Supplementary material for “Multiscale Fisher’s Independence Test for Multivariate Dependence”

BY S. GORSKY
Department of Mathematics and Statistics, University of Massachusetts Amherst, Amherst, Massachusetts 01003, U.S.A.
sgormy@umass.edu

AND L. MA
Department of Statistical Science, Duke University, Durham, North Carolina 27708, U.S.A.
li.ma@duke.edu

1. TECHNICAL PROOFS

In this section, we establish proofs for Theorem 2.1, Theorem 2.2, and Corollary 2.1 in the main paper. We will introduce a number of definitions and lemmas along the way that will be used to complete the proofs.

DEFINITION S1 (LEVEL-k CANONICAL MARGINAL PARTITION). \( P^k \) is a level-\( k \) canonical marginal partition if

\[ P^k = \left\{ \left[ \frac{l - 1}{2^k}, \frac{l}{2^k} \right) \right\}_{l \in \{1, \ldots, 2^k\}}. \]

To simplify the notations in the proofs, we let \( Z = (X, Y) \). More specifically, we set \( Z_1 = X_1, \ldots, Z_{D_x} = X_{D_x} \) and \( Z_{D_x+1} = Y_1, \ldots, Z_D = Y_{D_y} \). Thus, \( Z \) is a random vector that distributed according to the distribution \( F \), the joint sampling distribution of \( (X, Y) \).

Denote now any \( k \)-stratum as \( A_k \). As described in Section 2, we form \( D \)-dimensional cuboids by taking the Cartesian product of one interval from each of the \( D \) canonical marginal partitions of \( \Omega \). Given then \( k = (k_1, \ldots, k_D) \in \mathbb{N}_0^D \), we refer to the partition \( P^{k_1} \times \cdots \times P^{k_D} \) as the \( k \)-stratum of \( \Omega \).

For \( k = (k_1, \ldots, k_D) \in \mathbb{N}_0^D \), we refer to the partition \( P^{k_1} \times \cdots \times P^{k_D} \) as the \( k \)-stratum of \( \Omega \). Denote now any \( k \)-stratum as \( A^k \). As described in Section 2, we form \( D \)-dimensional cuboids by taking the Cartesian product of one interval from each of the \( D \) canonical marginal partitions of \( \Omega \). Given then \( k = (k_1, \ldots, k_D) \in \mathbb{N}_0^D \), we refer to the partition \( P^{k_1} \times \cdots \times P^{k_D} \) as the \( k \)-stratum of \( \Omega \). Denote now any \( k \)-stratum as \( A^k \). As described in Section 2, we form \( D \)-dimensional cuboids by taking the Cartesian product of one interval from each of the \( D \) canonical marginal partitions of \( \Omega \). Given then \( k = (k_1, \ldots, k_D) \in \mathbb{N}_0^D \), we refer to the partition \( P^{k_1} \times \cdots \times P^{k_D} \) as the \( k \)-stratum of \( \Omega \). Denote now any \( k \)-stratum as \( A^k \). As described in Section 2, we form \( D \)-dimensional cuboids by taking the Cartesian product of one interval from each of the \( D \) canonical marginal partitions of \( \Omega \). Given then \( k = (k_1, \ldots, k_D) \in \mathbb{N}_0^D \), we refer to the partition \( P^{k_1} \times \cdots \times P^{k_D} \) as the \( k \)-stratum of \( \Omega \). Denote now any \( k \)-stratum as \( A^k \). As described in Section 2, we form \( D \)-dimensional cuboids by taking the Cartesian product of one interval from each of the \( D \) canonical marginal partitions of \( \Omega \). Given then \( k = (k_1, \ldots, k_D) \in \mathbb{N}_0^D \), we refer to the partition \( P^{k_1} \times \cdots \times P^{k_D} \) as the \( k \)-stratum of \( \Omega \).

We say that \( X \) and \( Y \) are \( k \)-independent and write it as \( X \perp \perp_k Y \) if for any \( A \in A^k \),

\[ \text{pr} (X \in A_x, Y \in A_y) = \text{pr} (X \in A_x) \cdot \text{pr} (Y \in A_y). \]

Figure S1 illustrates the above notations in three dimensional space.

We say that \( X \) and \( Y \) are \( k \)-independent and write it as \( X \perp \perp_k Y \) if for any \( A \in A^k \),

\[ \text{pr} (X \in A_x, Y \in A_y) = \text{pr} (X \in A_x) \cdot \text{pr} (Y \in A_y). \]

Figure S1 illustrates the above notations in three dimensional space.

We say that \( X \) and \( Y \) are \( k \)-independent and write it as \( X \perp \perp_k Y \) if for any \( A \in A^k \),

\[ \text{pr} (X \in A_x, Y \in A_y) = \text{pr} (X \in A_x) \cdot \text{pr} (Y \in A_y). \]

Figure S1 illustrates the above notations in three dimensional space.
DEFINITION S3 ((i, j)-BLOCKS OF A CUBOID). For every cuboid $A \in A^k$, $i \in \{1, \ldots, D_x\}$ and $j \in \{1, \ldots, D_y\}$, one can partition $A$ into four blocks by dividing $A$ in the $(i, j)$th face, i.e.
Multiscale Fisher’s Independence Test

the side of $A$ spanned by the $i$th and $j$th dimensions, while keeping the other dimensions intact.

$$A = A_{ij}^0 \cup A_{ij}^1 \cup A_{ij}^{10} \cup A_{ij}^{11},$$

where for $a, b \in \{0, 1\},$

$$A_{ij}^{ab} = \prod_{d=1}^{D} \left\{ \begin{array}{ll}
\left[ \frac{2l_d - 2 + a}{2^{k_d + 1}}, \frac{2l_d - 1 + a}{2^{k_d + 1}} \right) & \text{if } d = i \\
\left[ \frac{2l_d - 2 + b}{2^{k_d + 1}}, \frac{2l_d - 1 + b}{2^{k_d + 1}} \right) & \text{if } d = D_x + j \\
\left[ \frac{l_d - 1}{2^{k_d + 1}}, \frac{l_d}{2^{k_d + 1}} \right) & \text{if } d \in \{ 1, \ldots, D \} \setminus \{ i, D_x + j \}
\end{array} \right..$$

Figure S3 illustrates Definition S3 in three dimensions.

![Diagram of A](image)

Fig. S3. $(i = 1, j = 1)$ blocks of $A$.

A 3D view of the $k$-cuboid $A$ where $D = 3, k = (1, 0, 2)$ and $l = (2, 1, 2)$, the same as in Fig. S1 and Fig. S2. The $i = 1$ dimension of $X$ corresponds to dimension 1 of $Z$ the $j = 1$ dimension of $Y$ corresponds to dimension 3 of $Z$. The blocks are:

$$A_{11}^{00} = \left[ \frac{2 \cdot 2 - 2 + 0}{2^{1+1}}, \frac{2 \cdot 2 - 1 + 0}{2^{1+1}} \right) \times \left( \frac{1 - 1}{2^0}, \frac{1}{2^0} \right) \times \\
\left[ \frac{2 \cdot 2 - 2 + 0}{2^{1+1}}, \frac{2 \cdot 2 - 1 + 0}{2^{1+1}} \right) = \left[ \frac{0.5}{2^{1+1}}, \frac{0.5}{2^{1+1}} \right),$$

$$A_{11}^{01} = \left[ \frac{2 \cdot 2 - 2 + 1}{2^{1+1}}, \frac{2 \cdot 2 - 1 + 1}{2^{1+1}} \right) \times \left( \frac{1 - 1}{2^0}, \frac{1}{2^0} \right) \times \\
\left[ \frac{2 \cdot 2 - 2 + 1}{2^{1+1}}, \frac{2 \cdot 2 - 1 + 1}{2^{1+1}} \right) = \left[ \frac{0.75}{2^{1+1}}, \frac{0.75}{2^{1+1}} \right),$$

$$A_{11}^{10} = \left[ \frac{2 \cdot 2 - 2 + 1}{2^{1+1}}, \frac{2 \cdot 2 - 1 + 1}{2^{1+1}} \right) \times \left( \frac{1 - 1}{2^0}, \frac{1}{2^0} \right) \times \\
\left[ \frac{2 \cdot 2 - 2 + 0}{2^{1+1}}, \frac{2 \cdot 2 - 1 + 0}{2^{1+1}} \right) = \left[ \frac{0.75}{2^{1+1}}, \frac{0.75}{2^{1+1}} \right),$$

$$A_{11}^{11} = \left[ \frac{2 \cdot 2 - 2 + 1}{2^{1+1}}, \frac{2 \cdot 2 - 1 + 1}{2^{1+1}} \right) \times \left( \frac{1 - 1}{2^0}, \frac{1}{2^0} \right) \times \\
\left[ \frac{2 \cdot 2 - 2 + 1}{2^{1+1}}, \frac{2 \cdot 2 - 1 + 1}{2^{1+1}} \right) = \left[ \frac{0.75}{2^{1+1}}, \frac{0.75}{2^{1+1}} \right).$$

