Mathematics

https://www.aimspress.com/journal/Math

AIMS Mathematics, 10(4): 7721–7766.
DOI: 10.3934/math.2025355
Received: 25 June 2024
Revised: 18 February 2025
Accepted: 19 February 2025
Published: 01 April 2025

Research article

Testing for correlation in Gaussian databases via local decision making

Ran Tamir[*]

Signal and Information Processing Laboratory, ETH Zurich, 8092 Zurich, Switzerland

* **Correspondence:** Email: tamir@isi.ee.ethz.ch; Tel: +41766034012.

Abstract: We propose a computationally efficient statistical test for detecting correlation between two Gaussian databases. The problem is formulated as a hypothesis test: under the null hypothesis, the databases are independent; under the alternative hypothesis, they are correlated but subject to an unknown row permutation. We derive bounds on both type I and type II error probabilities and demonstrate that the proposed test outperforms a recently introduced method across a broad range of parameter settings. The test statistic is based on a sum of dependent indicator random variables. To effectively bound the type I error probability, we introduce a novel graph-theoretic technique for bounding the k-th moments of such statistics.

Keywords: correlation detection; data privacy; de-anonymization; Gaussian database; hypothesis testing
Mathematics Subject Classification: 60F10, 62F03, 62F05, 62H20, 62H22, 68P15, 68P27

1. Introduction

With the thriving use of smart devices and the growing popularity of big data, various institutions have been collecting more and more personal data from users, which are then either published or sold for research or profit-making purposes. Although the published data are typically anonymized (i.e., the explicit attributes of the users, such as names and dates of birth are erased), there has been increasing concern about potential privacy leakage from anonymized data, approached from juridical [1] and corporate [2] points of view. These concerns are also expressed in the literature regarding successful practical de-anonymization attacks on real data; the reader is referred to [3–7] for a non-exhaustive list. We briefly elaborate on some of these studies. In [4], the task of identifying users from the statistics of their behavioral patterns, based on call data records, web browsing histories, or GPS trajectories, has been studied. It was demonstrated that a large fraction of users can be easily identified, given only histograms of their data; hence, these histograms can act as users' fingerprints. In [6], the authors presented a class of statistical de-anonymization attacks against high-dimensional micro-data, such as individual preferences, recommendations, transaction records, and so on. Their techniques are robust to perturbation in the data and tolerate some mistakes in the adversary's background knowledge. More studies regarding practical de-anonymization attacks on real data have been conducted with a specific focus on social networks; in this case, the reader is referred to [8–12] and to the references therein. For example, in [8], a novel de-anonymization attack that exploits group membership information that is available on social networking sites was introduced. The authors showed that information about the group memberships of a user (i.e., the groups of a social network to which a user belongs) is sufficient to uniquely identify this person or at least to significantly reduce the set of possible candidates. Although being extremely valuable, these lines of work does not provide a complete rigorous understanding of the conditions under which anonymized databases are inclined to privacy attacks; thus, more fundamental research is currently missing around these topics.

Two fundamental problems of statistical inference in database models are correlation detection and alignment recovery. Correlation detection between two databases is basically a hypothesis testing problem; under the null hypothesis, the databases are independent, and under the alternative hypothesis, a permutation exists, for which the databases are correlated. In this task, the main objective is to achieve the best trade-off between type I and type II error probabilities. In the problem of database alignment recovery, one assumes that the two databases are correlated and wants to estimate the underlying permutation. While these two problems are intimately related, this paper is devoted solely to correlation detection.

Correlation detection of databases with n sequences, each containing d independent Gaussian entries, has recently been studied in [13,14], while correlation detection of databases with independent entries and general distributions has lately been explored in [15]. In the present manuscript, we continue along this line of work and analyze a simple statistical test that solves the problem of correlation detection between two Gaussian databases relatively efficiently[I]. The tester first calculates a statistic which is based on local decisions for all n^2 pairs of sequences and then makes a final decision in favor of one of the hypotheses based on this statistic. Since the proposed statistical test is based on a sum of indicator random variables which are weakly dependent, the analysis of the two error probabilities, especially the type I probability of error, turns out to be highly non-trivial, and this is the main technical contribution of this work.

We compare the performances of the proposed statistical test with those of [13]. The main problem in [13] was to identify the smallest possible correlation coefficient between each consecutive entries of the databases, as a function of n and d, such that the sum of the type I and type II error probabilities can be driven to zero as n and d tend to infinity. While the detector that was proposed in [13] can also be used for non-asymptotic correlation values, we show that for a wide range of parameter choices, its performance is inferior to that of the new proposed statistical test. In addition to the comparison between the analytically derived theoretical guarantees, we also present Monte–Carlo simulation results for specific parameter values, which also demonstrate that the proposed statistical test outperforms the previous one.

