A general asymptotic framework for distribution-free graph-based two-sample tests

Bhaswar B. Bhattacharya

University of Pennsylvania, Philadelphia, USA

[Received August 2017. Final revision February 2019]

Summary. Testing equality of two multivariate distributions is a classical problem for which many non-parametric tests have been proposed over the years. Most of the popular two-sample tests, which are asymptotically distribution free, are based either on geometric graphs constructed by using interpoint distances between the observations (multivariate generalizations of the Wald–Wolfowitz runs test) or on multivariate data depth (generalizations of the Mann–Whitney rank test). The paper introduces a general notion of distribution-free graph-based two-sample tests and provides a unified framework for analysing and comparing their asymptotic properties. The asymptotic (Pitman) efficiency of a general graph-based test is derived, which includes tests based on geometric graphs, such as the Friedman–Rafsky test, the test based on the K-nearest-neighbour graph, the cross-match test and the generalized edge count test, as well as tests based on multivariate depth functions (the Liu–Singh rank sum statistic). The results show how the combinatorial properties of the underlying graph affect the performance of the associated two-sample test and can be used to validate and decide which tests to use in practice. Applications of the results are illustrated both on synthetic and on real data sets.

Keywords: Asymptotic efficiency; Distribution-free tests; Minimum spanning tree; Nearest neighbour graphs; Two-sample problem

1. Introduction

Let $F$ and $G$ be two continuous distribution functions in $\mathbb{R}^d$. Given independent and identically distributed (IID) samples

$$\mathcal{X}_{N_1} = \{X_1, X_2, \ldots, X_{N_1}\},$$
$$\mathcal{Y}_{N_2} = \{Y_1, Y_2, \ldots, Y_{N_2}\}$$

(1.1)

from two unknown distributions $F$ and $G$ respectively, the two-sample problem is to distinguish the hypotheses

$$H_0 : F = G \quad \text{versus} \quad H_1 : F \neq G.$$  

(1.2)

More precisely, $H_0$ is the collection of all distributions of mutually independent IID observations with sample size $N_1 + N_2$ from a distribution in $\mathbb{R}^d$, and $H_1$ is the collection of all distributions of mutually independent IID observations with sample size $N_1$ from some distribution $F$ in $\mathbb{R}^d$, and IID observations with sample size $N_2$ from some other distribution $G \neq F$ in $\mathbb{R}^d$.

There are many multivariate two-sample testing procedures, ranging from tests for parametric hypotheses such as Hotelling’s $T^2$-test, and the likelihood ratio test, to more general non-parametric procedures (Aslan and Zech, 2005; Baringhaus and Franz, 2004; Bickel, 1969;...
Biswas et al., 2014; Chwialkowski et al., 2015; Friedman and Rafsky, 1979; Gretton et al., 2012; Hall and Tajvidi, 2002; Henze, 1988; Rousson, 2002; Rosenbaum, 2005; Schilling, 1986; Weiss, 1960). In this paper, we consider multivariate two-sample tests, which are 

asymptotically distribution free, i.e. tests for which the asymptotic null distribution does not depend on the underlying (unknown) distribution of the data. As a result, these tests can be directly implemented as an asymptotically level \( \alpha \) test, making them practically convenient.

For univariate data, there are several celebrated non-parametric distribution-free tests such as the Wald–Wolfowitz (WW) runs test (Wald and Wolfowitz, 1940) and the Mann–Whitney rank test (Mann and Whitney, 1947) (see Lehmann (1975) for more on these tests). Many multivariate generalizations of these tests, which are asymptotically distribution free, have been proposed. Most of these tests can be broadly classified into two categories.

(a) Tests based on geometric graphs: these tests are constructed by using the interpoint distances between the observations. This includes the test based on the Euclidean minimal spanning tree by Friedman and Rafsky (FR) (1979) (generalization of the Wald and Wolfowitz (1940) runs test to higher dimensions) and tests based on nearest neighbour graphs (Henze, 1988; Schilling, 1986). This also includes Rosenbaum’s (2005) test-based minimum non-bipartite matching and the test based on the Hamiltonian path by Biswas et al. (2014) (both of which are distribution free in finite samples), and the recent tests of Chen and Friedman (CF) (2017). Refer to Maa et al. (1996) for theoretical motivations for using tests based on interpoint distances.

(b) Tests based on depth functions: the Liu and Singh (1993) rank sum tests are a class of multivariate two-sample tests that generalize the Mann–Whitney rank test by using the notion of data depth. This includes tests based on halfspace depth (HD) (Tukey, 1975) and simplicial depth (Liu, 1990, 1992), among others. For other generalizations of the Mann–Whitney test, refer to the survey by Oja (2010) and the references therein.

In this paper, we provide a general framework of graph-based two-sample tests, which includes all the tests that were discussed above. We begin with a few definitions: a subset \( S \subseteq \mathbb{R}^d \) is locally finite if \( S \cap C \) is finite, for all compact subsets \( C \subseteq \mathbb{R}^d \). A locally finite set \( S \subseteq \mathbb{R}^d \) is nice (with respect to a metric \( \rho \) in \( \mathbb{R}^d \) ) if all interpoint distances among the elements of \( S \) are distinct. Note that if, for example, \( S \) is a set of \( N \) IID points \( W_1, W_2, \ldots, W_N \) from some continuous distribution \( F \), then the distribution of \( W_1 - W_2 \), and hence \( \| W_1 - W_2 \| \), where \( \| \cdot \| \) denotes the Euclidean norm, does not have any point mass, and \( S \) is nice. For the same reason, in a set of \( N \) IID points from a continuous distribution ties occur with zero probability.

A graph functional \( \mathcal{G} \) in \( \mathbb{R}^d \) defines a graph for all finite subsets of \( \mathbb{R}^d \), i.e., given \( S \subseteq \mathbb{R}^d \) finite, \( \mathcal{G}(S) \) is a graph with vertex set \( S \). A graph functional is said to be undirected or directed if the graph \( \mathcal{G}(S) \) is respectively an undirected or directed graph with vertex set \( S \). We assume that \( \mathcal{G}(S) \) has no self-loops and multiple edges, i.e. no edge is repeated more than once in the undirected case, and no edge in the same direction is repeated more than once in the directed case. The set of edges in the graph \( \mathcal{G}(S) \) will be denoted by \( E(\mathcal{G}(S)) \). The cardinality of a finite set \( A \) is denoted by \( |A| \).

**Definition 1.** Let \( \mathcal{X}_{N_1} \) and \( \mathcal{Y}_{N_2} \) be IID samples of size \( N_1 \) and \( N_2 \) from densities \( f \) and \( g \) respectively, as in expression (1.1). The two-sample test statistic based on the graph functional \( \mathcal{G} \) is defined as

\[
T(\mathcal{G}(\mathcal{X}_{N_1} \cup \mathcal{Y}_{N_2})) = \frac{\sum_{i=1}^{N_1} \sum_{j=1}^{N_2} I[(X_i, Y_j) \in E(\mathcal{G}(\mathcal{X}_{N_1} \cup \mathcal{Y}_{N_2}))]}{|E(\mathcal{G}(\mathcal{X}_{N_1} \cup \mathcal{Y}_{N_2}))|}.
\] (1.3)
Denote by \( N = N_1 + N_2 \) and \( Z_N = \mathcal{X}_{N_1} \cup \mathcal{Y}_{N_2} = \{Z_1, Z_2, \ldots, Z_N\} \) the elements of the pooled sample, with \( Z_i \) labelled \( c_i = 1 \) if \( Z_i \in \mathcal{X}_{N_1} \) and \( c_i = 2 \) if \( Z_i \in \mathcal{Y}_{N_2} \). Then equation (1.3) can be rewritten as

\[
T\{G(Z_N)\} := \sum_{1 \leq i < j \leq N} \psi(c_i, c_j) \mathbf{1}[(Z_i, Z_j) \in E\{G(Z_N)\}] / |E\{G(Z_N)\}|
\]

where \( \psi(c_i, c_j) = 1(c_i = 1, c_j = 2) \). If \( G \) is an undirected graph functional, then statistic (1.4) counts the proportion of edges in the graph \( G(Z_N) \) with one end point in \( \mathcal{X}_{N_1} \) and the other end point in \( \mathcal{Y}_{N_2} \). If \( G \) is a directed graph functional, then statistic (1.4) is the proportion of directed edges with the outward end in \( \mathcal{X}_{N_1} \) and the inward end in \( \mathcal{Y}_{N_2} \). By conditioning on the graph \( G(Z_N) \), it is easy to see that, under the null \( H_0 \), \( E\{T(Z_N)\} = 2N_1N_2/(N(N-1)) \) or \( N_1N_2/(N(N-1)) \), depending on whether the graph functional is undirected or directed.

In this paper, the graph functionals are computed on the basis of the Euclidean distance in \( \mathbb{R}^d \), and the rejection region of statistic (1.3) will be based on its asymptotic null distribution in the usual limiting regime \( N \to \infty \), with

\[
\begin{align*}
N_1/(N_1 + N_2) & \to p \in (0, 1), \\
N_2/(N_1 + N_2) & \to q := 1 - p.
\end{align*}
\]

The test statistics that are considered in this paper will have \( N^{1/2} \)-fluctuations under \( H_0 \). Thus, depending on the type of alternative, the test based on statistic (1.3) will reject \( H_0 \) for large and/or small values of the standardized statistic

\[
R\{G(Z_N)\} = \sqrt{N(T\{G(Z_N)\} - \mathbb{E}[T\{G(Z_N)\}])}.
\]

### 1.1. Two-sample tests based on geometric graphs

Many popular multivariate two-sample test statistics are of the form (1.3) where the graph functional \( G \) is constructed by using the interpoint distances of the pooled sample.

#### 1.1.1. Wald–Wolfowitz runs test

The WW runs test is one of the earliest known non-parametric tests for the equality of two univariate distributions (Wald and Wolfowitz, 1940): let \( \mathcal{X}_{N_1} \) and \( \mathcal{Y}_{N_2} \) be IID samples of size \( N_1 \) and \( N_2 \) as in expression (1.1). A run in the pooled sample \( Z_N = \mathcal{X}_{N_1} \cup \mathcal{Y}_{N_2} \) is a maximal non-empty segment of adjacent elements with the same label when the elements in \( Z_N \) are arranged in increasing order. If the two distributions are different, the elements with labels 1 and 2 would be clumped together, and the total number of runs \( C(Z_N) \) in \( Z_N \) will be small. In contrast, for distributions which are equal or close, the different labels are jumbled up and \( C(Z_N) \) will be large. Thus, the WW test rejects \( H_0 \) for small values of \( C(Z_N) \).

Note that the number of runs in \( Z_N \) minus 1 equals the number of times that the sample label changes as we move along \( Z_N \) in increasing order. This implies that the WW test is a graph-based test (1.3):

\[
\frac{C(\mathcal{X}_{N_1} \cup \mathcal{Y}_{N_2}) - 1}{N - 1} = T\{P(\mathcal{X}_{N_1} \cup \mathcal{Y}_{N_2})\},
\]

where \( P(S) \) is the path with \(|S| - 1\) edges through the elements of \( S \) arranged in increasing order, for a finite set \( S \subset \mathbb{R} \).

Let \( R\{P(Z_N)\} \) be the standardized version of \( T\{P(Z_N)\} \) as in equation (1.6). Wald and Wolfowitz (1940) proved that \( R\{P(Z_N)\} \) is distribution free in finite samples, is asymptotically
normal under $H_0$ and is consistent under all fixed alternatives. The WW test often has low power in practice and has zero asymptotic efficiency, i.e. it is powerless against $O(N^{-1/2})$ alternatives (Mood, 1954).

1.1.2. Friedman–Rafsky test

Friedman and Rafsky (1979) generalized the WW runs test to higher dimensions by using the Euclidean minimal spanning tree of the pooled sample.