Lemma S1 establishes an equivalence between multivariate independence and a cascade of discretized multivariate independence relations:
**Lemma S1.**

$$X \perp Y \iff X \perp_k Y \text{ for all } k \in \mathbb{N}_0^D.$$  

**Proof.** $\Rightarrow$: Immediate. 

$\Leftarrow$: 

Let 

$$\mathcal{P}_x = \bigcup_{k(1, \ldots, D_x) \in \mathbb{N}_0^D} \mathcal{P}^1 \times \ldots \times \mathcal{P}^{D_x}.$$  

Therefore $\sigma(\mathcal{P}_x) = B([0, 1]^{D_x})$. 

For $A_x \in B([0, 1]^{D_x})$ let $[X \in A_x] = \{\omega : X(\omega) \in A_x\}$, then $\sigma(X) = \{[X \in A_x], A_x \in B([0, 1]^{D_x})\}$. 

Let 

$$\mathcal{Q}_x = \bigcup_{k(1, \ldots, D_x) \in (-1, 0, 1, \ldots, D_x)^D} \mathcal{Q}^{k_1} \times \ldots \times \mathcal{Q}^{k_{D_x}}$$  

so that $Q^{-1} = \emptyset$ and for all $k \geq 0$, $Q^k = \mathcal{P}^k$. 

Let $C_X = \{[X \in B_x], B_x \in \mathcal{Q}_x\}$. 

Hence 

$$\sigma(C_X) = \sigma\{X^{-1}(B_x), B_x \in \mathcal{Q}_x\} = \sigma\{X^{-1}(\mathcal{Q}_x)\} = X^{-1}\{\sigma(\mathcal{P}_x)\} = \sigma(X).$$ 

$C_X$ is a $\pi$-system: 

$$E, E' \in C_X \Rightarrow E = \left\{X \in \left(\emptyset \text{ or } \left[\frac{l_1 - 1}{2^{k_1}}, \frac{l_1}{2^{k_1}}\right]\right) \times \ldots \times \left(\emptyset \text{ or } \left[\frac{l_{D_x} - 1}{2^{k_{D_x}}}, \frac{l_{D_x}}{2^{k_{D_x}}}\right]\right)\right\}$$ 

and 

$$E' = \left\{X \in \left(\emptyset \text{ or } \left[\frac{l'_1 - 1}{2^{k'_1}}, \frac{l'_1}{2^{k'_1}}\right]\right) \times \ldots \times \left(\emptyset \text{ or } \left[\frac{l'_{D_x} - 1}{2^{k'_{D_x}}}, \frac{l'_{D_x}}{2^{k'_{D_x}}}\right]\right)\right\}$$ 

for some $l, l', k, k'$. Therefore: 

$$E \cap E' = \left\{X \in \left(\emptyset \text{ or } \left[\frac{l_1 - 1}{2^{k_1}}, \frac{l_1}{2^{k_1}}\right]\right) \times \ldots \times \left(\emptyset \text{ or } \left[\frac{l_{D_x} - 1}{2^{k_{D_x}}}, \frac{l_{D_x}}{2^{k_{D_x}}}\right]\right)\right\} \bigcap \left\{\left(\emptyset \text{ or } \left[\frac{l'_1 - 1}{2^{k'_1}}, \frac{l'_1}{2^{k'_1}}\right]\right) \times \ldots \times \left(\emptyset \text{ or } \left[\frac{l'_{D_x} - 1}{2^{k'_{D_x}}}, \frac{l'_{D_x}}{2^{k'_{D_x}}}\right]\right)\right\}$$ 

$$= \left\{X \in \left(\emptyset \text{ or } \left[\frac{l_1 - 1}{2^{k_1}}, \frac{l_1}{2^{k_1}}\right] \cap \left[\frac{l'_1 - 1}{2^{k'_1}}, \frac{l'_1}{2^{k'_1}}\right]\right) \times \ldots \times \left(\emptyset \text{ or } \left[\frac{l_{D_x} - 1}{2^{k_{D_x}}}, \frac{l_{D_x}}{2^{k_{D_x}}}\right] \cap \left[\frac{l'_{D_x} - 1}{2^{k'_{D_x}}}, \frac{l'_{D_x}}{2^{k'_{D_x}}}\right]\right)\right\} \in C_X$$ 

Similarly, define $\sigma(Y)$ and $C_Y$, another $\pi$-system, independent of $C_X$ and $\sigma(Y) = \sigma(C_Y)$. By the basic criterion, then, $\sigma(X) \perp \sigma(Y)$. 

The following lemma shows that if $X$ and $Y$ are $k$-independent, they are also independent on all coarser strata. 

**Lemma S2.** If $X \perp_k Y$ then $X \perp_{k'} Y$ for all $k' \leq k$. 

**Proof.** Let $F_X$ be the probability distribution of $X$ and $F_Y$ be the probability distribution of $Y$. Then $X \perp_k Y \iff F(A) = F_X(A_x) F_Y(A_y)$ for all $A \in \mathcal{A}^k$. 

Multiscale Fisher’s Independence Test

It is enough to show that $X \perp_{k} Y \Rightarrow X \perp_{k'} Y$ for any $k'$ such that (i) $k'_i = k_i - 1$ for some $i \in \{1, ..., D_x\}$ and (ii) $k'_d = k_d$ for all $d \in \{1, ..., D\} \setminus \{i\}$.

Hence, assuming $X \perp_{k} Y$ we get for $A \in \mathcal{A}^k$ that:

$$F(A) = F_0(A) - F_1(A)$$

$$= F(X_i \perp_{k} Y, Z | D_{d', d}) = F(X_i \perp_{k} Y, Z | D_{d', d})$$

As required. $\square$

In Lemma S6 we will show how to characterize the multivariate $k$-independence with a collection of univariate $(i, j)$ odds-ratios. However, before we state and prove Lemma S6 we develop additional notations and provide discretized versions of some basic results in probability.

Denote:

$$d \subset \{1, ..., D\}, i = d \cap \{1, ..., D_x\} \text{ and } j = \{j^*: j^* + D_x \in d \cap \{D_x + 1, ..., D\}\}$$

And

$$X_i = \{X_{i^*} : i^* \in i\}, \quad X_{(i)} = \{X_{i^*} : i^* \in \{1, ..., D_x\} \setminus i\}$$

$$Y_j = \{Y_{j^*} : j^* \in j\}, \quad Y_{(j)} = \{Y_{j^*} : j^* \in \{1, ..., D_y\} \setminus j\}$$

$$Z_d = X_i \times Y_j = \{Z_{d'} : d' \in d\}, \quad Z_{(d)} = X_{(i)} \times Y_{(j)} = \{Z_{d'} : d' \in \{1, ..., D\} \setminus d\}$$

$$A_{x,d} = \bigotimes_{d' \in d \setminus \{1, ..., D_x\}} A_{d'}, \quad A_{x,(d)} = \bigotimes_{d' \in \{1, ..., D_x\} \setminus d} A_{d'}$$

$$A_{y,d} = \bigotimes_{d' \in d \setminus \{1, D_x + 1, ..., D\}} A_{d'}, \quad A_{y,(d)} = \bigotimes_{d' \in \{D_x + 1, ..., D\} \setminus d} A_{d'}$$

$$A_d = A_{x,d} \times A_{y,d} = \bigotimes_{d' \in d} A_{d'}, \quad A_{(d)} = A_{x,(d)} \times A_{y,(d)} = \bigotimes_{d' \in \{1, ..., D\} \setminus d} A_{d'}$$

$$k_d = \{k_{d'} : d' \in d\}, \quad k_{(d)} = \{k_{d'} : d' \in \{1, ..., D\} \setminus d\}$$