[I] The proposed statistical test has a computational complexity of $O(n^2)$, while the generalized likelihood ratio (GLR) test has a computational complexity of $O(n^3)$. More on the GLR test can be found in Section 4.2.

1.1. Related work

Alignment recovery of correlated databases has been investigated from an information-theoretic point of view. The database alignment problem was originally introduced by Cullina et al. [16]. The discrete setting was studied in [16], which derived achievability and converse bounds in terms of a newly introduced measure of correlation, called the cycle mutual information. Exact recovery of the underlying permutation for correlated Gaussian databases was studied in [17], and a follow-up work extended the results to partial recovery [18]. A typicality-based framework for permutation estimation was investigated in [19]. Database matching under random column deletions and adversarial column deletions were researched, respectively, in [20] and [21]. Theoretical guarantees for the de-anonymization of random correlated databases without prior knowledge of the data distribution were proposed in [22]. The problem of database matching under noisy synchronization errors was recently studied in [23]. The Gaussian database alignment recovery problem is equivalent to a certain idealized tracking problem studied in [24, 25]. The recovery problem with two correlated random graphs has also been investigated in the past few years. A starting point of this problem proposed the correlated Erdős–Rényi graph model with dependent Bernoulli edge pairs [26]. A subsequent work studied the recovery problem in the Gaussian setting [27]. More recent papers investigated the corresponding detection problem for correlation between graphs [28, 29]. It has lately been proved [30, 31] that detecting whether Gaussian graphs are correlated is as difficult as recovering the node labels.

1.2. Organization of this paper

The continuation of this paper is organized as follows. In Section 2, we establish the notation conventions. In Section 3, we formalize the model and formulate our main objectives. In Section 4, we state some results from previous work and discuss the GLR test. In Section 5, we state the new proposed statistical test and then provide and discuss the main results of this work. In Section 6, we introduce the novel graph-theoretic concepts that have paved the way to handling the type I probability of error of the proposed test. In Section 7, we compare the performances of the proposed test and an existing one. Section 8 concludes the work and suggests some open avenues for future research. All the results of this work are proved in the appendices.

2. Notation conventions

Throughout the paper, random variables will be denoted by capital letters, and the specific values they take will be denoted by the corresponding lower case letters. Random vectors and their realizations will be denoted, respectively, by capital letters and the corresponding lowercase letters, both in bold font. For example, the random vector $\boldsymbol{X} = (X_1, X_2, \dots, X_d)$ (d-positive integer) may take a specific vector value $\boldsymbol{x} = (x_1, x_2, \dots, x_d)$ in $\mathbb{R}^d$. When used in the linear-algebraic context, these vectors should be thought of as column vectors, so when they appear with the superscript T, they will be transformed into row vectors by transposition. Thus, $\boldsymbol{x}^\mathsf{T}\boldsymbol{y}$ is the inner product of $\boldsymbol{x}$ and $\boldsymbol{y}$. The notation $||\boldsymbol{x}||$ will stand for the Euclidean norm of the vector $\boldsymbol{x}$. As is customary in probability theory, we write $\boldsymbol{X} = (X_1, \dots, X_d) \sim \mathcal{N}(\boldsymbol{0}_d, \boldsymbol{I}_d)$ (with $\boldsymbol{0}_d$ being a vector of d zeros and $\boldsymbol{I}_d$ being the $d \times d$ identity matrix) to denote that the probability density function of $\boldsymbol{X}$ is

$$P_{\boldsymbol{X}}(\boldsymbol{x}) = (2\pi)^{-d/2} \cdot \exp\left(-\frac{1}{2}\|\boldsymbol{x}\|^2\right), \quad \boldsymbol{x} \in \mathbb{R}^d. \tag{2.1}$$

For $\boldsymbol{X} = (X_1, \ldots, X_d)$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_d)$, we write

$$(\boldsymbol{X}, \boldsymbol{Y}) \sim \mathcal{N}^{\otimes d}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right) \tag{2.2}$$

to denote the fact that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are jointly Gaussian with independent and identically distributed (IID) pairs, where each pair (X_i, Y_i), $i \in \{1, \ldots, d\}$, is a Gaussian vector with a zero mean and the specified correlation ρ. Logarithms are taken to the natural base. The probability of an event $\mathcal{E}$ will be denoted by $\mathbb{P}\{\mathcal{E}\}$ and the indicator function by $\mathbb{1}\{\mathcal{E}\}$. The expectation operator will be denoted by $\mathbb{E}[\cdot]$. The set of all permutations of $\{1, 2, \ldots, n\}$ is denoted by $\mathcal{S}_n$