**Definition 2.** Given a nice finite set $S \subset \mathbb{R}^d$, a spanning tree of $S$ is a connected graph $T$ with vertex set $S$ and no cycles. The length $w(T)$ of $T$ is the sum of the Euclidean lengths of the edges of $T$. A minimum spanning tree (MST) of $S$, denoted by $T(S)$, is a spanning tree with the smallest length, i.e. $w(T(S)) \leq w(T)$ for all spanning trees $T$ of $S$.

Thus, $T$ defines a graph functional in $\mathbb{R}^d$ and, given $\mathcal{X}_{N_1}$ and $\mathcal{Y}_{N_2}$ as in expression (1.1), the FR test rejects $H_0$ for small values of

$$T\{T(\mathcal{X}_{N_1} \cup \mathcal{Y}_{N_2})\} = \frac{\sum_{i=1}^{N_1} \sum_{j=1}^{N_2} I[(X_i, Y_j) \in E(T(\mathcal{X}_{N_1} \cup \mathcal{Y}_{N_2}))]}{N - 1}. \quad (1.7)$$

This is precisely the WW test in $d = 1$ and is motivated by the same intuition that, when the two distributions are different, the number of edges across labels 1 and 2 is small.

Friedman and Rafsky (1979) calibrated equation (1.7) as a permutation test and showed that it has good power in practice for multivariate data. Later, Henze and Penrose (1999) proved that $R\{T(Z_N)\}$ is asymptotically normal under $H_0$ and is consistent under all fixed alternatives. Recently, Chen and Zhang (2015) used the FR and related graph-based tests in change-point detection problems and suggested new modifications of the FR test for high dimensional and object data (Chen and Friedman, 2017; Chen et al., 2019).

1.1.3. Test based on K-nearest-neighbour graphs

As in equation (1.7), a multivariate two-sample test can be constructed by using the K-nearest-neighbour ($K$-NN) graph of $Z_N$. This was originally suggested by Friedman and Rafsky (1979) and later studied by Schilling (1986) and Henze (1988).

**Definition 3.** Given a nice finite set $S \subset \mathbb{R}^d$, the (undirected) $K$-NN graph is a graph with vertex set $S$ with an edge $(a, b)$, for $a, b \in S$, if the Euclidean distance between $a$ and $b$ is among the $K$th smallest distances from $a$ to any other point in $S$ and/or among the $K$th smallest distances from $b$ to any other point in $S$. Denote the undirected $K$-NN of $S$ by $N_K(S)$.

Given $\mathcal{X}_{N_1}$ and $\mathcal{Y}_{N_2}$ as in expression (1.1), the $K$-NN statistic is

$$T\{N_K(\mathcal{X}_{N_1} \cup \mathcal{Y}_{N_2})\} = \frac{\sum_{i=1}^{N_1} \sum_{j=1}^{N_2} I[(X_i, Y_j) \in E(N_K(\mathcal{X}_{N_1} \cup \mathcal{Y}_{N_2}))]}{|E(N_K(\mathcal{X}_{N_1} \cup \mathcal{Y}_{N_2}))|}. \quad (1.8)$$

As before, when the two distributions are different, the number of edges across the two samples is small (Fig. 1), so the $K$-NN test rejects $H_0$ for small values of equation (1.8). Schilling (1986) considered the case where $K$ remains fixed with $N$ and showed that the test based on $K$-NNs is asymptotically normal under $H_0$ and consistent against fixed alternatives. (Statistic (1.8) is slightly different from the test in Schilling (1986), section 2, which can also be rewritten as graph-based test (1.3) by allowing multiple edges in the $K$-NN graph.) However, test statistic (1.8) makes sense even when $K = K_N \to \infty$, which we consider in Section 4.3.
1.1.4. Cross-match test

Rosenbaum (2005) proposed a distribution-free multivariate two-sample test based on minimum non-bipartite matching. For simplicity, assume that the total number of samples \( N \) is even; otherwise, add or delete a sample point to make it even.

Definition 4. Given a finite \( S \subset \mathbb{R}^d \) and a symmetric distance matrix \( D := ((d(a, b)))_{a \neq b \in S} \), a non-bipartite matching of \( S \) is a pairing of the elements \( S \) into \( N/2 \) non-overlapping pairs, i.e. a partition of \( S = \bigcup_{i=1}^{N/2} S_i \), where \( |S_i| = 2 \) and \( S_i \cap S_j = \emptyset \). The weight of a matching is the sum of the distances between the \( N/2 \) matched pairs. A minimum non-bipartite matching of \( S \) is a matching which has the minimum weight over all matchings of \( S \).

The non-bipartite matching defines a graph functional \( W \) as follows: for every finite \( S \subset \mathbb{R}^d \), \( W(S) \) is the graph with vertex set \( S \) and an edge \( \{a, b\} \) whenever there exists \( i \in [N/2] \) such that \( S_i = \{a, b\} \). Note that \( W(S) \) is a graph with \( N/2 \) pairwise disjoint edges. Given \( X_{N_1} \) and \( Y_{N_2} \) as in expression (1.1), the cross-match (CM) test rejects \( H_0 \) for small values of

\[
T\{W(X_{N_1} \cup Y_{N_2})\} = \frac{\sum_{i=1}^{N_1} \sum_{j=1}^{N_2} 1\{X_i, Y_j\} \in E\{W(X_{N_1} \cup Y_{N_2})\}}{N/2}.
\]

Like the WW test, but unlike the FR and the K-NN tests, the CM test is distribution free in finite samples under \( H_0 \). Rosenbaum (2005) implemented this as a permutation test and derived the asymptotic normal distribution under the null \( H_0 \).

1.2. Two-sample tests based on data depth

Many non-parametric two-sample tests are based on depth functions, which are multivariate generalizations of ranks (Liu, 1992; Oja, 2010). Given a distribution function \( F \) in \( \mathbb{R}^d \), a depth function \( D(\cdot, F) : \mathbb{R}^d \rightarrow [0, 1] \) is a function that provides a ranking of points in \( \mathbb{R}^d \). High depth corresponds to centrality, whereas low depth corresponds to outlyingness. The centre consists of the points that globally maximize the depth and is often considered as a multivariate median of \( F \). For \( X \sim F \) and \( Y \sim G \), two independent random variables in \( \mathbb{R}^d \),

\[
R_D(y, F) = \mathbb{P}\{X : D(X, F) \leq D(y, F)\}
\]

is a measure of the relative outlyingness of a point \( y \in \mathbb{R}^d \) with respect to \( F \). In other words, \( R_D(y, F) \) is the fraction of the \( F \)-population with depth at most that of the point \( y \). The de-
pendence on $D$ in the notation $R_D(\cdot, F)$ will be dropped when it is clear from the context. The \textit{quality index}

$$Q(F, G) := \int R(y, F) dG(y) = \mathbb{P}\{D(X, F) \leq D(Y, F) \mid X \sim F, Y \sim G\} \quad (1.11)$$

is the average fraction of $F$ with depth at most that of the point $y$, averaged over $y$ distributed as $G$. When $F = G$ and $D(X, F)$ has a continuous distribution, then $R(Y, F) \sim \text{Unif}[0, 1]$ and $Q(F, G) = \frac{1}{2}$ (Liu and Singh (1993), proposition 3.1).

\textbf{Definition 5.} Let $\mathcal{X}_{N_1}$ and $\mathcal{Y}_{N_2}$ be as in expression (1.1) and $F_{N_1}$ and $G_{N_2}$ be the empirical distribution functions respectively. The \textbf{Liu–Singh rank sum statistic} (Liu and Singh, 1993) is

$$Q(F_{N_1}, G_{N_2}) := \int R(y, F_{N_1}) dG_{N_2}(y) = \frac{1}{N_2} \sum_{j=1}^{N_2} R(Y_j, F_{N_1}), \quad (1.12)$$

the sample estimator of $Q(F, G)$.

The test rejects $H_0$ for small or large values of $\sqrt{N} \{Q(F, G) - \frac{1}{2}\}$ and can be rewritten as a graph-based test (1.3). To see this, note that

$$N_1 N_2 Q(F_{N_1}, G_{N_2}) = \sum_{j=1}^{N_2} \sum_{i=1}^{N_1} 1\{D(X_i, F_{N_1}) \leq D(Y_j, F_{N_1})\}. \quad (1.13)$$

Let $Z_N$ be the pooled sample with labelling as in expression (1.4). Construct a graph $\mathcal{G}_D(Z_N)$ with vertices $Z_N$ with a directed edge from $(Z_i, Z_j)$ whenever $D(Z_i, F_{N_1}) \leq D(Z_j, F_{N_1})$. Note that $\mathcal{G}_D(Z_N)$ is a complete graph with directions on the edges depending on the relative order of the depth of the two end points. Then, from equation (1.13),

$$N_1 N_2 Q(F_{N_1}, G_{N_2}) = \sum_{1 \leq i \neq j \leq N} \psi(c_i, c_j) 1\{D(Z_i, F_{N_1}) \leq D(Z_j, F_{N_1})\}$$

$$= \sum_{1 \leq i \neq j \leq N} \psi(c_i, c_j) 1([Z_i, Z_j] \in E\{\mathcal{G}_D(Z_N)\}],$$

with $\psi(\cdot, \cdot)$ as in expression (1.4). Since $|E\{\mathcal{G}_D(Z_N)\}| = N(N - 1)/2$, this implies that

$$Q(F_{N_1}, G_{N_2}) = \frac{N(N - 1)}{2N_1 N_2} T\{\mathcal{G}_D(Z_N)\},$$

where $T$ is defined in expression (1.4).

Thus, the Liu–Singh rank sum statistic based on a depth function $D$ is a graph-based test (1.3). Common depth functions include the Mahalanobis depth (MD), the HD, the simplicial depth and the projection depth, among others. In what follows, we recall the definitions for a few of these (see Liu (1992) for more on depth functions).

(a) \textbf{Mann–Whitney test:} this is one of the oldest non-parametric two-sample tests for univariate data (Lehmann, 1975; van der Vaart, 2000). Depending on the alternative, the Mann–Whitney rank test rejects $H_0$ for small or large values of $\sum_{i=1}^{N_1} \sum_{j=1}^{N_2} 1(X_i < Y_j)$. This is a distribution-free test, which corresponds to taking $D(x, F) = F(x)$ in equation (1.13).

(b) \textbf{HD:} Tukey (1975) suggested the following depth function:

$$\text{HD}(x, F) = \inf_{H_x} \int_{H_x} dF, \quad (1.14)$$
where the infimum is taken over all closed halfspace \( H_x \) with \( x \) on its boundary. The two-sample test based on the HD is obtained by using function (1.14) in statistic (1.13).

(c) \( \text{MD} \): given a distribution function \( F \), the MD of a point \( x \in \mathbb{R}^d \) is

\[
\text{MD}(x, F) = \frac{1}{1 + (x - \mu(F))'\Sigma^{-1}(F)(x - \mu(F))}
\]

where \( \mu(F) = \int x \, dF(x) \) and \( \Sigma(F) = \int (x - \mu(F))(x - \mu(F))' \, dF(x) \) are the mean and covariance of the distribution \( F \).

The Liu–Singh rank sum statistic is consistent against alternatives for which the quality index \( Q(F, G) \neq \frac{1}{2} \) (recall equation (1.10)), under mild conditions (Liu and Singh, 1993). A special version of the Liu–Singh rank sum statistic, with a reference sample, inherits the distribution-free property of the Mann–Whitney test in finite samples (Liu and Singh (1993), section 4). However, the Liu–Singh rank sum statistic defined above expression (1.12) is only asymptotically distribution free under the null hypothesis (Liu and Singh, 1993; Zuo and He, 2006). Zuo and He (2006) theorem 1, proved the asymptotic normality of the Liu–Singh rank sum statistic under general alternatives for depth functions satisfying certain regularity conditions.