DEFINITION S4 (CONDITIONAL $k$-INDEPENDENCE). We say that $X_i$ and $Y_j$ are $k$-independent conditional on $Z_{(d)}$ and write it as $X_i \perp_{k} Y_j | Z_{(d)}$ if for any $A \in \mathcal{A}^k$:

$$\text{pr}(X_i \in A_{x,d}, Y_j \in A_{y,d} | Z_{d} \in A_{(d)}) = \text{pr}(X_i \in A_{x,d} | Z_{d} \in A_{(d)}) \cdot \text{pr}(Y_j \in A_{y,d} | Z_{d} \in A_{(d)})$$

Or equivalently:

$$\text{pr}(X_i \in A_{x,d} | Y_j \in A_{y,d}, Z_{d} \in A_{(d)}) = \text{pr}(X_i \in A_{x,d} | Z_{d} \in A_{(d)})$$

When $k_{d'} = 0$ for some $d' \in \{1, ..., D\} \setminus d$, $\Omega_{Z_{d'}} = [0, 1]$ for those indices and hence our notation may be compacted. For example, if $k_{d'} = 0$ for all $d' \in \{1, ..., D\} \setminus d$:

$$X_i \perp_{k} Y_j | Z_{(d)} \iff X_i \perp_{k_d} Y_j$$

We next provide discretized versions of some basic results in probability:
LEMMA S3 (Contraction). For \( d \subset \{1, \ldots, D\} \) such that \( \{1, \ldots, D_x\} \subset d \), i.e. \( i = \{1, \ldots, D_x\} \), \( X_i = X \) and \( Y_{(j)} = Z_{(d)} \):

\[
\begin{align*}
X \perp_{k} Y_j \mid Z_{(d)} \\
X \perp_{k(d)} Y_{(j)}
\end{align*}
\Rightarrow X \perp_{k} Y
\]

Proof. Immediate from definition.

LEMMA S4 (Decomposition). For \( d \subset \{1, \ldots, D\} \) such that \( \{1, \ldots, D_x\} \subset d \), i.e. \( i = \{1, \ldots, D_x\} \) and \( X_i = X \):

\[
X \perp_{k} Y \Rightarrow \begin{cases} X \perp_{k} Y_j \mid Z_{(d)} \\
X \perp_{k(d)} Y_{(j)}
\end{cases}
\]

Proof. Immediate from definition.

LEMMA S5 (Weak Union). For \( d \subset \{1, \ldots, D\} \) such that \( \{1, \ldots, D_x\} \subset d \), i.e. \( i = \{1, \ldots, D_x\} \), \( X_i = X \), \( Y_j = Z_{d} \) and \( Y_{(j)} = Z_{(d)} \):

\[
X \perp_{k} Y \Rightarrow \begin{cases} X \perp_{k} Y_j \mid Z_{(d)} \\
X \perp_{k} Y_{(j)} \mid Z_{d}
\end{cases}
\]

Proof. Immediate from definition.

DEFINITION S5. For \( A_d \in \mathcal{P}^{k_d-1} \) such that \( A_d = [(l_d - 1)/2^{k_d-1}, l_d/2^{k_d-1}] \), define

\[
A_0^d = \left[ \frac{2l_d - 2}{2^{k_d}}, \frac{2l_d - 1}{2^{k_d+1}} \right] \in \mathcal{P}^{k_d}
\]

and

\[
A_1^d = \left[ \frac{2l_d - 1}{2^{k_d+1}}, \frac{2l_d}{2^{k_d+1}} \right] \in \mathcal{P}^{k_d}
\].

DEFINITION S6. Given \( k \in \mathbb{N}_0^D \) and \( i \in \{1, \ldots, D_x\} \), \( j \in \{1, \ldots, D_y\} \), let

\[
k[i, j] = \{k' \in \mathbb{N}_0^D : k'_i < k_i, k'_{D_x+j} < k_{D_x+j} \text{ and } k'_d \leq k_d \text{ for all } d \in \{1, \ldots, D\} \setminus \{i, D_x + j\} \}
\]

and

\[
\mathcal{A}^{k[i, j]} = \bigcup_{k' \in k[i, j]} \mathcal{A}^{k'}.
\]

DEFINITION S7. Given \( k \in \mathbb{N}_0^D \) and \( i \in \{1, \ldots, D_x\} \), \( j \in \{1, \ldots, D_y\} \) let

\[
[k_i, k_j](k_{<i,<j}) = \{k' \in \mathbb{N}_0^D : k'_i < k_i, k'_{D_x+j} < k_{D_x+j} \text{ and } k'_d = k_d \text{ for all } d \in \{1, \ldots, i - 1\} \cup \{D_x + 1, \ldots, D_x + j - 1\} \}
\]

and

\[
k'_d = 0 \text{ for all } d \in \{i + 1, \ldots, D_x\} \cup \{D_x + j + 1, \ldots, D\}.
\]
Accordingly, let
\[ A^{[k_i,k_j](k_{<i},k_{<j})} = \bigcup_{k' \in [k_i,k_j](k_{<i},k_{<j})} A^{k'}, \]
which denotes the totality of all cuboids in strata that are coarser than the stratum \( A^k \) along the margins \( i \) and \( j \) such that margins \( \{i + 1, \ldots, D_x\} \) and \( \{D_x + j + 1, \ldots, D\} \) are allowed any value in \([0, 1]\).

Lemma S6 ties the \( k \)-independence of the random vectors \( X \) and \( Y \) with the \((i, j)\) odds-ratios:

**LEMMA S6.** For any \( k > 0_D, \) where \( 0_D \) is a vector of \( 0 \)s of length \( D \)
\[ X \perp k\ Y \iff \theta_{ij}(A) = 1 \quad \text{for all } i \in \{1, \ldots, D_x\}, \quad j \in \{1, \ldots, D_y\}, \quad A \in A^{k'} \]
\[ \text{for all } k' \in \mathbb{N}_0^D \text{ such that } k' \leq k \text{ with } k'_i < k_i \text{ and } k'_Dx+j < k_{Dx+j}. \]

