DISTRIBUTION-FREE MULTISAMPLE TEST BASED ON OPTIMAL MATCHING WITH APPLICATIONS TO SINGLE CELL GENOMICS

DIVYANSH AGARWAL†, SOMABHA MUKHERJEE†, BHASWAR B. BHATTACHARYA, AND NANCY R. ZHANG

Abstract. In this paper we propose a nonparametric graphical test based on optimal matching, for assessing the equality of multiple unknown multivariate probability distributions. Our procedure pools the data from the different classes to create a graph based on the minimum non-bipartite matching, and then utilizes the number of edges connecting data points from different classes to examine the closeness between the distributions. The proposed test is exactly distribution-free (the null distribution does not depend on the distribution of the data) and can be efficiently applied to multivariate as well as non-Euclidean data, whenever the inter-point distances are well-defined. We show that the test is universally consistent, and prove a distributional limit theorem for the test statistic under general alternatives. Through simulation studies, we demonstrate its superior performance against other common and well-known multisample tests. In scenarios where our test suggests distributional differences across classes, we also propose an approach for identifying which class or group contributes to this overall difference. The method is applied to single cell transcriptomics data obtained from the peripheral blood, cancer tissue, and tumor-adjacent normal tissue of human subjects with hepatocellular carcinoma and non-small-cell lung cancer. Our method unveils patterns in how biochemical metabolic pathways are altered across immune cells in a cancer setting, depending on the tissue location. All of the methods described herein are implemented in the R package multicross.

There are k samples of multidimensional probability distributions

1. Introduction

Given K multivariate probability distributions $F_1, F_2, \ldots, F_K$, the *K-sample problem* is to test the hypotheses

$$H_0 : F_1 = \cdots = F_K \quad \text{versus} \quad H_1 : F_s \neq F_t, \quad \text{for some } 1 \leq s < t \leq K. \tag{1.1}$$

This is a classical problem in statistical inference which has been extensively studied in the parametric regime, where the distributions are assumed to have certain, low-dimensional functional forms. Parametric methods, however, often perform poorly for misspecified models and for high-dimensional problems, especially when the number of nuisance parameters is large. This necessitates the development of non-parametric methods, which make no distributional assumptions on the data, but are still powerful for a wide class of alternatives. Moreover, with the recent accumulation of high-dimensional and non-Euclidean data arising from genomics, social networks, bioinformatics, and finance, it is imperative to develop non-parametric methods which are computationally efficient, robust and applicable to the

Date: June 13, 2019.

†The first two authors contributed equally to the paper.

various kinds of modern data types. In this paper, we consider non-parametric tests for the K-sample problem, which are exactly *distribution-free*, that is, tests for which the null distribution does not depend on the underlying (unknown) distribution of the data. This property is particularly desirable, because such tests can be directly calibrated under the null irrespective of the distribution or type of the data, making them readily applicable in a wide range of problems.

Nonparametric testing of two multivariate distributions has a long history, which has spawned renewed interest in light of modern applications. For univariate data, there are several celebrated distribution-free two-sample tests such as the Kolmogorov-Smirnov maximum deviation test [42], the Wald-Wolfowitz runs test [44], and the Mann-Whitney rank-sum test [29] (see the textbook [15] for more on these tests). Efforts to generalize these methods to higher dimensions go back to Weiss [45] and Bickel [7]. Friedman and Rafsky [13] proposed the first computationally efficient 2-sample test, which applies to high-dimensional data. The Friedman-Rafsky test, which can be viewed as a generalization of the univariate runs test, computes the Euclidean minimal spanning tree (MST)[1] of the pooled sample, and rejects the null if the number of edges with endpoints in different samples is small. Variants of this test based on nearest-neighbor graphs were considered by Henze [20] and Schilling [41]. Recently, Chen and Friedman [9] suggested novel modifications of this method for high-dimensional and object data, and Chen et al. [10] proposed new and powerful tests to deal with the issue of unequal sample sizes. Asymptotic properties of these tests can be studied in a general asymptotic framework introduced in [6]. Another computationally efficient 2-sample test based on the concept of multivariate ranks, defined using optimal transport, was recently proposed by Ghosal and Sen [14].

Even though many of these methods can be effectively used in high-dimensional problems, none of them inherit the exact distribution-free property of the univariate tests. A breakthrough in this direction was made relatively recently by Rosenbaum [39], who proposed the *crossmatch test*, a multivariate two-sample test based on the minimum non-bipartite matching (Definition 2.1) of the pooled sample. This test is exactly distribution-free in finite samples and computationally efficient (the test statistic can be computed in time which is polynomial in both the number of samples and the dimension of the data), making it particular attractive for high-dimensional applications. This test has also found many interesting applications in causal inference, especially in assessing balance between covariates in a treatment group and a matched control group [11, 18, 19, 31]. More recently, Biswas et al. [8] proposed another two-sample test based on Hamiltonian cycles, which is also distribution-free in finite samples. However, unlike the minimum non-bipartite matching, computing the minimum weight Hamiltonian path is NP-hard, making this test computationally prohibitive beyond small sample sizes.

Here, we study nonparametric distribution-free tests for the general K-sample problem (1.1). As expected, this problem is well-understood in dimension 1. Mood [32] considered the K-sample generalization of the runs test, and Kruskal and Wallis [24, 25] derived the

[1]Given a finite set $S \subset \mathbb{R}^d$, the *minimum spanning tree* (MST) of S is a connected graph with vertex-set S and no cycles, which has the minimum weight, where the weight of a graph is the sum of the distances of its edges.

K-sample analogue of the Mann-Whitney test, both of which are distribution-free. Our interest is in devising efficient distribution-free methods, which are powerful for a wide range of alternatives in arbitrary dimensions. We are motivated by applications in high-throughput biological experiments, where the multisample problem often arises. For instance, it is not uncommon to examine the distribution of a set of high-dimensional features across various models or conditions. Recently, single cell technologies have made it possible to profile the expression of tens of thousands of genes across thousands of cells. The cells might belong to different subtypes, where the K types are characterized based on some functional or morphological parameter. In this setting, determining whether the expression of a set of genes, corresponding to a particular biochemical pathway or function, belong to the same or different distribution across the K groups can yield insights into cellular processes and the underlying biological system.

Even though the high-dimensional multisample problem manifests itself in various modern applications, methodological progress to address this problem has been limited. A nearest-neighbor based test for testing the equality of multiple distributions with categorical components was considered in [33]. Recently, Petrie [36] considered the direct generalization of the Friedman-Rafsky and crossmatch tests, which counts the number of edges across the different samples in the geometric graph (MST or matching) constructed on the pooled sample. However, this test tends to lose power with increases in dimension and/or the number of classes, and the mathematical properties of this test have not been investigated.

We propose a new graph-based multisample test based on optimal matchings. To compute the test statistic, we construct the minimum non-bipartite matching of the pooled sample, compute the $K \times K$ matrix of cross-counts (the (s,t)-th element of this matrix is the number of edges in the matching from sample s to sample t), then combine these counts using their Mahalanobis distance. We show that this test is exactly distribution-free under the null (Proposition 2.1), derive its asymptotic null distribution (Theorem 2.3), and demonstrate its consistency under general alternatives (Theorem 2.4). We also prove a conditional central limit theorem (CLT) of the entire vector of cross-counts under the alternative (Section 4). More precisely, we show that the cross-counts, centered by their means, conditional on the pooled sample and scaled appropriately, converge in distribution to a multivariate normal distribution under general alternatives. As a consequence, we obtain a distributional limit theorem for the proposed test statistic under the alternative, which to the best of our knowledge, is a new result even for the 2-sample case (where the proposed test statistic is equivalent to Rosenbaum's 2-sample crossmatch test). Therefore, this result adds to our theoretical understanding of the crossmatch and general matching-based tests, which includes the proposed method and the baseline generalizations considered in [36]. In Section 3 we compare the power of our test with other existing tests for various alternatives. Our experiments demonstrate that the proposed method outperforms other relevant parametric and non-parametric tests in a diverse set of simulation settings.

Lastly, we demonstrate the broad potential of our method on real data examples from the single cell biology domain, where we use our test to investigate whether the activities of biological pathways across K closely-related cell types are conserved. Single cell data is sparse

count data with many zeros, so it is difficult to transform it to conform to normal distribution assumptions. Moreover, gene expression has complex correlation structure and thus any parametric model would require many nuisance parameters. Therefore, our novel crossmatch-based method, being nonparametric and distribution free, is especially fitting. We show the utility of our proposed test in the comparison of the distribution of gene sets (as a proxy for examining biochemical pathways) across cell populations using single cell RNA-sequencing (scRNA-seq) data (Section 5). Our method successfully recapitulates known biology by detecting differential distribution of metabolic pathways which are known to be disparate across T cell subtypes. We also discovered pathways such as purine metabolism that differed across T cell subtypes, irrespective of whether the sequenced cells were obtained from blood, liver cancer tissue, lung cancer tissue, or tumor-adjacent normal tissue. Moreover, we ascertain that our method can be used to narrow down gene sets that are differentially distributed across cell types, thereby facilitating its use in single cell clustering and visualization algorithms. Altogether, as illustrated through real case studies, our proposed method combines statistical innovation to answer practical, scientifically relevant questions.

2. Multisample distribution-free tests based on optimal matching

Recall the K-sample hypothesis (1.1), and assume that for every $s \in [K] := \{1, 2, \ldots, K\}$, we are given N_s i.i.d. observations $\boldsymbol{X}^{(s)} := \{X_1^{(s)}, X_2^{(s)}, \ldots, X_{N_s}^{(s)}\}$ from the distribution F_s. In this section we describe a novel distribution-free, computationally efficient K-sample test, based on the minimum non-bipartite matching, which can be readily used for data in arbitrary metric spaces, such as high dimensional data, functional data, and object data.

We begin with the formal definition of a minimum non-bipartite matching. For simplicity, we will assume throughout that the total number of samples $N = \sum_{s=1}^{K} N_s := 2I$ is even; otherwise, we can add or delete a sample point to make it even.

Definition 2.1. 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 partition of the elements of S into $I = \frac{N}{2}$ non-overlapping sets of size 2 each, that is,

$$S = S_1 \bigcup S_2 \bigcup \cdots \bigcup S_I, \quad \text{where } |S_a| = 2 \text{ and } S_a \cap S_b = \varnothing, \text{ for } 1 \leq a \neq b \leq I.$$

The *weight* of a non-bipartite matching is the sum of the distances between the I matched pairs. A *minimum non-bipartite matching* of S is a matching which has the minimum weight over all matchings of S. (In case of multiple minimizers any one of them can be chosen.) The *minimum non-bipartite matching graph* $\mathscr{G}(S) = (V(\mathscr{G}(S)), E(\mathscr{G}(S)))$ is the graph with vertex set $V(\mathscr{G}(S)) = S$ and edge set $E(\mathscr{G}(S)) = \{S_1, S_2, \ldots, S_I\}$, consisting of the I disjoint pairs.

The 2-sample *cross-match* (CM) test proposed by Rosenbaum [39] rejects the null hypothesis in (1.1) for small values of

What then is the difference between Friedman Rafsky test and MCM?

$$R_{2,N} := \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \mathbf{1}\{(X_i^{(1)}, X_j^{(2)}) \in E(\mathscr{G}(\mathscr{X}))\}, \tag{2.1}$$

where $\mathscr{X} := \{\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}\}$ is the pooled sample. Note that the statistic $R_{2,N}$ counts the number of matched edges in the pooled sample with one end point in sample 1 and the other endpoint in sample 2 (the "cross-matches"), which is expected to be small when the two distributions are different.

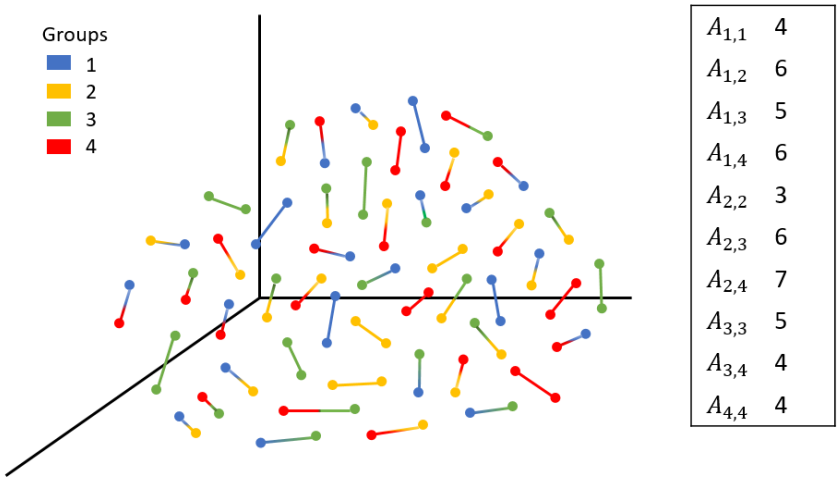


FIGURE 1. Illustration of a minimum non-bipartite matching for 100 points in 3-dimensions with 4 classes, and the different cross/pure counts.

Here, we consider two generalizations of the CM statistic when there are more than 2 samples, based on the minimum non-bipartite matching of the pooled sample. In this case, denoting the pooled sample by $\mathscr{X} := \{\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \ldots, \boldsymbol{X}^{(K)}\}$, we define, for $1 \le s \ne t \le K$, the (s,t)-*cross count* as the number of matched edges in the pooled sample with one endpoint in sample s and the other end-point in sample t, which is denoted by

$$a_{st}(\mathscr{G}(\mathscr{X})) := \sum_{i=1}^{N_s} \sum_{j=1}^{N_t} \mathbf{1}\{(X_i^{(s)}, X_j^{(t)}) \in E(\mathscr{G}(\mathscr{X}))\}. \tag{2.2}$$

For each $s \in [K]$, we also define the (s,s)-*pure count* as the number of matched edges in the pooled sample with both endpoints in sample s, which is denoted by

$$a_{ss}(\mathscr{G}(\mathscr{X})) := \frac{1}{2} \sum_{i=1}^{N_s} \sum_{j=1}^{N_s} \mathbf{1}\{(X_i^{(s)}, X_j^{(s)}) \in E(\mathscr{G}(\mathscr{X}))\}. \tag{2.3}$$

The matrix of cross/pure counts $\boldsymbol{A}_N(\mathscr{G}(\mathscr{X})) = (a_{st}(\mathscr{G}(\mathscr{X})))_{1\le s,t\le K}$ will be referred to as the *count matrix*. (Hereafter, we will drop the dependence on $\mathscr{G}(\mathscr{X})$ from $\boldsymbol{A}_N(\mathscr{G}(\mathscr{X}))$ and its elements $a_{st}(\mathscr{G}(\mathscr{X})$, whenever it is clear from the context.) Figure 1 shows the cross/pure counts for a sample of 100 points in 3-dimensions with 4 classes.

We show below in Proposition 2.1, that the joint distribution of the elements of the count matrix is exactly distribution-free under the null. Therefore, we can construct distribution-free tests for (1.1) by considering real-valued functions of the cross-matrix, as follows:

- The *multisample crossmatch* (MCM) test rejects the null in (1.1) for small values of the statistic

$$R_{K,N} := \sum_{1\le s<t\le K} a_{st}(\mathscr{G}(\mathscr{X})). \tag{2.4}$$

This direct generalization of the 2-sample cross-match statistic (recall (2.1)), which counts the total number of cross edges in the matching constructed using the pooled sample, was considered in [36]. Here, we derive its asymptotic properties, and use it as a baseline for empirical comparisons.

- There are many natural multivariate alternatives where the MCM test described above performs poorly, especially when the dimension is large and the number of groups is big. To circumvent this issue, we propose a new test statistic based on the Mahalanobis distance of the observed cross-counts, which rejects the null for large values of

Need explanation for this

$$S_{K,N} := (\underline{\boldsymbol{A}}_N - \mathbb{E}_{H_0}\underline{\boldsymbol{A}}_N)^\top \mathrm{Cov}_{H_0}^{-1}(\underline{\boldsymbol{A}}_N)\,(\underline{\boldsymbol{A}}_N - \mathbb{E}_{H_0}\underline{\boldsymbol{A}}_N)\,, \tag{2.5}$$

where $\underline{\boldsymbol{A}}_N$ is the vector of length $\binom{K}{2}$ corresponding to the cross-counts (the upper-triangular part of $\boldsymbol{A}_N$),[2] and $\mathbb{E}_{H_0}(\underline{\boldsymbol{A}}_N)$ and $\mathrm{Cov}_{H_0}(\underline{\boldsymbol{A}}_N)$ denote the mean and the covariance matrix of $\underline{\boldsymbol{A}}_N$ under the null hypothesis, respectively (exact formulas are given below in Proposition 2.2 and the invertibility of $\mathrm{Cov}_{H_0}(\underline{\boldsymbol{A}}_N)$ is proved in Lemma C.1). We refer to this test as the *multisample Mahalanobis crossmatch* (MMCM) test.[3] Note that adjusting by the sample covariance matrix brings the cross-counts in the same scale, which makes $S_{K,N}$ a more appropriate measure of the centrality of the empirical cross-counts, leading to significant power improvements when K becomes large.

2.1. **Exact Null Distribution.** The following proposition shows that the joint distribution of the elements of the count matrix is distribution-free under the null, that is, it does not depend on the unknown distribution $F_1 = \cdots = F_K$.

Proposition 2.1. *Let $\boldsymbol{A}_N = ((a_{st}))_{1\le s,t\le K}$ be as defined in (2.2) and (2.3). Then*

$$\mathbb{P}_{H_0}\left(\boldsymbol{A}_N = \boldsymbol{b}\middle|\mathscr{X}\right) = \frac{1}{\binom{N}{N_1,\ldots,N_K}} \cdot \frac{2^{\sum_{1\le s<t\le K} b_{st}} I!}{\prod_{1\le s\le t\le K} b_{st}!}, \quad \text{for} \quad \boldsymbol{b} = ((b_{st}))_{1\le s,t\le K} \in \mathscr{B}, \tag{2.6}$$

where $\mathscr{B}$ is the set of all symmetric $K\times K$ matrices $\boldsymbol{b} = ((b_{st}))$ with non-negative integer entries, satisfying $2b_{ss} + \sum_{t\neq s} b_{st} = N_s$, for all $s \in [K]$. As a consequence, the statistics $R_{K,N}$ and $S_{K,N}$ defined above, are distribution-free under H_0.

[2] More formally, $\underline{\boldsymbol{A}}_N := (a_{12},\ldots,a_{1K},a_{23},\ldots,a_{2K},\ldots,a_{K-2,K-1},a_{K-2,K},a_{K-1,K})^\top$, the vector obtained by concatenating the rows of $\boldsymbol{A}_N$ in the upper triangular part.

[3] When $K = 2$ (the two-sample problem), the tests based on $R_{2,N}$ and $S_{2,N}$ are equivalent: In this case, the vector $\underline{\boldsymbol{A}}_N$ has only 1 element which is the number of cross-matches a_{11}, and (2.5) simplifies to $S_{2,N} = \frac{(R_{2,N}-\mathbb{E}_{H_0}(R_{2,N}))^2}{\mathrm{Var}_{H_0}(R_{2,N})}$, which is the square of the standardized CT statistic (2.1).

Proof. Note that

$$2a_{ss} + \sum_{t \neq s} a_{st} = N_s, \quad \text{for each } s \in [K],$$

since all edges in the graph $\mathscr{G}(\mathscr{X})$ are disjoint, and each (s,s)-edge has both endpoints in sample s and each (s,t)-edge, where $s \neq t$, has one of its endpoints in sample s. Therefore, the distribution of the cross-count matrix $\boldsymbol{A}_N$ is supported on the set $\mathscr{B}$ defined above.

Now, given $\boldsymbol{b} \in \mathscr{B}$ and the pooled sample $\mathscr{X}$, there are

$$\frac{2^{\sum_{1 \le s < t \le K} b_{st}} I!}{\prod_{1 \le s \le t \le K} b_{st}!}$$

ways of forming the classes $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(K)}$, such that $a_{st} = b_{st}$ for all $s, t \in [K]$. This comes from first assigning the I matched pairs such that there are b_{st} pairs corresponding to the (s,t)-counts, for $1 \le s \le t \le K$, in $\frac{I!}{\prod_{1 \le s \le t \le K} b_{st}!}$ ways, and then, for each of the b_{st} cross-matched pairs, assigning either one of the end points s or t in $2^{b_{st}}$ ways, for $1 \le s < t \le K$. Since, the random vector $(X_1^{(1)}, \ldots, X_{N_1}^{(1)}, \ldots, X_1^{(K)}, \ldots, X_{N_K}^{(K)})$ is exchangeable under H_0, each of these classifications has probability $\binom{N}{N_1, \ldots, N_K}^{-1}$. Hence, (2.6) follows.