1.3. Properties of graph-based tests

A test function \( \phi_N \) for the testing problem (1.1) is said to be asymptotically exact level \( \alpha \), if \( \lim_{N \to \infty} \mathbb{E}_{H_0}(\phi_N) = \alpha \). An exact level \( \alpha \) test function \( \phi_N \) is said to be consistent against the alternative \( H_1 \), if \( \lim_{N \to \infty} \mathbb{E}_{H_1}(\phi_N) = 1 \), i.e. the power of the test \( \phi_N \) converges to 1 in the usual asymptotic regime (1.5).

To describe the asymptotic properties of the tests above, we assume that the distributions \( F \) and \( G \) have densities \( f \) and \( g \) with respect to Lebesgue measure respectively; and, under the alternative \( H_1 \), \( f \) and \( g \) differ on a set of positive Lebesgue measure. Under this assumption, all the tests that were considered in Sections 1.1 and 1.2 have the following common properties.

(a) \textit{Asymptotically distribution free}: the standardized test statistic (1.6) converges to \( \mathcal{N}(0, \sigma_1^2) \) under \( H_0 \), where the limiting variance \( \sigma_1^2 \) depends on the graph functional \( \mathcal{G} \), but not on the unknown distributions. Therefore, under the null, the asymptotic distribution of the test statistic (1.6) is \textit{distribution free}. This was proved for the FR test by Henze and Penrose (1999), for the test based on the \( K \)-NN graph by Henze (1988), and by Liu and Singh (1993) for depth-based tests. Finally, recall that the CM test is exactly distribution free in finite samples (Rosenbaum, 2005).

(c) \textit{Consistent against fixed alternatives}: for a geometric graph functional \( \mathcal{G} \) in \( \mathbb{R}^d \), the test function with rejection region

\[
\{ \mathcal{R}(\mathcal{G}(Z_N)) < \sigma_1 z_\alpha \},
\]

where \( z_\alpha \) is the \( \alpha \)th quantile of the standard normal distribution, is asymptotically exact level (size) \( \alpha \) (see Lehmann and Romano (2005) for definitions of level and size of a test) and consistent against all fixed alternatives, for the testing problem (1.1). This was proved for the MST by Henze and Penrose (1999) and for the \( K \)-NN graph (when \( K = O(1) \)) by Schilling (1986) and later by Henze (1988). Recently, Arias-Castro and Pelletier (2016) showed that the CM test is consistent against general alternatives. Similarly, the Liu–Singh rank sum statistic is asymptotically size \( \alpha \) and consistent against alternatives for which the quality index \( Q(F, G) \neq \frac{1}{2} \). (For depth-based tests, the rejection region can be of the form \( \{ |\mathcal{R}(\mathcal{G}(Z_N))| \geq \sigma_1 z_{1 - \alpha/2} \} \), depending on the alternative.)
Even though the basic asymptotic properties of the tests based on geometric graphs and those based on data depth are quite similar, their algorithmic complexities are very different.

(a) Tests based on geometric graphs such as the MST, the $K$-NN graph and the CM test, can be computed in polynomial time with respect to both the number of data points $N$ and the dimension $d$. For instance, the MST and the $K$-NN graph of a set of $N$ points in $\mathbb{R}^d$ can be computed easily in $O(dN^2)$ time, and the non-bipartite matching can be computed in $O(dN^3)$ time (Papadimitriou and Steiglitz, 1982).

(b) In contrast, depth-based tests are generally difficult to compute when the dimension is large because the computation time is polynomial in $N$ but exponential in $d$. In fact, computing the Tukey depth of a set of $N$ points in $\mathbb{R}^d$ is computationally hard if both $N$ and $d$ are parts of the input (Johnson and Preparata, 1978) and it is even difficult to approximate (Amaldi and Kann, 1995). For more details on algorithms to compute various depth functions, refer to the survey of Rousseeuw and Hubert (2015) and the references therein.

1.4. **Summary of results**

The general notion of graph-based two-sample tests (1.3) that was introduced above provides a unified framework for analysing and statistically comparing their asymptotic properties. We derive the limiting power of these tests against local alternatives, i.e. alternatives which shrink towards the null as the sample size grows to $\infty$. If statistic (1.6) is asymptotically normal under the null and the test based on statistic (1.16) is consistent, it can be expected to have non-trivial power (greater than the level of the test) against $O(N^{-1/2})$ local alternatives. This is formalized by using the notion of Pitman efficiency of a test (see equation (2.2) below) and can be used to compare the asymptotic performances of different tests. Here is a summary of the results obtained.

(a) In Section 3 the asymptotic (Pitman) efficiency of a general graph-based test is derived. This result can be used as a black box to derive the efficiency of any such a two-sample test (theorem 1 and corollary 1). The results that are obtained show how combinatorial properties of the underlying graph affect the performance of the associated test, which can be effectively used to construct and analyze new tests. The results are illustrated through simulations and compared with other parametric tests in Section 5.

(b) As a consequence of the general result the asymptotic efficiency of the tests that were described in Sections 1.1 and 1.2 can be derived.

(i) It is shown that the FR test has zero asymptotic efficiency, i.e. it is powerless against any $O(N^{-1/2})$ local alternatives (theorem 2). In fact, this phenomenon extends to a large class of random geometric graphs that exhibit local spatial dependence. This can be formalized by using the notion of stabilization of geometric graphs (Penrose and Yukich, 2001), which includes the MST, the $K$-NN graph (where $K = O(1)$ is fixed) among others (theorem 3).

(ii) Our general theorem can be used to compute the asymptotic efficiency of the $K$-NN-based test, when $K = K_N \to \infty$, as well (proposition 1). Here, the Pitman efficiency can be non-zero, when $K$ grows with $N$ sufficiently fast. This is reinforced in simulations, which, combined with the computational efficiency of the $K$-NN test (the running time is polynomial in $K$, the sample size $N$ and the dimension $d$), makes this test particularly attractive, both theoretically and in applications.

(iii) The Pitman efficiency of tests based on data depth (1.12) is computed in Section 4.5. These tests have non-trivial local power and, hence, non-zero asymptotic efficiencies, for many $O(N^{-1/2})$ alternatives (theorem 3). However, as mentioned earlier, these tests become computationally expensive as the dimension increases.
(iv) Recently, Chen and Friedman (2017) proposed a modification of the test statistic (1.6), which is especially powerful when the sample size is small and the dimension is large. Our general framework can be modified to include these tests as well, and we derive their limiting power against local alternatives (theorem 5).

(c) Finally, the performance of the FR test and the test based on the HD are compared on the sensorless drive diagnosis data set (Bayer et al., 2013) (Section 6). All codes used in the paper can be downloaded from http://www-stat.wharton.upenn.edu/~bhaswar/Graph_Based_Two_Sample_Codes.zip, and the data that are analysed in the paper and the programs that were used to analyse them can be obtained also from https://rss.onlinelibrary.wiley.com/hub/journal/14679868/series-b-datasets

In Section 7 we study the performance of these tests when sample sizes are small and the dimension is comparable with the sample size. In this case, classical parametric tests, which involve computation of the sample covariance matrix, often break down, and the non-parametric tests start to dominate (Biswas et al., 2014; Chen and Friedman, 2017). This is validated by the simulations in Section 7, where the tests based on geometric graphs, such the FR and the CF tests, outperform the other tests, in the high dimensional regime. We conclude with a discussion about the performances of the various tests, and which tests to use in practice (Section 8).

2. Preliminaries

Begin by recalling some definitions and notation: for \( x \in \mathbb{R} \), \( x_+ = \max\{x, 0\} \) and, for a vector \( z \in \mathbb{R}^s \), \( \|z\| = (\sum_{i=1}^s z_i^2)^{1/2} \) is the Euclidean norm of \( z \).

To quantify the notion of local alternatives, let \( \Theta \subseteq \mathbb{R}^p \) and \( \{P_\theta\}_{\theta \in \Theta} \) be a parametric family of distributions in \( \mathbb{R}^d \) with density \( f(\cdot | \theta) \), with respect to Lebesgue measure, indexed by a \( p \)-dimensional parameter \( \theta \in \Theta \). Throughout, we shall assume that the distributions in \( \{P_\theta\}_{\theta \in \Theta} \) have a common support \( \mathcal{K} \subseteq \mathbb{R}^d \), which does not depend on \( \theta \). To compute the asymptotic efficiency of tests, certain smoothness conditions are required on \( f(\cdot | \theta) \). The standard technical condition is to assume that the family \( \{P_\theta\}_{\theta \in \Theta} \) is quadratic mean differentiable (see Lehmann and Romano (2005), definition 12.2.1, for details). The quadratic mean differentiable assumption implies differentiability in norm of the square root of the density, which holds for most standard families of distributions, including exponential families in natural form.

**Assumption 1.** A parametric family of distributions \( \{P_\theta\}_{\theta \in \Theta} \), with \( \Theta \subseteq \mathbb{R}^p \), is said to satisfy assumption 1 at \( \theta = \theta_1 \), if \( P_\theta \) is quadratic mean differentiable at \( \theta = \theta_1 \) and, for \( V_1 \sim P_{\theta_1} \) and all \( h \in \mathbb{R}^p \), \( \mathbb{E}(|\langle h, \eta(V_1, \theta_1) \rangle|^4) < \infty \), where \( \eta(\cdot, \theta) := \nabla_\theta f(\cdot | \theta) / f(\cdot | \theta) \) is the score function.

Let \( \mathcal{X}_{N_1} \) and \( \mathcal{Y}_{N_2} \) be samples from \( P_{\theta_1} \) and \( P_{\theta_2} \) as in expression (1.1) respectively. For \( h \in \mathbb{R}^p \), consider the testing problem

\[
H_0 : \theta_2 - \theta_1 = 0 \quad \text{versus} \quad H_1 : \theta_2 - \theta_1 = h / \sqrt{N}.
\]

Note that the tests are still carried out in the non-parametric set-up assuming no knowledge of the distributions of the two samples. However, the efficiency is computed by assuming a parametric form for the unknown distributions. (Another choice of alternatives for computing the efficiency of a non-parametric test is to consider \( H_0 : g = f \), versus \( H_1 : g = (1 - \delta / \sqrt{N}) f + (\delta / N) g' \), for some density \( g' \) in \( \mathbb{R}^d \) and \( \delta > 0 \). Then, under mild integrability assumptions, the densities that are associated with the alternative are contiguous, and the efficiency results that are derived in this paper easily extend to this case, as well. We chose formulation (2.1) because it yields slightly cleaner asymptotic expansions and formulae.)
3. Asymptotic efficiency of graph-based two-sample tests

This section describes the main results about the asymptotic efficiencies of general graph-based two-sample tests. There are two cases, depending on whether the graph functional is directed or undirected. To begin, let \( \mathcal{G} \) be a directed graph functional in \( \mathbb{R}^d \). Denote by \( E\{\mathcal{G}(S)\} \) the set of edges in \( \mathcal{G}(S) \), and by \( E^\pm\{\mathcal{G}(S)\} \) the set of pairs of vertices with edges in both directions (i.e. the set of ordered pairs of vertices \((x, y)\) such that both \((x, y)\) and \((y, x)\) belong to \( E\{\mathcal{G}(S)\}\)).