**Proof.** \( \Rightarrow: \)
Assume \( X \perp k\ Y \) for some \( k \in \mathbb{N}_0^D. \)
Let \( i \in \{1, \ldots, D_x\} \) and \( j \in \{1, \ldots, D_y\}. \)
By applying the weak union lemma twice we get that \( X_i \perp_{(k_i,k_{Dx+j})} Y_j | Z_{\{i,Dx+j\}}. \)
For \( k' \) such that \( k'_i = k_i - 1, \) \( k'_{Dx+j} = k_{Dx+j} - 1, \) \( k'_D = k_D \) for all \( d \in \{1, \ldots, D\} \setminus \{i, D_x + j\} \) and \( A \in A^{k'} \) we get that \( A_{ij}^{00}, A_{ij}^{11}, A_{ij}^{01} \in A^k. \) So \( \theta_{ij}(A) = 1 \) by the definitions of conditional independence and the \((i, j)\) odds-ratios.
By Lemma S2 we get that indeed \( \theta_{ij}(A) = 1 \) for \( A \in A^{[k_i,k_j]} \).
\( \Leftarrow: \)
First, it suffices to show that:
\[ X \perp k\ Y \iff \theta_{ij}(A) = 1 \]
\[ \text{for all } i \in \{1, 2, \ldots, D_x\}, \quad j \in \{D_x + 1, \ldots, D\}, \quad A \in A^{[k_i,k_j](k_{<i},k_{<j})} \]
Since then we may rely on the opposite direction to get that \( \theta_{ij}(A) = 0 \) for all \( A \in A^{[k_i,k_j]} \).
Examine
\[ X_i \perp_{k \{i+1, \ldots, D_x\},k_{Dx+j}} Y_j | X_{\{1, \ldots, i-1\}}, Y_{\{1, \ldots, j-1\}} \]
To see that the above is true, let \( k' \in \mathbb{N}_0^D \) such that \( k'_d = k_d \) for all \( d \in \{1, \ldots, i\} \cup \{D_x + 1, \ldots, D_x + j\} \)
and \( k'_d = 0 \) for all \( d \in \{i + 1, \ldots, D_x\} \cup \{D_x + j + 1, \ldots, D\}. \)
For a given \( A \in A^{k'} \) let \( x, y \) be a pair of univariate random variables whose joint distribution is given by \( G := F_{X_i,Y_j|x_{\{1, \ldots, i-1\}},y_{\{1, \ldots, j-1\}}} \in A_{x_{\{1, \ldots, i-1\}},y_{\{1, \ldots, j-1\}}} \in A_{x_{\{1, \ldots, i\}},y_{\{1, \ldots, j\}}}. \) By assumption:
\[ 1 = \theta_{ij}(A) = \frac{F(A_{ij}^{00}) \cdot F(A_{ij}^{11})}{F(A_{ij}^{01}) \cdot F(A_{ij}^{10})} = \frac{G(A_{ij}^{00} \times A_{Dx+j}^{00}) \cdot G(A_{ij}^{11} \times A_{Dx+j}^{11})}{G(A_{ij}^{01} \times A_{Dx+j}^{01}) \cdot G(A_{ij}^{10} \times A_{Dx+j}^{10})} \]
Since the above holds for every \( A \in A^{[k_i,k_j](k_{<i},k_{<j})} \) we may apply Theorem 2 from Ma and Mao (2019) and conclude that \( x \perp_{(k_i,k_{Dx+j})} y \) which is equivalent to \( X_i \perp_{k_{\{1, \ldots, i, D_x+1, \ldots, D_x+j\}}} Y_j | X_{\{1, \ldots, i-1\}}, Y_{\{1, \ldots, j-1\}}. \)
Next, examine:

\[
\begin{align*}
X_1 & \perp_{k(1,D_x+1)} Y_1 \\
X_1 & \perp_{k(1,D_x+1,D_x+2)} Y_2 \mid Y_1 \\
X_1 & \perp_{k(1,D_x+1,D_x+2,D_x+3)} Y_3 \mid Y_{\{1,2\}} \\
& \vdots \\
X_1 & \perp_{k(1,D_x+1,\ldots,D_x+D-1)} Y_{D_y-1} \mid Y_{\{1,\ldots,D_y-2\}} \\
X_1 & \perp_{k(1,D_x+1,\ldots,D_x+D)} Y_{D_y} \mid Y_{\{1,\ldots,D_y-1\}} \\
\end{align*}
\]

\((*)\)

\[
\begin{align*}
X_2 & \perp_{k(1,2,D_x+1)} Y_1 \mid X_1 \\
X_2 & \perp_{k(1,2,D_x+1,D_x+2)} Y_2 \mid X_1, Y_1 \\
& \vdots \\
X_2 & \perp_{k(1,2,D_x+1,\ldots,D_x+D)} Y_{D_y} \mid X_1, Y_{\{1,\ldots,D_y-1\}} \\
& \vdots \\
\end{align*}
\]

\((***)\)

\[
\begin{align*}
X_{D_x} & \perp_{k(1,\ldots,D_x,D_x+1)} Y_1 \mid X_{\{1,\ldots,D_x-1\}} \\
X_{D_x} & \perp_{k(1,\ldots,D_x,D_x+1,D_x+2)} Y_2 \mid X_{\{1,\ldots,D_x-1\}}, Y_1 \\
& \vdots \\
X_{D_x} & \perp_k Y_{D_y} \mid X_{\{1,\ldots,D_x-1\}}, Y_{\{1,\ldots,D_y-1\}} \\
\end{align*}
\]

Each of the above rows is obtained by the previous argument. Applying the contraction lemma recursively from top to bottom to each of the rows in \((*)\) shows that \(X_1 \perp_{\{1,D_x,\ldots,D_y\}} Y\). Further applying the contraction lemma to the latter result and the rows of \((***)\) shows that

\[
X_{\{1,2\}} \perp_{\{1,2,D_x,\ldots,D_y\}} Y.
\]

A similar application of the contraction lemma to the previous results and all the rows up to \((***)\) shows that \(X \perp_k Y\). \(\Box\)

**Remark SI.** For every \(i \in \{1, \ldots, D_x\}\) and \(j \in \{1, \ldots, D_y\}\), and for each of the above conditions \(X_i \perp_{k(1,\ldots,i-1,D_x+1,\ldots,D_x+j)} Y_j \mid X_{\{1,\ldots,i-1\}}, Y_{\{1,\ldots,j-1\}}\) there are \((2^{k_i} - 1)(2^{k_j} - 1)\) one degrees of freedom tests required, each repeated due to the conditioning

\[
\prod_{s=1}^{i-1} 2^{k_s} \cdot \prod_{t=D_x+1}^{D-1} 2^{k_t}
\]

times. Summing over \(j\), we get that for each \(i\), there are

\[
(2^{k_i} - 1) \cdot \prod_{s=1}^{i-1} 2^{k_s} \cdot (2^{\sum_{t=D_x+1}^{D} k_t} - 1)
\]

one degrees of freedom tests. Summing those over \(i\) we get that overall we need to perform

\[
(2^{\sum_{s=1}^{D_x} k_s} - 1)(2^{\sum_{t=D_x+1}^{D} k_t} - 1)
\]

one degrees of freedom tests.

Under \(H_1\) there are

\[
2^{\sum_{s=1}^{D_x} k_s} \cdot 2^{\sum_{t=D_x+1}^{D} k_t} - 1
\]
degrees of freedom, under $H_0$ there are

$$(2\sum_{s=1}^{D_x} k_s - 1) + (2\sum_{t=D_x+1}^{D} k_t - 1)$$

degrees of freedom, and thus we need

$$(2\sum_{s=1}^{D_x} k_s - 1)(2\sum_{t=D_x+1}^{D} k_t - 1)$$

degrees of freedom to identify a difference between the null and the alternative. It follows from the proof that we may indeed use

$$(2\sum_{s=1}^{D_x} k_s - 1)(2\sum_{t=D_x+1}^{D} k_t - 1)$$

1-degree of freedom independence tests to do so.

**Proof (Theorem 1).** The proof of Theorem 1 is immediate from Lemmas S1 and S6. \(\square\)

**Proof (Theorem 2).** Let $A$ be a cuboid in resolution $r$, $i \in \{1, \ldots, D_x\}$, $j \in \{1, \ldots, D_y\}$ and $p_{ij}(A)$ the $p$-value that is determined by the table \(\{n(A_{ij}^{00}), n(A_{ij}^{01}), n(A_{ij}^{10}), n(A_{ij}^{11})\}\). For any $r > R^*$, whether or not $A$ is selected in the MULTIFIT procedure for testing is determined by the $p$-values observed on the collection of all potential ancestral cuboids of $A$. Without loss of generality we let $R^* = 0$ to simplify notation. The general case requires trivial changes to the proof. A cuboid $A$ will be selected in MULTIFIT if there exists a sequence of nested cuboids, or a lineage, $A_0 \subset A_1 \subset \cdots \subset A_r$ of resolution $0, 1, \ldots, r$ respectively such that each $A_{k+1}$ is a child cuboid of $A_k$ in the $(i_k, j_k)$-face, and moreover, the $p$-value of the $(i_k, j_k)$-table of $A_k$ is less than the threshold $p^*$. As such, the event that a cuboid $A$ is in $C^{(r)}$ is in the $\sigma$-algebra generated by the $2 \times 2$-table counts $n(\bar{A}_{ij})$ for all $(i, j)$ pairs and all sets $\bar{A}$ that can be an ancestor cuboid of $A$ along some lineage.

