 |
| max-ratio | 100 | 0.060 | 0.045 | 0.052 | 250 | 0.051 | 0.057 | 0.070 |
| avgₙ      | 100 | 0.016 | 0.301 | 0.929 | 250 | 0.022 | 0.864 | 1.000 |
| avgₚ      | 200 | 0.027 | 0.740 | 0.999 | 500 | 0.027 | 0.999 | 1.000 |

6 Real ADNI study

We also applied the proposed LaPM approach on the real ADNI samples to find the optimal kernel pair and the associations between the genetic variants and multivariable brain MRI scans. In contrast to the previous simulation setup, we utilized all 741 subjects of the ADNI study, using 141 SNPs within gene FLJ16124 as the predictors, and 119 ROIs as the responses. To select the optimal kernel, we randomly split the 741 subjects into 400 training samples and 341 test samples, and we set $b = 100$, $B = 10^4$ for computing $p$-value where the significance level was 0.05. The same distance-induced kernels class $K$ (22 kernel pairs) discussed in the simulation was again considered here. Table 2 displays the top five kernel pairs selected by the LaPM method, the optimal kernel pair was the ($L_2$ distance,$L_2$ distance), and the $p$-value was 0.0239. The result is consistent to the findings from Hua and Ghosh (2014), that the SNP located on gene FLJ16124 shows the strongest association to brain MRI regions in the neuroimaging genomewide association study. In addition, the second to fifth kernel pairs in Table 2 were insignificant because their $p$-values were all greater than 0.05, one possible reason would be the number of test samples are too small. Furthermore, the $p$-value of avgₚ was 0.125, where we used the entire 741 subjects, applied first PC for the phenotype, and considered the linear, quadratic and IBS kernels for 141 genotype effects.

| Rank | Kernel pair ($k,l$) | $p$-value |
|------|--------------------|-----------|
| 1    | $L_2$, $L_2$       | 0.0231    |
| 2    | Quadratic, Quadratic | 0.1180    |
| 3    | Quadratic, Linear   | 0.1805    |
| 4    | Linear, Quadratic   | 0.1764    |
| 5    | Linear, Linear      | 0.1988    |

Table 2: Applying LaPM on real ADNI dataset: Top five kernel pairs among 22 kernel pairs with the statistic on 391 test samples, and the corresponding $p$-values based on $10^4$ permutations.
7 Conclusion and discussion

In this work, we have introduced a strategy to select the optimal kernel combination for test of independence against local alternatives. Our proposed method using the quadratic time estimator of HSIC achieved the highest power when compared to other existing methods in our experiments, and the optimal kernel pair selected from our proposed LaPM approach could potentially explain the underlying structure of input variables. Furthermore, we adopted permutations to evaluate the test power when the asymptotic null distribution is not a Normal or Chi-Square distribution.

The permutation-based approach was utilized to approximate the asymptotic null distribution in this work, and while the study of asymptotic distribution is worthy of further studies, we leave it for future work as it is beyond the scope of the current article. Finally, based on our simulations, the simple kernel pair (i.e., \((k^*, l^*)=\text{(linear,linear)}\) and \((\text{quadratic,linear})\)) outperformed other candidate kernel combinations, which suggests the importance of further explorations in the kernel structures by examining an objective function combined with kernel complexity penalization.

References

ADNI (2003). Alzheimer’s disease neuroimaging initiative. http://www.loni.ucla.edu/ADNI/.

Anderson, N. H., P. Hall, and D. Titterington (1994). Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. Journal of Multivariate Analysis.

Bakirov, N. K. and G. J. Szekelyo (2008). Brownian covariance and central limit theorem for stationary sequences. Theory of Probability & Its Applications.

Bartlett, P. L. and S. Mendelson (2003). Rademacher and gaussian complexities: Risk bounds and structural results. JMLR.

Casella, G. and R. L. Berger (1990). Statistical inference. Duxbury Press Belmont, CA.

Gretton, A., K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola (2012). A kernel two-sample test. JMLR.

Gretton, A., O. Bousquet, A. Smola, and B. Scholkopf (2005). Measuring statistical dependence with hilbert-schmidt norms. In Algorithmic Learning Theory.

Gretton, A., K. Fukumizu, C. Teo, L. Song, B. Schölkopf, and A. J. Smola (2008). A kernel statistical test of independence. In NIPS.

Gretton, A., R. Herbrich, and A. Hyvarinen (2005). Kernel methods for measuring independence. JMLR.

Gretton, A., B. Sriperumbudur, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, and K. Fukumizu (2012). Optimal kernel choice for large-scale two-sample tests. In NIPS.
Hua, W.-Y. and D. Ghosh (2014). Equivalence of kernel machine regression and kernel distance covariance for multidimensional trait association studies. arXiv preprint arXiv:1402.2679.

Liu, D., X. Lin, and D. Ghosh (2007). Semiparametric Regression of Multidimensional Genetic Pathway Data: Least-Squares Kernel Machines and Linear Mixed Models. Biometrics.

Lyons, R. et al. (2013). Distance covariance in metric spaces. The Annals of Probability.

McDiarmid, C. (1989). On the method of bounded differences. Surveys in combinatorics.

Sejdinovic, D., A. Gretton, K. Fukumizu, and B. K. Sriperumbudur (2012). Hypothesis testing using pairwise distances and associated kernels. In ICML.

Sejdinovic, D., B. Sriperumbudur, A. Gretton, K. Fukumizu, et al. (2013). Equivalence of distance-based and rkhs-based statistics in hypothesis testing. The Annals of Statistics.

Serfling, R. J. (1980). Approximation Theorems of Mathematical Statistics. Wiley Series in Probability and Statistics.

Smola, A., A. Gretton, L. Song, and B. Schölkopf (2007). A hilbert space embedding for distributions. In Algorithmic Learning Theory.

Sriperumbudur, B. K., K. Fukumizu, A. Gretton, G. R. Lanckriet, and B. Schölkopf (2009). Kernel choice and classifiability for RKHS embeddings of probability distributions. In NIPS.

Szekely, G. J., M. L. Rizzo, and N. K. Bakirov (2007). Measuring and testing dependence by correlation of distances. The Annals of Statistics.

Vounou, M., T. E. Nichols, and G. Montana (2010). Discovering Genetic Associations With High-dimensional Neuroimaging Phenotypes: A Sparse Reduced-rank Regression Approach. Neuroimage.

Wessel, J. and N. J. Schork (2006). Generalized Genomic DistanceVBased Regression Methodology for Multilocus Association Analysis. The American Journal of Human Genetics.

Wu, M. C., A. Maity, S. Lee, E. M. Simmons, Q. E. Harmon, X. Lin, S. M. Engel, J. J. Molldrem, and P. M. Armistead (2013). Kernel machine snp-set testing under multiple candidate kernels. Genetic Epidemiology.