3. Settings and problem formulation

Let $\boldsymbol{X}^n$ be a database comprising the n feature vectors $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ and let $\boldsymbol{Y}^n$ be a second database comprising the n feature vectors $\{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n\}$. Each feature vector contains d entries. We consider the following binary hypothesis testing problem. Under the *null hypothesis* H_0, the Gaussian databases $\boldsymbol{X}^n$ and $\boldsymbol{Y}^n$ are generated independently with $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{\text{IID}}{\sim} \mathcal{N}(\boldsymbol{0}_d, \boldsymbol{I}_d)$ and $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n \overset{\text{IID}}{\sim} \mathcal{N}(\boldsymbol{0}_d, \boldsymbol{I}_d)$ (Figure 1). Let us use $\mathbb{P}_0$ to denote the probability distribution that governs $(\boldsymbol{X}^n, \boldsymbol{Y}^n)$ under H_0. Under the *alternative hypothesis* H_1, the databases $\boldsymbol{X}^n$ and $\boldsymbol{Y}^n$ are correlated with some unknown permutation $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_n) \in \mathcal{S}_n$ and some known correlation coefficient $\rho \in (0, 1)$ (Figure 2). If $\rho \in (-1, 0)$, then one considers the negation of $\boldsymbol{Y}^n$. Let us use $\mathbb{P}_{1|\boldsymbol{\sigma}}$ to denote the probability distribution that governs $(\boldsymbol{X}^n, \boldsymbol{Y}^n)$ under H_1, for some permutation $\boldsymbol{\sigma} \in \mathcal{S}_n$. In summary,

$$\begin{aligned} \mathsf{H}_0: &\quad (\boldsymbol{X}_1, \boldsymbol{Y}_1), \ldots, (\boldsymbol{X}_n, \boldsymbol{Y}_n) \overset{\text{IID}}{\sim} \mathcal{N}^{\otimes d}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) \\ \mathsf{H}_1: &\quad (\boldsymbol{X}_1, \boldsymbol{Y}_{\sigma_1}), \ldots, (\boldsymbol{X}_n, \boldsymbol{Y}_{\sigma_n}) \overset{\text{IID}}{\sim} \mathcal{N}^{\otimes d}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right), \end{aligned} \tag{3.1}$$

for some permutation $\boldsymbol{\sigma} \in \mathcal{S}_n$.

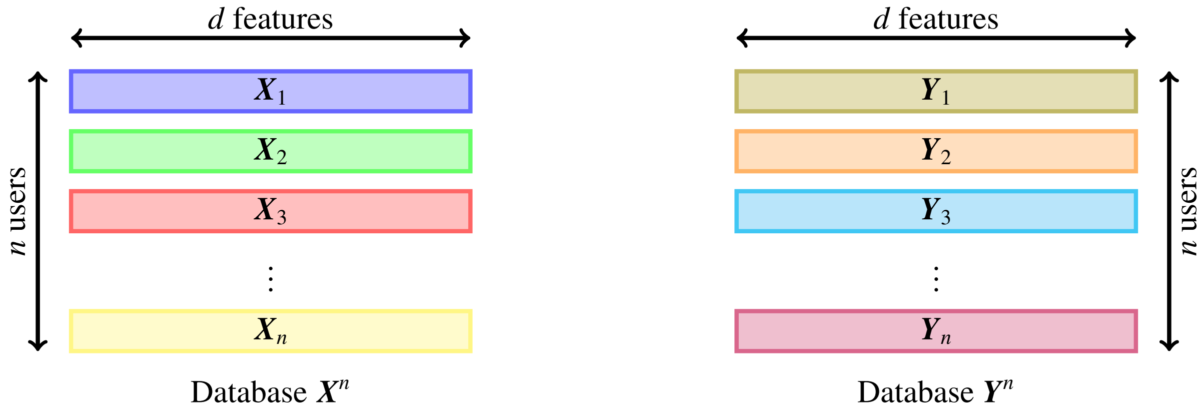


Figure 1. Under the null hypothesis H_0 in the detection problem, the feature vectors $\{\boldsymbol{X}_i\}_{i=1}^n$ and $\{\boldsymbol{Y}_i\}_{i=1}^n$ are generated independently according to $\mathcal{N}(\boldsymbol{0}_d, \boldsymbol{I}_d)$.