Suppose the resolution-$r$ cuboid $A$ is in the $A^k$ stratum with

$$|k| = \sum_{d=1}^{D} k_d = r.$$ 

Also, let

$$r_x = \sum_{d=1}^{D_x} k_d, r_y = \sum_{d=D_x+1}^{D} k_d.$$ 

Any potential ancestor cuboid of $A$, denoted by $\bar{A}$, is the union of several sets in $A^k$, and thus the $(i, j)$-table of $\bar{A}$, $n(\bar{A}_{ij})$, for every $(i, j)$ pair, is determined exactly if we know the counts in all sets in $A^k$.

For any positive integer $\rho$, we denote the collection of all level-$\rho$ marginal partitions of $\Omega_X$ as

$$\tilde{P}_x^\rho = \left\{ P^{k_1} \times \ldots \times P^{k_{D_x}} : \sum_{d=1}^{D_x} k_d = \rho \right\}$$

and the collection of all level-$\rho$ marginal partitions of $\Omega_Y$ as

$$\tilde{P}_y^\rho = \left\{ P^{k_{D_x+1}} \times \ldots \times P^{k_D} : \sum_{d=D_x+1}^{D} k_d = \rho \right\}$$
Consider the following sequence of nested marginal partitions on \( \Omega_X \), \( \mathcal{P}^1_x \subset \mathcal{P}^2_x \subset \cdots \subset \mathcal{P}^r_x \subset \mathcal{P}^{r+1}_x \subset \cdots \subset \mathcal{P}^{r+D_x}_x \), where \( \mathcal{P}^i_x \subset \mathcal{P}^{i+1}_x + \mathcal{P}^{i+2}_x + \cdots + \mathcal{P}^{r+D_x}_x \), such that we first divide \( \Omega_{X_1} \) \( k_1 \) times to get \( \mathcal{P}^{k_1}_x \), then divide \( \Omega_{X_{k_2}} \) \( k_2 \) times to get \( \mathcal{P}^{k_{k_1}+1}_x \), \( \mathcal{P}^{k_{k_1}+2}_x \), and so on until dividing \( \Omega_{X_{k_D}} \) \( k_D \) times to get \( \mathcal{P}^{k_{k_{D-1}}+1}_x \), \( \mathcal{P}^{k_{k_{D-1}}+2}_x \), and so on. Then divide \( \Omega_{X_i} \) once to get \( \mathcal{P}^{r-x+1}_x \), and finally divide each of the other \( D_x - 1 \) dimensions once in any order to get \( \mathcal{P}^{r+2}_x \), \( \mathcal{P}^{r+D_x}_x \).

In exactly the same manner, we can construct a sequence of nested marginal partitions on \( \Omega_Y \), \( \mathcal{P}^1_y \subset \mathcal{P}^2_y \subset \cdots \subset \mathcal{P}^r_y \subset \mathcal{P}^{r+1}_y \subset \cdots \subset \mathcal{P}^{r+D_y}_y \), such that we first divide \( \Omega_{Y_1} \) \( k_{D_y} \) times to get \( \mathcal{P}^{k_{D_y}+1}_y \), \( \mathcal{P}^{k_{D_y}+2}_y \), and so on until dividing \( \Omega_{Y_{k_D}} \) \( k_D \) times to get \( \mathcal{P}^{r-y-k_D+1}_y \), \( \mathcal{P}^{r-y-k_D+2}_y \), and so on. Then divide \( \Omega_{Y_i} \) once to get \( \mathcal{P}^{r-y+1}_y \), and finally divide each of the other \( D_y - 1 \) dimensions once in any order to get \( \mathcal{P}^{r+1}_y \), \( \mathcal{P}^{r+D_y}_y \).

Under these two marginal partition sequences, we have \( A \in \mathcal{P}^{r-x+1}_x \times \mathcal{P}^{r-y+1}_y = A^k \), whereas the four child cuboids of \( A \) with respect to the \((i, j)\)-face are in the two strata \( \mathcal{P}^{r-x+1}_x \times \mathcal{P}^{r-y+1}_y \) and \( \mathcal{P}^{r-x+1}_x \times \mathcal{P}^{r-y+1}_y \). Moreover, any \((i, j)\)-face of any ancestral cuboid of \( A \) are formed by unions of sets that are not in the strata \( \mathcal{P}^{r-x+i}_x \times \mathcal{P}^{r-y+j}_y \) for \( i = 1, 2, \ldots, D_x \) and \( j = 1, 2, \ldots, D_y \).

Now, by Theorem 3 in Ma and Mao (2019), conditional on the \( X \) and \( Y \) marginal values of the observations, the counts in any \((i, j)\)-table of \( A \) are determined from the central hypergeometric distribution given the row and column totals under the null hypothesis of independence whether or not one conditions on the event that \( A \) is selected for testing in the MULTIFIT procedure. This completes the proof.

Proof (Corollary 1). This corollary follows immediately since the \( p \)-value on the \((i, j)\)-table of a cuboid \( A \) is determined from the central hypergeometric distribution given the row and column totals of that table, which due to Theorem 2, is the actual sampling distribution of the table given the row and column totals under the null hypothesis of independence whether or not one conditions on the event that \( A \) is selected for testing in the MULTIFIT procedure.

Proof (Theorem 3). We adopt here a similar strategy to what was used in the proof of Theorem 5 in Ma and Mao (2019). We recall that the null distribution of the \( 2 \times 2 \) contingency table on a cuboid \( A \) arising from an i.i.d. sample conditional on its margins is the central hypergeometric distribution, as discussed in Section 2.3, and the same table when the null does not hold is distributed according to the noncentral hypergeometric distribution. Zhou (2019) developed a normal approximation for the noncentral hypergeometric distribution and accompanying results on which we rely.

Under the conditions in the theorem’s statement, we can take \( R_{\text{max}} = R^* \), that is, exhaustive testing. Then according to our Strategy 1 for multiple testing adjustment, or the holistic approach with Bonferroni’s correction

\[
\alpha_{ij}(A) = \alpha / \left( \sum_{\rho = 0}^{R_{\text{max}}} D_x \cdot D_y \cdot 2^\rho \cdot \left( \rho + D - 1 \right) \right)
\]

is the table-specific threshold. According to our Strategy 2 for multiple testing adjustment, or the resolution-specific approach with exhaustive testing up to resolution \( R_{\text{max}} \) and utilizing Bonferroni’s correction.
roni’s correction within resolution as well as between resolutions

\[ \alpha_{ij}(A) = \alpha_j \left( D_x \cdot D_y \cdot 2^{r} \cdot \left( \frac{r + D - 1}{D - 1} \right), R_{\text{max}} \right) \]

is the table-specific threshold. The selection of Bonferroni’s correction here ensures that the proof is valid for any correction that is less conservative, e.g. Holm’s correction that we use in our implementation.

We start by proving the theorem in a simplified case where we assume that \( R^* = R_{\text{max}} \) is fixed but large enough such that \( r \leq R^* \). Under both instances \( \alpha_{ij}(A) \) is constant when \( R^* \) is constant. Let then \( \theta_{ij}(A) = \theta \). Without loss of generality, let us assume that \( \theta > 1 \).

By Theorem 2.2 in Zhou (2019), we have that given \( n(A_{ij}^{0}), n(A_{ij}^{0}), n(A) \),

\[ Z_{n,\theta,i,j}(A) = \frac{n(A_{ij}^{0}) - E_{\theta}\{n(A_{ij}^{0}) \mid n(A_{ij}^{0}), n(A_{ij}^{0}), n(A)\}}{\text{var}_{\theta}^{1/2}\{n(A_{ij}^{0}) \mid n(A_{ij}^{0}), n(A_{ij}^{0}), n(A)\}} \rightarrow L N(0, 1). \]

\( p_{ij}(A) \), the \( p \)-value for the table determined by \( n(A_{ij}^{0}) \) when the margins \( \{n(A_{ij}^{0}), n(A_{ij}^{0}), n(A)\} \) are given, is computed for the two sided version of Fisher’s exact test by summing the probabilities of all tables that are more extreme than the given one. That is, those with a smaller probability of occurrence compared to the table for which the \( p \)-value is computed according to the central hypergeometric distribution.