Note that the RHS of (2.6) does not depend on $\mathscr{X}$, which implies $\mathbb{P}_{H_0}(\boldsymbol{A}_N = \boldsymbol{b} | \mathscr{X}) = \mathbb{P}_{H_0}(\boldsymbol{A}_N = \boldsymbol{b})$. Therefore, the statistics $R_{K,N}$ and $S_{K,N}$, which are functions of the matrix $\boldsymbol{A}_N$, are distribution-free under H_0. □

The mean and covariances of this distribution, which are required for computing the MMCM statistic, can be easily derived:

Proposition 2.2. *Let $\underline{\boldsymbol{A}}_N$ be as in (2.5). The entries of the mean vector of $\mathbb{E}_{H_0} \underline{\boldsymbol{A}}_N$ are given by:*

$$\mathbb{E}_{H_0}(a_{st}) = \begin{cases} \frac{N_s N_t}{N-1} & \text{if} \quad s < t, \\ \frac{N_s(N_s-1)}{2(N-1)} & \text{if} \quad s = t. \end{cases} \tag{2.7}$$

The entries of the covariance matrix $\mathrm{Cov}_{H_0}(\underline{\boldsymbol{A}}_N)$ are as follows:

- *If $1 \le s_1 \neq s_2 \le K$,* $\mathrm{Var}_{H_0}(a_{s_1 s_2}) = \frac{N_{s_1} N_{s_2} (N_{s_1}-1)(N_{s_2}-1)}{(N-1)(N-3)} + \frac{N_{s_1} N_{s_2}}{N-1} \left(1 - \frac{N_{s_1} N_{s_2}}{N-1}\right)$.
- *If $1 \le s_1 \neq s_2 \neq s_3 \le K$,* $\mathrm{Cov}_{H_0}(a_{s_1 s_2}, a_{s_1 s_3}) = \frac{N_{s_1}(N_{s_1}-1) N_{s_2} N_{s_3}}{(N-1)(N-3)} - \frac{N_{s_1}^2 N_{s_2} N_{s_3}}{(N-1)^2}$.
- *If $1 \le s_1 \neq s_2 \neq s_3 \neq s_4 \le K$,* $\mathrm{Cov}_{H_0}(a_{s_1 s_2}, a_{s_3 s_4}) = \frac{2 N_{s_1} N_{s_2} N_{s_3} N_{s_4}}{(N-1)^2 (N-3)}$.

The proof of the proposition is given in Appendix A.1. It follows by a direct combinatorial analysis and observing that, under the permutation null distribution, all possible $\binom{N}{N_1, N_2, \ldots, N_K}$ relabelings of the data are equally likely.

2.2. **Asymptotic Null Distribution.** In theory, the exact cutoff for the MMCM test can be obtained using the quantiles of the distribution in (2.6). Another alternative is to use the exchangeability of the data under H_0, and perform a permutation test. However, both these approaches are computationally cumbersome when the sample size increases. In this case, it is more convenient to use rejection regions based on the asymptotic null distribution. This

is derived in the theorem in the usual limiting regime where $N \to \infty$ such that

$$\left(\frac{N_1}{N}, \frac{N_2}{N}, \ldots, \frac{N_K}{N}\right) \to (p_1, p_2, \ldots, p_K) \in (0,1)^K, \tag{2.8}$$

where $\sum_{s=1}^K p_s = 1$.

Theorem 2.3. *Under the null H_0,*

$$\mathrm{Cov}_{H_0}^{-\frac{1}{2}}(\underline{\boldsymbol{A}}_N)\,(\underline{\boldsymbol{A}}_N - \mathbb{E}_{H_0}\underline{\boldsymbol{A}}_N) \xrightarrow{D} N_{\binom{K}{2}}(0, \mathrm{I}). \tag{2.9}$$

This implies, under H_0, the MMCM statistic $S_{K,N} \xrightarrow{D} \chi^2_{\binom{K}{2}}$, as $N \to \infty$, and the test with rejection region

$$\left\{S_{K,N} > \chi^2_{\binom{K}{2},1-\alpha}\right\}, \tag{2.10}$$

is asymptotically level α.[4]

The proof of the theorem is given in Appendix A.2. Note that, by Proposition 2.2, the elements of $\mathrm{Cov}_{H_0}(\underline{\boldsymbol{A}}_N)/N$ has a non-degenerate limit, that is, there is a $\binom{K}{2} \times \binom{K}{2}$ matrix $\boldsymbol{\Gamma}$, such that $\mathrm{Cov}_{H_0}(\underline{\boldsymbol{A}}_N)/N \to \boldsymbol{\Gamma}$. An application of the Slutsky's theorem and (2.9) then implies,

$$\frac{\underline{\boldsymbol{A}}_N - \mathbb{E}_{H_0}\underline{\boldsymbol{A}}_N}{\sqrt{N}} \xrightarrow{D} N_{\binom{K}{2}}(0, \boldsymbol{\Gamma}),$$

and the test with rejection region

$$\left\{\frac{(\underline{\boldsymbol{A}}_N - \mathbb{E}_{H_0}\underline{\boldsymbol{A}}_N)^\top \boldsymbol{\Gamma}^{-1} (\underline{\boldsymbol{A}}_N - \mathbb{E}_{H_0}\underline{\boldsymbol{A}}_N)}{N} > \chi^2_{\binom{K}{2},1-\alpha}\right\}, \tag{2.11}$$

is also asymptotically level α.

Note that Theorem 2.3 gives the asymptotic normality of entire cross-count vector, which implies the normality of any linear function of the cross-count vector, in particular, the MCM statistic (2.4) as well. More formally, (2.9) implies, under H_0,

$$Q_{K,N} := \frac{R_{K,N} - \mathbb{E}_{H_0}(R_{K,N})}{\sqrt{\mathrm{Var}_{H_0}(R_{K,N})}} \xrightarrow{D} N(0,1), \tag{2.12}$$

where, by Proposition 2.2, $\mathbb{E}_{H_0}(R_{K,N}) = \frac{\sum_{s<t} N_s N_t}{N-1}$ and

$$\mathrm{Var}_{H_0}(R_{K,N}) = \frac{G_1}{N-1}\left(1 - \frac{G_1}{N-1}\right) + \frac{G_1^2 - G_1 - 2G_2}{(N-1)(N-3)},$$

with $G_1 := \sum_{1 \le s < t \le K} N_s N_t$ and $G_2 := \frac{1}{2}\sum_{s=1}^K N_s(N - N_s)(N - N_s - 1)$. Therefore, the asymptotically level α test has rejection region $\{Q_{K,N} < z_\alpha\}$, where z_α is the α-th quantile of the standard normal distribution.

Remark 2.1. (Non-Euclidean data) As the conditional (permutation) null distribution is same as the unconditional distribution (by Proposition 2.1), it follows from the proof of

[4]For $p \ge 1$, $N_p(\boldsymbol{\mu}, \Sigma)$ denotes the multivariate normal distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^p$ and covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$. Moreover, χ^2_n denotes the chi-squared distribution with n degrees of freedom and $\chi^2_{n,1-\alpha}$ denotes the $(1-\alpha)$-th quantile of the χ^2_n distribution.

Theorem 2.3 that the asymptotic null distributions for $S_{K,N}$ and $Q_{K,N}$ obtained above hold verbatim for non-Euclidean spaces, as long as a similarity measure on the sample space can be defined. This is one of the highlights of tests based on inter-point distances, which makes them readily applicable for combinatorial and object data [9, 10]. Maa et al. [28] provided theoretical motivations for using tests based on inter-point distances, by showing that, under mild conditions, two multivariate distributions are equivalent if and only if the distributions of inter-point distances within each distribution and between the distributions are equivalent.

2.3. **Consistency.** In this section, we show the consistency of the tests discussed above. To this end, assume that the K distributions $F_1, F_2, \ldots, F_K$ have densities $f_1, f_2, \ldots, f_K$, respectively, with respect to the Lebesgue measure on $\mathbb{R}^d$. A test is said to be *universally consistent* for the hypothesis (1.1) if the power of the test converges to 1, whenever there exists $1 \le s \ne t \le K$ such that $f_s \ne f_t$ on a set of positive Lebesgue measure. Recently, Arias-Castro and Pelletier [3] showed that the 2-sample CM test is universally consistent. Their arguments can be easily adapted to show the universal consistency of the MCM and MMCM tests:

Theorem 2.4. *In the usual limiting regime* (2.8), $\frac{1}{N}\boldsymbol{A}_N \to \boldsymbol{H} = ((h_{st}))_{1\le s,t\le N}$ *almost surely, where*

$$h_{st} = \begin{cases} p_s p_t \int_{\mathbb{R}^d} \frac{f_s(z) f_t(z)}{\sum_{a=1}^K p_a f_a(z)}\, \mathrm{d}z & \text{if } s \ne t, \\ \frac{p_s^2}{2} \int_{\mathbb{R}^d} \frac{f_s^2(z)}{\sum_{a=1}^K p_a f_a(z)}\, \mathrm{d}z & \text{otherwise.} \end{cases} \tag{2.13}$$

This implies that the MCM test with rejection $\{Q_{K,N} < z_\alpha\}$ *and the MMCM test with rejection region* (2.11) *are universally consistent.*

The proof of the theorem is given in Appendix A.3. The limit in (2.13) implies that the MCM statistic (recall (2.4))

$$R_{K,N} \xrightarrow{a.s.} \sum_{1\le s<t\le K} h_{st} := H(f_1, f_2, \ldots, f_K) := \tfrac{1}{2} - \mathrm{tr}(\boldsymbol{H}). \tag{2.14}$$

The consistency of the MCM and the MMCM tests then follows from the fact that

$$H(f_1, f_2, \ldots, f_K) \le H(f, f, \ldots, f),$$

and equality holds if and only if $f_1 = f_2 = \cdots = f_K$ outside a set of Lebesgue measure 0 (details given Appendix A.3).

Remark 2.2. (Henze-Penrose divergence) In the case $K = 2$, the limiting constant h_{12} equals $1 - \delta(f_1, f_2)$, where

$$\delta(f_1, f_2) = \int \frac{p_1^2 f_1^2(x) + p_2^2 f_2^2(x)}{p_1 f_1(x) + p_2 f_2(x)} \mathrm{d}x,$$

is the well-known *Henze-Penrose divergence* between probability measures [16]. This quantity appears as the almost sure limit of a large class of graph-based 2-sample tests, which includes the Friedman-Rafsky test [21], the nearest-neighbor based tests [20], and the CM test [3], and has an interesting interpretation in terms of treatment-control assignment, using the

propensity score [19]. For general K, the limit h_{st} in (2.13) is a multi-sample generalization of the Henze-Penrose integral, which aggregated over $1 \leq s < t \leq K$ (as in the RHS of (2.14)), is a global measure of dissimilarity between the densities $f_1, f_2, \ldots, f_K$.

3. Power Comparisons

In this section, we illustrate the effectiveness of the tests described above by comparing their power with various other parametric and non-parametric tests, for several alternative hypotheses across different dimensions and number of groups. In Section 3.1 we illustrtate the advantage of using optimal matchings by comparing the performance of the MCM and MCMM tests with the multisample Friedman-Rafsky test (a natural generalization of the 2-sample Friedman-Rafsky test [13], where the optimal matching is replaced with the minimum spanning tree (MST)). In Section 3.2 we compare our matching based tests with other relevant parametric tests. Finally, in Section 3.3 we present an extensive comparison of the empirical power of the MCM and MCMM tests, across increasing dimensions (d) and group sizes (K). Additional simulations are given in Appendix D. Throughout, the nominal level of the tests are chosen to be 0.05.

3.1. Comparison with the MST.

Here, we compare the performance of the optimal matching based tests described above, with the test based on the MST. As in (2.4), a natural extension of the 2-sample Friedman-Rafsky test based on the MST, is the *multi-sample Friedman-Rafsky* test (MFRT), which rejects the null hypothesis for small values of the statistic:

$$T_{K,N} := \sum_{1 \leq s < t \leq K} \sum_{i=1}^{N_s} \sum_{j=1}^{N_t} \mathbf{1}\{(X_i^{(s)}, X_j^{(t)}) \in E(\mathcal{T}(\mathscr{X}))\}, \tag{3.1}$$

where $\mathcal{T}(\mathscr{X})$ is the minimum spanning tree of the pooled sample $\mathscr{X} := \{\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \ldots, \boldsymbol{X}^{(K)}\}$. The MFRT, unlike the MCM and MMCM statistics, is not distribution-free under the null, however, it can be easily calibrated as a permutation test. We compare the power of this test with the MCM and the MMCM tests in the following two scenarios. All the tests are calibrated using 500 permutations, and the empirical power is calculated over 500 iterations.

- *Normal Location*: Here, we consider the family $\{N_d(\boldsymbol{\mu}, \mathrm{I}) : \boldsymbol{\mu} \in \mathbb{R}^d\}$. Figure 2(a) shows the empirical power of the MFRT, the MCM test, and the MCMM test, when $K = 3$ and the data consists of 100 samples each from $N_d(\mathbf{0}, \mathrm{I})$, $N_d(0.3 \cdot \mathbf{1}, \mathrm{I})$, and $N_d(0.6 \cdot \mathbf{1}, \mathrm{I})$, respectively. Figure 2(b) shows the empirical power of tests when $K = 4$ and 100 samples each are drawn from $N_d(\mathbf{0}, \mathrm{I})$, $N_d(0.25 \cdot \mathbf{1}, \mathrm{I})$, $N_d(0.5 \cdot \mathbf{1}, \mathrm{I})$ and $N_d(0.75 \cdot \mathbf{1}, \mathrm{I})$, respectively. In both cases, the dimension d varies from 2 to 2000.
- *Spherical Normal Scale*: Here, we consider the family $\{N_d(\mathbf{0}, \sigma^2 \mathrm{I}) : \sigma > 0\}$. Figure 2(c) shows the empirical power of the MFRT, the MCM test, and the MCMM test, when $K = 3$ and the data consists of 100 samples each from $N_d(\mathbf{0}, \mathrm{I})$, $N_d(\mathbf{0}, 1.5 \cdot \mathrm{I})$, and $N_d(\mathbf{0}, 2 \cdot \mathrm{I})$, respectively. Figure 2(d) shows the empirical power of tests when $K = 4$ and 100 samples each are drawn from $N_d(\mathbf{0}, \mathrm{I})$, $N_d(\mathbf{0}, 1.25 \cdot \mathrm{I})$, $N_d(\mathbf{0}, 1.5 \cdot \mathrm{I})$ and $N_d(\mathbf{0}, 1.75 \cdot \mathrm{I})$, respectively. As before, the dimension d varies from 2 to 2000.

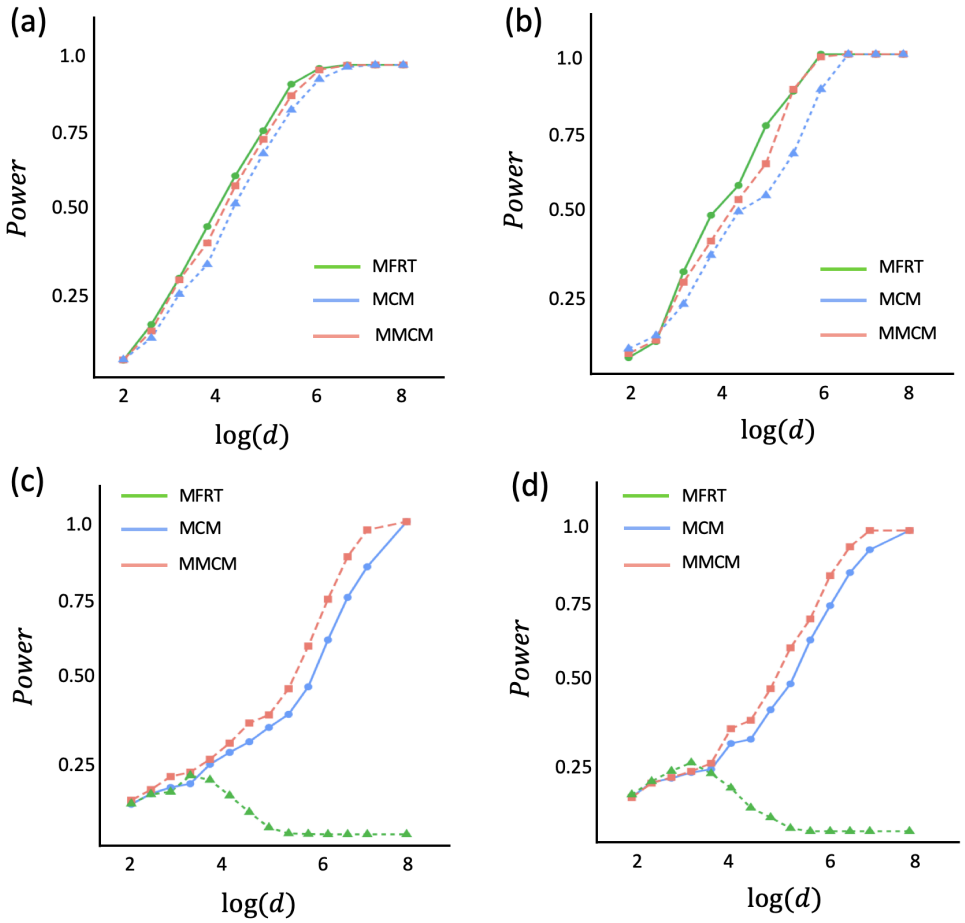
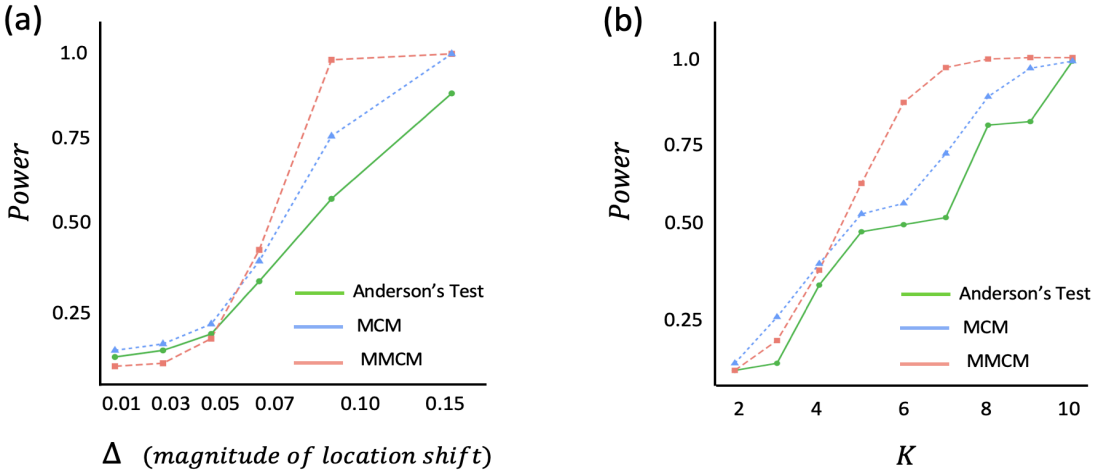


FIGURE 2. Power of the different tests across increasing dimension in (a) the normal location family with $K = 3$ groups, (b) the normal location family with $K = 4$ groups, (c) the spherical normal scale family with $K = 3$ groups, and (d) the spherical normal scale family with $K = 4$ groups.

The plots show that for location alternatives, the 3 tests have very similar power, with the MFRT performing marginally better than the MCMM, which is marginally better than the MCM. However, in the scale problem, the MCM and the MMCM tests drastically outperforms the MFRT. Here the power of the MFRT goes down to zero as the dimension increases, whereas the MCM and MCMM both have power improving with dimension and eventually going up to 1, illustrating the benefits of the distribution-free property of optimal matchings in high-dimensional problems.

3.2. **Comparison with Parametric Tests.** Here, we compare the performance of the MCM and the MMCM tests with baseline parametric tests, for relatively low-dimensional problems (where the corresponding parametric tests are applicable). As before, we consider location and scale alternatives in the normal family. Here, use the asymptotic distributions derived above to choose the cutoffs of the tests.



What is this?

FIGURE 3. Power of the different tests in the normal location family when the number of classes (a) $K = 7$, and (b) K varying between 2 and 10.