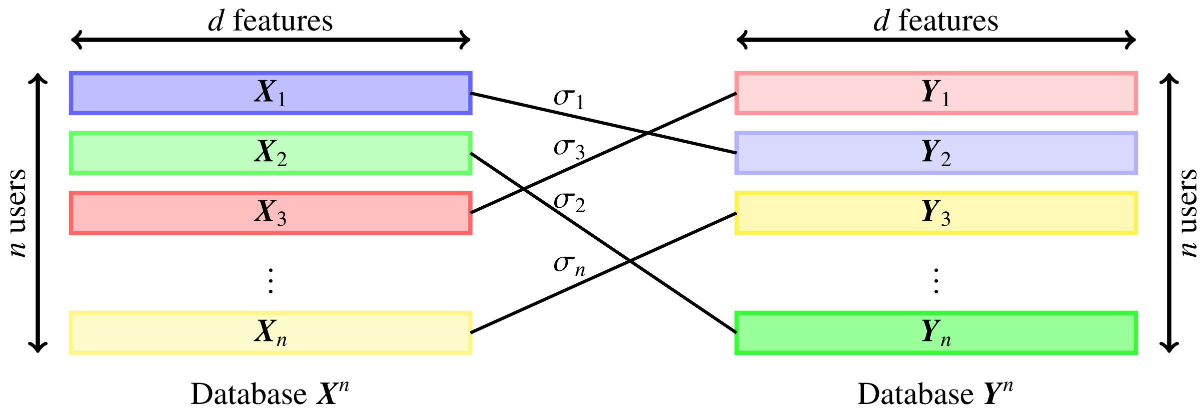


Figure 2. Under the alternative hypothesis H_1 in the detection problem, the collection of pairs $(\boldsymbol{X}_1, \boldsymbol{Y}_{\sigma_1}), \ldots, (\boldsymbol{X}_n, \boldsymbol{Y}_{\sigma_n})$ is independent. Each pair of feature vectors $(\boldsymbol{X}_i, \boldsymbol{Y}_{\sigma_i})$ is jointly Gaussian with a known correlation $\rho \in (0, 1)$, under an unknown permutation $\boldsymbol{\sigma} \in \mathcal{S}_n$.

We would like to consider a problem of correlation detection. Given the databases $\boldsymbol{X}^n$ and $\boldsymbol{Y}^n$ (and not the permutation $\boldsymbol{\sigma}$), a *statistical test* $\phi : \mathbb{R}^{d\times n} \times \mathbb{R}^{d\times n} \to \{0, 1\}$ decides whether the null hypothesis or the alternative hypothesis occurred.

For a given n and d, the type I probability of error of a test ϕ is

$$P_{\text{FA}}(\phi) \triangleq \mathbb{P}_0 \{\phi(\boldsymbol{X}^n, \boldsymbol{Y}^n) = 1\}, \tag{3.2}$$

and the type II probability of error is

$$P_{\text{MD}}(\phi) \triangleq \max_{\boldsymbol{\sigma}\in\mathcal{S}_n} \mathbb{P}_{1|\boldsymbol{\sigma}} \{\phi(\boldsymbol{X}^n, \boldsymbol{Y}^n) = 0\} \tag{3.3}$$

where in (3.2), FA stands for false alarm, and in (3.3), MD stands for missed detection.

For the specific statistical test that is proposed in Section 5, which solves the testing problem defined above, our main objective is to find theoretical guarantees on the two error probabilities as functions of the parameters ρ, n, and d.

4. The sum and GLR tests

4.1. The sum test

The testing problem between databases with IID Gaussian entries has already been considered in [13, 14]. For this specific testing problem, the following statistical test was proposed in [13]. The statistic is defined by

$$T \triangleq \sum_{i=1}^{n} \sum_{j=1}^{n} \boldsymbol{X}_i^{\mathsf{T}} \boldsymbol{Y}_j, \tag{4.1}$$

and the sum test is defined by comparing T with a threshold

$$\phi_{\text{Sum}}(\boldsymbol{X}^n, \boldsymbol{Y}^n) = \begin{cases} 0 & T < t \\ 1 & T \geq t. \end{cases} \tag{4.2}$$

The following result was proved in [13, Section IV].