We can utilize the form \( Z_{n,\theta,i,j} \) to write the probability of rejection for the test on the \((i, j)\)-table of \( A \). If \( n(A_{ij}^{0}) \) is less than or equal to the mode of the hypergeometric distribution with the parameters \( \{n(A_{ij}^{0}), n(A_{ij}^{0}), n(A)\} \) we have

\[ \text{pr}\{p_{ij}(A) < \alpha_{ij}(A) \mid \theta_{ij}(A) = \theta, n_{r_x,0}, n_{0,r_y}\} \]

\[ > \text{pr} \left[ Z_{n,1,i,j}(A) > F_{A,n}^{-1}\{\alpha_{ij}(A)\} \mid \theta_{ij}(A) = \theta, n_{r_x,0}, n_{0,r_y}\right] \]

where \( F_{A,n} \) denotes the exact cdf of \( Z_{n,1,i,j} \) given the marginal totals. Else, if \( n(A_{ij}^{0}) \) is greater than the mode of the hypergeometric distribution with the parameters \( \{n(A_{ij}^{0}), n(A_{ij}^{0}), n(A)\} \)

\[ \text{pr}\{p_{ij}(A) < \alpha_{ij}(A) \mid \theta_{ij}(A) = \theta, n_{r_x,0}, n_{0,r_y}\} \]

\[ > \text{pr} \left[ Z_{n,1,i,j}(A) > F_{A,n}^{-1}\{1 - \alpha_{ij}(A)\} \mid \theta_{ij}(A) = \theta, n_{r_x,0}, n_{0,r_y}\right]. \]

Without loss of generality we continue to work with the second case.

Assume now that indeed \( \theta_{ij}(A) = \theta \neq 1 \). Then

\[ \lim_{n \to \infty} \text{pr} \left[ Z_{n,1,i,j}(A) > F_{A,n}^{-1}\{1 - \alpha_{ij}(A)\} \mid \theta_{ij}(A) = \theta, n_{r_x,0}, n_{0,r_y}\right] \]

\[ = \lim_{n \to \infty} \text{pr} \left[ Z_{n,\theta,i,j}(A) > c_n F_{A,n}^{-1}\{1 - \alpha_{ij}(A)\} - c_n d_n \mid \theta_{ij}(A) = \theta, n_{r_x,0}, n_{0,r_y}\right] \]

where

\[ c_n = \frac{\text{var}_{\theta}^{1/2}\{n(A_{ij}^{0}) \mid n(A_{ij}^{0}), n(A_{ij}^{0}), n(A)\}}{\text{var}_{\theta}^{1/2}\{n(A_{ij}^{0}) \mid n(A_{ij}^{0}), n(A_{ij}^{0}), n(A)\}} \]

and

\[ d_n = \frac{E_{\theta}\{n(A_{ij}^{0}) \mid n(A_{ij}^{0}), n(A_{ij}^{0}), n(A)\} - E_{1}\{n(A_{ij}^{0}) \mid n(A_{ij}^{0}), n(A_{ij}^{0}), n(A)\}}{\text{var}_{\theta}^{1/2}\{n(A_{ij}^{0}) \mid n(A_{ij}^{0}), n(A_{ij}^{0}), n(A)\}}. \]
By Corollary 2.1 in Zhou (2019) we have $1/\sqrt{\max(\theta, \theta^{-1})} \leq c_n \leq 1/\sqrt{\min(\theta, \theta^{-1})}$ for all $n$ and $d_n \sim \sqrt{n}$ with $F^\infty$ probability 1.

By the above normal approximation $F_{A,n}^{-1}\{1 - \alpha_{ij}(A)\} \to \Phi^{-1}\{1 - \alpha_{ij}(A)\}$. Therefore $c_n\Phi^{-1}\{1 - \alpha_{ij}(A)\} - c.nd_n \to -\infty$ with $F^\infty$ probability 1, and so with $F^\infty$ probability 1,

$$\lim_{n\to\infty} \Pr\{p_{ij}(A) < \alpha_{ij}(A) \mid n_{r_i,0}, n_{0,r_j}\} = 1.$$  

This completes the proof for the case where $R^*$ is constant since the null of independence is rejected whenever $p_{ij}(A) < \alpha_{ij}(A)$ for some $A, i$ and $j$.

Assume now that $R_{\max} = R^* = o(\log n)$. Under both the holistic and resolution-specific strategies for multiple testing we have $\sqrt{n} \cdot \alpha_{ij,n}(A) \to \infty$ as $n \to \infty$.

Since $c.nd_n = O(\sqrt{n})$ with $F^\infty$ probability 1, we need to show that $c_nF_{A,n}^{-1}\{1 - \alpha_{ij}(A)\} = o(\sqrt{n})$ with $F^\infty$ probability 1 in order to establish consistency. Rewrite:

$$c_nF_{A,n}^{-1}\{1 - \alpha_{ij,n}(A)\} = c_n\left[F_{A,n}^{-1}\{1 - \alpha_{ij,n}(A)\} - \Phi^{-1}\{1 - \alpha_{ij,n}(A)\}\right] + c_n\Phi^{-1}\{1 - \alpha_{ij,n}(A)\}$$

Examine $c_n[F_{A,n}^{-1}\{1 - \alpha_{ij,n}(A)\} - \Phi^{-1}\{1 - \alpha_{ij,n}(A)\}]$: by Theorem 2.3 of Zhou (2019), we have with $F^\infty$ probability 1,

$$\Pr\{\Phi[F_{A,n}^{-1}\{1 - \alpha_{ij,n}(A)\}] - \{1 - \alpha_{ij,n}(A)\} < \gamma/\sqrt{n}\}$$

for some positive constant $\gamma$ and large enough $n$. Therefore

$$\Phi^{-1}\{1 - \alpha_{ij,n}(A) - \gamma/\sqrt{n}\} < F_{A,n}^{-1}\{1 - \alpha_{ij,n}(A)\} < \Phi^{-1}\{1 - \alpha_{ij,n}(A) + \gamma/\sqrt{n}\}.$$  

Since $1 - \Phi(x) \approx e^{-x^2/2}/x$ as $x \to \infty$ and $\sqrt{n} \cdot \alpha_{ij,n}(A) \to \infty$ as $n \to \infty$, we have

$$\Phi^{-1}\{1 - \alpha_{ij,n}(A) - \gamma/\sqrt{n}\} - \Phi^{-1}\{1 - \alpha_{ij,n}(A)\} \to 0$$  

and

$$\Phi^{-1}\{1 - \alpha_{ij,n}(A) + \gamma/\sqrt{n}\} - \Phi^{-1}\{1 - \alpha_{ij,n}(A)\} \to 0$$

Hence,

$$|F_{A,n}^{-1}\{1 - \alpha_{ij,n}(A)\} - \Phi^{-1}\{1 - \alpha_{ij,n}(A)\}| \to 0.$$  

Examine $c_n\Phi^{-1}\{1 - \alpha_{ij,n}(A)\}$: since $1 - \Phi(x) \approx e^{-x^2/2}/x$ as $x \to \infty$ and $\sqrt{n} \cdot \alpha_{ij,n}(A) \to \infty$ as $n \to \infty$, we have $\Phi^{-1}\{1 - \alpha_{ij,n}(A)\} = o(\sqrt{n})$. Therefore, we have with $F^\infty$ probability 1 that $c_nF_{A,n}^{-1}\{1 - \alpha_{ij}(A)\} = o(\sqrt{n})$. That is, we have with probability 1 $p_{ij}(A) < \alpha_{ij}(A)$ for some $A, i$ and $j$ and therefore the null of independence is rejected with probability 1.  

$\square$
2. **Pseudo-code for the MultiFIT Procedure**

*Algorithm S1.* MultiFIT procedure for testing multivariate independence.

Let \( C(r) \) be the collection of all cuboids of resolution \( r \) for \( r = 0, 1, \ldots, R^* \), and let \( C(r) = \emptyset \) for \( r = R^* + 1, \ldots, R_{\text{max}} \).