Example 1. (*Normal Location*) Here, we compare the MCM and the MMCM tests with the Anderson's test for Gaussian location alternatives [2].[5] As is common in parametric tests, since Anderson's statistic requires inverting certain sample covariance matrices, the test, unlike the MCM and the MMCM, is inapplicable for large dimensions, specifically if $d \geq \frac{N}{K(K-1)}$. When the dimension is much smaller than this threshold and there is a moderate sample size, Anderson's test performs well (this is expected because the test is specifically designed for such alternatives). However, when the dimension increases and comes closer to the boundary, Anderson's test start to lose power. In fact, we see below that for dimensions close to this threshold, that non-parametric matching based methods outperform Anderson's test, for relatively small sample sizes.

- Figure 3(a) shows the empirical power (over 500 iterations) of the Anderson's test, the MCM test, and the MCMM test, for $K = 7$ classes in dimension $d = 16$. The horizontal axis shows the magnitude of separation Δ, and the data consists of 100 samples from each of the following 7 distributions: for $1 \leq s \leq 7$, the s-th distribution corresponds to $N_{16}((s-1)\Delta \cdot \mathbf{1}, \mathrm{I})$. Each simulation was repeated for 6 values of Δ: $0.01, 0.03, 0.05, 0.07, 0.1, 0.15$.
- Figure 3(b) shows the empirical power when we vary both the number of classes K and the dimension d. Here, K varies from 2 to 10 (shown in the horizontal axis), and for each K the dimension d is chosen just below the dimension threshold of Anderson's test, and the s-th distribution corresponds to $N_d(\frac{s-1}{10} \cdot \mathbf{1}, \mathrm{I})$, for $1 \leq s \leq K$.

In both the cases, we observe that the power of the MCM and the MMCM tests are noticeably better than that of Anderson's test.

[5]This is a standard multisample method for testing difference of normal means. In the case where all the K sample sizes are equal, the Anderson's test constructs, a vector $V^{(s)}$ of length $(K-1)d$ formed by appending $K-1$ linearly independent contrasts based on the s-th observations from each of the K classes, for each $1 \leq s \leq N/K$. Then the test statistic is based on a Hoteling's T^2 statistic constructed from the vectors $V^{(1)}, \ldots, V^{(N/K)}$. The reader is referred to [2] for the precise description of this test.

Example 2. (*Normal Scale*) Here, we compare the MCM and the MMCM test with the likelihood ratio test (LRT) for the equality of covariance matrices, when the means are unknown,[6] in the Normal scale family. This test performs well for small dimensions, especially when the sample sizes across the classes are equal. However, we see below that even in relatively small dimensions, the LRT performs poorly when the sample sizes become unbalanced, but the matching based tests continue to have significant power.

- Figure 4(a) shows the empirical power (over 500 iterations) of the LRT, the MCM test, and the MCMM test, for $K = 4$ classes in dimension $d = 20$. The horizontal axis shows the magnitude of separation Δ, and the data consists of 80, 95, 110 and 125 samples from the following 4 distributions: for $1 \leq s \leq 4$, the s-th distribution corresponds to $N_{20}\left(\mathbf{0}, (1 + (s-1)\Delta)\mathrm{I}\right)$. Each simulation was repeated for 10 values of Δ: $0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85$.
- Figure 4(b) shows the empirical power (over 500 iterations) when we vary the sample size difference δ among the classes from 8 to 18. The horizontal axis shows this sample size difference δ. For each δ, the data consists of $80, 80+\delta, 80+2\delta$ and $80+3\delta$ samples from the the following 4 distributions respectively: for $1 \leq s \leq 4$, the s-th distribution corresponds to $N_{20}\left(\mathbf{0}, \frac{s+1}{2}\mathrm{I}\right)$.

In the first case, we observe that the power of all the tests increase to 1, but the MCM and the MMCM tests dominate that of the LRT by a significant margin for smaller separations. In the second case, however, with increase in the difference of sample size across the classes, the power of the LRT decreases to 0, while the power of the MCM and MMCM tests remain stable at high values, illustrating the robustness of these methods even for low-dimensional problems.

3.3. **Comparison between the MCM and the MMCM Tests.** In this section, we compare the finite-sample power of the MCM and the MMCM tests in various examples. Here, we consider 3 distributional models: (1) the normal location family $\{N_d(\boldsymbol{\mu}, \mathrm{I}) : \boldsymbol{\mu} \in \mathbb{R}^d\}$, (2) the spherical normal scale family $\{N_d(\mathbf{0}, \sigma^2\mathrm{I}) : \sigma > 0\}$, and (3) the equi-correlated normal scale family $\{N_d(\mathbf{0}, (1-\rho)\mathrm{I}) + \rho\mathbf{1}\mathbf{1}^\top : 0 \leq \rho < 1\}$. A typical simulation instance looks as follows. We generate samples from K different d-dimensional distributions from an underlying distributional model. The difference between the K distributions is quantified by a separation parameter Δ (which depends on location parameter/spherical scale parameter/correlation parameter, depending on the underlying distributional model). For each of the distributional models we consider two scenarios: (1) the *fixed class scenario* where the number of classes K is fixed and we perform a two-way power comparison with Δ versus d, and (2) the *fixed dimension scenario*, where we fix the dimension d, and perform a two-way power comparison with Δ versus K. In all the simulations, the power is calculated over 100 iterations. Additional simulations, comparing the MCM and the MCMM tests, in the lognormal family

[6]In this case, the LRT statistic rejects for small values of $\lambda = \exp(-\frac{1}{2}\sum_{i=1}^{K} N_i \log|\boldsymbol{S}_i^{-1}\boldsymbol{S}|)$, where $\boldsymbol{S}_i$ is the covariance matrix of the i-th sample, for $1 \leq i \leq K$, and $\boldsymbol{S}$ is the pooled sample covariance matrix. Under the null hypothesis of equality of covariance matrices, $-2\log\lambda$ has an asymptotic chi-squared distribution with $\frac{1}{2}d(d+1)(K-1)$ degrees of freedom.

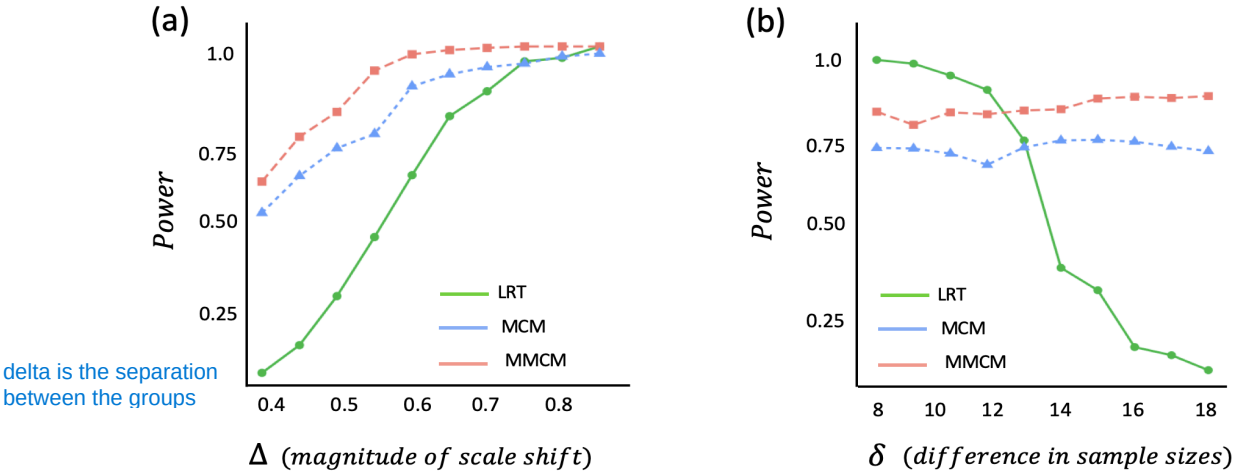


FIGURE 4. Power of the different tests for the spherical normal scale problem for (a) varying scale shift, and (b) varying difference in sample size

Δ ↓	Dimension	5	10	50	100	200	300	500
.04	MCM	.26	.30	**.16**	**.19**	**.36**	.41	.57
	MCMM	**.29**	**.44**	.10	.17	.25	**.44**	**.76**
.06	MCM	.43	.51	.31	**.43**	**.41**	.52	.92
	MCMM	**.59**	**.69**	**.46**	.32	.37	**.70**	**1.0**
.08	MCM	.49	.61	.39	.48	.50	.85	1.0
	MCMM	**.67**	**.81**	**.65**	**.64**	**.65**	**.99**	**1.0**
.10	MCM	.66	.70	.77	.60	.97	.99	1.0
	MCMM	**.81**	**.84**	**.90**	**.79**	**.99**	**1.0**	**1.0**
.12	MCM	.81	.97	.81	.87	1.0	1.0	1.0
	MCMM	**.94**	**1.0**	**.93**	**1.0**	**1.0**	**1.0**	**1.0**

(a)

Δ ↓	Groups	4	6	8	10
.04	MCM	**.06**	.42	.45	.80
	MCMM	.04	**.53**	**.78**	**.97**
.05	MCM	**.11**	.61	.85	1.0
	MCMM	.07	**.78**	**.99**	**1.0**
.07	MCM	**.19**	.77	1.0	1.0
	MCMM	.13	**0.96**	**1.0**	**1.0**
.09	MCM	.47	.93	1.0	1.0
	MCMM	**.53**	**1.0**	**1.0**	**1.0**
.10	MCM	.55	1.0	1.0	1.0
	MCMM	**.70**	**1.0**	**1.0**	**1.0**

(b)

TABLE 1. Power of the MCM and the MMCM tests in the normal location family with (a) the number of classes $K = 6$ fixed, and (b) the dimension $d = 150$ fixed. (The higher power in each case is in bold.)

are given in Appendix D. Overall, we observe that MCM and MCMM tests are comparable for small dimension and group sizes, but the MCMM test outperforms the MCM as the dimension, groups, or separation increases.

- *Normal Location*: Here, we consider samples from the following K distributions: $N_d((s-1)\Delta\mathbf{1}, \mathrm{I})$, for $1 \le s \le K$. Table 1(a) shows the fixed class scenario, where we take $K = 6$ groups and vary the dimension d from 5 to 500, and Δ from 0.04 to 0.12. Table 1(b) shows the fixed dimension scenario, where the dimension $d = 150$ is fixed, the number of groups K varies along $4, 6, 8, 10$, and Δ varies from 0.04 to 0.10. In both cases, the sample sizes were taken in equal increments of 50, starting from 50.
- *Spherical Normal Scale*: Here, we consider samples from the following K distributions: $N_d(\mathbf{0}, (1+(s-1)\Delta)\mathrm{I})$, for $1 \le s \le K$. Table 2(a) shows the fixed class scenario, with

$K = 6$ and dimension d varying from 5 to 500, and Δ varying from 0.05 to 0.4. Table 2(b) shows the fixed dimension scenario, where the $d = 150$ is fixed, and K varies along $4, 6, 8, 10$, and Δ varies from 0.05 to 0.4. As before, in both cases, the sample sizes were taken in equal increments of 50, starting from 50.

$\Delta\downarrow$	Dimension	5	10	50	100	200	300	500
.15	MCM	**.12**	**.16**	.22	.33	.51	.55	.71
	MMCM	.06	.13	**.28**	**.37**	**.67**	**.86**	**.92**
.20	MCM	**.13**	**.24**	.41	.46	.68	.79	.94
	MMCM	.12	.19	**.59**	**.78**	**.95**	**.99**	**1.0**
.25	MCM	.21	.31	.47	.60	.81	.88	1.0
	MMCM	**.23**	**.38**	**.85**	**.98**	**1.0**	**1.0**	**1.0**
.30	MCM	.22	.43	.70	.91	.98	1.0	1.0
	MMCM	**.25**	**.53**	**.99**	**1.0**	**1.0**	**1.0**	**1.0**
.35	MCM	.24	.38	.75	.87	1.0	1.0	1.0
	MMCM	**.30**	**.54**	**.99**	**1.0**	**1.0**	**1.0**	**1.0**
.40	MCM	.29	.50	.86	1.0	1.0	1.0	1.0
	MMCM	**.49**	**.85**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**

(a)

$\Delta\downarrow$	Groups	4	6	8	10
.15	MCM	**.27**	.68	.91	1.0
	MMCM	.18	**.91**	**1.0**	**1.0**
.20	MCM	.48	.89	.99	1.0
	MMCM	**.48**	**1.0**	**1.0**	**1.0**
.25	MCM	.66	.96	1.0	1.0
	MMCM	**.89**	**1.0**	**1.0**	**1.0**
.30	MCM	.87	1.0	1.0	1.0
	MMCM	**.98**	**1.0**	**1.0**	**1.0**
.35	MCM	.95	1.0	1.0	1.0
	MMCM	**1.0**	**1.0**	**1.0**	**1.0**
.40	MCM	1.0	1.0	1.0	1.0
	MMCM	**1.0**	**1.0**	**1.0**	**1.0**

(b)

TABLE 2. Power of the MCM and the MMCM tests in the spherical normal scale family with (a) the number of classes $K = 6$ fixed, and (b) the dimension $d = 150$ fixed.

- *Equi-correlated Normal Scale*: Here, we consider samples from the following K distributions: $N_d(0, (1-\rho_s)\mathrm{I} + \rho_s \mathbf{1}\mathbf{1}^\top)$, where $\rho_s := (s-1)\frac{\Delta}{K-1}$, for $1 \le s \le K$. Table 3(a) shows the fixed class scenario, with $K = 6$ and dimension d varying from 5 to 500, and Δ varying from 0.15 to 0.4. The sample sizes are taken to be $50, 100, 150, 200, 250$ and 300. Table 3(b) shows the fixed dimension scenario, where the $d = 150$ is fixed, and K varies along $4, 6, 8, 10$, and Δ varies from 0.15 to 0.4, as before. The sample sizes are taken in equal increments from 50 to 200 when $K = 4$, from 50 to 300 when $K = 6$, from 50 to 260 when $K = 8$, and from 50 to 230 when $K = 10$.

In all the simulations above (and those in Appendix D), we observe that the power of both the MCM and the MMCM tests improve with increasing separation and dimension. For smaller dimensions and separations, the power of both the tests are comparable, however, the MCMM test quickly gains power and performs noticeably better than the MCM, for higher dimensions and larger separations. This leads to our preference for using the MMCM over the MCM especially in high dimensions, which is often the case for real life datasets.

4. Distribution Under the Alternative

In this section we will prove a central limit theorem for the vector of cross-counts, and, as a corollary, derive the asymptotic distribution of the cross-match (2.1), the MCM (2.4) and MMCM (2.5) statistics, under general alternatives. We begin with an alternative way to describe the joint distribution of the data $\{\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \ldots, \boldsymbol{X}^{(K)}\}$:

$\Delta\downarrow$	Dimension	5	10	50	100	200	300	500
.15	MCM	**.10**	.11	.16	.22	.34	**.38**	.35
	MMCM	.08	**.11**	**.19**	**.27**	**.36**	.37	**.39**
.20	MCM	**.07**	**.10**	**.17**	.24	.25	.35	.43
	MMCM	.06	.09	.15	**.25**	**.38**	**.49**	**.64**
.25	MCM	**.09**	**.16**	.25	.27	.39	.44	.50
	MMCM	.02	.09	**.29**	**.42**	**.55**	**.66**	**.78**
.30	MCM	.08	**.17**	.27	.32	.56	.67	.71
	MMCM	**.14**	.13	**.36**	**.48**	**.78**	**.79**	**.93**
.35	MCM	**.12**	**.19**	.45	.46	.60	.65	.88
	MMCM	.06	.15	**.50**	**.70**	**.81**	**.90**	**.99**
.40	MCM	**.15**	.26	.51	.70	.77	.84	.95
	MMCM	.12	**.26**	**.70**	**.91**	**.98**	**1.0**	**1.0**

(a)

$\Delta\downarrow$	Groups	4	6	8	10
.15	MCM	**.26**	.19	.29	.31
	MMCM	.20	**.26**	**.32**	**.34**
.20	MCM	**.43**	.30	.29	.18
	MMCM	.37	**.35**	**.42**	**.26**
.25	MCM	.48	.40	.34	**.32**
	MMCM	**.54**	**.47**	**.43**	.28
.30	MCM	**.61**	.53	.43	.28
	MMCM	.59	**.63**	**.58**	**.47**
.35	MCM	.66	.63	.47	.43
	MMCM	**.78**	**.80**	**.76**	**.71**
.40	MCM	.87	.72	.67	.57
	MMCM	**.99**	**.95**	**.90**	**.83**

(b)

TABLE 3. Power of the MCM and the MMCM tests in the equi-correlated normal scale family with (a) the number of classes $K = 6$ fixed, and (b) the dimension $d = 150$ fixed.

- Let $Z_1, Z_2, \ldots, Z_N$ be i.i.d. from the density $\phi_N := \sum_{s=1}^K \frac{N_s}{N} f_s$ in $\mathbb{R}^d$, where $f_1, f_2, \ldots, f_K$ are the densities (with respect to the Lebesgue measure on $\mathbb{R}^d$) of the distributions $F_1, F_2, \ldots, F_K$, respectively.
- Given $\mathcal{Z}_N = (Z_1, Z_2, \ldots, Z_N)$, assign a random label $L_j \in [K] := \{1, 2, \ldots, K\}$ to Z_j, independently for each $1 \le j \le N$, where

$$\mathbb{P}(L_j = s | Z_j) = \frac{\frac{N_s}{N} f_s(Z_j)}{\phi_N(Z_j)}, \quad \text{for all } s \in [K]. \tag{4.1}$$

- Denote by $\eta_s := \sum_{i=1}^N \mathbf{1}\{L_i = s\}$, the number of elements labelled s. Then it is easy to verify that the joint distribution of $(\{Z_j : L_j = 1\}, \{Z_j : L_j = 2\}, \ldots, \{Z_j : L_j = K\})$ conditional on $(\eta_1, \cdots, \eta_K) = (N_1, N_2, \ldots, N_K)$ is same as the joint distribution of the data $\left(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \ldots, \boldsymbol{X}^{(K)}\right)$ (see Lemma B.1).[7]

It is also often convenient (because of the independence of the labelings) to work with unconditional distribution of $(\{Z_j : L_j = 1\}, \{Z_j : L_j = 2\}, \ldots, \{Z_j : L_j = K\})$, which we will refer to as the *bootstrap alternative distribution*.

Now, define the $K \times K$ matrix $\boldsymbol{B}_N = (b_{st})_{1 \le s,t \le K}$ as follows:

$$b_{st} = \begin{cases} \sum_{1 \le i \ne j \le N} e(Z_i, Z_j) \mathbf{1}\{L_i = s, L_j = t\} & \text{if } s \ne t, \\ \frac{1}{2} \sum_{1 \le i \ne j \le N} e(Z_i, Z_j) \mathbf{1}\{L_i = L_j = s\} & \text{if } s = t. \end{cases} \tag{4.2}$$

where $e(x, y) := \mathbf{1}\{(x, y) \in E(\mathscr{G}(\mathcal{Z}_N \cup \{x, y\}))\}$. Moreover, for notational convenience, denote $\boldsymbol{\eta} := (\eta_1, \ldots, \eta_K)$ and $\boldsymbol{N} := (N_1, \ldots, N_K)$. Then, for $1 \le s, t \le K$ the cross/pure counts

[7]Note that under the null, (4.1) simplifies to $P(L_j = s | Z_j) = \frac{N_s}{N}$, and the procedure described above, is precisely the way to generate the permutation null distribution.

a_{st} (recall (2.2) and (2.3)) can be re-written in terms of the $\mathcal{Z}_N$ and the labelings as follows:

$$a_{st} \stackrel{D}{=} b_{st} \Big| \{\boldsymbol{\eta} = \boldsymbol{N}\}, \tag{4.3}$$

Therefore, the conditional mean of the pure/cross-counts under the bootstrap alternative distribution is

$$\nu_N(s,t) := \mathbb{E}_{H_1}(b_{st}|\mathcal{Z}_N) = \begin{cases} \sum_{1\le i\ne j\le N} e(Z_i,Z_j) h_{st}^{(N)}(Z_i,Z_j), & \text{if } s \ne t \\ \\ \frac{1}{2}\sum_{1\le i\ne j\le N} e(Z_i,Z_j) h_{ss}^{(N)}(Z_i,Z_j), & \text{if } s = t, \end{cases} \tag{4.4}$$

where $h_{st}^{(N)}(x,y) = \frac{N_s}{N}\frac{N_t}{N}\frac{f_s(x)f_t(y)}{\phi_N(x)\phi_N(y)}$. We denote the matrix of these conditional expectations by $\boldsymbol{\mu}_N = ((\mu_N(s,t)))_{1\le s,t\le K}$. Note the expressions in the RHS above is permutation invariant and a function of the pooled sample (forgetting the labels). For example, for $s \ne t$,

$$\nu_N(s,t)\Big|\{\boldsymbol{\eta} = \boldsymbol{N}\} \stackrel{D}{=} \sum_{1\le a,b\le K} \sum_{i=1}^{N_a}\sum_{j=1}^{N_b} h_{st}^{(N)}(X_i^{(a)}, X_j^{(b)}) e(X_i^{(a)}, X_j^{(b)}) =: \mu_N(s,t), \tag{4.5}$$

which can be computed from the pooled data at a known alternative point $(f_1, f_2, \ldots, f_K)$. As usual, denote by $\underline{\boldsymbol{\mu}}_N$ the vector obtained by concatenating the rows of the matrices $\boldsymbol{\mu}_N := ((\mu_N(s,t)))$ in the upper triangular part. Note that $\underline{\boldsymbol{\mu}}_N$ can be thought of as the conditional mean of the vector $\underline{\boldsymbol{A}}_N$ given the pooled sample, where the randomness comes only from the labeling of the classes.