Proposition 1. (Sum test) *Let $t = \sqrt{\gamma}\frac{dn}{2}$ with $\gamma \in (0, 4\rho^2)$. The error probabilities of the sum test ϕ_{Sum} for the binary hypothesis testing problem* (3.1) *are upper-bounded by*

$$P_{FA}(\phi_{Sum}) \leq \exp\left(-\frac{d}{2}G_{FA}(\gamma)\right) \tag{4.3}$$

$$P_{MD}(\phi_{Sum}) \leq \exp\left(-\frac{d}{2}G_{MD}(\gamma)\right), \tag{4.4}$$

where,

$$G_{FA}(\gamma) \triangleq \sqrt{1+\gamma} - 1 - \ln\left(\frac{1+\sqrt{1+\gamma}}{2}\right), \tag{4.5}$$

$$G_{MD}(\gamma) \triangleq \frac{1}{1-\rho^2}\left(\sqrt{(1-\rho^2)^2+\gamma} - \sqrt{\rho^2\gamma}\right) - 1 - \ln\left(\frac{1-\rho^2+\sqrt{(1-\rho^2)^2+\gamma}}{2}\right). \tag{4.6}$$

4.2. The structure and complexity of the GLR test

The GLR test is known to be an asymptotically uniformly most powerful (UMP)[II] test among all invariant tests [32, Section 6.4.2]. Although there are no general results regarding non-asymptotic properties of the GLR test, it is known to be optimal in special cases [33]. Since the sum test from Section 4.1, as well as the count test (which will be presented in the following section) can be proved to be invariant tests, it makes sense to also briefly present the GLR test and discuss its computational complexity. We first derive a simplified statistic for the GLR test. Since only the alternative hypothesis is composite, the GLR test is given by

$$\frac{\max_{\sigma\in\mathcal{S}_n} p_{1|\sigma}(\boldsymbol{x}^n, \boldsymbol{y}^n)}{p_0(\boldsymbol{x}^n, \boldsymbol{y}^n)} \gtrless \lambda, \tag{4.7}$$

for some $\lambda \in \mathbb{R}$, or, more explicitly, by

$$\frac{\max_{\sigma\in\mathcal{S}_n} \prod_{k=1}^n c_1 \exp\left\{-\frac{1}{2(1-\rho^2)}(\|\boldsymbol{x}_k\|^2 - 2\rho\boldsymbol{x}_k^\mathsf{T}\boldsymbol{y}_{\sigma_k} + \|\boldsymbol{y}_{\sigma_k}\|^2)\right\}}{\prod_{k=1}^n c_0 \exp\left\{-\frac{1}{2}(\|\boldsymbol{x}_k\|^2 + \|\boldsymbol{y}_k\|^2)\right\}} \gtrless \lambda. \tag{4.8}$$

By simple manipulations,

$$\max_{\sigma\in\mathcal{S}_n}\left\{-\frac{1}{1-\rho^2}\sum_{k=1}^n(\|\boldsymbol{x}_k\|^2 - 2\rho\boldsymbol{x}_k^\mathsf{T}\boldsymbol{y}_{\sigma_k} + \|\boldsymbol{y}_{\sigma_k}\|^2)\right\} + \sum_{k=1}^n(\|\boldsymbol{x}_k\|^2 + \|\boldsymbol{y}_k\|^2) \gtrless \lambda', \tag{4.9}$$

which is equivalent to

$$\max_{\sigma\in\mathcal{S}_n}\sum_{k=1}^n 2\rho\boldsymbol{x}_k^\mathsf{T}\boldsymbol{y}_{\sigma_k} - \sum_{k=1}^n(\rho^2\|\boldsymbol{x}_k\|^2 + \rho^2\|\boldsymbol{y}_k\|^2) \gtrless \lambda''. \tag{4.10}$$

[II] A UMP test minimizes the type II probability of error among all tests with a constrained type I probability of error. For example, the Neyman-Pearson test is UMP when testing between two simple hypotheses.

Thus, the GLR test requires us to find the solution of an assignment problem of size n with the scores given by the empirical correlations. Relying on the Kuhn–Munkres algorithm [34–36] (also known as the Hungarian algorithm) and its modifications [37], this assignment problem can be solved with a running time of $O(n^3)$. It is important to note that this computational complexity holds in general, but when the optimal assignment is relatively easy to find (e.g., for a diagonal score matrix), the running time of the Hungarian algorithm is reduced to $O(n^2)$. In our settings, this holds when ρ is close to 1.

While the GLR test has high chances of performing well in the studied testing problem, it has two main disadvantages. First, as explained above, its computational complexity grows cubically in n, which is discouraging from a practical point of view; second, the GLR test does not lend itself to a tractable analytical analysis (specifically, the type I error event). These facts motivate us to propose an alternative statistical test that will be computationally efficient (relative to the GLR test) and analytically tractable, on the one hand, and perform at least as well as the existing sum test, on the other hand. The count test, which is the main theme of the next section, fulfills all these requirements.