For \( x \in \mathbb{R}^d \), let \( d^\downarrow \{x, \mathcal{G}(S)\} \) be the out-degree of the vertex \( x \) in the graph \( \mathcal{G}(S \cup \{x\}) \), i.e. the number of outgoing edges \((x, y)\), where \( y \in S \cup \{x\} \), in the graph \( \mathcal{G}(S \cup \{x\}) \). Similarly, let \( d^\uparrow \{x, \mathcal{G}(S)\} \) be the in-degree of the vertex \( x \) in the graph \( \mathcal{G}(S \cup \{x\}) \), i.e. the number of incoming edges \((y, x)\), where \( y \in S \cup \{x\} \), in the graph \( \mathcal{G}(S \cup \{x\}) \). Moreover, let \( d \{x, \mathcal{G}(S)\} = d^\downarrow \{x, \mathcal{G}(S)\} + d^\uparrow \{x, \mathcal{G}(S)\} \) be the total degree of the vertex \( x \) in the graph \( \mathcal{G}(S \cup \{x\}) \). Define the scaled in-degree and the scaled out-degree of a vertex as

\[
\lambda^\uparrow \{x, \mathcal{G}(S)\} = \frac{Nd^\uparrow \{x, \mathcal{G}(S^\downarrow)\}}{|E\{\mathcal{G}(S^\downarrow)\}|},
\]
\[
\lambda^\downarrow \{x, \mathcal{G}(S)\} = \frac{Nd^\downarrow \{x, \mathcal{G}(S^\uparrow)\}}{|E\{\mathcal{G}(S^\uparrow)\}|},
\]

where \( S^\downarrow = S \cup \{x\} \). Also, let

\[
T_2^\uparrow \{\mathcal{G}(S)\} = \sum_{x \in S} \binom{d^\uparrow \{x, \mathcal{G}(S)\}}{2},
\]
\[
T_2^\downarrow \{\mathcal{G}(S)\} = \sum_{x \in S} \binom{d^\downarrow \{x, \mathcal{G}(S)\}}{2}
\]

be the number of outward 2-stars and inward 2-stars in \( \mathcal{G}(S) \) respectively. Finally, let \( T_2^\pm \{\mathcal{G}(S)\} \) be the number of 2-stars in \( \mathcal{G}(S) \) with different directions on the two edges.

For an undirected graph functional \( \mathcal{G} \), denote by \( d \{x, \mathcal{G}(S)\} \) the degree of the vertex \( x \) in the graph \( \mathcal{G}(S \cup \{x\}) \). As in expressions (3.1) and (3.2), let

\[
\lambda \{x, \mathcal{G}(S)\} = \frac{Nd \{x, \mathcal{G}(S^\downarrow)\}}{|E\{\mathcal{G}(S^\downarrow)\}|},
\]
\[
T_2 \{\mathcal{G}(S)\} = \sum_{x \in S} \binom{d \{x, \mathcal{G}(S)\}}{2}
\]

be the number of 2-stars in \( \mathcal{G}(S) \).
Intuitively, the functions \( \lambda^\uparrow \{ x, \mathcal{G}(S) \}, \lambda^\downarrow \{ x, \mathcal{G}(S) \} \) and \( \lambda \{ x, \mathcal{G}(S) \} \) measure the relative position of the point \( x \) in set \( S \cup \{ x \} \). For example, in the FR test or the test based on the \( K \)-NN graph, large values of \( \lambda \) correspond to points near the centre of the data cloud, whereas small values correspond to outliers. Similarly, for depth-based tests, small or large values of \( \lambda \) generally correspond to points which are outliers. In fact, for such tests, this is directly related to the relative outlyingness of the point \( x \), as defined in equation (1.10) (see observation G.1 in the on-line supporting information).

### 3.1. Asymptotic efficiency for directed graph functionals

Let \( \mathcal{G} \) be a directed graph functional in \( \mathbb{R}^d \) and \( \{ \mathbb{P}_{\theta} \}_{\theta \in \Theta} \) a parametric family of distributions satisfying assumption 1 at \( \theta = \theta_1 \in \Theta \subseteq \mathbb{R}^p \). To derive the asymptotic efficiency of a general graph-based test various assumptions are required on the graph functional \( \mathcal{G} \).

**Assumption 2.** (variance condition). The pair \( (\mathcal{G}, \mathbb{P}_{\theta_1}) \) is said to satisfy the variance condition with parameters \( (\beta_0, \beta_0^+, \beta_1^+, \beta_4^+) \) if, for \( \mathcal{V}_n \coloneqq \{ V_1, V_2, \ldots, V_N \} \) IID from \( \mathbb{P}_{\theta_1} \), there are finite non-negative constants \( \beta_0, \beta_0^+, \beta_1^+, \beta_4^+ \) (which do not depend on \( \mathbb{P}_{\theta_1} \)) such that

(a) \( N |E\{ \mathcal{G}(\mathcal{V}_N) \}| \to^P \beta_0 \) and \( N |E^+\{ \mathcal{G}(\mathcal{V}_N) \}|/|E\{ \mathcal{G}(\mathcal{V}_N) \}|^2 \to^P \beta_0^+ \),

(b) \( NT_1^+\{ \mathcal{G}(\mathcal{V}_N) \}/|E\{ \mathcal{G}(\mathcal{V}_N) \}| \to^P \beta_1^+ \) and \( NT_2^+\{ \mathcal{G}(\mathcal{V}_N) \}/|E\{ \mathcal{G}(\mathcal{V}_N) \}| \to^P \beta_4^+ \),

(c) \( NT_2^+\{ \mathcal{G}(\mathcal{V}_N) \}/|E\{ \mathcal{G}(\mathcal{V}_N) \}| \to^P \beta_4^+ \).

The asymptotic efficiency will be derived by using Le Cam’s third lemma (Lehmann and Romano (2005), corollary 12.3.2), for which the joint normality of \( \mathcal{V}_N \) is required. For this the following two conditions are required.

**Assumption 3.** (covariance condition). The pair \( (\mathcal{G}, \mathbb{P}_{\theta_1}) \) is said to satisfy the covariance condition if for \( \mathcal{V}_N \coloneqq \{ V_1, V_2, \ldots, V_N \} \) IID from \( \mathbb{P}_{\theta_1} \) the following holds hold.

(a) There are functions \( \lambda^\uparrow, \lambda^\downarrow : \mathcal{K} \to \mathbb{R} \), such that, for almost all \( z \in \mathcal{K} \),

\[
\lambda^\uparrow(z) = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E} \{ \lambda^\uparrow \{ z, \mathcal{G}(\mathcal{V}_N) \} \},
\lambda^\downarrow(z) = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E} \{ \lambda^\downarrow \{ z, \mathcal{G}(\mathcal{V}_N) \} \}.
\]

(b) For \( h \in \mathbb{R}^p \) as in expression (2.1),

\[
\frac{1}{N} \sum_{i=1}^{N} \langle h, \eta(V_i, \theta_1) \rangle \lambda^\uparrow \{ V_i, \mathcal{G}(\mathcal{V}_N) \} \to^P \int \langle h, \nabla f(z|\theta_1) \rangle \lambda^\uparrow(z) dz,
\]

and the same holds for \( \lambda^\downarrow \).

Next, define \( \Delta\{ \mathcal{G}(S) \} \coloneqq \max_{x \in \mathcal{G}(S)} d \{ V_i, \mathcal{G}(S) \} \) the total maximum degree of the graph \( \mathcal{G}(S) \), where \( \|N\| \coloneqq \{ 1, 2, \ldots, N \} \).

**Assumption 4.** (normality condition). The pair \( (\mathcal{G}, \mathbb{P}_{\theta_1}) \) is said to satisfy the normality condition if for \( \mathcal{V}_N \coloneqq \{ V_1, V_2, \ldots, V_N \} \) IID from \( \mathbb{P}_{\theta_1} \) either one of the following holds:

(a) \( \text{(condition 1)} \) the maximum degree \( \Delta\{ \mathcal{G}(\mathcal{V}_N) \} = O_P(1) \);

(b) \( \text{(condition 2)} \) the maximum degree \( \Delta\{ \mathcal{G}(\mathcal{V}_N) \} \to \infty \) and \( N \Delta\{ \mathcal{G}(\mathcal{V}_N) \}/|E\{ \mathcal{G}(\mathcal{V}_N) \}| = O_P(1) \).

**Remark 1.** Assumption 4 implies that either
(a) the graph has bounded degree (condition 1) or
(b) the maximum degree $\Delta\{G(V_N)\}$ is of the same order as the average degree $|E\{G(V_N)\}|/N$ (condition 2), i.e. the graph is ‘approximately’ regular.

This ensures that the conditional standard deviation $\sqrt{\text{var}[T\{G(Z_N)\}|Z_N]}=\Theta(1/\sqrt{N})$, justifying the scaling in statistic (1.6). It is possible to consider a slightly more general class of statistics (instead of $1/\sqrt{N}$) which renormalizes $T\{G(Z_N)\} - \mathbb{E}[T\{G(Z_N)\}]$ by $\sqrt{\text{var}[T\{G(Z_N)\}|Z_N]}$ itself. The asymptotic efficiency of such a statistic can be similarly derived, after the covariance condition has been rescaled and the normality condition has been modified appropriately. We have decided to scale by $1/\sqrt{N}$ instead, because this leads to more interpretable conditions (assumption 4), which are easier to apply in our examples, and includes all known graph-based two-sample tests in the literature.

The following result gives the asymptotic efficiency of a graph-based test, if the graph functional satisfies the above conditions. For this, define $r:=2pq$, where $p$ and $q$ are defined in expression (1.5).

**Theorem 1.** Let $G$ be a directed graph functional and $\{P_{\theta}\}_{\theta \in \Theta}$ a parametric family of distributions in $\mathbb{R}^d$ satisfying assumption 1 at $\theta = \theta_1 \in \Theta \subseteq \mathbb{R}^p$. Suppose that the pair $(G, P_{\theta_1})$ satisfies the variance assumption 2 with parameters $(\beta_0, \beta_0^+, \beta_1^+, \beta_1^+, \beta_1^-, \beta_1^-)$, the covariance assumption 3 and the normality assumption 4. Then the asymptotic efficiency of the test statistic (1.6) for the testing problem (2.1) is

$$AE(G) := \frac{(r/2)\{p \int (h, \nabla f(z|\theta_1))\lambda^+(z)dz - q \int (h, \nabla f(z|\theta_1))\lambda^-(z)dz\}}{\sqrt{r\{(\beta_0 - 1)/2 + g\beta_1^+ + p\beta_1^- - (r/2)(\beta_0/2 + \beta_0^+ + \beta_1^+ + \beta_1^- - 2)\}}}$$

whenever the denominator above is strictly positive.

The proof of theorem 1 is given in the on-line appendix A. The efficiency formula in theorem 1 shows how combinatorial properties of the underlying graph affect the performance of the associated test, through the functions $\lambda^+$ and $\lambda^-$. Moreover, the formula holds (for tests based on geometric graphs) for any distance function $\rho$ in $\mathbb{R}^d$ as long as the pooled sample $Z_N$ is nice with respect to $\rho$ (i.e. all pairwise distances are unique) and the pair $(G, P_{\theta_1})$ satisfies the assumptions of theorem 1. In our applications in Section 4, $\rho$ will be the Euclidean metric, but the result continues to hold for other natural distance functions like $L_p$ and the MD.

**Remark 2 (limiting null distribution).** The proof of theorem 1 also gives the limiting null distribution of test statistic (1.6), unifying several known results in the literature. Note that the variance assumption 2 and the normality assumption 4 naturally extend to the pair $(G, P_f)$, where $P_f$ is the probability measure induced by $f$. The proof of theorem 1 shows that, if the pair $(G, P_f)$ satisfies the variance assumption 2 and the normality assumption 4, then under the null hypothesis $(f = g) \rightarrow D N(0, \sigma_f^2)$, where $\sigma_f$ is the denominator of the formula in theorem 1. In particular, this gives a unified proof for the asymptotic null distribution of the FR test (Henze and Penrose (1999), theorem 1), the K-NN test (Schilling (1986), theorem 3.1), the CM test (Rosenbaum (2005), proposition 2) and the Liu–Singh rank sum statistic (Liu and Singh (1993), theorem 6.2).

### 3.2. Asymptotic efficiency for undirected graph functionals

Every undirected graph functional $G$ can be modified to a directed graph functional $G_+$ in a natural way: for $S \subset \mathbb{R}^d$ finite, $G_+(S)$ is obtained by replacing every edge in $G(S)$ with two edges,
one in each direction. The asymptotic efficiency of the test based on \( G \) can then be derived by applying theorem 1 to the directed graph functional \( G \). The following assumptions are the analogue of assumptions 2–4 for undirected graph functionals.