In the theorem below we show that the vector $\underline{\boldsymbol{A}}_N$ (recall that this is the vector obtained by concatenating the rows of $\boldsymbol{A}_N$ in the upper triangular part, as defined in (2.5)) centered by the corresponding vector of conditional means $\underline{\boldsymbol{\mu}}_N$ and scaled appropriately, converges in distribution to a $\binom{K}{2}$-dimensional multivariate normal in the usual asymptotic regime (2.8). The proof of the theorem is given in Appendix B.

Theorem 4.1. *Under general alternatives, in the usual asymptotic regime* (2.8),

$$\frac{\underline{\boldsymbol{A}}_N - \underline{\boldsymbol{\mu}}_N}{\sqrt{N}} \stackrel{D}{\longrightarrow} N_{\binom{K}{2}}(0, \boldsymbol{\Gamma}_{f_1,f_2,\ldots,f_k}), \tag{4.6}$$

where the covariance matrix $\boldsymbol{\Gamma}_{f_1,f_2,\ldots,f_k}$ *is as in Definition B.1 (in Appendix B).*

The joint normality of the vector $\underline{\boldsymbol{A}}_N$ implies the normality of linear functions of $\underline{\boldsymbol{A}}_N$, in particular the MCM statistic (recall (2.4)), which can be re-written as $R_{K,N} = \mathbf{1}^\top \underline{\boldsymbol{A}}_N$. Therefore,

$$\frac{R_{K,N} - \mathbf{1}^\top \underline{\boldsymbol{\mu}}_N}{\sqrt{N}} \stackrel{D}{\longrightarrow} N_{\binom{K}{2}}(\mathbf{0}, \mathbf{1}^\top \boldsymbol{\Gamma}_{f_1,f_2,\ldots,f_K} \mathbf{1}).$$

Similarly, for the MMCT statistic, (4.6) implies

$$\frac{(\underline{\boldsymbol{A}}_N - \underline{\boldsymbol{\mu}}_N)^\top \boldsymbol{\Gamma}^{-1}_{f_1,f_2,\ldots,f_K} (\underline{\boldsymbol{A}}_N - \underline{\boldsymbol{\mu}}_N)}{N} \stackrel{D}{\longrightarrow} \chi^2_{\binom{K}{2}}.$$

Even though the general expression for the covariance matrix $\boldsymbol{\Gamma}_{f_1,f_2,\ldots,f_k}$ (Definition B.1 in Appendix B) can be complicated, it simplifies nicely for the case $K = 2$. To this end, recall the 2-sample cross match statistic $R_{2,N}$ from (2.1). Let $\mathscr{X}$ denote the pooled sample. Then,

by (4.5),

$$\mathbb{E}_{H_1}(R_{2,N}|\mathscr{X}) = \sum_{1\le a,b\le 2}\sum_{i=1}^{N_a}\sum_{j=1}^{N_b}\frac{N_1}{N}\frac{N_2}{N}\frac{f_1(X_i^{(a)})f_2(X_i^{(b)})}{\phi_N(X_i^{(a)})\phi_N(X_i^{(b)})}e(X_i^{(a)},X_j^{(b)}). \tag{4.7}$$

We now have the following result for the 2-sample cross match test, which is a straightforward calculation from (4.6) above.

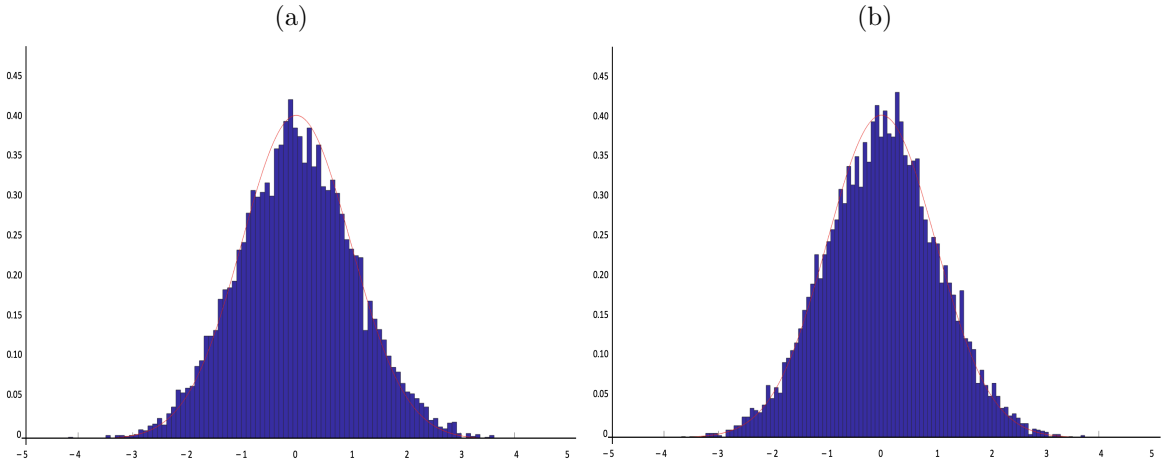


FIGURE 5. Histogram of the centered CM statistic (over 1000 iterations) and the predicted normal density (the red curve) when (a) $f_1 = N_{10}(\mathbf{0}, \mathrm{I})$ and $f_2 = N_{10}(\mathbf{1}, \mathrm{I})$ (normal location alternatives) and (b) $f_1 = N_{10}(\mathbf{0}, \mathrm{I})$ and $f_2 = N_{10}(\mathbf{0}, 2\mathrm{I})$ (normal scale alternatives).

Corollary 4.2. *For 2-sample cross match statistic $R_{2,N}$ as in (2.1), as $N \to \infty$,*

$$\frac{R_{2,N} - \mathbb{E}_{H_1}(R_{2,N}|\mathscr{X})}{\sqrt{N}} \xrightarrow{D} N(0, \gamma^2_{f_1,f_2}),$$

where $\mathbb{E}_{H_1}(R_{2,N}|\mathscr{X})$ is as in (4.7) and

$$\gamma^2_{f_1,f_2} := p_1p_2\left\{\int \frac{f_1(z)f_2(z)(p_1^2f_1^2(z)+p_2^2f_2^2(z))}{\phi(z)^3}\mathrm{d}z - \left(\int \frac{f_1(z)f_2(z)(p_2f_2(z)-p_1f_1(z))}{\phi(z)^2}\mathrm{d}z\right)^2\right\},$$

with $\phi := p_1f_1 + p_2f_2$.

The results above add to our mathematical understanding of the alternative properties of matching-based tests, which up till now have been largely unexplored. Figure 5 shows the histogram of the centered CM statistic computed using 600 samples from one distribution and 400 samples from another, repeated over 10000 iterations, for normal location and scale alternatives, and density of the corresponding limiting normal distribution (the red curve), as predicted by corollary above. The plots validate the asymptotic results and show that the normal approximation is quite accurate even for moderate sample sizes.

5. Application to Single Cell RNA Sequencing Data

In this section we apply the tests described above to single cell transcriptomics data obtained from the peripheral blood, cancer tissue and tumor-adjacent normal tissue of human subjects with hepatocellular carcinoma and non-small-cell lung cancer. Our goal is to investigate how biochemical metabolic pathways change across immune cells in a cancer environment, depending on the location of the tissue. We begin with a short background on biological pathways and the single cell RNA sequencing data.

Every tissue in the human body comprises of numerous different cell types, and each cell in turn contains tens of thousands of genes. The function of a tissue or an organ is rarely driven by a single unique gene, and analogously, complex disorders of organ dysfunction affect multiple genes. Therefore, to understand complex diseases, a systems biology approach examines sets or functional modules of related genes, called biological pathways. Because diseases such as cancer result from different combinations of perturbed gene activities, grouping genes into functional sets can often provide deeper insights into the underlying biological system. Indeed, the activity of certain pathways, particularly those that regulate cellular metabolism, has been found to be a strong predictor of complex phenotypes and response to treatment, both at the level of cells as well as that of individual patients [1, 12].

A *biological pathway* can be defined as a collection of molecules that coordinate to perform a specific action or change in the cell. This change could involve production of a new molecule, movement, growth or a physical transformation, or even cell death. While the activity of a particular pathway can be understood qualitatively based on the phenotypic changes in a cell, its quantitative estimation relies on the relative proportion of RNA molecules produced. The use of gene expression as a proxy for activity rests on the notion that the amount of mRNA molecules produced represent the economic resources of the cell [27]. Despite an understanding of how individual molecules in a biological pathway orchestrate a particular cellular function, it remains unclear whether the distribution of certain gene modules, and by proxy the corresponding pathway activities, are shared across cell types [35]. It has recently been shown that vastly different cell types contain similar ratios of metabolic enzymes, and tightly control the amounts of specialized proteins produced [26], highlighting previously unappreciated similarities among cell types. Nonetheless, the extent to which the distribution of relative mRNA molecules for genes in a given pathway might be consistent across different cell types remains elusive. Are there certain pathways which maintain a similar activity across cell types? This question is of fundamental significance because if true, it suggests a widespread design principle of cell biology.

With the advent of single cell RNA sequencing, it is now possible to study distinct but closely related cell populations [46] and examine the aforementioned question. Single cell RNA-sequencing (scRNA-seq) allows us to measure gene expression information from tens of thousands of individual cells, unraveling the cellular heterogeneity of a tissue in unprecedented detail. The resulting data can be thought of as a $c \times g$ matrix, $\boldsymbol{\eta} = ((\eta_{ab}))_{1\le a\le c, 1\le b\le g}$, where c corresponds to the number of cells, and g refers to the number of genes, and each entry η_{ab} corresponds to the number of RNA molecules detected for a given gene a in some cell b.

The high-dimensional, multisample (corresponding to multiple cell-types) nature of a typical scRNA-seq experiment makes this a fitting application of the multi-sample crossmatch test.

5.1. **Data Overview and Study Setup.** We apply our method on scRNA-seq data generated from purified T cell populations found in three tissue locations: (a) peripheral blood (hereafter referred to as blood), (b) tumor-infiltrating immune cells (hereafter referred to as tumor), and (c) normal tissue adjacent to the tumor from the same organ (hereafter referred to as adj. normal). We examined the following two datasets, where T cells extracted from each location were assigned a particular subtype based on flow cytometry and expression of known canonical cell surface proteins.[8]

(1) Non-Small-Cell Lung Cancer (NSCLC [17]) dataset: Here, the T cell subtypes found at each location were: $CD8^+$ Cytotoxic, $CD4^+$ Naive, $CD4^+$ Regulatory T cells (denoted by T_{reg}), and $CD4^+$ Naive Helper.
(2) Hepatocellular Carcinoma (HCC [47]) dataset: Here, the T cell subtypes profiled were: $CD8^+$ Cytotoxic, $CD4^+$ Naive, and $CD4^+$ Regulatory T cells (T_{reg}).

Recall that single cell data is extremely sparse count data with complex correlation structure. Nonetheless, the two datasets are comparable in terms of the sequencing protocol used, data generation, and the technical quality of the data (Table 4). In both the datasets we used deep sequencing was performed, ensuring our ability to detect genes with low expression. The summary of the number of cells sequenced for each cell type is provided in Table 5.

	NSCLC	HCC
Mean reads per cell	1,040,000	1,100,000
Median genes per cell	2,859	2,702
Total number of T cells profiled	12,210	4,794

TABLE 4. Summary characteristics of the two scRNA-seq datasets: Non-Small-Cell Lung Cancer (NSCLC) and Hepatocellular Carcinoma (HCC). For scRNA-seq it has been shown that with half a million reads per cell, most genes expressed can be detected, and that one million reads are sufficient to estimate the mean and variance of gene expression [38].

The metabolic state of T cells is implicated in diseases such as cancer, wherein the tumor microenvironment enforces dysfunctional T cell metabolism, thereby negatively affecting their anticancer functionality [4]. Thus, for each of the two cancer datasets and in the three tissue locations described above (namely, Adj. Normal, Tumor, and Blood), we examine whether the distribution of gene sets that correspond to biological metabolic pathways are consistent across T cell subtypes (therefore, the number of different T cell subtypes corresponds to the number of classes K). We specifically focus on 86 metabolic pathways described in the Kyoto Encyclopedia of Genes and Genomes (KEGG [23]). We obtained a list of genes corresponding to each pathway, and subsequently ascribed genes into 86 subsets, each subset corresponding

[8]Both the datasets used are open access, and available in the Gene Expression Omnibus (GEO). The raw sequencing data for T cells for the Hepatocellular Carcinoma dataset can be obtained from the GEO entry GSE98638 and the European Genome-phenome Archive database entry EGAS00001002072. The single cell sequencing data corresponding to the Non-Small-Cell Lung cancer case study can be found at GSE99254 and EGAS00001002430.

Non-small-cell Lung Cancer ($K = 4$ groups)				
Tissue Type	$CD8^+$ Cytotoxic	$CD4^+$ Naive	$CD4^+$ T_{reg}	$CD4^+$ Helper
Adj. Normal (2115 cells)	934	655	288	238
Tumor (5835 cells)	2182	1591	1170	892
Blood (4260 cells)	1323	1254	1011	672

Hepatocellular Carcinoma ($K = 3$ groups)			
Tissue Type	$CD8^+$ Cytotoxic	$CD4^+$ Naive	$CD4^+$ T_{reg}
Adj. Normal (997 cells)	412	406	179
Tumor (2170 cells)	563	515	549
Blood (1627 cells)	777	606	787

TABLE 5. The number of cells sequenced for various T cell subtypes in each of the two cancer settings. These correspond to the sample sizes of the different groups in the K-sample hypothesis testing problem. (Adj. Normal denotes tumor-adjacent normal tissue from the same organ, T_{reg} stands for regulatory T cell, and CD abbreviates *cluster of differentiation* or *classification determinant*, a protocol used for immunophenotyping cells.)

to a metabolic pathway. Then, for each one of the 3 different tissue locations and for each one of the 86 pathways, we tested the null hypothesis (using the MCMM test (2.5) described above) that the multivariate distribution of the genes belonging to that particular pathway is alike across the K different T cell subtypes (recall that $K = 4$ in the Non-Small-Cell Lung Cancer dataset, and $K = 3$ in the Hepatocellular Carcinoma dataset). In each of the cases the corresponding sample sizes for the K groups are given in the rows of Table 5. The number of genes in a given metabolic pathway ranged approximately between 30 and 110. To account for multiple-hypothesis testing, the resulting p-values are adjusted using the Benjamini-Hochberg (BH) correction procedure [5].

This study design allowed us to understand which metabolic pathways change in distribution across distinct, but closely-related, T cell subtypes. In particular, assessing the distribution of gene sets that belong to a particular metabolic pathway across the T cell subtypes allows us to address the following questions: (1) Which pathways have a similar distribution across T cell subtypes in a given tissue? (2) Are there pathways that have a stable and comparable distribution across T cell subtypes in a normal/healthy tissue, but a heterogenous or perturbed distribution in a tumor? (3) For pathways that have a disparate distribution across the T cell subtypes, which subtypes show the most distinct distribution?

5.2. **Comparing Pathway Distributions Based on Tissue Location.** The following is the outcome of the BH-corrected MCMM tests for assesing metabolic pathway distributions across the different T cell subtypes in the two datasets:

- In the NSCLC dataset, we found that of the 86 pathways examined, our test did not reject the null hypothesis for merely 35 pathways in the Tumor tissue, compared to 56 and 74 pathways in the Blood and Adj. Normal tissues, respectively.
- In HCC dataset, the set of pathways for which we failed to reject the null were remarkably alike to that for NSCLC. Specifically, 41, 63 and 76 metabolic pathways

were undistinguishable across T cell subtypes in the Tumor, Blood and Adj. Normal tissues, respectively.

The fact that majority of the metabolic pathways do not show evidence for dissimilar distribution across cell types based on our test indicates that the T cell subtypes might be more similar than previously appreciated in terms of how they regulate their basic metabolic machinery. Further, for each pair of tissue location, we computed the overlap in the pathways for which the null hypothesis was accepted or rejected. As expected, we found that the concordance between the results was substantially higher for Blood and Adj. Normal, than either of these tissues had with the Tumor tissue (this is seen from the off-diagonal values in the tables in Figure 6).

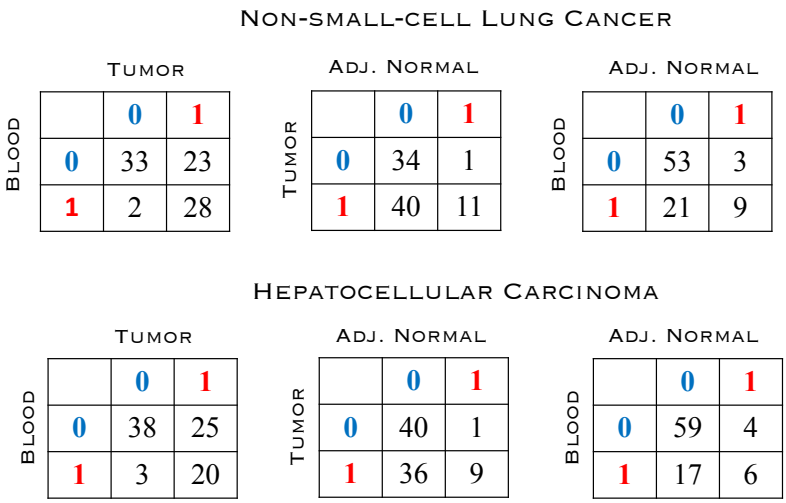


FIGURE 6. A set of 2×2 tables showing how many of the 86 hypothesis (corresponding to the pathways) were accepted/rejected for each of 3 pairs of tissue locations. Here, **0** stands for pathways whose distribution are either similar/stable (null hypothesis accepted), and **1** stands for pathways whose distributions are perturbed/heterogenous (null hypothesis rejected), across the T cell subtypes in two types of solid organ malignancies.

Interestingly, we found that 8 metabolic pathways were differentially distributed across the T cell subtypes in each tissue examined in NSCLC whereas 5 pathways exhibited this pattern in the HCC dataset (Figure 7). The *purine metabolism* pathway was a common pathway shared by both datasets which showed evidence for heterogenous distribution among the T cell subtypes in every tissue. This discovery suggests that purine metabolism is fundamentally different even among the closely related T cell subtypes. In order to find out which T cell subtype contributed most to this difference for the metabolic pathways that emerged as being heterogeneously distributed in all tissues, we employed a class selection procedure described below. To select a single class, we looked at all the pairwise comparisons where the null

was rejected, and then identified the class that was common across all the cases of rejection (Figure 7c-d).

We found that $CD4^+$ regulatory T cells (T_{regs}) were the strongest contributors as to why a pathway, such as purine metabolism, was detected as being differentially distributed. It is important to note that, purine metabolism regulates the balance of proinflammatory and immunosuppressive molecules produced by T cells, and the activation of the purinergic receptor P2X7 has been shown to inhibit the immunosuppressive functions specifically in T_{regs} [40]. Thus, the fact that our test discovered purine metabolism as being differentially distributed across T cell subtypes in both studies, and specifically in T_{regs}, showcases its ability to unearth meaningful biological phenomena.