For \( r = 0 \) to \( r = R_{\text{max}} \)

- For each \( A \in C(r) \)
  - For \( i = 1 \) to \( i = D_x \)
    - For \( j = 1 \) to \( j = D_y \)
      - Apply Fisher’s exact test on the \((i,j)\)-table of \( A \) and record the \( p \)-value
      - If \( R^* \leq r < R_{\text{max}} \)
        - If the \((i,j)\)-table of \( A \) has a \( p \)-value smaller than a threshold \( p^* \)
          - Add the four half cuboids of \( A \) into \( C(r+1) \)

Output the list of \( p \)-values

Apply multiple testing adjustment on these \( p \)-values.
### 3. Scenarios for the Power Study

| Scenario     | # of Data Points | Max Res | Simulation Setting                                           |
|--------------|------------------|---------|-------------------------------------------------------------|
| Sine         | 300              | 4       | \(X_1 = Z, Y_1 = Z', X_2 = U,\)                           |
|              |                   |         | \(Y_2 = \sin(5\pi X_2) + 4\epsilon\)                    |
| Circular     | 300              | 4       | \(X_1 = Z, Y_1 = Z', \theta \sim \text{Uniform}(-\pi, \pi)\) |
|              |                   |         | \(X_2 = \cos(\theta) + \epsilon, Y_2 = \sin(\theta) + \epsilon'\) |
| Checkerboard | 1500             | 5       | \(W \sim \text{Multi-Bern}(\{1, 2, 3, 4, 5\}, (1/5, 1/5, 1/5, 1/5, 1/5))\) |
|              |                   |         | \(V_1 \sim \text{Multi-Bern}(\{1, 3, 5\}, (1/3, 1/3, 1/3))\) |
|              |                   |         | \(V_2 \sim \text{Multi-Bern}(\{2, 4\}, (1/2, 1/2))\)      |
|              |                   |         | \(X_1 = Z, Y_1 = Z', X_2 = W + \epsilon,\)              |
|              |                   |         | \(Y_2 = \begin{cases} V_1 + \epsilon', & \text{if } W \text{ is odd} \\ V_2 + \epsilon', & \text{if } W \text{ is even} \end{cases} \) |
| Linear       | 300              | 4       | \(X_1 = Z, Y_1 = Z', X_2 = U,\)                           |
|              |                   |         | \(Y_2 = X_2 + 3\epsilon\)                                 |
| Parabolic    | 300              | 4       | \(X_1 = Z, Y_1 = Z', X_2 = U,\)                           |
|              |                   |         | \(Y_2 = (X_2 - 0.5)^2 + 0.75\epsilon\)                    |
| Local        | 1000             | 6       | \(X_1 = Z, Y_1 = Z', X_2 = Z''\)                         |
|              |                   |         | \(Y_2 = \begin{cases} X_2 + \epsilon/6 & \text{if } 0 < X_2, Z'' < 0.7 \\ Z'', & \text{otherwise} \end{cases} \) |

Six simulation scenarios. In all cases, \(Z, Z', Z'', Z''\) are i.i.d \(N(0, 1)\). At each noise level \(l = 1, 2, ..., 20\), \(\epsilon, \epsilon'\) and \(\epsilon''\) are i.i.d \(N\{0, (l/20)^2\}\), and \(U \sim \text{Uniform}(0, 1)\). The maximal resolution is the algorithm’s default: \(\lfloor \log_2(n/10) \rfloor\) where \(n\) is the number of data points.
## Table S2. ‘Spread’ Simulation Scenarios (part 1)

| Scenario     | # of Data Points | Max Res | Simulation Setting |
|--------------|------------------|---------|-------------------|
| Sine         | 300              | 4       | $X_1 = U, X_2 = U$ |
|              |                  |         | $Y_1 = \sin(5\pi X_1) + \cos(5\pi X_2) + 4\epsilon$ |
|              |                  |         | $Y_2 = -\sin(5\pi X_2) + \cos(5\pi X_1) + 4\epsilon'$ |
| Circular     | 300              | 4       | $Z_1, Z_2 \sim \text{Beta}(0.9, 0.27)$ |
|              |                  |         | $b_1, b_2, b_3, b_4 \sim \text{Multi-Bern}(\{-1, 1\}, (0.5, 0.5))$ |
|              |                  |         | $X_1 = b_1 Z_1 + \epsilon, X_2 = b_2 Z_2 + \epsilon'$ |
|              |                  |         | $Y_1 = b_3 (1 - (0.2b_1 Z_1 + 0.8b_2 Z_2)^2)^{1/2} + \epsilon''$ |
|              |                  |         | $Y_2 = b_4 (1 - (0.8b_1 Z_1 + 0.2b_2 Z_2)^2)^{1/2} + \epsilon'''$ |
| Checkerboard | 1500             | 5       | $W, W' \sim \text{Multi-Bern}(\{1, 2, 3, 4, 5\}, (0.2, 0.2, 0.2, 0.2, 0.2))$ |
|              |                  |         | $X_1 = W + 0.75^2 \epsilon, X_2 = W' + 0.75^2 \epsilon'$ |
|              |                  |         | $m_1 \sim \text{Multi-Bern}(\{1, 2\}, (0.2, 0.8))$ |
|              |                  |         | $m_2 \sim \text{Multi-Bern}(\{1, 2\}, (0.8, 0.2))$ |
|              |                  |         | $V_1, V'_1 \sim \text{Multi-Bern}(\{1, 3, 5\}, (1/3, 1/3, 1/3))$ |
|              |                  |         | $V_2, V'_2 \sim \text{Multi-Bern}(\{2, 4\}, (0.5, 0.5))$ |
|              |                  |         | $Y_1 = \begin{cases} V'_1 + 0.75^2 \epsilon'' & \text{if } x_{m_1} \text{ is odd} \\ V_2 + 0.75^2 \epsilon'' & \text{if } x_{m_1} \text{ is even} \end{cases}$ |
|              |                  |         | $Y_2 = \begin{cases} V'_2 + 0.75^2 \epsilon''' & \text{if } x_{m_2} \text{ is odd} \\ V'_2 + 0.75^2 \epsilon''' & \text{if } x_{m_2} \text{ is even} \end{cases}$ |

$\epsilon, \epsilon', \epsilon''$ and $\epsilon'''$ are i.i.d $\mathcal{N}(0, (l/20)^2)$, and $U \sim \text{Uniform}(0, 1)$. The maximal resolution is the algorithm’s default: $\lfloor \log_2(n/10) \rfloor$ where $n$ is the number of data points.
Table S3. ‘Spread’ Simulation Scenarios (part 2)

| Scenario   | # of Data Points | Max Res | Simulation Setting |
|------------|------------------|---------|--------------------|
| Linear     | 300              | 4       | $X_1, X_2 \sim U(0, 1)$  
                        $Y_1 = X_1 - 2X_2 + 6\epsilon, Y_2 = -X_1 + X_2 + 6\epsilon'$ |
| Parabolic  | 300              | 4       | $X_1, X_2 \sim \text{Uniform}(0, 1)$  
                        $Y_1 = (X_1 - 0.5)^2 - (X_2 - 0.5)^2 + 0.75\epsilon$  
                        $Y_2 = (X_2 - 0.5)^2 - (X_1 - 0.5)^2 + 0.75\epsilon'$ |
| Local      | 1000             | 6       | A spread linear signal scaled to (0,0.7) for all margins:  
                        $x_1, x_2 \sim U(0, 0.7)$  
                        $y_1 = 0.7(x_1 - x_2 + 1.4)/2.1 + 1/6 \cdot \epsilon$  
                        $y_2 = 0.7(-x_1 + x_2 + 0.7)/1.4 + 1/6 \cdot \epsilon'$  
                        Embedded within a small portion of the space:  
                        $X_1 = Z, X_2 = Z'$  
                        $Y_1 = \begin{cases} y_1 & \text{if } 0 < X_1, Z'' < 0.7 \\ Z'', \text{ otherwise} \end{cases}$  
                        $Y_2 = \begin{cases} y_3 & \text{if } 0 < X_2, Z'''' < 0.7 \\ Z'', \text{ otherwise} \end{cases}$ |

$Z, Z', Z''', Z''''$ are i.i.d $N(0, 1)$. At each noise level $l = 1, 2, \ldots, 20$, $\epsilon$, $\epsilon'$ and $\epsilon''$ are i.i.d $N\{0, (l/20)^2\}$, and $U \sim \text{Uniform}(0, 1)$. The maximal resolution is the algorithm’s default: $\lfloor \log_2(n/10) \rfloor$ where $n$ is the number of data points.
4. Numerical Validation of Level Control through Simulations

To demonstrate that the multiscale Fisher’s independence test properly controls the level we executed 500 simulations with the default tuning parameters for various sample sizes. The underlying data \( \{X_1\}, \{X_2\}, \{Y_1\} \) and \( \{Y_2\} \) are drawn independently from a standard normal distribution for \( i \in \{1, \ldots, n\} \) with \( n \in \{100, 200, \ldots, 2000\} \). Figure S4 shows the estimated level for the multiscale Fisher’s independence test with different variations for the independence tests on each table and multiple testing adjustment options on Fisher’s exact test with mid-\( p \) corrected \( p \)-values. See Agresti and Gottard (2007) for a discussion on the mid-\( p \) correction.