5. The count test

5.1. Motivation

In order to motivate the new proposed statistical test, let us shortly consider the sum test in Section 4.1. Note that its statistic can also be written as follows:

$$T = \left(\sum_{i=1}^{n} \boldsymbol{X}_i\right)^{\mathsf{T}} \left(\sum_{j=1}^{n} \boldsymbol{Y}_j\right), \tag{5.1}$$

which means that one has to first sum up all the n feature vectors in each database to get the vectors $\hat{\boldsymbol{X}}$ and $\hat{\boldsymbol{Y}}$, and then test whether these vectors are correlated or not. The meaning of transforming each database into a single vector is obvious; one wastes the original structure of the problem and, as a consequence, valuable information is lost. One has to note that, nonetheless there is a significant advantage to such a procedure: The computational complexity of the sum test is linear in n. As opposed to the sum test, we would like to take advantage of the fact that each database comprises n sequences. We note the following simple observation: The inner product between two independent vectors $\boldsymbol{X}, \boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{0}_d, \boldsymbol{I}_d)$ is typically close to zero, while the inner product between the vectors

$$(\boldsymbol{X}, \boldsymbol{Y}) \sim \mathcal{N}^{\otimes d}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right) \tag{5.2}$$

concentrates around ρd. This means that one can first test in a local manner for each pair of feature vectors and in a second step, count how many of them seem to be correlated. Ideally, under the null hypothesis, the number of detected correlated pairs is close to zero, while under the alternative hypothesis, this number should be close to n times the probability of local detection, which, in turn, should be close to one. Thus, on the basis of the count of local detections, one should be able to distinguish very well between the two hypotheses. It should be noted that the computational complexity of this testing procedure is $O(n^2)$, which is higher than that of the sum test, but still much lower than the complexity of the GLR test ($O(n^3)$).

5.2. The proposed statistical test

According to the explanation above, an appropriate statistic is given by

$$M(\xi) \triangleq \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{1}\left\{\boldsymbol{X}_i^{\mathsf{T}} \boldsymbol{Y}_j \geq \xi\right\}, \tag{5.3}$$

where ξ is some predefined threshold. While a test based on $M(\xi)$ can be analyzed, it turns out that by first normalizing the feature vectors before calculating the inner products, one can propose a test which lends itself to a much tighter analysis. We will elaborate more on this point in Section 6. In the meanwhile, we continue as follows. Denote the normalized vectors

$$\tilde{\boldsymbol{X}}_i = \frac{\boldsymbol{X}_i}{\|\boldsymbol{X}_i\|}, \quad i \in \{1, 2, \ldots, n\}, \tag{5.4}$$

and

$$\tilde{\boldsymbol{Y}}_j = \frac{\boldsymbol{Y}_j}{\|\boldsymbol{Y}_j\|}, \quad j \in \{1, 2, \ldots, n\}, \tag{5.5}$$

and define the statistic

$$N(\theta) \triangleq \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{1}\left\{\tilde{\boldsymbol{X}}_i^{\mathsf{T}} \tilde{\boldsymbol{Y}}_j \geq \theta\right\}, \tag{5.6}$$

where $\theta \in (0, 1)$ is the *local* threshold.

In order to present our statistical test, we need some definitions. For two feature vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$, correlated according to a given $\rho \in (0, 1)$, the *local detection probability* is defined by

$$P \equiv P(d, \rho, \theta) \triangleq \mathbb{P}\left\{\tilde{\boldsymbol{X}}^{\mathsf{T}} \tilde{\boldsymbol{Y}} \geq \theta\right\}, \tag{5.7}$$

while for two independent feature vectors (i.e., $\rho = 0$), the *local false-alarm probability* is defined by

$$Q \equiv Q(d, \theta) \triangleq P(d, 0, \theta). \tag{5.8}$$

Under H_1, note that the expectation of $N(\theta)$ is given by

$$\begin{aligned}
\Delta_{n,d} \triangleq \mathbb{E}_{1|\sigma}[N(\theta)] &= \mathbb{E}_{1|\sigma}\left[\sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{1}\left\{\tilde{\boldsymbol{X}}_i^{\mathsf{T}} \tilde{\boldsymbol{Y}}_j \geq \theta\right\}\right] && (5.9)\\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{P}_{1|\sigma}\left\{\tilde{\boldsymbol{X}}_i^{\mathsf{T}} \tilde{\boldsymbol{Y}}_j \geq \theta\right\} && (5.10)\\
&= nP + n(n-1)Q. && (5.11)
\end{aligned}$$

Thus, the proposed test compares $N(\theta)$ with a threshold as follows:

$$\phi_{\text{Count}}(\boldsymbol{X}^n, \boldsymbol{Y}^n) = \begin{cases} 0 & N(\theta) < \beta\Delta_{n,d} \\ 1 & N(\theta) \geq \beta\Delta_{n,d}, \end{cases} \tag{5.12}$$

where $\beta \in (0, 1)$ is the *global* threshold.