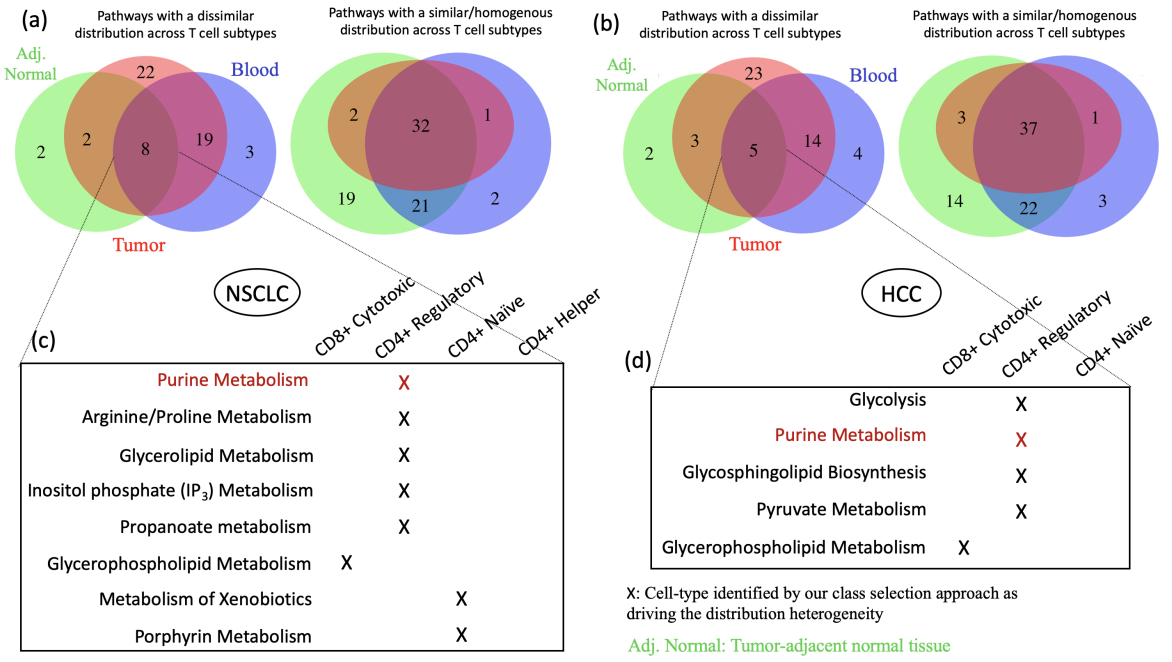


FIGURE 7. (a), (b) Venn diagrams showing the overlap among the metabolic pathways that demonstrated evidence for differential and similar distributions across the T cell subtypes in NSCLC and HCC, respectively. For the pathways that were heterogenous among the T cell classes across all 3 tissue types, we used the *class selection* procedure to identify which T cell subtype had the most disparate distribution (c), (d). Differential distribution for the purine metabolism pathway, driven largely by T_{regs}, was observed in both NSCLC and HCC.

5.3. **Differentially Distributed Pathways as Biological and Algorithmic Features.** In the NSCLC dataset, we found that 51 pathways showed evidence for differential distribution across the tumor-infiltrating T cell subtypes, whereas 45 pathways exhibited this pattern in the tumor-infiltrating T cells in the HCC dataset. We observed that if a metabolic pathway demonstrated an intra-T cell-type heterogenous distribution in Blood and Adj. Normal, that heterogeneity was preserved in the Tumor tissue (in the first and third Venn diagrams

in Figure 7, the intersection of the green (Adj. Normal) and the blue circles (Blood) is completely contained in the orange circle corresponding to Tumor). In other words, rejecting the null for a particular pathway for T cells in the Adj. Normal and Blood tissues was useful in prognosticating that pathway's behavior in the tumor-infiltrating T cells. On the other hand, we found that certain pathways that demonstrated a dissimilar distribution across T cell subtypes in the Tumor and Adj. Normal tissues showed evidence for homogenous distribution in blood (the intersection of the green and orange circles minus the blue circle in the first and third Venn diagrams in Figure 7). Specifically, this pattern was true for 2 pathways (sphingolipid metabolism and glycophospholipid synthesis) in the NSCLC dataset, and for 3 pathways (phenylalanine metabolism, oxidative phosphorylation and O-glycan biosynthesis) in the HCC dataset. This difference can be attributed, at least partially, to the organ-specific function of these pathways. For instance, ceramide, a central molecule in the sphingolipid metabolism, regulates endothelial permeability and airway smooth muscle function in the lungs [43], and increased sphingolipid metabolism is a hallmark of lung cancer [34]. Similarly, phenylalanine hydroxylase, the main enzyme in the phenylalanine metabolism pathway is active exclusively in the liver [30], and otherwise inactive in the blood. Hence, in light of the organ-specific roles of these pathways, it is rather reassuring that our test appropriately rejects the null hypothesis and captures their heterogenous distribution among the T cell subtypes based on the tissue location.

Clustering and cell type identification are crucial steps in single cell data analysis. scRNA-seq analysis pipelines often first identify highly variable genes in the dataset, and subsequently use those genes as an input to the t-distributed stochastic neighbor embedding (tSNE) algorithm. Even with this approach, however, single cell analysis pipelines typically fail to resolve the different T cell populations, and rarely identify the different immune cell subtypes in any reliable or discrete fashion. We wondered whether in practice, one might be able to utilize our test to identify differentially distributed pathways and then use this information to improve the tSNE-based clustering approach. To investigate whether the pathways that emerge as being differentially distributed can serve as meaningful features in identifying cell types, we focussed on the pathways that were consistently identified as being differentially distributed across the T cells in all three tissue types in NSCLC. To our surprise, we found that using the subset of genes corresponding to the pathways that our analysis identified as being heterogeneously distributed improves the clustering results (Figure 8(a)). Moreover, the clustering results converge with, and act as an indirect validation for our class selection approach because the latter identified $CD4^+$ T_{regs} as the cell type with the dissimilar distribution for metabolic pathways such as IP_3 metabolism, purine metabolism, and arginine and proline metabolism. When we used the genes comprising these pathways as the input to tSNE (Fig. 8a), we observed that indeed the $CD4^+$ T_{regs} became more easily visually distinguishable compared to using the genes in other pathways which had a similar distribution across the T cell subtypes (Figure 8(b)).

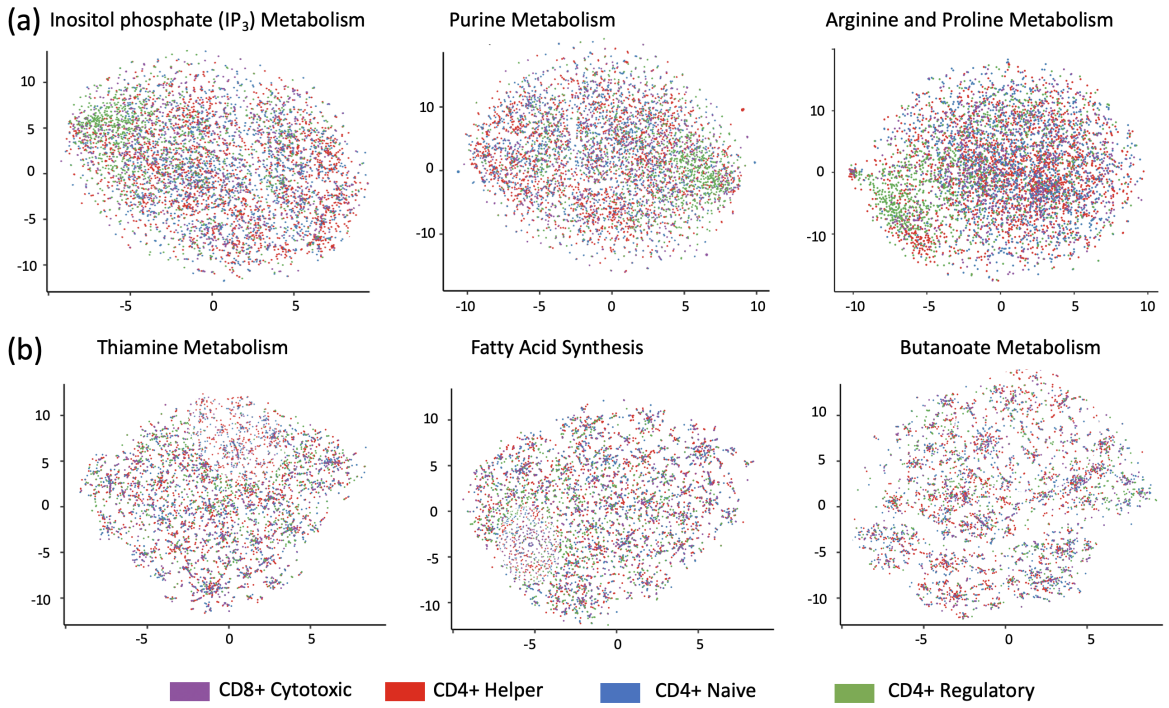


FIGURE 8. Results of the tSNE analysis performed using a specific pathway to cluster the 4 different tumor-infiltrating T cell subtypes in NSCLC: (a) pathways that were identified as differentially distributed successfully manage to segregate $CD4^+$ T_{regs} as predicted by our class selection procedure, whereas (b) pathways that were identified as being similar across the T cell subtypes produce a more patchy clustering wherein the subtypes are visually indistinguishable. All tSNE plots were generated with *perplexity* = 30.

6. Discussion

This paper introduces a novel graph-based nonparametric test for comparing multiple multivariate distributions. Using optimal matching as the basis, we demonstrate that our test, in addition to other multisample generalizations of Rosenbaum's crossmatch test [39], is distribution-free, computationally efficient, and consistent for general alternatives, making it particularly attractive for modern high-dimensional statistical applications. We also obtain a joint central limit theorem for the entire matrix of cross-counts, and, hence, derive a distributional limit theorem for the test statistics, under general alternatives. Our numerical experiments demonstrate that the proposed method outperforms other non-parametric graph-based multisample methods as well as commonly used parametric tests, in a variety of simulation settings. Lastly, we showcase the utility of this test in the field of single cell transcriptomics where we used our test to address an important question in signal transduction and cell biology. Our multisample procedure uncovered revealing patterns about how closely-related, key immune cells in the body (namely, T cell subtypes), might alter their metabolic machinery in solid organ malignancies, particularly depending on the tissue location. We envision that this test and the underlying theoretical intuitions described in

this work will find broad applicability in future research that examines hypothesis testing in the multisample, multivariate framework. Furthermore, our test opens up new paradigms of investigation for practitioners to assess its properties and its feasibility in being adapted to other algorithms, as we have demonstrated here through its usage as a pre-processing step in tSNE-based data visualization.

References

[1] A. Amadoz, C. Çubuk, D. Crespo, J. Carbonell-Caballero, M. R. Hidago, F. Salavert, J. Dopazo, and I. Medina, Actionable pathways: interactive discovery of therapeutic targets using signaling pathway models, *Nucleic Acids Research*, 44(W1):W212–W216, 05 2016.

[2] T. W. Anderson, A test for equality of means when covariance matrices are unequal, *Ann. Math. Statist.*, 34(2):671–672, 06 1963.

[3] E. Arias-Castro and B. Pelletier, On the consistency of the crossmatch test, *Journal of Statistical Planning and Inference*, 171:184 – 190, 2016.

[4] G. R. Bantug, L. Galluzzi, G. Kroemer, and C. Hess, The spectrum of T cell metabolism in health and disease. *Nature Reviews Immunology*, 18(1):19–34, 2017.

[5] Y. Benjamini and Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[6] B. B. Bhattacharya, Two-sample tests based on geometric graphs: Asymptotic distribution and detection Thresholds, arXiv:1512.00384, 2018.

[7] P. J. Bickel, A distribution free version of the Smirnov two sample test in the p-variate case, *Annals of Mathematical Statistics*, Vol. 40, 1–23, 1969.

[8] M. Biswas, M. Mukhopadhyay, and A. K. Ghosh, A distribution-free two-sample run test applicable to high-dimensional data, *Biometrika*, 101(4):913–926, 10 2014.

[9] H. Chen and J. H. Friedman, A new graph-based two-sample test for multivariate and object data, *Journal of the American Statistical Association*, 112(517):397–409, 2017.

[10] H. Chen, X. Chen, X. and Y. Su, A weighted edge-count two sample test for multivariate and object data, *Journal of the American Statistical Association*, Vol. 113 (523), 1146–1155, 2018.

[11] H. Chen and D. S. Small, New multivariate tests for assessing covariate balance in matched observational studies, arXiv:1609.03686, 2019.

[12] D. Fey, M. Halasz, D. Dreidax, S. P. Kennedy, J. F. Hastings, N. Rauch, A. G. Munoz, R. Pilkington, M. Fischer, F. Westermann, W. Kolch, B. N. Kholodenko, and D. R. Croucher, Signaling pathway models as biomarkers: Patient-specific simulations of jnk activity predict the survival of neuroblastoma patients, *Science Signaling*, 8(408):ra130–ra130, 2015.

[13] J. H. Friedman and L. C. Rafsky, Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests, *Ann. Statist.*, 7(4):697–717, 07 1979.

[14] P. Ghosal and B. Sen, Multivariate ranks and quantiles using optimal transportation and applications to goodness-of-fit testing, arXiv:1905.05340, 2019.

[15] J. D. Gibbons and S. Chakraborty, *Nonparametric Statistical Inference*, Fourth Edition. Marcel Dekker Inc., 2003.

[16] L. Györfi and T. Nemetz, f-dissimilarity: A general class of separation measures of several probability measures, *In Topics in Information Theory. Colloq. Math. Soc. János Bolyai*, Vol. 16, 309–321, 1975.

[17] X. Guo, Y. Zhang, L. Zheng, C. Zheng, J. Song, Q. Zhang, B. Kang, Z. Liu, L. Jin, R. Xing, R. Gao, L. Zhang, M. Dong, X. Hu, X. Ren, D. Kirchhoff, H. G. Roider, T. Yan, and Z. Zhang, Global characterization of t cells in non-small-cell lung cancer by single-cell sequencing, *Nature medicine*, 24(7):978–985, 2018.

[18] R. Heller, S. T. Jensen, P. R. Rosenbaum, and D. S. Small, Sensitivity analysis for the cross-match test, with applications in genomics, *Journal of the American Statistical Association*, Vol. 105 (491), 1005–1013, 2010.

[19] R. Heller, P. R. Rosenbaum, and D. S. Small, Using the crossmatch test to appraise covariate balance in matched pairs, *The American Statistician*, Vol. 64, 299–309, 2010.

[20] N. Henze, A multivariate two-sample test based on the number of nearest neighbor type coincidences, *Ann. Statist.*, 16(2):772–783, 06 1988.

[21] N. Henze and M. D. Penrose, On the multivariate runs test, *The Annals of Statistics*, Vol. 27 (1), 290–298, 1999.

[22] L. Holst, Two conditional limit theorems with applications, *The Annals of Statistics*, 7(3):551–557, 1979.

[23] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, KEGG: new perspectives on genomes, pathways, diseases and drugs, *Nucleic Acids Research*, 45(D1):D353–D361, 11 2016.

[24] W. H. Kruskal. A nonparametric test for the several sample problem, *Ann. Math. Statist.*, 23(4):525–540, 12 1952.

[25] W. H. Kruskal and W. A. Wallis, Use of ranks in one-criterion variance analysis, *Journal of the American Statistical Association*, 47(260):583–621, 1952.

[26] J.-B. Lalanne, J. C. Taggart, M. S. Guo, L. Herzel, A. Schieler, and G.-W. Li, Evolutionary convergence of pathway-specific enzyme expression stoichiometry, *Cell*, 173(3):749 – 761.e38, 2018.

[27] M. Lynch and G. K. Marinov, The bioenergetic costs of a gene, *Proceedings of the National Academy of Sciences*, 112(51):15690–15695, 2015.

[28] J.-F. Maa, D. K. Pearl, and R. Bartoszyński, Reducing multidimensional two-sample data to one-dimensional interpoint comparisons, *The Annals of Statistics*, Vol. 24 (3), 1069–1074, 1996.

[29] H. B. Mann and D. R. Whitney, On a test of whether one of two random variables is stochastically larger than the other, *Annals of Mathematical Statistics*, Vol. 18(1), 50–60, 1947.

[30] D. E. Matthews, An overview of phenylalanine and tyrosine kinetics in humans, *The Journal of Nutrition*, 137(6):1549S–1555S, 2007.

[31] M. D. McHugh, J. Berez, and D. S. Small, Hospitals with higher nurse staffing had lower odds of readmissions penalties than hospitals with lower staffing, *Health Affairs*, Vol. 32, 1740–1747, 2013.

[32] A. M. Mood, The distribution theory of runs, *The Annals of Mathematical Statistics*, Vol. 11 (4), 367–392, 1940.

[33] D. Nettleton, and T. Banerjee, Testing the equality of distributions of random vectors with categorical components, *Comput. Statist. Data Anal.*, Vol. 37, 195–208, 2001.

[34] B. Ogretmen. Sphingolipid metabolism in cancer signaling and therapy, *Nature Reviews Cancer*, 18(1):33–50, 2017.

[35] J. M. Peregrín-Alvarez, C. Sanford, and J. Parkinson, The conservation and evolutionary modularity of metabolism, *Genome Biology*, 10(6):R63, Jun 2009.

[36] A. Petrie, Graph-theoretic multisample tests of equality in distribution for high dimensional data, *Comput. Stat. Data Anal.*, 96(C):145–158, Apr. 2016.

[37] M. Raič, A multivariate Berry–Esseen theorem with explicit constants, arXiv:1802.06475, 2018.

[38] S. Rizzetto, A. A. Eltahla, P. Lin, R. Bull, A. R. Lloyd, J. W. K. Ho, V. Venturi, and F. Luciani, Impact of sequencing depth and read length on single cell rna sequencing data of t cells, *Scientific Reports*, 7(1):12781, 2017.

[39] P. R. Rosenbaum. An exact distribution-free test comparing two multivariate distributions based on adjacency, *Journal of the Royal Statistical Society B*, 67:515–530, 2005.

[40] U. Schenk, M. Frascoli, M. Proietti, R. Geffers, E. Traggiai, J. Buer, C. Ricordi, A. M. Westendorf, and F. Grassi, Atp inhibits the generation and function of regulatory t cells through the activation of purinergic p2x receptors, *Science Signaling*, 4(162):ra12–ra12, 2011.

[41] M. F. Schilling, Multivariate two-sample tests based on nearest neighbors, *Journal of the American Statistical Association*, 81(395):799–806, 1986.

[42] N. Smirnov, On the estimation of the discrepancy between empirical curves of distribution for two independent samples, *Bulletin de Universite de Moscow, Serie internationale (Mathematiques)*, Vol. 2, 3–14, 1939.
[43] S. Uhlig and E. Gulbins, Sphingolipids in the lungs, *American journal of respiratory and critical care medicine*, 178(11):1100–1114, 2008.
[44] A. Wald and J. Wolfowitz, On a test whether two samples are from the same population, *Annals of Mathematical Statistics*, 11(2):147–162, 06 1940.
[45] L. Weiss, Two-sample tests for multivariate distributions, *The Annals of Mathematical Statistics*, Vol. 31, 159–164, 1960.
[46] A. R. Wu, J. Wang, A. M. Streets, and Y. Huang, Single-cell transcriptional analysis, *Annual Review of Analytical Chemistry*, 10(1):439–462, 2017. PMID: 28301747.
[47] C. Zheng, L. Zheng, J.-K. Yoo, H. Guo, Y. Zhang, X. Guo, B. Kang, R. Hu, J. Y. Huang, Q. Zhang, Z. Liu, M. Dong, X. Hu, W. Ouyang, J. Peng, and Z. Zhang, Landscape of infiltrating t cells in liver cancer revealed by single-cell sequencing, *Cell*, 169(7):1342 – 1356.e16, 2017.

Appendix A. Proofs from Section 2

In this section we present the proofs of the results from Section 2. The proof of Proposition 2.2, which computes the mean and the variance of $\underline{\boldsymbol{A}}_N$ is given in Section A.1. The proof of the asymptotic null distribution in Theorem 2.3 is described in Section A.2, and the proof of the consistency (Theorem 2.4) is in Section A.3.