The results confirm the theoretical guarantees that the level can be controlled at any given level \( \alpha \). In fact, the procedure appears to be a bit conservative in controlling the level. Note, although \( X \) and \( Y \) are independent under the null hypothesis, the dependency structure between the different margins of \( X \) is arbitrary, as well as the dependency structure between the different margins of \( Y \). Therefore, this simulation is not exhaustive. However, we repeated the estimation of the level under various dependency structures for the margins, and the results are consistent with these that are shown here.
Fig. S4. Estimated level for MultiFIT versus sample size. Fourteen different variants of the method are examined: the holistic strategy with Holm’s correction (black solid), the holistic strategy with Holm’s and mid-p correction (black dashes), the resolution specific strategy with Holm’s correction (black dotted), the resolution specific strategy with Holm’s and mid-p correction (black dot-dash), the early stopping strategy with Holm’s correction (black long-dashes), the early stopping strategy with Holm’s and mid-p correction (black two-dash), the holistic strategy with the approximation keeps up to 10 most significant $p$-values at each resolution and Holm’s correction (thick grey solid), the holistic strategy with the approximation keeps up to 10 most significant $p$-values at each resolution and Holm’s and mid-p correction (thick grey dashes), the resolution specific strategy with the approximation keeps up to 10 most significant $p$-values at each resolution and Holm’s correction (grey solid), the resolution specific strategy with the approximation keeps up to 10 most significant $p$-values at each resolution and Holm’s and mid-p correction (grey dashes), the holistic strategy with the approximation keeps up to 100 most significant $p$-values at each resolution and Holm’s correction (grey dotted), the holistic strategy with the approximation keeps up to 100 most significant $p$-values at each resolution and Holm’s and mid-p correction (grey dot-dash), the resolution specific strategy with the approximation keeps up to 100 most significant $p$-values at each resolution and Holm’s correction (grey long dashes), the resolution specific strategy with the approximation keeps up to 100 most significant $p$-values at each resolution and Holm’s and mid-p correction (grey two dashes).
5. **Scaling: Comparison of Scenarios for the Multiscale Fisher’s Independence Test**

Fig. S5. Computational scalability: run-time vs. sample size for the six simulation scenarios from Table S1 when fitted with the multiscale Fisher’s independence test in different dimensionalities with $D_x = D_y = d$, $R^* = 1$ and $l = 3$. ‘Null’ scenario (solid), ‘parabolic’ (dashes), ‘sine’ (dotted), ‘checkerboard’ (dot-dash), ‘circle’ (long dashes) and ‘local’ (two-dash). In all cases the ‘linear’ scenario requires the most computations and the ‘null’ scenario the least.
6. **Scaling: Comparison of the Multiscale Fisher’s Independence Test to Other Near Linear Time Tests**

Figure S6 shows a comparison between the multiscale Fisher’s independence test and the kernel partial correlation methods from Deb et al. (2020).

![Figure S6](image_url)

Fig. S6. Computational scalability: run-time vs. sample size. Left panel: ‘linear’ scenario; Right panel: ‘null’ scenario. Two variants of MultIFIT were tested: full method on data with $D_x = D_y = 2$ (black solid) and on data with $D_x = D_y = 10$ (dark grey long dashes), MultIFIT with early stopping on data with $D_x = D_y = 2$ (black dashes) and on data with $D_x = D_y = 10$ (black dotted); MultIFIT was run with $R^* = 1$ and $p^* = \{D_x D_y \log_2(n)\}^{-1}$. Two variants of the kernel partial correlation method were computed: the unconditional version of graph-based kernel partial correlation using directed K-NN graph on data with $D_x = D_y = 2$ (grey dashes) and on data with $D_x = D_y = 10$ (grey solid); the unconditional version of graph-based kernel partial correlation using directed K-NN graph on data with $D_x = D_y = 2$ (grey dot-dash) and on data with $D_x = D_y = 10$ (grey two-dash). The multiscale Fisher’s independence test does not require permutations. The kernel partial correlation method requires permutations for level control and the reported time is for a single permutation.
7. Sensitivity analysis for the parameters $p^*$ and $R^*$ of the multiscale Fisher’s independence test

In Fig. S7 and Fig. S8 we demonstrate the effect of our main tuning parameters, $p^*$ and $R^*$, on the estimated power: $R^* = 1$ (solid), $R^* = 2$ (dashes), $R^* = 3$ (dotted), $R^* = 4$ (dot-dash). We evaluate the power for the simulation scenarios that are detailed in Table S1 over the grid $p^* \times R^*$ where $p^* = \{0.005, 0.01, 0.05, 0.1, 0.5\}$ and $R^* = \{1, 2, 3, 4\}$. In general, higher $p^*$ and higher $R^*$ values entail more testing. Under some scenarios, where the dependency structure is ‘more local’, higher values are needed to ensure better power. When the dependency structures are ‘more global’ we see that the selection of the tuning parameters makes little difference to the estimated power.
Fig. S7. Power versus noise level for different specifications of MultiFIT. Estimated power at 20 noise levels for the multiscale Fisher’s independence test with different $p^*$ and $R^*$ values under the six scenarios from Table S1.
Fig. S8. Power versus noise level for different specifications of MultiFIT. $p^* = 0.005$ (solid), $p^* = 0.01$ (dashes), $p^* = 0.05$ (dotted), $p^* = 0.1$ (dot-dash), $p^* = 0.5$ (long dashes). Estimated power at 20 noise levels for the multiscale Fisher’s independence test with different $p^*$ and $R^*$ values under the six scenarios from Table S1.
8. Software

For the MultiFIT procedure we used our R package MultiFit on the comprehensive R archive network, or CRAN. For the Heller-Heller-Gorfine test (Heller et al., 2013) we used the HHG package on CRAN. For distance covariance (Székely and Rizzo, 2009) we used the energy package on CRAN. For Brownian distance covariance (Pfister et al., 2018) we used the dHSIC package on CRAN.

REFERENCES

Agresti, A. and Gottard, A. (2007). Nonconservative exact small-sample inference for discrete data. *Computational Statistics & Data Analysis*, 51(12):6447 – 6458.

Deb, N., Ghosal, P., and Sen, B. (2020). Measuring association on topological spaces using kernels and geometric graphs.

Heller, R., Heller, Y., and Gorfine, M. (2013). A consistent multivariate test of association based on ranks of distances. *Biometrika*, 100:503–510.

Ma, L. and Mao, J. (2019). Fisher exact scanning for dependency. *Journal of the American Statistical Association*, 114(525):245–258.

Pfister, N., Bühlmann, P., Schölkopf, B., and Peters, J. (2018). Kernel-based tests for joint independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):5–31.

Székely, G. J. and Rizzo, M. L. (2009). Brownian distance covariance. *The Annals of Applied Statistics*, 3(4):1236–1265.

Zhou, Q. (2019). Asymptotics of multivariate contingency tables with fixed marginals. *Journal of Statistical Planning and Inference*, 198:165–170.