**Assumption 5.** Let \( G \) be an undirected graph functional and assume that \( V_N := \{V_1, V_2, \ldots, V_N\} \) are IID with density \( P_{\theta_1} \). The pair \((G, P_{\theta_1})\) is said to satisfy assumption 5 if the following conditions hold.

(a) \((\gamma_0, \gamma_1)\) undirected variance condition: there exist finite non-negative constants \( \gamma_0 \) and \( \gamma_1 \) such that

\[
\frac{N}{|E\{G(V_N)\}|} \xrightarrow{P} \gamma_0, \\
\frac{N|T_2\{G(V_N)\}|}{|E\{G(V_N)\}|^2} \xrightarrow{P} \gamma_1.
\]

(b) Undirected covariance condition: there exists a function \( \lambda: \mathcal{K} \to \mathbb{R} \), such that, for almost all \( z \in \mathcal{K} \), \( \lambda(z) := \lim_{N \to \infty} \mathbb{E}[\lambda(z, G(V_N))] \), and \( \lambda(z) = 0 \) otherwise. Moreover, for \( h \in \mathbb{R}^p \) as in expression (2.1),

\[
\frac{1}{N} \sum_{i=1}^{N} \langle h, \eta(V_i, \theta_1) \rangle \lambda(V_i, G(V_N)) \xrightarrow{P} \int \langle h, \nabla f(z|\theta_1) \rangle \lambda(z)dz.
\]

(c) Normality condition: the pair \((G_+, P_{\theta_1})\) satisfies the normality condition as in assumption 4.

If \( G \) satisfies the above condition, then \( G_+ \) satisfies assumptions 2–4, and applying theorem 1 for \( G_+ \) we obtain the following corollary.

**Corollary 1.** Let \( G \) be an undirected graph functional and \( \{P_{\theta}\}_{\theta \in \Theta} \) a parametric family of distributions in \( \mathbb{R}^p \) satisfying assumption 1 at \( \theta = \theta_1 \in \Theta \). If the pair \((G, P_{\theta_1})\) satisfies assumption 5, then the asymptotic efficiency of test statistic (1.6) is

\[
AE(G) = \frac{(r/2)(p-q) \int \langle h, \nabla f(z|\theta_1) \rangle \lambda(z)dz}{\sqrt{r\{\gamma_0(1-r) + (\gamma_1 - 2)(1-2r)\}}},
\]

whenever the denominator above is strictly positive.

**Remark 3.** Note that the numerator in equation (3.8) is 0 when \( p = q = \frac{1}{2} \), i.e. when the two sample sizes \( N_1 \) and \( N_2 \) are asymptotically equal. This is because, conditionally on the graph, the variables \( \{\psi(c_i, c_j) : (Z_i, Z_j) \in E\{G(Z_N)\}\} \) are pairwise independent when \( p = q \), and, as a result, test statistic (1.6) does not correlate with the likelihood ratio. Therefore, depending on the value of the denominator in equation (3.8) the following two cases arise.

(a) If the graph functional is non-sparse, i.e. \( |E\{G(V_N)\}|/N \to \infty \), then \( \gamma_0 = 0 \) in expression (3.6) and the denominator in equation (3.8) is 0, since \( r = 2pq = \frac{1}{2} \). Therefore, non-sparse undirected graph functionals do not have a non-degenerate distribution at the \( N^{1/2} \)-scale when \( p = q \), and corollary 1 does not apply. This degeneracy is well known in the graph colouring literature: for example, the limiting distribution of the test statistic under the null hypothesis in this case follows from Bhattacharya et al. (2017), theorem 1.3.

(b) If the graph is sparse, i.e. \( \gamma_0 > 0 \), then the denominator of equation (3.8) is non-zero when
\[ p = q. \] This shows that tests based on sparse graph functionals cannot have non-zero efficiencies when the two-sample sizes are asymptotically equal.

4. Applications

In this section we compute the asymptotic efficiencies of the tests that were described in Section 1, under the Euclidean distance, using theorem 1 and corollary 1. Extensions and generalizations which can be used to construct locally efficient tests are also discussed.

4.1. Friedman–Rafsky test

Let \( T \) be the MST functional as in definition 2 and consider the two-sample test based on \( T \). The following theorem shows that this test has zero asymptotic efficiency, under the Euclidean distance.

**Theorem 2.** Let \( \{P_\theta\}_{\theta \in \Theta} \) be a parametric family of distributions in \( \mathbb{R}^d \) satisfying assumption 1 at \( \theta = \theta_1 \in \Theta \). Then the asymptotic efficiency of the FR test (1.7), under the Euclidean distance, for the testing problem (2.1), is 0, i.e. \( \text{AE}(T) = 0. \)

The FR test has zero asymptotic efficiency because the function \( \lambda(z, T(x)) \) does not depend on \( z \) and, hence, the numerator in equation (3.8) is 0. This is a consequence of the following well-known result: for \( \mathcal{V}_N = \{V_1, V_2, \ldots, V_N\} \) IID \( \mathbb{P}_{\theta_1} \) and \( z \in \mathcal{K}, \lim_{N \to \infty} E[\lambda(z, T(N))] = 2 \) (refer to Henze and Penrose (1999), proposition 1). The Efron–Stein inequality (Efron and Stein, 1981) can then be used to show that \( \lambda(z) = 2 \). The details of the proof are given in the on-line supplementary materials.

4.2. Tests based on stabilizing graphs

The proof of theorem 2 uses the fact that the MST graph functional has local dependence, i.e. addition or deletion of a point affects only the edges that are incident on the neighbourhood of that point. This holds for many other random geometric graphs and was formalized by Penrose and Yukich (2003) by using the notion of stabilization.

Let \( \mathcal{G} \) be a graph functional defined for all locally finite subsets of \( \mathbb{R}^d \). (The \( K\)-NN graph can be naturally extended to locally finite infinite points sets. Aldous and Steele (1992) extended the MST graph functional to locally finite infinite point sets by using Prim’s algorithm.) For \( S \subset \mathbb{R}^d \) locally finite and \( x \in \mathbb{R}^d \), let \( E\{x, \mathcal{G}(S)\} \) be the set edges incident on \( x \) in \( \mathcal{G}(S) \). Note that \( |E\{x, \mathcal{G}(S)\}| = d\{x, \mathcal{G}(S)\} \), the (total) degree of the vertex \( x \) in \( \mathcal{G}(S \cup \{x\}) \).

**Definition 6.** Given \( S \subset \mathbb{R}^d \) and \( y \in \mathbb{R}^d \) and \( a \in \mathbb{R} \), denote by \( y + S = \{y + z : z \in S\} \) and \( aS = \{az : z \in S\} \). A graph functional \( \mathcal{G} \) is said to be translation invariant if the graphs \( \mathcal{G}(x + S) \) and \( \mathcal{G}(S) \) are isomorphic for all points \( x \in \mathbb{R}^d \) and all locally finite \( S \subset \mathbb{R}^d \). A graph functional \( \mathcal{G} \) is scale invariant if \( \mathcal{G}(aS) \) and \( \mathcal{G}(S) \) are isomorphic for all points \( a \in \mathbb{R} \) and all locally finite \( S \subset \mathbb{R}^d \).

Let \( \mathcal{P}_\lambda \) be the Poisson process of intensity \( \lambda \geq 0 \) in \( \mathbb{R}^d \), and \( \mathcal{P}_\lambda^x := \mathcal{P}_\lambda \cup \{x\} \), for \( x \in \mathbb{R}^d \). Penrose and Yukich (2003) defined stabilization of graph functionals over homogeneous Poisson processes as follows.

**Definition 7** (Penrose and Yukich, 2003). A translation and scale invariant graph functional \( \mathcal{G} \) stabilizes \( \mathcal{P}_\lambda \) if there is a random but almost surely finite variable \( R \) such that
\[ E\{0, \mathscr{G}(P_0^0)\} = E[0, \mathscr{G}\{P_0^0 \cap B(0, R) \cup A\}], \]  
for all finite \(A \subset \mathbb{R}^d \setminus B(0, R)\), where \(B(0, R)\) is the (Euclidean) ball of radius \(R\) with centre at the point \(0 \in \mathbb{R}^d\).

**Remark 4.** Informally, stabilization ensures the insertion of a point (or finitely many points) ‘far’ from the origin 0 does not effect the degree of 0 in the graph \(\mathscr{G}(P_0^0)\), i.e. it has only a ‘local effect’. Many graph functionals such as the MST, the K-NN, the Delaunay graph and the Gabriel graph are stabilizing (Penrose and Yukich, 2003).

The following theorem shows that tests based on stabilizing graph functionals have zero asymptotic efficiency, under the Euclidean distance. The proof is given in the on-line supplementary materials.

**Theorem 3.** Let \(\{P_\theta\}_{\theta \in \Theta}\) be a parametric family of distributions in \(\mathbb{R}^d\) satisfying assumption 1 at \(\theta = \theta_1 \in \Theta\), and \(\mathscr{G}\) be a translation and scale invariant graph functional which stabilizes \(P_1\). If the pair \((\mathscr{G}, P_{\theta_1})\) satisfies assumption 4 and

\[
\sup_{N \in \mathbb{N}} \sup_{z \in \mathbb{R}^d} \mathbb{E}\{d(z, \mathscr{G}(Z_N))\} < \infty, \tag{4.2}
\]

for some \(s > 4\), then the asymptotic efficiency of the two-sample test based on \(\mathscr{G}\) (1.6), under Euclidean distance, for the testing problem (2.1), is 0, i.e. \(AE(\mathscr{G}) = 0\).

Theorem 3 can be used to rederive theorem 2 and to compute the asymptotic efficiency of the K-NN test.

(a) **MST:** by Penrose and Yukich (2003), lemma 2.1, the MST graph functional \(T\) stabilizes \(P_1\). Moreover, the degree of a vertex in the MST of a set of points in \(\mathbb{R}^d\) is bounded by a constant \(B_d\), depending only on the dimension \(d\) (Aldous and Steele (1992), lemma 4). Therefore, the normality condition 1 in assumption 4 and the moment condition (4.2) are trivially satisfied. Theorem 3 then implies that \(AE(T) = 0\), thus rederiving theorem 2.

(b) **K-NN:** by Penrose and Yukich (2001), lemma 6.1, the K-NN graph functional \(\mathcal{N}_K\), where \(K = O(1)\) is fixed with \(N\), stabilizes \(P_1\). Condition 1 in assumption 4 and the moment condition (4.2) are trivially satisfied, since \(\mathcal{N}_K\) is a bounded degree graph functional. This implies that \(AE(\mathcal{N}_K) = 0\), showing that the test based on the K-NN graph has no asymptotic local power, when \(K = O(1)\).