A.1. **Proof of Proposition 2.2.** Since the conditional distribution of $\boldsymbol{A}_N$ given $\mathscr{X}$ equals the unconditonal distribution of $\boldsymbol{A}_N$ under the null, we can assume that the edges of $\mathscr{G}(\mathscr{X})$ are fixed, say $\{(1,2),(3,4),\ \ldots,(N-1,N)\}$ without loss of generality. For notational convenience, let us define $\boldsymbol{\eta} := (\eta_1,\ldots,\eta_K)$ and $\boldsymbol{N} := (N_1,\ldots,N_K)$. Following the notations in subsection A.2, it follows that:

$$\mathbb{E}_{H_0}(a_{st}) = \sum_{j=1}^{I} \mathbb{P}_{H_0}\left(\{L_{2j-1},L_{2j}\} = \{s,t\} \big| \boldsymbol{\eta} = \boldsymbol{N}\right). \tag{A.1}$$

By an easy sampling without replacement argument, for each $j \le I$,

$$\mathbb{P}_{H_0}\left(\{L_{2j-1},L_{2j}\} = \{s,t\} \big| \boldsymbol{\eta} = \boldsymbol{N}\right) = \begin{cases} \frac{2N_sN_t}{N(N-1)} & \text{if } s<t, \\ \\ \frac{N_s(N_s-1)}{N(N-1)} & \text{if } s=t. \end{cases}$$

The result in (2.7) now follows from (A.1) on observing that $I = N/2$.

Next, note that for $1 \le s_1 \ne s_2 \le K$,

$$\begin{aligned}\mathrm{Var}_{H_0}(a_{s_1s_2}) = &\sum_{j=1}^{I} \mathrm{Var}_{H_0}\left(\mathbf{1}\left\{\{L_{2j-1},L_{2j}\} = \{s_1,s_2\}\right\} \big| \boldsymbol{\eta} = \boldsymbol{N}\right) \\ &+ \sum_{j_1 \ne j_2} \mathrm{Cov}_{H_0}\left(\mathbf{1}\left\{\{L_{2j_1-1},L_{2j_1}\} = \{s_1,s_2\}\right\}, \mathbf{1}\left\{\{L_{2j_2-1},L_{2j_2}\} = \{s_1,s_2\}\right\} \big| \boldsymbol{\eta} = \boldsymbol{N}\right).\end{aligned}$$

First, note that

$$\sum_{j=1}^{I} \mathrm{Var}_{H_0}\left(\mathbf{1}\left\{\{L_{2j-1},L_{2j}\} = \{s_1,s_2\}\right\} \big| \boldsymbol{\eta} = \boldsymbol{N}\right) = I\left[\frac{2N_{s_1}N_{s_2}}{N(N-1)}\right]\left[1 - \frac{2N_{s_1}N_{s_2}}{N(N-1)}\right]$$

$$= \frac{N_{s_1}N_{s_2}}{N-1}\left[1 - \frac{2N_{s_1}N_{s_2}}{N(N-1)}\right]. \tag{A.2}$$

Next, we have

$$\begin{aligned}
&\sum_{j_1\neq j_2} \mathrm{Cov}_{H_0}\left(\mathbf{1}\left\{\{L_{2j_1-1}, L_{2j_1}\} = \{s_1, s_2\}\right\}, \mathbf{1}\left\{\{L_{2j_2-1}, L_{2j_2}\} = \{s_1, s_2\}\right\} \middle| \boldsymbol{\eta} = \boldsymbol{N}\right) \\
&= \sum_{j_1\neq j_2} \mathbb{P}_{H_0}\left(\{L_{2j_1-1}, L_{2j_1}\} = \{L_{2j_2-1}, L_{2j_2}\} = \{s_1, s_2\} \middle| \boldsymbol{\eta} = \boldsymbol{N}\right) - \sum_{j_1\neq j_2}\left(\frac{2N_{s_1}N_{s_2}}{N(N-1)}\right)^2 \\
&= I(I-1)\frac{4N_{s_1}N_{s_2}(N_{s_1}-1)(N_{s_2}-1)}{N(N-1)(N-2)(N-3)} - I(I-1)\left(\frac{2N_{s_1}N_{s_2}}{N(N-1)}\right)^2 \\
&= \frac{N_{s_1}N_{s_2}(N_{s_1}-1)(N_{s_2}-1)}{(N-1)(N-3)} - \frac{N_{s_1}N_{s_2}}{N-1}\left[\frac{N_{s_1}N_{s_2}}{N-1} - \frac{2N_{s_1}N_{s_2}}{N(N-1)}\right]
\end{aligned} \tag{A.3}$$

Adding (A.2) and (A.3) gives the expression for $\mathrm{Var}_{H_0}(a_{s_1s_2})$ given in Proposition 2.2. Next, take $1 \le s_1 \neq s_2 \neq s_3 \le K$. Then,

$$\begin{aligned}
\mathbb{E}_{H_0}\left(a_{s_1s_2}a_{s_1s_3}\right) &= \sum_{j_1\neq j_2} \mathbb{P}_{H_0}\left(\{L_{2j_1-1}, L_{2j_1}\} = \{s_1, s_2\}, \{L_{2j_2-1}, L_{2j_2}\} = \{s_1, s_3\} \middle| \boldsymbol{\eta} = \boldsymbol{N}\right) \\
&= I(I-1)\frac{4N_{s_1}(N_{s_1}-1)N_{s_2}N_{s_3}}{N(N-1)(N-2)(N-3)} \\
&= \frac{N_{s_1}(N_{s_1}-1)N_{s_2}N_{s_3}}{(N-1)(N-3)}.
\end{aligned} \tag{A.4}$$

The expression for $\mathrm{Cov}_{H_0}(a_{s_1s_2}, s_{s_1,s_3})$ now follows from (A.4) on observing that:

$$\left(\mathbb{E}_{H_0}a_{s_1s_2}\right)\left(\mathbb{E}_{H_0}a_{s_1s_3}\right) = \frac{N_{s_1}^2N_{s_2}N_{s_3}}{(N-1)^2}.$$

Finally, take $1 \le s_1 \neq s_2 \neq s_3 \neq s_4 \le K$. In this case,

$$\begin{aligned}
\mathbb{E}_{H_0}\left(a_{s_1s_2}a_{s_3s_4}\right) &= \sum_{j_1\neq j_2} \mathbb{P}_{H_0}\left(\{L_{2j_1-1}, L_{2j_1}\} = \{s_1, s_2\}, \{L_{2j_2-1}, L_{2j_2}\} = \{s_3, s_4\} \middle| \boldsymbol{\eta} = \boldsymbol{N}\right) \\
&= I(I-1)\frac{4N_{s_1}N_{s_2}N_{s_3}N_{s_4}}{N(N-1)(N-2)(N-3)} \\
&= \frac{N_{s_1}N_{s_2}N_{s_3}N_{s_4}}{(N-1)(N-3)}.
\end{aligned} \tag{A.5}$$

The expression for $\mathrm{Cov}_{H_0}(a_{s_1s_2}, s_{s_3,s_4})$ now follows from (A.5) on observing that:

$$\left(\mathbb{E}_{H_0}a_{s_1s_2}\right)\left(\mathbb{E}_{H_0}a_{s_3s_4}\right) = \frac{N_{s_1}N_{s_2}N_{s_3}N_{s_4}}{(N-1)^2}.$$

This completes the proof of Proposition 2.2. □

A.2. **Proof of Theorem 2.3.** Note that it suffices to prove (2.9). The remaining assertions in Theorem 2.3 is an immediate consequence of (2.9). The proof of (2.9) proceeds in two-steps: (1) $\frac{1}{\sqrt{N}}(\underline{\boldsymbol{A}}_N - \mathbb{E}_{H_0}(\underline{\boldsymbol{A}}_N)) \xrightarrow{D} N(0, \boldsymbol{\Gamma})$, for some non-negative definite matrix $\boldsymbol{\Gamma}$, and (2) $\frac{1}{N}\mathrm{Cov}_{H_0}(\underline{\boldsymbol{A}}_N) \to \boldsymbol{\Gamma}$, and $\boldsymbol{\Gamma}$ is invertible.

We begin with the proof of (1): Denote the pooled sample

$$\mathscr{X} = (X_1^{(1)}, \ldots, X_{N_1}^{(1)}, \ldots, X_1^{(K)}, \ldots, X_{N_K}^{(K)})$$

(forgetting the labels) as $\mathcal{Z}_N := (Z_1, Z_2, \ldots, Z_N)$. Under the null H_0, $Z_1, Z_2, \ldots, Z_N$ are i.i.d. F (the unknown null distribution). Let $L_1, \ldots, L_N$ be i.i.d. random variables, taking value in $\{1, 2, \ldots, K\}$, independent of $Z_1, \ldots, Z_N$, such that

$$\mathbb{P}(L_1 = s) = \frac{N_s}{N}, \quad \text{for all } s \in [K]. \tag{A.6}$$

For each $s \in [K]$, define $\eta_s = \sum_{i=1}^N \mathbf{1}\{L_i = s\} \sim \text{Bin}(N, \frac{N_s}{N})$ and for each $x, y \in \mathbb{R}^d$, let $e(x, y) := \mathbf{1}\{(x, y) \in E(\mathscr{G}(\mathcal{Z}_N \cup \{x, y\}))\}$. Define the $K \times K$ matrix $\boldsymbol{B}_N = (b_{st})_{1 \le s,t \le K}$ as follows:

$$b_{st} = \begin{cases} \sum_{1 \le i \ne j \le N} e(Z_i, Z_j) \mathbf{1}\{L_i = s, L_j = t\} & \text{if } s \ne t, \\ \\ \frac{1}{2} \sum_{1 \le i \ne j \le N} e(Z_i, Z_j) \mathbf{1}\{L_i = L_j = s\} & \text{if } s = t. \end{cases}$$

Under H_0, it follows from Lemma B.1 that the conditional distribution of $\boldsymbol{B}_N$ given $(\eta_1, \ldots, \eta_K) = (N_1, \ldots, N_K)$ is same as the distribution of $\boldsymbol{A}_N$. Therefore, it suffices to derive the limiting distribution of $\boldsymbol{B}_N | \{(\eta_1, \ldots, \eta_K) = (N_1, \ldots, N_K)\}$.

To this end, note that conditional on $\mathcal{Z}_N$, the matching graph $\mathscr{G}(Z_N)$ is fixed and since the $I = N/2$ matched edges in the graph are disjoint, the samples in $\mathcal{Z}_N$ can be (re)-labelled $1, 2, \ldots, N$ such that $\{(1, 2), (3, 4), \ldots, (N-1, N)\}$ are the I matched edges. Then the elements of the matrix $\boldsymbol{B}_N$ can be written as

$$b_{st} \overset{D}{=} \sum_{j=1}^I \left(\mathbf{1}\{L_{2j-1} = s, L_{2j} = t\} + \mathbf{1}\{L_{2j-1} = t, L_{2j} = s\} \right).$$

Moreover, $\eta_s = \sum_{j=1}^I \mathbf{1}\{L_{2j-1} = s\} + \sum_{j=1}^I \mathbf{1}\{L_{2j} = s\}$. Therefore, conditional on $\mathcal{Z}_N$ the $\binom{K}{2} + K$ vector $\boldsymbol{V}_N := (\underline{\boldsymbol{B}}_N, \eta_1, \ldots, \eta_K)'$ can be written as the sum of I i.i.d. random vectors. This implies, under H_0, as $N \to \infty$, by the multivariate CLT,

$$\frac{\boldsymbol{V}_N - \mathbb{E}_{H_0}(\boldsymbol{V}_N)}{\sqrt{I}} \Big| \mathcal{Z}_N \overset{D}{\to} N_{\binom{K}{2}+K}(0, \boldsymbol{\Gamma}_0), \tag{A.7}$$

where

$$\boldsymbol{\Gamma}_0 := \text{Cov}(\underline{\boldsymbol{B}}, \bar{\eta}_1, \bar{\eta}_2, \ldots, \bar{\eta}_K),$$

where

- $\boldsymbol{B} = ((\mathbf{1}\{\bar{L}_1 = s, \bar{L}_2 = t\}))_{1 \le s,t \le K}$, and
- $\bar{\eta}_s = \mathbf{1}\{\bar{L}_1 = s\} + \mathbf{1}\{\bar{L}_2 = s\}$, and
- $\bar{L}_1, \bar{L}_2$ are i.i.d. random variables taking value s with probability p_s, for $1 \le s \le K$. (This is the limit of the random variable L_1 defined in (A.6).)

As the RHS in (A.7) does not depend on the conditioning event, the unconditional limit is also the same:

$$\frac{\boldsymbol{V}_N - \mathbb{E}_{H_0}(\boldsymbol{V}_N)}{\sqrt{I}} \overset{D}{\to} N_{\binom{K}{2}+K}(0, \boldsymbol{\Gamma}_0).$$

Then by [22, Theorem 2], there exists a $\binom{K}{2} \times \binom{K}{2}$ matrix $\mathbf{\Gamma}$ such that, under H_0,

$$\frac{1}{\sqrt{N}}\left(\underline{\boldsymbol{A}}_N - \mathbb{E}_{H_0}(\underline{\boldsymbol{A}}_N)\right) \stackrel{D}{=} \frac{1}{\sqrt{N}}\left(\underline{\boldsymbol{B}}_N - \mathbb{E}_{H_0}(\underline{\boldsymbol{B}}_N)\right) | \{(\eta_1, \ldots, \eta_K) = (N_1, \ldots, N_K)\}$$
$$\stackrel{D}{\to} N_{\binom{K}{2}}(0, \mathbf{\Gamma}), \tag{A.8}$$

which completes the proof of (1).

To see (2) note that the fourth moments of the elements of $\frac{\underline{\boldsymbol{A}}_N - \mathbb{E}_{H_0}(\underline{\boldsymbol{A}}_N)}{\sqrt{N}}$ are bounded. Therefore, by uniform integrability, $\frac{1}{N}\mathrm{Cov}_{H_0}(\underline{\boldsymbol{A}}_N) \to \mathbf{\Gamma}$. The invertibility of $\mathrm{Cov}_{H_0}(\underline{\boldsymbol{A}}_N)$ (and hence $\mathbf{\Gamma}$) follows from Lemma C.1 (in Appendix C). The result in (2.9) then follows from (A.8) and an application of the Slutsky's theorem.

A.3. **Proof of Theorem 2.4.** The entry-wise almost sure limit of $\frac{1}{N}\boldsymbol{A}_N$ as in (2.13) is a direct consequence of [3, Proposition 1] (by choosing $\phi = \sum_{s=1}^K p_s f_s$ and $\phi_N = \frac{1}{N}\sum_{s=1}^K N_s f_s$ in [3, Proposition 1]).

Now, to prove consistency we need show that the test statistics have different limits under the null and the alternative. To this end, we have the following lemma.

Lemma A.1. *Let $H(f_1, f_2, \ldots, f_K)$ be as defined in (2.14). Then*

$$H(f_1, f_2, \ldots, f_K) \le H(f, f, \ldots, f),$$

and equality holds if and only if $f_1 = f_2 = \cdots = f_K$ outside a set of Lebesgue measure 0.

Proof. It follows from the Cauchy-Schwarz inequality, that for every $s \in [K]$,

$$\int_{\mathbb{R}^d} \frac{p_s^2 \left(f_s(z)\right)^2}{\sum_{u=1}^K p_u f_u(z)} \mathrm{d}z \ge \frac{\left(\int_{\mathbb{R}^d} p_s f_s(z) \mathrm{d}z\right)^2}{\int_{\mathbb{R}^d} \sum_{u=1}^K p_u f_u(z) \mathrm{d}z} = p_s^2. \tag{A.9}$$

This implies, $H(f_1, f_2, \ldots, f_K) = \frac{1}{2} - \mathrm{tr}(\boldsymbol{H}) \le \frac{1}{2} - \frac{1}{2}\sum_{s=1}^K p_s^2 = H(f, f, \ldots, f)$, as required.

Now, note that equality holds in (A.9) if and only if $p_s f_s = c_s \sum_{u=1}^K p_u f_u$, for some constant c_s, almost everywhere. Integrating both sides of the last relation, gives $c_s = p_s$, that is, $f_s = \sum_{u=1}^K p_u f_u$, almost everywhere. Therefore, equality holds if and only if $f_1 = f_2 = \cdots = f_K$, outside a set of Lebesgue measure 0. □

Now, suppose there exists $1 \le s \ne t \le K$ such that $f_s \ne f_t$ on a set of positive Lebesgue measure. By (2.13) and Lemma A.1, it follows that

$$\frac{R_{K,N} - \mathbb{E}_{H_0}(R_{K,N})}{N} \xrightarrow{a.s.} H(f_1, f_2, \ldots, f_k) - H(f, f, \ldots, f) < 0,$$

and hence, $\frac{1}{\sqrt{N}}(R_{K,N} - \mathbb{E}_{H_0}(R_{K,N})) \xrightarrow{a.s.} -\infty$. Now, since the matrix $\mathrm{Cov}_{H_0}(\underline{\boldsymbol{A}}_N)$ scales with N (by Proposition 2.2), $\lim_{N\to\infty} \frac{1}{N}\mathrm{Var}_{H_0}(R_{K,N}) < \infty$. Hence,

$$\lim_{N\to\infty} Q_{K,N} = \lim_{N\to\infty} \frac{R_{K,N} - \mathbb{E}_{H_0}(R_{K,N})}{\sqrt{\mathrm{Var}_{H_0}(R_{K,N})}} \xrightarrow{a.s.} -\infty, \tag{A.10}$$

which implies the limiting power of the MCM test $\lim_{N\to\infty} \mathbb{P}_{H_1}(Q_{K,N} < z_\alpha) = 1$, proving universal consistency.

For the MMCM test note that under the alternative, $\frac{1}{N}(\underline{\boldsymbol{A}}_N - \mathbb{E}_{H_0}\underline{\boldsymbol{A}}_N) \xrightarrow{a.s.} \Delta_0 \in \mathbb{R}^{\binom{K}{2}}$, where sum of the entries of γ_0 is $H(f_1, f_2, \ldots, f_K) - H(f, f, \ldots, f)$, that is, Δ_0 is non-zero. Now, since $\frac{1}{N}\mathrm{Cov}_{H_0}(\underline{\boldsymbol{A}}_N) \to \boldsymbol{\Gamma}$, where $\boldsymbol{\Gamma}$ is positive definite (by Proposition 2.2 and Lemma C.1),

$$\frac{1}{N} S_{K,N} \xrightarrow{a.s.} \Delta_0^\top \boldsymbol{\Gamma}^{-1} \Delta_0 > 0,$$

that is, $S_{K,N} \xrightarrow{a.s.} \infty$ under the alternative, proving universal consistency of the MMCM test.

Appendix B. Proof of Theorem 4.1

Recall the alternative way to describe the joint distribution of the data described in Section 4: Choose $\mathcal{Z}_N := (Z_1, Z_2, \ldots, Z_N)$ i.i.d. from the density $\phi_N = \sum_{s=1}^K \frac{N_s}{N} f_s$ in $\mathbb{R}^d$. Then given $\mathcal{Z}_N = (Z_1, Z_2, \ldots, Z_N)$, assign a random label $L_j \in [K] := \{1, 2, \ldots, K\}$ to Z_j, independently for each $1 \le j \le N$, as in (4.1). Then it is easy to verify that the following fact, which is proved in Appendix C.2.

Lemma B.1. *The joint distribution of* $(\{Z_j : L_j = 1\}, \{Z_j : L_j = 2\}, \ldots, \{Z_j : L_j = K\})$ *conditional on* $(\eta_1, \cdots, \eta_K) = (N_1, N_2, \ldots, N_K)$ *is same as the joint distribution of the data* $\left(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \ldots, \boldsymbol{X}^{(K)}\right)$, *where* $\eta_s := \sum_{i=1}^N \mathbf{1}\{L_i = s\}$, *the number of elements labelled* s, *for* $1 \le s \le K$.

The proof of Theorem 4.1 has two steps: (1) Computing the conditional covariance matrix $\boldsymbol{R}(\mathcal{Z}_N) := \mathrm{Cov}_{H_1}((\underline{\boldsymbol{B}}_N^\top, \eta_1, \ldots, \eta_{K-1})|\mathcal{Z}_N)$ (as usual, $\underline{\boldsymbol{B}}_N$ denotes the vectorized upper triangular part of the matrix $\boldsymbol{B}_N$ defined in (4.2)), under the bootstrap alternative distribution (which is the unconditional distribution of $(\{Z_j : L_j = 1\}, \{Z_j : L_j = 2\}, \ldots, \{Z_j : L_j = K\})$), and show that scales with N (Section B.1), and (2) deriving the asymptotic normality of $\underline{\boldsymbol{A}}_N$ from the joint distribution of the vector $(\underline{\boldsymbol{B}}_N^\top, \eta_1, \ldots, \eta_{K-1})^\top$ under the bootstrap alternative distribution (Section B.2).