Our proposed statistical test may be presented in a simple graphical way. One has to draw a table with n rows and n columns. In this table, Row i stands for $\boldsymbol{X}_i$ and Column j for $\boldsymbol{Y}_j$. The cell (i, j) is filled with a dot only if $\tilde{\boldsymbol{X}}_i^{\mathsf{T}} \tilde{\boldsymbol{Y}}_j \geq \theta$. The dots are then counted and one rejects the null if the total number of dots exceeds a threshold of $\beta[nP + n(n-1)Q]$. Typical tables for $n = 10$ under H_0 and under H_1 with the identity permutation are given in Figure 3.

$\boldsymbol{X}_{10}$					●					
$\boldsymbol{X}_9$										
$\boldsymbol{X}_8$										
$\boldsymbol{X}_7$										
$\boldsymbol{X}_6$										
$\boldsymbol{X}_5$										
$\boldsymbol{X}_4$		●								
$\boldsymbol{X}_3$										
$\boldsymbol{X}_2$									●	
$\boldsymbol{X}_1$										
	$\boldsymbol{Y}_1$	$\boldsymbol{Y}_2$	$\boldsymbol{Y}_3$	$\boldsymbol{Y}_4$	$\boldsymbol{Y}_5$	$\boldsymbol{Y}_6$	$\boldsymbol{Y}_7$	$\boldsymbol{Y}_8$	$\boldsymbol{Y}_9$	$\boldsymbol{Y}_{10}$

$\boldsymbol{X}_{10}$										●
$\boldsymbol{X}_9$									●	
$\boldsymbol{X}_8$				●				●		
$\boldsymbol{X}_7$							●			
$\boldsymbol{X}_6$						●				
$\boldsymbol{X}_5$					●					
$\boldsymbol{X}_4$				●						
$\boldsymbol{X}_3$			●							
$\boldsymbol{X}_2$										
$\boldsymbol{X}_1$	●						●			
	$\boldsymbol{Y}_1$	$\boldsymbol{Y}_2$	$\boldsymbol{Y}_3$	$\boldsymbol{Y}_4$	$\boldsymbol{Y}_5$	$\boldsymbol{Y}_6$	$\boldsymbol{Y}_7$	$\boldsymbol{Y}_8$	$\boldsymbol{Y}_9$	$\boldsymbol{Y}_{10}$

Figure 3. Examples of typical tables under H_0 (left) and H_1 with the identity permutation.

5.3. Theoretical guarantees

While the two error probabilities of the sum test (4.1)-(4.2) were upper-bounded using standard large deviation techniques, i.e., Chernoff's bound, this cannot be similarly done with the new proposed test, since calculating the moment-generating function of the double summation in (5.6) seems to be hopeless. Thus, alternative tools from large deviation theory have to be invoked. We start with the type II probability of error, which is given explicitly by

$$P_{\text{MD}}(\phi_{\text{Count}}) = \mathbb{P}_{1|\sigma} \left\{ \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{1} \left\{ \tilde{\boldsymbol{X}}_i^{\mathsf{T}} \tilde{\boldsymbol{Y}}_j \geq \theta \right\} \leq \beta[nP + n(n-1)Q] \right\}. \tag{5.13}$$

The main difficulty in analyzing this lower tail probability is the statistical dependencies between the indicator random variables. Nevertheless, an existing large deviation result by Janson [38, Theorem 10] provides the appropriate tool to handle the situation at hand. Other recently treated settings where tools from [38] proved useful can be found in [39, Appendixes B and I] and [40, Appendix A.10].

Regarding the type I probability of error, the situation is somewhat more complicated, since no general large deviation bounds exist to assess upper tail probabilities of the form

$$P_{\text{FA}}(\phi_{\text{Count}}) = \mathbb{P}_0 \left\{ \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{1} \left\{ \tilde{\boldsymbol{X}}_i^{\mathsf{T}} \tilde{\boldsymbol{Y}}_j \geq \theta \right\} \geq \beta[nP + n(n-1)Q] \right\}. \tag{5.14}$$