### 4.3. Test based on the K-nearest-neighbour graph

The result in the previous section shows that the test based on the K-NN graph has zero Pitman efficiency, when \(K = O(1)\) is fixed with the \(N\). But what happens when \(K = K_N \rightarrow \infty\) with \(N\)? In this case, the nearest neighbour graph is no longer stabilizing, and theorem 3 does not apply. However, we can directly invoke corollary 1 to compute the asymptotic efficiency. For this, let \(S \subset \mathbb{R}^d\) be a finite set and \(z \in \mathbb{R}^d\) be a fixed point and \(K = K_N \rightarrow \infty\). It follows from Linderman et al. (1975), lemma 1, that \(d(z, \mathcal{N}_{K_N}(S)) \leq C_d K_N\), where \(C_d\) is a constant depending only on the dimension \(d\). This implies, for \(V_N = \{V_1, V_2, \ldots, V_N\}\) IID \(f(\cdot|\theta_1)\), that the maximum degree \(\Delta(\mathcal{N}_{K_N}(V_N)) = \max_{i \in [N]} d\{V_i, \mathcal{N}_{K_N}(V_N)\} \leq C_d K_N\). Moreover, each vertex in the graph \(\mathcal{N}_{K_N}(V_N)\) has degree at least \(K_N\), which means that the total number of edges

\[
|E(\mathcal{N}_{K_N}(V_N))| = \frac{1}{2} \sum_{i=1}^{N} d\{V_i, \mathcal{N}_{K_N}(V_N)\} \geq \frac{K_N N}{2}.
\]
Hence,

$$\frac{N \Delta \{ N_{K_N} (\mathcal{V}_N) \}}{|E \{ N_{K_N} (\mathcal{V}_N) \}|} \leq C_d = O(1),$$

(4.3)
i.e. condition 2 in assumption 4 is satisfied. Therefore, assuming that there are functions $\eta_0, \eta_1 : \mathcal{K} \rightarrow \mathbb{R}$ such that

$$\eta_0(z) := \lim_{N \rightarrow \infty} \frac{1}{K_N} \mathbb{E} \left[ d \{ z, N_{K_N} (\mathcal{V}_N) \} \right],$$

$$\eta_1(z) := \lim_{N \rightarrow \infty} \frac{1}{K_N^2} \mathbb{E} \left[ \frac{d \{ z, N_{K_N} (\mathcal{V}_N) \}^2}{2} \right],$$

(4.4)
for almost all $z \in \mathcal{K}$, the asymptotic efficiency of the $K$-NN test can be derived by using corollary 1. (Note that both the limits in expression (4.4) are finite, since $\max_{z \in \mathbb{R}^d} d \{ z, N_{K_N} (\mathcal{V}_N) \} \leq C_d K_N$, almost surely.)

**Proposition 1.** Let $\{ \mathbb{P}_\theta \}_{\theta \in \Theta}$ be a parametric family of distributions in $\mathbb{R}^d$ satisfying assumption 1 at $\theta = \theta_1 \in \Theta$. Then the asymptotic efficiency of the $K$-NN test (where $K = K_N \rightarrow \infty$), for the testing problem (2.1), is

$$\text{AE}(N_{K_N}) = \frac{r(p-q) \int \eta_0(z) \nabla f(z|\theta_1)dz}{\sqrt{r \left[ 4 \int \eta_1(z) f(z|\theta_1)dz \{ \int \eta_0(z) f(z|\theta_1)dz \}^2 - 2 \right] (1-2r)}},$$

(4.5)
whenever the denominator above is strictly positive, where $\eta_0(\cdot)$ and $\eta_1(\cdot)$ are as defined in expression (4.4).

The proof of this result, which entails verifying the conditions in corollary 1, is given in the on-line supplementary materials. Note that formula (4.5) has a couple of degeneracies:

(a) when $p = q$, both the numerator and the denominator in formula (4.5) are 0, and the result does not apply (recall the discussion in remark 3);

(b) when

$$\frac{\int \eta_1(z) f(z|\theta_1)dz}{\{ \int \eta_0(z) f(z|\theta_1)dz \}^2} = \frac{1}{2},$$

the denominator in formula (4.5) is 0. This happens when $K_N = N - o(N)$, i.e. the graph $N_{K_N} (\mathcal{V}_N)$ is ‘nearly complete’ (it has $(N^2) - o(N^2)$ edges) and $\eta_0(z) = 1$ and $\eta_1(z) = \frac{1}{2}$, for all $z \in \mathbb{R}^d$. This is expected because, in the extreme case where $K_N = N - 1$, $N_{K_N} (\mathcal{V}_N)$ is the complete graph and the statistic (1.3) is non-random and, hence, powerless.

Proposition 1 has several interesting consequences. To begin with, note that, when $K = O(1)$ is fixed, then $\eta_0(z)$ does not depend on $z$ and the right-hand side of formula (4.5) is 0, as shown earlier in theorem 3. The situation, however, is different when $K = K_N \rightarrow \infty$ grows with $N$. Even though the exact dependence of $d \{ z, N_{K_N} (\mathcal{V}_N) \}$ on $z$ and $K$ appears to be quite delicate, the right-hand side in formula (4.5) is expected to be non-zero, when $K = K_N \rightarrow \infty$ sufficiently fast, for instance, when $K_N \asymp N^\alpha$, for some $\alpha \in (0, 1]$. This is validated by the simulation results in Section 5, where we observe that the local power of the $K$-NN test increases with $K$ and eventually dominates other parametric and non-parametric tests. This makes the $K$-NN test (when $K$
grows with \( N \) desirable, both theoretically (non-zero Pitman efficiency) and in applications (easy computation and good finite sample power).

### 4.4. Cross-match test

Let \( W \) be the minimum non-bipartite matching graph functional as in definition 4. It is unknown whether \( W \) is stabilizing (Penrose and Yukich, 2003), and so theorem 3 cannot be applied to compute the asymptotic efficiency of the CM test. However, in this case, corollary 1 can be used directly to derive the following corollary.

**Corollary 2.** Let \( \{ \mathbb{P}_\theta \}_{\theta \in \Theta} \) be a parametric family of distributions in \( \mathbb{R}^d \) satisfying assumption 1 at \( \theta = \theta_1 \in \Theta \). Then the asymptotic efficiency of the CM test (1.9), under the Euclidean distance, for the testing problem (2.1), is 0, i.e. \( \text{AE}(W) = 0 \).

**Proof.** Let \( N \) be even and \( \forall \theta \in \{ V_1, V_2, \ldots, V_N \} \) IID \( \mathbb{P}_{\theta_1} \). In this case, \( |E\{W(V_N)\}| = N/2 \) and \( |T_2\{W(V_N)\}| = 0 \). Therefore, \( (W, \mathbb{P}_{\theta_1}) \) satisfies the (2, 0) undirected variance condition (3.6). The normality condition 1 holds, since \( d\{V_i, W(V_N)\} = 1 \) for all \( V_i \in V_N \). This also implies that the undirected covariance condition holds with the constant function \( \lambda(z) = 2 \). The result then follows by corollary 1.

### 4.5. Depth-based tests

Let \( Z_N = \mathcal{D}_{N_1} \cup \mathcal{D}_{N_2} \) be the pooled sample, and \( F_{N_1} \) the empirical distribution of \( \mathcal{D}_{N_1} \). The two-sample test based on a depth function \( D \) (1.12) rejects for large values of \( \left| E\{G_D(Z_N)^2\} \right| \), where the graph \( G_D(Z_N) \) has vertex set \( \mathcal{D}_{N_1} \) with a directed edge \( (Z_i, Z_j) \) whenever \( D(Z_i, F_{N_1}) \leq D(Z_j, F_{N_1}) \). If the depth function \( D(X, F) \), where \( X \sim F \), has a continuous distribution then \( G_D(Z_N) \) is a complete graph \( \left| E\{G_D(Z_N)\} = N(N - 1)/2 \right| \) with directions on the edges depending on the relative ordering of the depth of the two end points.

**Definition 8.** Let \( F \) be a distribution function in \( \mathbb{R}^d \) with empirical distribution function \( F_{N_1} \). A depth function \( D \) is said to be good with respect to \( F \) if

- \( (a) \) for \( X \sim F \), the distribution of \( D(X, F) \) is continuous,
- \( (b) \) \( \mathbb{P}\{Y_1 \leq D(Y, F) \leq Y_2\} \leq C|Y_1 - Y_2| \), for some constant \( C \) and any \( Y_1, Y_2 \in [0, 1] \), and
- \( (c) \) \( \sup_{x \in \mathbb{R}^d} |D(x, F_{N_1}) - D(x, F)| = o(1) \) almost surely and in expectation.

The standard depth functions, like that discussed in Section 1.2, satisfy the above conditions, for any continuous distribution function \( F \). The following result gives the asymptotic efficiencies of tests based on such depth functions. For this, let \( \{ \mathbb{P}_\theta \}_{\theta \in \Theta} \) be a parametric family of distributions in \( \mathbb{R}^d \) satisfying assumption 1 at \( \theta = \theta_1 \in \Theta \subseteq \mathbb{R}^p \). Moreover, let \( F_{\theta_1} \) be the distribution function of \( \mathbb{P}_{\theta_1} \).

**Theorem 4.** The asymptotic efficiency of the two-sample test (1.12) based on a good depth function \( D \) (with respect to \( F_{\theta_1} \)), for the testing problem (2.1), is

\[
\text{AE}(G_D) = -\sqrt{6r} \int \langle h, \nabla f(x|\theta_1) \rangle R(x, F_{\theta_1}) \, dx,
\]

where \( R(x, F_{\theta_1}) \) is as defined in expression (1.10).

### 4.6. Chen–Friedman test

Recently, Chen and Friedman (2017) proposed a modification of test statistic (1.6), which improves on the finite sample power of the FR and the K-NN tests, especially when the sample
Consider the parametric family \( \Theta \) of distributions in \( \mathbb{R}^d \) that are similar to the proof of theorem 1. For this, recall that, for \( \lambda \in \Theta \), the non-central \( \chi^2 \)-distribution with \( d \) degrees of freedom and non-centrality parameter \( \theta^T \theta \). The theorem is proved in the on-line appendix H.

**Theorem 5.** Let \( G \) be an undirected graph functional and \( \{P_\theta\}_{\theta \in \Theta} \) a parametric family of distributions in \( \mathbb{R}^p \) satisfying assumption 1 at \( \theta = \theta_1 \in \Theta \). If the pair \((G, P_{\theta_1})\) satisfies assumption 5, then the limiting power of the CF test, for the testing problem (2.1), is given by

\[
\lim_{N \to \infty} \mathbb{P}_{\theta_1 + h/\sqrt{N}} [S\{G(Z_N)\} > \chi^2_{2,1-\alpha}] = \mathbb{P} \left\{ \chi^2_{2,1-\alpha} > \chi^2_{2,1-\alpha} \right\},
\]

where

\[
\mu = \left( -p^2 q \int \langle h, \nabla f(z|\theta_1) \rangle \lambda(z) dz \right),
\]

\[
\Lambda = \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{12} & \lambda_{22} \end{pmatrix},
\]

with \( \lambda_{11} := p^2 \{(1-p^2)\gamma_0 + r(\gamma_1 - 2)\}, \lambda_{12} := -p^2 q^2 (\gamma_0 + 2\gamma_1 - 8) \) and \( \lambda_{22} := q^2 \{(1-q^2)\gamma_0 + r(\gamma_1 - 2)\}, \) whenever \( \Lambda \) is invertible.

### 5. Finite sample local power

In this section, we compare the power of the various tests against local alternatives in simulations. The CF test is computed by using the R package gTests (https://cran.r-project.org/web/packages/gTests/index.html) and the depth functions are computed by using the package fda.usc (https://cran.r-project.org/web/packages/fda.usc/index.html). Throughout our simulations the level of significance is set at \( \alpha = 0.05 \).

**5.1. Example 1 (normal location)**

Consider the parametric family \( P_{\theta} \sim N(\theta, I) \), for \( \theta \in \mathbb{R}^d \). Table 1 shows the empirical power (out
Table 1. Power of the various tests for the normal location problem across increasing dimensions when the respective means differ by $2 \cdot 1/\sqrt{N}$

| Dimension | FR  | HD  | MD  | CF  | $T^2$ |
|-----------|-----|-----|-----|-----|------|
| 4         | 0.11| 0.06| 0.03| 0.04| 0.35 |
| 10        | 0.06| 0.04| 0.04| 0.03| 0.49 |
| 20        | 0.1 | 0.05| 0.04| 0.04| 0.68 |
| 30        | 0.08| 0.11| 0.06| 0.07| 0.88 |
| 50        | 0.09| 0.07| 0.03| 0.06| 0.96 |
| 100       | 0.07| 0.06| 0.04| 0.09| 1    |
| 200       | 0.15| 0.08| 0.08| 0.09| 1    |
| 300       | 0.22| 0.04| 0.11| 0.13| 1    |

Fig. 2. Power of the various tests for the normal location problem in dimension $d = 10$ when the means differ by $\delta \cdot 1/\sqrt{N}$, as a function of $\delta$: $T^2$, FR, CF, HD, MD of 100 repetitions) of the FR test based on the MST, the test based on the HD, the test based on the MD, the CF test based on the MST and Hotelling's $T^2$-test, with $N_1 = 1000$ samples from $\mathbb{P}_0$ and $N_2 = 500$ samples from $\mathbb{P}_{2 \cdot 1/\sqrt{N}}$, across increasing dimensions. (Here, $N = N_1 + N_2 = 1500.$) Fig. 2 shows the empirical power (out of 100 repetitions) in dimension $d = 10$ of these tests, based on $N_1 = 1000$ samples from $\mathbb{P}_0$ and $N_2 = 500$ samples from $\mathbb{P}_{\delta \cdot 1/\sqrt{N}}$, over a grid of 20 values of $\delta$ in $[0, 3]$ (smoothed out by using the loess function in R). Table 1 and Fig. 2 show that Hotelling’s $T^2$-test, which is the most powerful test in this case, has the highest power. The powers of the FR test and CF test improve slightly with dimension but are generally low,
as predicted by the results above. In this case, the tests based on depth functions (the HD test and the MD test) also have low power (see remark G.1 in the on-line supplementary material).