B.1. Computing the Joint Conditional Covariance Matrix.

Given the K densities $f_1, f_2, \ldots, f_K$, we will begin by defining the matrix $\Gamma_{f_1, f_2, \ldots, f_K}$ in Theorem 4.1. To this end, let $\phi = \sum_{s=1}^K p_s f_s$ and for each $s, t \in [K]$, define the function $h_{st} : \mathbb{R}^d \times \mathbb{R}^d \mapsto [0, 1]$ as:

$$h_{st}(x, y) := \frac{p_s p_t f_s(x) f_t(y)}{\phi(x)\phi(y)},$$

and set $\bar{h}_{st}(x, y) := h_{st}(x, y) + h_{st}(y, x)$.

Definition B.1. (Defining the matrix $\boldsymbol{\Gamma}_{f_1, f_2, \ldots, f_K}$) Throughout, let $Z \sim \phi = \sum_{s=1}^K p_s f_s$. To begin with, let $\boldsymbol{Q}$ be a square matrix of dimension $\binom{K}{2} + K - 1$, partitioned as:

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{Q}_{11} & \boldsymbol{Q}_{12} \\ \boldsymbol{Q}_{12}^\top & \boldsymbol{Q}_{22} \end{bmatrix}, \tag{B.1}$$

where $\boldsymbol{Q}_{11}$, $\boldsymbol{Q}_{12}$, and $\boldsymbol{Q}_{22}$ have dimensions $\binom{K}{2} \times \binom{K}{2}$, $\binom{K}{2} \times (K-1)$ and $(K-1) \times (K-1)$, respectively, and their elements are defined as follows:

- The elements of the matrix $\boldsymbol{Q}_{11}$ will be denoted by $q_{11}((s,t),(u,v))$, for $1 \le s < t \le K$ and $1 \le u < v \le K$, which is defined as:

$$q_{11}((s,t),(u,v)) := \begin{cases} \frac{1}{2}\mathbb{E}\left[\bar{h}_{st}(Z,Z)(1-\bar{h}_{st}(Z,Z))\right] & \text{if } (s,t)=(u,v), \\ -\frac{1}{2}\mathbb{E}\left[\bar{h}_{st}(Z,Z)\bar{h}_{uv}(Z,Z)\right] & \text{otherwise.} \end{cases} \tag{B.2}$$

- The elements of the matrix $\boldsymbol{Q}_{12}$ will be denoted by $q_{12}((s,t),u)$, for $1 \le s < t \le K$ and $1 \le u \le K-1$, which is defined as:

$$q_{12}((s,t),u) = \begin{cases} \frac{1}{2}\mathbb{E}\left[\bar{h}_{st}(Z,Z)\left(1-\frac{2p_u f_u(Z)}{\phi(Z)}\right)\right] & \text{if } u \in \{s,t\}, \\ -\mathbb{E}\left[\bar{h}_{st}(Z,Z)\left(\frac{p_u f_u(Z)}{\phi(Z)}\right)\right] & \text{otherwise.} \end{cases} \tag{B.3}$$

- The elements of the matrix $\boldsymbol{Q}_{22}$ will be denoted by $q_{22}(s,t)$, for $1 \le s,t \le K-1$, which is defined as:

$$q_{22}(s,t) := \begin{cases} p_s(1-p_s) & \text{if } s=t, \\ -p_s p_t & \text{otherwise.} \end{cases} \tag{B.4}$$

Finally, define

$$\boldsymbol{\Gamma}_{f_1,f_2,\ldots,f_K} := \boldsymbol{Q}_{11} - \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{12}^{\top}. \tag{B.5}$$

To compute the limit of the covariance matrix $\boldsymbol{R}(\mathcal{Z}_N) := \mathrm{Cov}_{H_1}((\underline{\boldsymbol{B}}_N^{\top}, \eta_1, \ldots, \eta_{K-1})|\mathcal{Z}_N)$, we need the following lemma from [3]. Recall $\phi := \sum_{i=1}^{s} p_s f_s$.

Lemma B.2. [3, Proposition 1] *Let $Z_1, Z_2, \ldots, Z_N$ be i.i.d. from the density $\phi_N = \sum_{s=1}^{K} \frac{N_s}{N} f_s$, and $g : \mathbb{R}^d \times \mathbb{R}^d \mapsto [0,1]$ be a symmetric, measurable function, such that almost any $z \in \mathbb{R}^d$ is a Lebesgue continuity point of $\phi(\cdot)g(z,\cdot)$. Then, as $N \to \infty$,*

$$\frac{1}{N}\sum_{1 \le i < j \le N} e(Z_i, Z_j)g(Z_i, Z_j) \xrightarrow{P} \frac{1}{2}\,\mathbb{E}g(Z,Z),$$

where $e(x,y) := \mathbf{1}\{(x,y) \in E(\mathscr{G}(\mathcal{Z}_N \cup \{x,y\})\}$ and $Z \sim \phi$.

The following lemma uses the above result to show that the conditional covariance matrix of $(\underline{\boldsymbol{B}}_N^{\top}, \eta_1, \ldots, \eta_{K-1})$ divided by N, converges to a deterministic limit in probability.

Lemma B.3. *Let $\boldsymbol{R}(\mathcal{Z}_N) := \mathrm{Cov}_{H_1}((\underline{\boldsymbol{B}}_N^{\top}, \eta_1, \ldots, \eta_{K-1})|\mathcal{Z}_N)$ be the conditional covariance matrix under the bootstrap alternative distribution. Then*

$$\frac{1}{N}\boldsymbol{R}(\mathcal{Z}_N) \xrightarrow{P} \boldsymbol{R} := \begin{bmatrix} \boldsymbol{Q}_{11} & \boldsymbol{Q}_{12} \\ \boldsymbol{Q}_{12}^{\top} & \boldsymbol{R}_{22} \end{bmatrix}, \tag{B.6}$$

where the matrices $\boldsymbol{Q}_{11}$ and $\boldsymbol{Q}_{12}$ are as defined in (B.2) and (B.3), respectively, and $\boldsymbol{R}_{22} = ((r_{22}(s,t)))_{1 \le s,t \le K-1}$, where

$$r_{22}(s,t) := \begin{cases} \mathbb{E}\left[\frac{p_s f_s(Z)}{\phi(Z)}\left(1-\frac{p_s f_s(Z)}{\phi(Z)}\right)\right] & \text{if } s=t, \\ -\mathbb{E}\left[\frac{p_s p_t f_s(Z) f_t(Z)}{\phi(Z)^2}\right] & \text{if } s \ne t, \end{cases} \tag{B.7}$$

with $Z \sim \phi := \sum_{s=1}^K p_s f_s$, *as before.*

Proof. Let $\phi_N = \sum_{s=1}^K \frac{N_s}{N} f_s$, and recall that $\eta_s = \sum_{j=1}^N \mathbf{1}\{L_j = s\}$, for $s \in [K]$. This implies, for $1 \le s, \neq t \in [K-1]$,

$$\mathrm{Var}_{H_1}(\eta_s|\mathcal{Z}_N) = \sum_{j=1}^N \frac{\frac{N_s}{N} f_s(Z_j)}{\phi_N(Z_j)} \left(1 - \frac{\frac{N_s}{N} f_s(Z_j)}{\phi_N(Z_j)}\right),$$

and $\mathrm{Cov}_{H_1}(\eta_s, \eta_t|\mathcal{Z}_N) = -\sum_{j=1}^N \frac{\frac{N_s}{N}\frac{N_t}{N} f_s(Z_j) f_t(Z_j)}{\phi_N(Z_j)^2}$. Hence, by the law of large numbers and the dominated convergence theorem, as $N \to \infty$,

$$\frac{1}{N}\mathrm{Cov}_{H_1}((\eta_1, \eta_2, \ldots, \eta_K)^\top|\mathcal{Z}_N) \xrightarrow{P} \boldsymbol{R}_{22}, \tag{B.8}$$

where $\boldsymbol{R}_{22}$ is as defined in (B.7).

Next, define $h_{st}^{(N)}(x,y) := \frac{\frac{N_s}{N}\frac{N_t}{N} f_s(x) f_y(t)}{\phi_N(x)\phi_N(y)}$, and let $\bar{h}_{st}^{(N)}(x,y) := h_{st}^{(N)}(x,y) + h_{st}^{(N)}(x,y)$. Now, for each $1 \le s < t \le K$, the conditional variance of b_{st} (recall (4.2)) is,

$$\begin{aligned}\frac{1}{N}\mathrm{Var}_{H_1}(b_{st}|\mathcal{Z}_N) &= \frac{1}{N}\sum_{1\le i<j\le N} e(Z_s, Z_t)\bar{h}_{st}^{(N)}(Z_s, Z_t)(1 - \bar{h}_{st}^{(N)}(Z_s, Z_t)) \\ &\xrightarrow{P} \frac{1}{2}\mathbb{E}\left[\bar{h}_{st}(Z,Z)(1 - \bar{h}_{st}(Z,Z))\right], \end{aligned} \tag{B.9}$$

by Proposition B.2, as $\bar{h}_{st}^{(N)} \to \bar{h}_{st}$ uniformly. Similarly, for any two distinct pairs (s,t) and (u,v) with $1 \le s < t \le K$ and $1 \le u < v \le K$,

$$\frac{1}{N}\mathrm{Cov}_{H_1}(b_{st}, b_{uv}|\mathcal{Z}_N) \xrightarrow{P} -\frac{1}{2}\mathbb{E}\left[\bar{h}_{st}(Z,Z)\bar{h}_{uv}(Z,Z)\right]. \tag{B.10}$$

Combining (B.9) and (B.10) gives

$$\frac{1}{N}\mathrm{Cov}_{H_1}(\underline{\boldsymbol{B}}_N^\top|\mathcal{Z}_N) \xrightarrow{P} \boldsymbol{Q}_{11}, \tag{B.11}$$

where $\boldsymbol{Q}_{11}$ is as defined in (B.2).

Finally, for each $1 \le s, t \le K$ and $1 \le u \in K-1$,

$$\mathrm{Cov}_{H_1}(b_{st}, \eta_u|\mathcal{Z}_N) = \sum_{1\le s<t\le N} e(Z_i, Z_j)\psi_{\{(s,t),u\}}^{(N)}(Z_i, Z_j),$$

where

$$\psi_{\{(s,t),u\}}^{(N)}(Z_i, Z_j) = \begin{cases} \bar{h}_{st}^{(N)}(Z_i, Z_j)\left[1 - \frac{\frac{N_u}{N} f_u(Z_i)}{\phi_N(Z_i)} - \frac{\frac{N_u}{N} f_u(Z_j)}{\phi_N(Z_j)}\right] & \text{if } u \in \{s,t\}), \\ -\bar{h}_{st}^{(N)}(Z_i, Z_j)\left[\frac{\frac{N_u}{N} f_u(Z_i)}{\phi_N(Z_i)} + \frac{\frac{N_u}{N} f_u(Z_j)}{\phi_N(Z_j)}\right] & \text{if } u \notin \{s,t\}. \end{cases}$$

Then by Lemma B.2, as $N \to \infty$,

$$\frac{1}{N}\mathrm{Cov}_{H_1}(b_{st}, \eta_u|\mathcal{Z}_N) \xrightarrow{P} q_{12}((s,t),u), \tag{B.12}$$

where $q_{12}((s,t),u)$ is as in (B.3).

The result in (B.6) now follows by combining (B.8), (B.11), and (B.12), completing the proof of the lemma. □

B.2. **The Joint Central Limit Theorem of the Cross-Counts.** We now have all the tools necessary for proving Theorem 4.1. Towards this, define

$$V_N := \frac{1}{\sqrt{N}} \left(\underline{\boldsymbol{B}}_N^\top - \mathbb{E}_{H_1}(\underline{\boldsymbol{B}}_N^\top | \mathcal{Z}_N),\ \eta_1 - \mathbb{E}_{H_1}(\eta_1 | \mathcal{Z}_N),\ \cdots,\ \eta_{K-1} - \mathbb{E}_{H_1}(\eta_{K-1} | \mathcal{Z}_N) \right)^\top,$$

a vector of length $\binom{K}{2} + K - 1$. Define $U_N := \overline{\boldsymbol{R}}(\mathcal{Z}_N)^{-\frac{1}{2}} V_N$, where $\overline{\boldsymbol{R}}(\mathcal{Z}_N) = \frac{1}{N}\boldsymbol{R}(\mathcal{Z}_N)$, so that $\mathbb{E}_{H_1}(U_N|\mathcal{Z}_N) = 0$ and $\mathrm{Cov}_{H_1}(U_N|\mathcal{Z}_N) = \mathrm{I}$, under the bootstrap alternative distribution.

For each $(a, b) \in [N]^2$, define the $K \times K$ matrix $\boldsymbol{C}_{ab} = ((C_{ab}(s,t)))_{1 \le s \ne t \le K}$, where,

$$C_{ab}(s,t) := \mathbf{1}\left\{ \{L_a, L_b\} = \{s,t\} \right\},$$

for $s \ne t$ and zero otherwise. Let $\underline{\boldsymbol{C}}_{ab}$ be the vector of length $\binom{K}{2}$ obtained by concatenating the rows of $\boldsymbol{C}_{ab}$ in the upper triangular part. Now, for each $(a, b) \in [N]^2$, define:

$$Y_{ab} := \left(\underline{\boldsymbol{C}}_{ab}^\top,\ \mathbf{1}\{L_a = 1\} + \mathbf{1}\{L_b = 1\},\ \cdots,\ \mathbf{1}\{L_a = K-1\} + \mathbf{1}\{L_b = K-1\} \right)^\top,$$

and let $\overline{Y}_{ab} := \frac{1}{\sqrt{N}} \overline{\boldsymbol{R}}(\mathcal{Z}_N)^{-\frac{1}{2}} (Y_{ab} - \mathbb{E}_{H_1}(Y_{ab}|\mathcal{Z}_N))$. Further, define $S(\mathcal{Z}_N) := \{\{a, b\} \subset [N] : e(Z_a, Z_b) = 1\}$. Then, it is easy to see that:

$$U_N = \sum_{\{a,b\} \in S(\mathcal{Z}_N)} \overline{Y}_{ab}.$$

Note that, under the bootstrap alternative distribution, given $\mathcal{Z}_N$, the collection $\{\overline{Y}_{ab}\}_{\{a,b\} \in S(\mathcal{Z}_N)}$ is independent, so by an application of the multivariate Berry-Essen theorem [37, Theorem 1.1], we get:

$$\sup_{A \in \mathcal{C}} \left| \mathbb{P}(U_N \in A | \mathcal{Z}_N) - \Phi_{\binom{K}{2}+K-1}(A) \right| \le L(K) \sum_{\{a,b\} \in S(\mathcal{Z}_N)} \mathbb{E}\left(||\overline{Y}_{ab}||^3 \big| \mathcal{Z}_N \right), \tag{B.13}$$

where $\mathcal{C}$ denotes the class of all measurable convex subsets of $\mathbb{R}^{\binom{K}{2}+K-1}$, $\Phi_{\binom{K}{2}+K-1}(\cdot)$ the standard normal distribution function in dimension $\binom{K}{2} + K - 1$, and $L(K)$ is a constant depending only on K.

Lemma B.4. *Let $\overline{Y}_{ab}$ be as defined above. Then $\sum_{\{a,b\} \in S(\mathcal{Z}_N)} \mathbb{E}\left(||\overline{Y}_{ab}||^3 \big| \mathcal{Z}_N \right) \xrightarrow{P} 0$, as $N \to \infty$.*

Proof. Note that every entry of the vector $Y_{ab} - \mathbb{E}_{H_1}(Y_{ab}|\mathcal{Z}_N)$ is bounded in absolute value by 2. Hence, $\|Y_{ab} - \mathbb{E}_{H_1}(Y_{ab}|\mathcal{Z}_N)\| \le \sqrt{2(K-1)(K+2)}$. Consequently, denote the operator norm of matrix by $||\cdot||_{\mathrm{op}}$

$$\|\overline{Y}_{ab}\| \le \frac{1}{\sqrt{N}} \left\| \overline{\boldsymbol{R}}(\mathcal{Z}_N)^{-\frac{1}{2}} \right\|_{\mathrm{op}} \left\| Y_{ab} - \mathbb{E}\left(Y_{ab} \big| \mathcal{Z}_N \right) \right\| \le \left(\frac{2(K-1)(K+2)}{N} \right)^{\frac{1}{2}} \left\| \overline{\boldsymbol{R}}(\mathcal{Z}_N)^{-\frac{1}{2}} \right\|_{\mathrm{op}}.$$

Then by the Cauchy-Schwarz inequality,

$$\sum_{\{a,b\} \in S(\boldsymbol{Z})} \mathbb{E}\left(||\overline{Y}_{ab}||^3 \big| \mathcal{Z}_N \right) \quad \le \quad \sqrt{\frac{2(K-1)^3(K+2)^3}{N}} \left\| \overline{\boldsymbol{R}}(\mathcal{Z}_N)^{-\frac{1}{2}} \right\|_{\mathrm{op}}^3$$

$$\leq \sqrt{\frac{2(K-1)^3(K+2)^3}{N}} \left[\mathrm{tr}\left(\overline{\boldsymbol{R}}(\mathcal{Z}_N)^{-\frac{1}{2}}\right)\right]^3.$$

The RHS above converges to zero in probability, because by Lemma B.3 $\overline{\boldsymbol{R}}(\mathcal{Z}_N)$ converges in probability. □

The lemma combined with (B.13) shows that, under the bootstrap alternative distribution, the vector $U_N|\mathcal{Z}_N$ converges in distribution to $N_{\binom{K}{2}+K-1}(0, \mathrm{I})$, and by Lemma B.3 $V_N|\mathcal{Z}_N$ converges in distribution to $N_{\binom{K}{2}+K-1}(0, \boldsymbol{R})$, where $\boldsymbol{R}$ is as defined in (B.6). Hence, for every vector $t \in \mathbb{R}^{\binom{K}{2}+K-1}$,

$$\mathbb{E}\left(e^{it^\top V_N}\Big|\mathcal{Z}_N\right) \xrightarrow{P} \mathbb{E}\left(e^{it^\top W}\right), \tag{B.14}$$

where $W \sim N_{\binom{K}{2}+K-1}(0, \boldsymbol{R})$. Next, define

$$B_N := \frac{1}{\sqrt{N}}\left(\mathbf{0}^\top,\ \mathbb{E}_{H_1}(\eta_1|\mathcal{Z}_N) - N_1,\ \cdots,\ \mathbb{E}_{H_1}(\eta_{K-1}|\mathcal{Z}_N) - N_{K-1}\right)^\top,$$

where the $\mathbf{0}$ here denotes a vector of all zeros of length $\binom{K}{2}$. Now, by the usual central limit theorem, under the bootstrap alternative distribution, as $N \to \infty$, $B_N \xrightarrow{D} N_{\binom{K}{2}}(0, \boldsymbol{\Psi})$, where

$$\boldsymbol{\Psi} := \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{M} \end{bmatrix},$$

with the elements of $\boldsymbol{M}$ will be denoted by $m(s,t)$, for $1 \leq s,t \leq K-1$, and

$$m(s,t) := \begin{cases} \mathrm{Var}_{Z\sim\phi}\left[\frac{p_s f_s(Z)}{\phi(Z)}\right] & \text{if } s = t, \\ \mathrm{Cov}_{Z\sim\phi}\left[\frac{p_s f_s(Z)}{\phi(Z)}, \frac{p_t f_t(Z)}{\phi(Z)}\right] & \text{if } s \neq t. \end{cases} \tag{B.15}$$

Now, recalling the definitions of the matrix $\boldsymbol{Q}$ (from (B.1)) and the matrix $\boldsymbol{R}$ (from (B.6)), it is easy to see that $\boldsymbol{Q} = \boldsymbol{R} + \boldsymbol{\Psi}$. Hence, by (B.14) and Lemma C.2 (putting $A_N = V_N$, $\boldsymbol{C}_N = \mathcal{Z}_N$ and $f_N(\boldsymbol{C}_N) = B_N$), gives

$$\frac{1}{\sqrt{N}}\left(\underline{\boldsymbol{B}}_N^\top - \mathbb{E}_{H_1}(\underline{\boldsymbol{B}}_N^\top|\mathcal{Z}_N),\ \Delta_N^\top\right)^\top = V_N + B_N$$
$$\xrightarrow{D} N_{\binom{K}{2}}(\mathbf{0}, \boldsymbol{R}+\boldsymbol{\Psi}) \overset{D}{=} N(\mathbf{0}, \boldsymbol{Q}).$$

where $\Delta_N := (\eta_1 - N_1,\ \cdots,\ \eta_{K-1} - N_{K-1})^\top$. Therefore, by Lemma B.1 and (B.5), the distribution of $\frac{1}{\sqrt{N}}(\underline{\boldsymbol{B}}_N^\top - \mathbb{E}_{H_1}(\underline{\boldsymbol{B}}_N^\top|\mathcal{Z}_N))$ conditional on $\Delta_N = \mathbf{0}$, converges to $N_{\binom{K}{2}}(\mathbf{0}, \boldsymbol{Q}_{11} - \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{12}^\top)$, as required.

Appendix C. Proofs of Technical Lemmas

Here, we collect the proofs of the different technical lemmas, required in the proofs above. In Section C.1, we show the invertibility of the matrix $\mathrm{Cov}_{H_0}(\underline{\boldsymbol{A}}_N)$. The proof of Lemma B.1 is given in Section C.2. Other technical lemma used in the proof of Theorem B are proved in Section C.3.

C.1. **Invertibility of the Count Matrix Under the Null.** In order for the MCMM statistic to be well-defined, we need to make sure the matrix $\mathrm{Cov}_{H_0}(\underline{\boldsymbol{A}}_N)$ (recall Proposition 2.2) is invertible. This is proved in the following lemma:

Lemma C.1. *The matrix* $\mathrm{Cov}_{H_0}(\underline{\boldsymbol{A}}_N)$ *is invertible.*

Proof. For simplicity, we assume that the sample sizes N_s are even, for all $1 \le s \le K$. For $1 \le r < s \le K$, define a $K \times K$ matrices $\boldsymbol{a}_{rs} = ((a_{rs}(u,v)))_{1\le u,v\le K}$ as follows:

$$\begin{aligned}
&a_{rs}(r,s) = a_{rs}(s,r) = \min\{N_r, N_s\},\\
&a_{rs}(u,v) = 0, && \text{for all } \{u,v\} \ne \{r,s\} \text{ and } u \ne v,\\
&a_{rs}(r,r) = \tfrac{1}{2}(N_r - \min\{N_r, N_s\}),\\
&a_{rs}(s,s) = \tfrac{1}{2}(N_s - \min\{N_r, N_s\}),\\
&a_{rs}(u,u) = \tfrac{1}{2}N_u, && \text{for all } u \ne \{r,s\}.
\end{aligned}$$

Clearly, $\boldsymbol{a}_{rs} \in \mathscr{B}$ (recall Proposition 2.1), which implies, by (2.6), $\mathbb{P}_{H_0}(\boldsymbol{A}_N = \boldsymbol{a}_{rs}) > 0$, for all $1 \le r < s \le K$. Now, as in (2.5), denote by $\underline{\boldsymbol{a}}_{rs}$ the vector of length $\binom{K}{2}$ obtained by concatenating the rows of $\boldsymbol{a}_{rs}$ in the upper triangular part. The argument above shows that $\mathbb{P}_{H_0}(\underline{\boldsymbol{A}}_N = \underline{\boldsymbol{a}}_{rs}) > 0$, for all $1 \le r < s \le K$. Moreover, also note that $\mathbb{P}_{H_0}(\underline{\boldsymbol{A}}_N = \boldsymbol{0}) > 0$, where $\boldsymbol{0}$ denotes the vector of length $\binom{K}{2}$ with all entries 0. Also, note that the vectors $\{\underline{\boldsymbol{a}}_{rs}\}_{1\le r<s\le K}$, each of which has only have one non-zero element corresponding to the element $a_{rs}(r,s)$, form a basis of $\mathbb{R}^{\binom{K}{2}}$.

Now, suppose that $\mathrm{Cov}_{H_0}(\underline{\boldsymbol{A}}_N)$ is singular, whence there exists a non-zero vector $\eta \in \mathbb{R}^{\binom{K}{2}}$ such that $\mathrm{Cov}_{H_0}(\underline{\boldsymbol{A}}_N)\eta = 0$. This implies that $\mathrm{Var}_{H_0}(\eta^\top \underline{\boldsymbol{A}}_N) = 0$, and hence,

$$\mathbb{P}_{H_0}\left(\eta^\top \underline{\boldsymbol{A}}_N = \mathbb{E}_{H_0}(\eta^\top \underline{\boldsymbol{A}}_N)\right) = 1. \tag{C.1}$$

The fact $\mathbb{P}_{H_0}(\underline{\boldsymbol{A}}_N = 0) > 0$, now implies that $\mathbb{E}_{H_0}(\eta^\top \underline{\boldsymbol{A}}_N) = 0$ (otherwise, assuming $\mathbb{E}_{H_0}(\eta^\top \underline{\boldsymbol{A}}_N) \ne 0$, leads to, by (C.1), $\mathbb{P}_{H_0}(\underline{\boldsymbol{A}}_N = 0) \le \mathbb{P}_{H_0}(\eta^\top \underline{\boldsymbol{A}}_N = 0) = 0$, which is a contradiction). Again, since $\mathbb{P}_{H_0}(\underline{\boldsymbol{A}}_N = \underline{\boldsymbol{a}}_{rs}) > 0$, it follows that $\eta^\top \underline{\boldsymbol{a}}_{rs} = 0$, for all $1 \le r < s \le K$. This implies, since the vectors $\{\underline{\boldsymbol{a}}_{rs}\}_{1\le r<s\le K}$ form a basis of $\mathbb{R}^{\binom{K}{2}}$, $\eta = \boldsymbol{0}$, which is a contradiction. □

C.2. **Proof of Lemma B.1.** For notational convenience, we will prove the result only for the case $K = 2$. The proof for general K follows similarly. We begin with a few notations: Let Π denote the set of all permutations σ of $[N]$, such that $\sigma(a) < \sigma(b)$, for all $1 \le a < b \le N_1$, and $\sigma(a) < \sigma(b)$, for all $N_1 + 1 \le a < b \le N_1 + N_2 = N$. Moreover, for a vector $x \in \mathbb{R}^d$ and a $\mathbb{R}^d$-valued random variable X, we denote by $\{X \le x\}$ the event $\{X \in \{y \in \mathbb{R}^d : y \le x\}\}$.[9]

Now, considering the sets $\{Z_i : L_i = 1\}$ and $\{Z_i : L_i = 2\}$ as vectors with the indices arranged in increasing order, it follows that

$$(\{Z_i : L_i = 1\}, \{Z_i : L_i = 2\}) = (Z_{\pi(1)}, \cdots, Z_{\pi(N)}),$$

where π is a random permutation of $[N]$, such that $\pi(1) < \pi(2) \cdots < \pi(\eta_1)$ and $\pi(\eta_1 + 1) < \pi(\eta_1 + 2) \cdots < \pi(\eta_1 + \eta_2)$, where $L_{\pi(a)} = 1$, for all $1 \le a \le \eta_1$, and $L_{\pi(a)} = 2$, for all

[9]For any two vectors $u = (u_1, u_2, \ldots, u_d) \in \mathbb{R}^d$ and $v = (v_1, v_2, \ldots, v_d) \in \mathbb{R}^d$, we write $u \le v$, if $u_a \le v_a$, for all $1 \le a \le d$.

$\eta_1 + 1 \le a \le \eta_1 + \eta_2$. Then, for every $z_1, \cdots, z_N \in \mathbb{R}^d$,

$$\begin{aligned}
&\mathbb{P}\left(Z_{\pi(1)} \le z_1, \cdots, Z_{\pi(N)} \le z_N \middle| \eta_1 = N_1 \right) \\
&= \frac{1}{\mathbb{P}(\eta_1 = N_1)} \sum_{\sigma \in \Pi} \mathbb{P}\left(Z_{\sigma(1)} \le z_1, \cdots, Z_{\sigma(N)} \le z_N,\ \pi = \sigma,\ \eta_1 = N_1 \right) \\
&= \frac{1}{\mathbb{P}(\eta_1 = N_1)} \sum_{\sigma \in \Pi} \mathbb{P}\left(\bigcap_{a=1}^{N_1} \{ Z_{\sigma(a)} \le z_a,\ L_{\sigma(a)} = 1 \} \bigcap \bigcap_{a=N_1+1}^{N_1+N_2} \{ Z_{\sigma(a)} \le z_a,\ L_{\sigma(a)} = 2 \} \right) \\
&= \frac{|\Pi|}{\mathbb{P}(\eta_1 = N_1)} \prod_{a=1}^{N_1} \mathbb{P}\left(Z_a \le z_a,\ L_a = 1 \right) \prod_{a=N_1+1}^{N_1+N_2} \mathbb{P}\left(Z_a \le z_a,\ L_a = 2 \right) \\
&= \frac{\binom{N}{N_1}}{\mathbb{P}(\eta_1 = N_1)} \prod_{a=1}^{N_1} \frac{N_1}{N} F_1(z_a) \prod_{a=N_1+1}^{N_1+N_2} \frac{N_2}{N} F_2(z_a) \\
&= \prod_{a=1}^{N_1} F_1(z_a) \prod_{a=N_1+1}^{N_1+N_2} F_2(z_a) \qquad \left(\text{using } \mathbb{P}(\eta_1 = N_1) = \tbinom{N}{N_1} \left(\tfrac{N_1}{N}\right)^{N_1} \left(\tfrac{N_2}{N}\right)^{N_2}\right) \\
&= \mathbb{P}(X_1^{(1)} \le z_1, \cdots, X_{N_1}^{(1)} \le z_{N_1}, X_1^{(2)} \le z_{N_1+1}, \cdots, X_{N_2}^{(2)} \le z_{N_1+N_2}),
\end{aligned}$$

which completes the proof of the lemma. □

C.3. **Missing Details in the Proof of Theorem 4.1.** Here, we provide the proof of a lemma used in the proof of Theorem 4.1.

Lemma C.2. *Let $\{X_N\}_{N \ge 1}$ be a sequence of $\mathbb{R}^p$-valued random vectors, for some $p \ge 1$, and $\boldsymbol{C}_N$ be a sequence random variable, such that $\mathbb{E}(e^{it^\top X_N} | \boldsymbol{C}_N) \xrightarrow{P} a$, for some real number a and some vector $t \in \mathbb{R}^p$. Moreover, suppose that f_N is a sequence of deterministic functions with codomain $\mathbb{R}^p$, such that $\mathbb{E}(e^{it^\top f_N(\boldsymbol{C}_N)}) \to b$, for some real number b. Then,*

$$\lim_{N \to \infty} \mathbb{E}\left(e^{it^\top (A_N + f_N(\boldsymbol{C}_N))} \right) = ab.$$

Proof. Note that,

$$\begin{aligned}
\left| \mathbb{E}\left(e^{it^\top (A_N + f_N(\boldsymbol{C}_N))} \right) - ab \right| &= \left| \mathbb{E}\left[e^{it^\top f_N(\boldsymbol{C}_N)} \mathbb{E}\left(e^{it^\top A_N} \middle| \boldsymbol{C}_N \right) \right] - ab \right| \\
&\le \left| \mathbb{E}\left[e^{it^\top f_N(\boldsymbol{C}_N)} \left(\mathbb{E}\left(e^{it^\top A_N} \middle| \boldsymbol{C}_N \right) - a \right) \right] \right| + |a| \left| \mathbb{E}\left[e^{it^\top f_N(\boldsymbol{C}_N)} - b \right] \right| \\
&\le \mathbb{E}\left| \mathbb{E}\left(e^{it^\top A_N} \middle| \boldsymbol{C}_N \right) - a \right| + |a| \left| \mathbb{E}\left(e^{it^\top f_N(\boldsymbol{C}_N)} - b \right) \right|.
\end{aligned}$$

The first term in the last expression goes to 0 by hypothesis and the dominated convergence theorem, while the last term goes to 0 by hypothesis, completing the proof. □

Appendix D. Additional Simulations

In this section, we present simulations comparing the empirical powers of the MCM and the MMCM tests for location, spherical scale, and equi-correlation scale changes in the log normal family. As before, in all the simulations, the power is calculated over 100 iterations,

the tests are implemented using the permutation distribution, and the nominal level is chosen to be 0.05.

$\Delta\downarrow$	Dimension	5	10	50	100	200	300	500
.06	MCM	**.13**	.17	**.49**	.65	.81	.88	.90
	MMCM	.10	**.18**	.42	**.68**	**.93**	**.98**	**.99**
.08	MCM	**.15**	.19	.52	.68	.87	.95	1.0
	MMCM	.12	**.20**	**.70**	**.92**	**.98**	**1.0**	**1.0**
.10	MCM	**.19**	.26	.79	.85	.96	1.0	1.0
	MMCM	.18	**.31**	**.94**	**1.0**	**1.0**	**1.0**	**1.0**
.12	MCM	**.50**	.69	1.0	1.0	1.0	1.0	1.0
	MMCM	.49	**.84**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**

(a)

$\Delta\downarrow$	Groups	4	6	8	10
.04	MCM	**.13**	**.22**	.49	.71
	MMCM	.10	.16	**.63**	**.99**
.06	MCM	**.21**	.35	.85	1.0
	MMCM	.14	**.41**	**.98**	**1.0**
.08	MCM	**.37**	.92	1.0	1.0
	MMCM	.31	**1.0**	**1.0**	**1.0**
.10	MCM	.52	1.0	1.0	1.0
	MMCM	**.56**	**1.0**	**1.0**	**1.0**

(b)

TABLE 6. Power of the MCM and the MMCM tests in the lognormal location family with (a) the number of classes $K = 6$ fixed, and (b) the dimension $d = 150$ fixed.

- *Lognormal Location*: Here, we consider samples from the following K log-normal distributions: $\exp(N_d((s-1)\Delta\cdot\mathbf{1}, \mathrm{I}))$, for $1 \le s \le K$. Table 6(a) shows the fixed class scenario, where we take $K = 6$ groups and vary the dimension d from 5 to 500, and Δ from 0.06 to 0.12. Table 6(b) shows the fixed dimension scenario, where the dimension $d = 150$ is fixed, the number of groups K varies along $4, 6, 8, 10$, and Δ varies from 0.04 to 0.10. In both cases, the sample sizes were taken in equal increments of 50, starting from 50.

$\Delta\downarrow$	Dimension	5	10	50	100	200	300	500
.15	MCM	**.15**	**.21**	.44	.70	.94	.99	1.0
	MMCM	.13	.17	**.66**	**.99**	**1.0**	**1.0**	**1.0**
.20	MCM	.16	.22	.70	.96	1.0	1.0	1.0
	MMCM	**.18**	**.35**	**.96**	**1.0**	**1.0**	**1.0**	**1.0**
.25	MCM	**.17**	.23	.87	.97	1.0	1.0	1.0
	MMCM	.10	**.37**	**.98**	**1.0**	**1.0**	**1.0**	**1.0**
.30	MCM	.18	.27	.91	1.0	1.0	1.0	1.0
	MMCM	**.21**	**.46**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
.35	MCM	.13	.42	1.0	1.0	1.0	1.0	1.0
	MMCM	**.20**	**.51**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
.40	MCM	**.34**	.73	1.0	1.0	1.0	1.0	1.0
	MMCM	.25	**.96**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**

(a)

$\Delta\downarrow$	Groups	4	6	8	10
.15	MCM	.39	.91	.99	1.0
	MMCM	**.48**	**1.0**	**1.0**	**1.0**
.20	MCM	.75	.99	1.0	1.0
	MMCM	**.86**	**1.0**	**1.0**	**1.0**
.25	MCM	.93	1.0	1.0	1.0
	MMCM	**.99**	**1.0**	**1.0**	**1.0**
.30	MCM	.98	1.0	1.0	1.0
	MMCM	**1.0**	**1.0**	**1.0**	**1.0**
.35	MCM	.99	1.0	1.0	1.0
	MMCM	**1.0**	**1.0**	**1.0**	**1.0**
.40	MCM	1.0	1.0	1.0	1.0
	MMCM	**1.0**	**1.0**	**1.0**	**1.0**

(b)

TABLE 7. Power of the MCM and the MMCM tests in the lognormal spherical scale family with (a) the number of classes $K = 6$ fixed, and (b) the dimension $d = 150$ fixed.

- *Spherical Lognormal Scale*: Here, we consider samples from the following K log-normal distributions: $\exp(N_d(\mathbf{0}, (1 + (s-1)\Delta)\mathrm{I}))$, for $1 \le s \le K$. Table 7(a) shows

$\Delta \downarrow$	Dimension	5	10	50	100	200	300	500
.15	MCM	**.08**	.09	**.14**	**.16**	**.23**	**.31**	**.32**
	MMCM	.05	**.10**	.11	.14	.18	.23	.27
.20	MCM	**.09**	.11	**.16**	.13	.24	.25	.34
	MMCM	.08	**.11**	.13	**.22**	**.25**	**.27**	**.39**
.25	MCM	**.07**	**.11**	**.17**	**.26**	.30	.36	.43
	MMCM	.02	.08	.12	.25	**.33**	**.45**	**.53**
.30	MCM	.06	.14	.15	.26	.38	.45	.52
	MMCM	**.10**	**.15**	**.23**	**.36**	**.54**	**.68**	**.71**
.35	MCM	.07	.15	.28	.32	.45	.57	.72
	MMCM	**.08**	**.16**	**.28**	**.43**	**.69**	**.77**	**.89**
.40	MCM	**.20**	.22	.37	.45	.72	.75	.84
	MMCM	.16	**.24**	**.48**	**.63**	**.95**	**.99**	**1.0**

(a)

$\Delta \downarrow$	Groups	4	6	8	10
.15	MCM	**.20**	.14	.12	.11
	MMCM	.14	**.15**	**.12**	**.11**
.20	MCM	**.28**	.18	.13	.14
	MMCM	.22	**.19**	**.18**	**.16**
.25	MCM	.31	.31	**.27**	**.21**
	MMCM	**.37**	**.32**	.19	.16
.30	MCM	.41	.35	.29	.23
	MMCM	**.42**	**.40**	**.31**	**.30**
.35	MCM	.47	.39	.38	.33
	MMCM	**.55**	**.51**	**.48**	**.45**
.40	MCM	.63	.52	.49	.47
	MMCM	**.79**	**.77**	**.68**	**.63**

(b)

TABLE 8. Power of the MCM and the MMCM tests in the lognormal equi-correlated scale family with (a) the number of classes $K = 6$ fixed, and (b) the dimension $d = 150$ fixed.

the fixed class scenario, with $K = 6$ and dimension d varying from 5 to 500, and Δ varying from 0.15 to 0.4. As before, in this case, the sample sizes were taken in equal increments of 50, starting from 50. Table 7(b) shows the fixed dimension scenario, where the $d = 150$ is fixed, and K varies along $4, 6, 8, 10$, and Δ varies from 0.15 to 0.4, as well. Here, the sample sizes are taken in equal increments from 50 to 200 when $K = 4$, from 50 to 300 when $K = 6$, from 50 to 260 when $K = 8$, and from 50 to 230 when $K = 10$.

- *Equi-correlated Lognormal Scale*: Here, we consider samples from the following K log-normal distributions: $\exp(N_d(0, (1-\rho_s)\mathrm{I} + \rho_s \mathbf{1}\mathbf{1}^\top))$, where $\rho_s := (s-1)\frac{\Delta}{K-1}$, for $1 \le s \le K$. Table 8(a) shows the fixed class scenario, with $K = 6$ and dimension d varying from 5 to 500, and Δ varying from 0.15 to 0.4. The sample sizes are taken to be $50, 100, 150, 200, 250$ and 300. Table 8(b) shows the fixed dimension scenario, where the $d = 150$ is fixed, and K varies along $4, 6, 8, 10$, and Δ varies from 0.15 to 0.4, as before. The sample sizes are taken in equal increments from 50 to 200 when $K = 4$, from 50 to 300 when $K = 6$, from 50 to 260 when $K = 8$, and from 50 to 230 when $K = 10$.

GRADUATE GROUP IN GENOMICS AND COMPUTATIONAL BIOLOGY, MEDICAL SCIENTIST TRAINING PROGRAM, PERELMAN SCHOOL OF MEDICINE, UNIVERSITY OF PENNSYLVANIA
E-mail address: Divyansh.Agarwal@pennmedicine.upenn.edu

DEPARTMENT OF STATISTICS, THE WHARTON SCHOOL, UNIVERSITY OF PENNSYLVANIA
E-mail address: {somabha,bhaswar,nzh}@wharton.upenn.edu