5.2. Example 2 (spherical normal)
Consider the parametric family $\mathbb{P}_\sigma \sim N(0, \sigma^2 I)$, for $\sigma > 0$. As before, Table 2 shows the empirical power (out of 100 repetitions) of the various tests based on $N_1 = 1000$ samples from $\mathbb{P}_1$ and $N_2 = 500$ samples from $\mathbb{P}_{1+2/\sqrt{N}}$ across increasing dimensions, and Fig. 3 shows the empirical power (out of 100 repetitions) in dimension $d = 10$ of the various tests, based on $N_1 = 1000$ samples from $\mathbb{P}_1$ and $N_2 = 500$ samples from $\mathbb{P}_{1+\delta/\sqrt{N}}$, over a grid of 20 values of $\delta$ in $[0, 3]$.

Table 2. Power of the various tests across increasing dimensions in the spherical normal problem when the standard deviations differ by $2/\sqrt{N}$

| Dimension | $T^2$ | CovTest | $\mathrm{CF}$ |
|-----------|-------|---------|--------------|
| 4         | 0.07  | 0.03    | 0.05         |
| 10        | 0.07  | 0.12    | 0.11         |
| 20        | 0.24  | 0.06    | 0.28         |
| 30        | 0.34  | 0.08    | 0.18         |
| 50        | 0.59  | 0.09    | 0.65         |
| 100       | 0.72  | 0.22    | 0.13         |
| 200       | 0.95  | 0.51    | 0.05         |
| 300       | 1     | 0.71    | 0.08         |

Fig. 3. Power of the various tests for the normal scale problem in dimension $d = 10$ when the standard deviations differ by $\delta/\sqrt{N}$, as a function of $\delta$: $T^2$, CovTest, $\mathrm{FR}$, $\mathrm{CF}$, HD, MD.
Here, the HD test performs very well across dimensions. The MD test also performs well for small to moderate dimensions but starts to lose power for higher dimensions. In contrast, the power of the FR and the CF tests are small in low dimension; however, quite interestingly, the power increases substantially with dimension, paralleling the HD test, with the CF test generally more powerful than the FR test. This supports the findings in Chen and Friedman (2017) where the FR and CF tests also exhibit high power as the dimension increases, in finite sample simulations. It is phenomena like this that make tests based on geometric graphs, such as the FR test and the CF test, particularly attractive for modern statistical applications.

This remarkable blessing of dimensionality can be mathematically explained as follows: even though tests based on geometric graphs have no power in the $O(N^{-1/2})$ scale, the detection threshold of these tests for the spherical normal problem (and more general scale alternatives) is expected to be around $\Theta(N^{-1/2+1/d})$. (This has been proved recently by Bhattacharya (2018) for the test based on the K-NN graph.) Note that this threshold becomes increasingly close to the parametric detection rate of $N^{-1/2}$ as $d$ increases, and, as a result, these tests attain high power as the dimension increases for scale problems.

Table 2 and Fig. 3 also show the power of the Hotelling $T^2$-test which, as expected, performs poorly for the scale problem, and CovTest, which is the parametric likelihood ratio test for testing the equality of two normal covariance matrices. This rejects for large values $N \log |\hat{\Sigma}_0| - N \log |\hat{\Sigma}_1| - N \log |\hat{\Sigma}_2|$, where $\hat{\Sigma}_0$, $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ are the maximum likelihood estimators of the covariance matrix of the whole data, the sample $\mathcal{F}_{N_1}$ and the sample $\mathcal{Y}_{N_2}$, respectively. CovTest performs quite poorly, as already observed in Chen and Friedman (2017) and Friedman and Rafsky (1979), which is expected because it must estimate an increasing number of parameters as the dimension increases and does not take into account the spherical structure of the covariance matrix.

### 5.3. Example 3 (log-normal location)

Consider the parametric family $\mathbb{P}_\theta \sim \exp\{N(\theta, I)\}$ for $\theta \in \mathbb{R}^d$, where the exponent is taken coordinatewise. As before, Table 3 shows the empirical power (out of 100 repetitions) of the various tests based on $N_1 = 1000$ samples from $\mathbb{P}_0$ and $N_2 = 500$ samples from $\mathbb{P}_d \cdot 1/\sqrt{N}$ across increasing dimensions, and Fig. 4 shows the empirical power (out of 100 repetitions) in dimension $d = 10$ of the various tests, based on $N_1 = 1000$ samples from $\mathbb{P}_0$ and $N_2 = 500$ samples from $\mathbb{P}_d \cdot 1/\sqrt{N}$, over a grid of 20 values of $\delta$ in $[0, 3]$. Changing the normal mean changes the log-normal distribution both in location and in scale. In this case, the HD test is powerless (see remark G.3 in the on-line supplementary information), but the test based on the MD performs very well, outperforming

| Dimension | $FR$ | $HD$ | $MD$ | $CF$ | $T^2$ |
|-----------|------|------|------|------|------|
| 4         | 0.05 | 0.02 | 0.11 | 0.06 | 0.17 |
| 10        | 0.05 | 0.03 | 0.34 | 0.03 | 0.24 |
| 20        | 0.08 | 0.09 | 0.45 | 0.08 | 0.48 |
| 30        | 0.19 | 0.07 | 0.65 | 0.13 | 0.51 |
| 50        | 0.37 | 0.06 | 0.83 | 0.47 | 0.81 |
| 100       | 0.58 | 0.07 | 0.91 | 0.58 | 0.94 |
| 200       | 0.74 | 0.07 | 1    | 0.61 | 1    |
| 300       | 0.73 | 0.05 | 1    | 0.53 | 1    |
Hotelling’s $T^2$ when the dimension increases. The FR and the CF tests have low power in small dimensions, but the power improves with dimension, for reasons that are similar to that in the spherical normal problem (recall example 2 above).

The examples show that, when the dimension is large, the FR and the CF tests can effectively detect two distributions, unless the alternative is location only. Moreover, both these tests can be computed very efficiently, which makes them especially useful in applications. Moreover, proposition 1 suggests that the test based on the $K$-NN graph can be powerful against $O(N^{-1/2})$ alternatives, when $K$ grows with $N$ sufficiently fast. We illustrate this result in the following example.

5.4. Example 4 (dependence on $K$ in the $K$-NN test)

To understand how the power of the $K$-NN test depends on $K$, we consider the log-normal location family $P_\theta \sim \exp\{N(\theta, \Sigma)\}$, where $\theta \in \mathbb{R}^{10}$ and $\Sigma$ is known.

(a) Fig. 5(a) shows the empirical power (out of 100 repetitions) in the independent case ($\Sigma = I$) of the $K$-NN test for various values of $K$, and the power of Hotelling’s $T^2$-test, the HD test and the MD test, based on $N_1 = 1000$ samples from $P_0$, and $N_2 = 800$ samples from $P_{\delta 1/\sqrt{N}}$, over a grid of 20 values of $\delta$ in $[0, 3]$. For small values of $K$, the $K$-NN test has low power. However, as $K$ increases, the power increases, and eventually it dominates all the other tests.

(b) Fig. 5(b) shows the empirical power of the various tests when $\Sigma = I + 11'$: a rank 1 perturbation of the identity matrix. In this case, the co-ordinates of the log-normal distribution are dependent. In this case, even though the overall power of all the tests is much lower, the $K$-NN test dominates all the other tests, for $K$ sufficiently large.
Fig. 5. Power in the log-normal family in dimension 10 where the means of the respective normal distributions differ by $\delta \cdot \frac{1}{\sqrt{N}}$, as a function of $\delta$ (in (a) $\mathcal{P}_{\theta} \sim \exp \{N(\theta, I)\}$, the log-normal distribution has independent co-ordinates, and in (b) $\mathcal{P}_{\theta} \sim \exp \{N(\theta, \Sigma)\}$, where $\Sigma = I + 11'$, the co-ordinates are dependent): ---, 5-NN; --, 20-NN; ---, 100-NN; ---, 600-NN; ---, 900-NN; ---, $T^2$; ---, HD; ---, MD
More simulations showing the power of the $K$-NN test are given in the on-line appendix I. These experiments show that the $K$-NN test is powerful against local alternatives when $K$ grows with $N$ (especially when $K = \alpha N$, for some $\alpha \in (0, 1)$), which supports the result in proposition 1 and illustrates the advantage of using dense geometric graphs. Although the computation cost for the $K$-NN test increases with $K$, it is always polynomial in $N$, $K$ and the dimension $d$, making it far more efficient than tests based on depth functions. This makes the $K$-NN test desirable, both theoretically (non-trivial efficiency) and in applications (easy computation and good finite sample power).

6. Application to sensorless drive diagnosis data set

In this section we compare the performances of the tests based on the MST (1.7) and the HD (1.14) on the sensorless drive diagnosis data set. (The data can be freely downloaded from the University of California, Irvine’s, machine learning repository: https://archive.ics.uci.edu/ml/datasets/Dataset+for+Sensorless+Drive+Diagnosis.) This data set was used by Bayer et al. (2013) for sensorless diagnosis of an autonomous electric drive train, which is composed of a synchronous motor and several attached components, like bearings, axles and a gear box. Damage to the drive causes severe disturbances and increases the risk of encountering breakdown costs. Monitoring the condition for such applications usually requires additional sensors. Sensorless drive diagnosis instead directly uses the phase currents of the motor for determining the performance of the entire drive unit.

In the data collected, the drive train had intact and defective components, and current signals were measured with a current probe and an oscilloscope on two phases under 11 different operating conditions; this means by different speeds, load moments and load forces. Thus, the data set consists of 11 different classes; each class consists of 5319 data points of dimension 48. The 48 features were extracted by using empirical mode decomposition of the measured signals. The first three intrinsic mode functions of the two phase currents and their residuals were used and broken down into subsequences. For each of this subsequences, the statistical features mean, standard deviation, skewness and kurtosis were calculated.

To detect the defects two-sample tests were performed on the $55 = \binom{11}{2}$ pairs of data sets. The goal is to investigate which of the tests can successfully detect the defect, i.e. reject the null hypothesis. Table 4 shows the $p$-values of the tests based on the MST and the HD (com-

| Pairs | Test | PCA1         | PCA2         | PCA3         | PCA48         |
|-------|------|--------------|--------------|--------------|--------------|
| (1, 2) | HD   | $1.498 \times 10^{-14}$ | $1.877 \times 10^{-14}$ | $1.204 \times 10^{-12}$ | 0             |
|       | MST  | 0.055        | 0.0256       | 8.766 $\times 10^{-5}$ | 1.615 $\times 10^{-7}$ |
| (1, 6) | HD   | $2.903 \times 10^{-11}$ | 9.47 $\times 10^{-5}$ | 0.0011       | 5.598 $\times 10^{-8}$ |
|       | MST  | 0.397        | 0.618        | 0.124        | 8.85 $\times 10^{-9}$ |
| (2, 6) | HD   | 0.038        | 0.266        | 0.166        | 0.002        |
|       | MST  | 0.676        | 0.481        | 0.457        | 0.0005       |
| (4, 9) | HD   | $1.366 \times 10^{-7}$ | 0.0117       | 0.3743       | 3.577 $\times 10^{-9}$ |
|       | MST  | 0.374        | 0.841        | 0.222        | 1.345 $\times 10^{-5}$ |
puted by using the mdepth.TD function in the R package fda.usc) for four such pairs. For demonstration the data were projected onto the first principal component PCA1, the first two principal components PCA2 and the first three principal components PCA3, and PCA48 is the whole data set. The results show that, in all the four pairs, the HD test has smaller $p$-values than the MST test for the first three principal components. In particular, the HD test performs significantly better in one dimension (PCA1). Even for higher principal components the HD test rejects at the 5% level more often than does the MST, illustrating that it is more sensitive to detecting local changes compared with the MST, as shown in Sections 4.1 and 1.2. Both tests reject at the 5% level for the whole data set, supporting the hypotheses that sensorless drive diagnosis is possible for the four pairs of defects that were considered above.

7. Finite sample power in the high dimensional regime

We conclude with a simulation study of the finite sample power of the tests that were described above, when the sample size is comparable with the dimension. In this high dimensional regime the performances of the graph-based tests are quite different. We illustrate the performances of the various tests in the three examples from Section 5. As before, the level of significance is set at $\alpha = 0.05$.

(a) $P_\theta \sim N(\theta, I)$, for $\theta \in \mathbb{R}^d$: the top part of Table 5 shows the empirical power (out of 100 repetitions) of the FR test based on the MST, the HD test, the MD test, the CF test based on the MST and Hotelling’s $T^2$-test, with $N_1 = 60$ samples from $P_0$ and $N_2 = 40$ samples from $P_{2.1/N}$ (here $N = N_1 + N_2 = 100$), for dimensions 10, 30, 50, 70 and 100. Here, the parametric Hotelling $T^2$-test has the highest power for dimensions up to 70 but is degenerate for dimension $d = 100$. The HD test has low power across dimensions. The MD test is powerless in low dimensions and degenerate for higher dimensions. Among the non-parametric tests, the FR test and the CF test are the most powerful as the dimension increases. For instance, in dimension $d = 100$, the FR and the CF tests dominate all the other tests.

Table 5. Power of the various tests across increasing dimensions with samples sizes $N_1 = 60$ and $N_2 = 40$ for the normal location problem with mean difference $0.2 \cdot 1$ and the spherical normal problem with scale difference $0.2$

| Dimension | FR  | HD  | MD  | CF  | $T^2$ | CovTest |
|-----------|-----|-----|-----|-----|------|---------|
| $N_1 = 60$ samples from $P_0$ and $N_2$ samples from $P_{0.2.1/N}$ in the normal location problem |
| 10     | 0.12 | 0.06 | 0.05 | 0.05 | 0.46 |
| 30     | 0.28 | 0.09 | 0.05 | 0.19 | 0.7  |
| 50     | 0.24 | 0.12 | 0.05 | 0.19 | 0.78 |
| 70     | 0.41 | 0.21 | 0.31 | 0.72 |
| 100    | 0.4  | 0.18 | 0.39 |      |

$N_1 = 60$ samples from $P_1$ and $N_2$ samples from $P_{1+2/\sqrt{N}}$ in the spherical normal problem

| Dimension | FR  | HD  | MD  | CF  | $T^2$ | CovTest |
|-----------|-----|-----|-----|-----|------|---------|
| 10     | 0.05 | 0.2  | 0.37 | 0.21 | 0.06 | 0.01   |
| 30     | 0.17 | 0.67 | 0.28 | 0.36 | 0.09 | 0.06   |
| 50     | 0.14 | 0.85 | 0.57 | 0.05 |      |        |
| 70     | 0.17 | 0.96 | 0.61 | 0.06 |      |        |
| 100    | 0.29 | 0.99 | 0.84 |      |      |        |
Table 6. Power of the various tests across increasing dimensions with samples sizes \( N_1 = 60 \) and \( N_2 = 40 \) for the log-normal location–scale problem, where the corresponding normal means differ by 0.2·1

| Dimension | FR | HD | MD | CF | \( T^2 \) |
|-----------|----|----|----|----|----------|
| 10        | 0.1 | 0.06 | 0.3 | 0.13 | 0.26 |
| 30        | 0.26 | 0.11 | 0.45 | 0.31 | 0.5 |
| 50        | 0.4 | 0.12 | — | 0.34 | 0.59 |
| 70        | 0.41 | 0.13 | — | 0.55 | 0.46 |
| 100       | 0.43 | 0.17 | — | 0.6 | — |

(b) \( \mathbb{P}_\sigma \sim N(0, \sigma I) \), for \( \sigma > 0 \): the bottom part of Table 5 shows the empirical power (out of 100 repetitions) of the various tests, based on \( N_1 = 60 \) samples from \( \mathbb{P}_1 \) and \( N_2 = 40 \) samples from \( \mathbb{P}_{1+2/\sqrt{N}} \), across the various dimensions. As expected, Hotelling’s \( T^2 \)-test and CovTest have no power in this case. The MD test has reasonable power for low dimensions but is powerless for dimensions greater than 50. The FR test has reasonable power which increases with the dimension. The HD test and the CF test are the two most powerful tests in this case, both of which have power improving with dimension. The HD test is slightly better than the CF test; however, the computation cost of the HD test is much higher.

(c) \( \mathbb{P}_\theta \sim \exp\{N(\theta, I)\} \) for \( \theta \in \mathbb{R}^d \), where the exponent is taken coordinate-wise: Table 6 shows the empirical power (out of 100 repetitions) of the various tests, based on \( N_1 = 60 \) samples from \( \mathbb{P}_0 \) and \( N_2 = 40 \) samples from \( \mathbb{P}_{2\cdot1/\sqrt{N}} \), across various dimensions. Here, the MD test and the HD test are powerless. The FR test has good power as the dimension increases; however, the CF test dominates all the other tests for moderate to high dimensions.

The experiments above show that tests based on geometric graphs, such as the FR and the CF tests, shine in moderate to high dimensions, making these procedures very useful for modern statistical applications.

8. Discussion

The asymptotic efficiency and finite sample results of the various graph-based two-sample tests obtained above illustrate the strengths and weaknesses of existing methods and show us how combinatorial properties of the underlying graph affect the performance of the associated two-sample test, which can help us to decide which test to use in practice.

Theoretical results show that tests based on sparse geometric graphs are powerless against \( O(N^{-1/2}) \) alternatives, when the dimension is fixed and the sample size is large. However, these tests exhibit good power in finite sample simulations which improves with increasing dimension, especially if the two distributions differ in scale. This is because the detection thresholds in scale problems for tests based on geometric graphs becomes increasingly close to the parametric detection rate of \( O(N^{-1/2}) \) as the dimension increases (recall the discussion in example 2). This blessing of dimensionality facilitates the application of these tests in modern statistical problems. Moreover, the asymptotic efficiency of these tests can be improved by increasing the density of the underlying geometric graph. For instance, the test based on the \( K \)-NN graph, where \( K \) grows polynomially with \( N \), has non-trivial Pitman efficiency (proposition 1) and often dominates the other tests (both parametric and non-parametric) in finite sample settings. Even
though the computational cost increases with $K$, it is still polynomial in the sample $N$, the number of nearest neighbours $K$ and the dimension $d$, which makes the $K$-NN test the front-runner for practical applications in the fixed dimension, large sample size regime. Another test which performs reasonably well in this regime is the test based on the MD; however, this becomes computationally unstable in large dimensions.

In the case that the dimension is comparable with the sample size, the situation is quite different. Here, simulation results in Section 7 show that tests based on geometric graphs, such as the FR and the CF tests, are the overall winner, reinforcing findings in Chen and Friedman (2017). The test based on the HD also performs particularly well in detecting pure scale changes, but it does not have good power in the other cases.

Acknowledgements

The author is indebted to his advisor Persi Diaconis for introducing him to graph-based tests and for his encouragement and support. The author thanks Ery Arias-Castro, Sourav Chatterjee, Probal Chaudhuri, Jerry Friedman, Shirshendu Ganguly, Anil Ghosh, Susan Holmes and David Siegmund for helpful comments. The author thanks Hao Chen and Nelson Ray for their help with the data sets. The author also thanks the Editor, the Associate Editor and the referees for providing many thoughtful comments, which greatly improved the quality of the paper.

References

Aldous, D. and Steele, J. M. (1992) Asymptotics for Euclidean minimal spanning trees on random points. *Probab. Theory Relat. Fields*, 92, 247–258.

Amaldi, E. and Kann, V. (1995) The complexity and approximability of finding maximum feasible subsystems of linear relations. *Theor. Comput. Sci.*, 147, 181–210.

Arias-Castro, E. and Pelletier, B. (2016) On the consistency of the crossmatch test. *J. Statist. Planng Inf.*, 171, 184–190.

Aslan, B. and Zech, G. (2005) New test for the multivariate two-sample problem based on the concept of minimum energy. *J. Statist. Comput Simuln*, 75, 109–119.

Baringhaus, L. and Franz, C. (2004) On a new multivariate two-sample test. *J. Multiv. Anal.*, 88, 190–206.

Bayer, C., Enge-Rosenblatt, O., Bator, M. and Mönks, U. (2013) Sensorless drive diagnosis using automated feature extraction, significance ranking and reduction. In *Proc. 18th Conf. Emerging Technologies and Factory Automation*, pp. 1–4.

Bhattacharya, B. B. (2018) Asymptotic distribution and detection thresholds for two-sample tests based on geometric graphs. *Preprint arXiv:1512.00384*. University of Pennsylvania, Philadelphia.

Bhattacharya, B. B., Diaconis, P. and Mukherjee, S. (2017) Universal Poisson and Normal limit theorems in graph coloring problems with connections to extremal combinatorics. *Ann. Appl. Probab.*, 27, 337–394.

Bickel, P. J. (1969) A distribution free version of the Smirnov two sample test in the $p$-variate case. *Ann. Math. Statist.*, 40, 1–23.

Biswas, M., Mukhopadhyay, M. and Ghosh, A. K. (2014) A distribution-free two-sample run test applicable to high-dimensional data. *Biometrika*, 101, 913–926.

Chen, H., Chen, X. and Su, Y. (2019) A weighted edge-count two sample test for multivariate and object data. *J. Am. Statist. Ass.*, 114, 1146–1155.

Chen, H. and Friedman, J. H. (2017) A new graph-based two-sample test for multivariate and object data. *J. Am. Statist. Ass.*, 112, 397–409.

Chen, H. and Zhang, N. (2015) Graph-based change-point detection. *Ann. Statist.*, 43, 139–176.

Chwialkowski, K., Ramdas, A., Sejdinovic, D. and Gretton, A. (2015) Fast two-sample testing with analytic representations of probability measures. In *Proc. A. Conf. Neural Information Processing Systems* (eds C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama and R. Garnett), pp. 1972–1980.

Efron, B. and Stein, C. (1981) The jackknife estimate of variance. *Ann. Statist.*, 9, 586–596.

Friedman, J. H. and Rafsky, L. C. (1979) Multivariate generalizations of the Wollowitz and Smirnov two-sample tests. *Ann. Statist.*, 7, 697–717.

Gretton, A., Borgwardt, K., Rasch, M., Scholkopf, B. and Smola, A. (2012) A kernel two-sample test. *J. Mach. Learn. Res.*, 16, 723–773.

Hall, P. and Tajvidi, N. (2002) Permutation tests for equality of distributions in high-dimensional settings *Biometrika*, 89, 359–374.
Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'A general asymptotic framework for distribution-free graph-based two-sample tests: supplementary materials'.