Hence, the best we could do is to upper-bound this probability using a generalization of Chebyshev's inequality as follows:

$$P_{\text{FA}}(\phi_{\text{Count}}) \leq \frac{\mathbb{E}_0\left[\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{1}\left\{\tilde{\boldsymbol{X}}_i^{\mathsf{T}}\tilde{\boldsymbol{Y}}_j \geq \theta\right\}\right)^{\alpha}\right]}{(\beta[nP+n(n-1)Q])^{\alpha}}, \tag{5.15}$$

where $\alpha > 0$ may be optimized to yield the tightest bound. In order to derive (or at least, to upper-bound) the α-th moment in (5.15), we consider two different cases. On the one hand, when $\alpha \in (0,1]$, we are able to upper-bound the α-th moment using similar techniques as in the proof of the lower bound on the error exponent of the typical random code [41, p. 6229, right column]. On the other hand, for a general $\alpha > 1$, we are not able to handle the α-th moment; hence, instead, we have considered only the k-th integer moments of orders $k \geq 2$. For the special case $k = 2$, the calculation is relatively straightforward, but for a general $k \geq 3$, the situation is somewhat more complicated. Thus, new techniques that rely on graph-counting considerations are developed in Appendix C to prove Theorem 2 (which is stated in Section 6).

In the following result, which is proved in Appendix B, we present upper bounds on the type I and type II error probabilities for a general set of parameters. Let us denote the Stirling numbers of the second kind as follows:

$$\mathsf{S}(k,\ell) = \frac{1}{\ell!}\sum_{i=0}^{\ell}(-1)^i\binom{\ell}{i}(\ell-i)^k, \tag{5.16}$$

and define the numbers

$$\mathsf{B}(k) = \sum_{\ell=1}^{k}\mathsf{S}(k,\ell)\cdot\frac{k^{2\ell}}{\ell!}. \tag{5.17}$$

Theorem 1. (Type I and type II probability bounds) *Let $n, d \in \mathbb{N}$ and $\rho \in (0,1)$ be given. Let $P = P(d,\rho,\theta)$ and $Q = Q(d,\theta)$ as defined above. Then, for any $\beta \in (0,1)$ and $\theta \in (0,1)$*

$$\begin{aligned} P_{FA}(\phi_{Count}) \leq \min\Bigg\{ &\min_{\alpha\in[0,1]} \frac{(n^2Q)^{\alpha}\cdot\mathbb{1}\{n^2Q>1\}+n^2Q\cdot\mathbb{1}\{n^2Q\leq 1\}}{(\beta[nP+n(n-1)Q])^{\alpha}}, \frac{n^2(n^2-1)Q^2+n^2Q}{(\beta[nP+n(n-1)Q])^2}, \\ &\inf_{k\geq 3}\frac{k(k+1)\mathsf{B}(k)\left[(n^2Q)^k\cdot\mathbb{1}\{n^2Q>1\}+n^2Q\cdot\mathbb{1}\{n^2Q\leq 1\}\right]}{(\beta[nP+n(n-1)Q])^k}\Bigg\}, \end{aligned} \tag{5.18}$$

and,

$$P_{MD}(\phi_{Count}) \leq \exp\left\{-\min\left((1-\beta)^2\frac{nP+n(n-1)Q}{16Qn+2}, (1-\beta)\frac{n}{12}\right)\right\}. \tag{5.19}$$

Regarding the asymptotic behavior of the bounds. One benefit of Proposition 1 (Section 4.1) with respect to Theorem 1 is the clear conclusion that when d tends to infinity, the error probabilities vanish exponentially. In order to understand the asymptotic behavior of the upper bounds in Theorem 1, we first have to handle the quantities $Q(d,\theta)$ and $P(d,\rho,\theta)$. $Q(d,\theta)$ is the probability of an independent pair of Gaussian vectors having an angle larger than θ, and thus, thanks to symmetry, $Q(d,\theta)$ equals the probability that a uniformly distributed vector on a d-dimensional hyper-sphere of a unit norm will fall within a d-dimensional hyper-spherical cap with a half-angle $\varphi = \arccos(\theta)$. This probability is given by the expression [44]

$$Q(d,\theta) = \frac{1}{2}\frac{B\left(\sin^2(\varphi); \frac{d-1}{2}, \frac{1}{2}\right)}{B\left(\frac{d-1}{2}, \frac{1}{2}\right)} = \frac{1}{2}\frac{B\left(1-\theta^2; \frac{d-1}{2}, \frac{1}{2}\right)}{B\left(\frac{d-1}{2}, \frac{1}{2}\right)}, \tag{5.20}$$

where the complete beta function is defined by

$$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}\mathrm{d}t, \quad a,b > 0, \tag{5.21}$$

and the incomplete beta function by

$$B(x;a,b) = \int_0^x t^{a-1}(1-t)^{b-1}\mathrm{d}t, \quad x \in [0,1],\ a,b > 0. \tag{5.22}$$

The following result, which is proved in Appendix D, shows that $Q(d,\theta)$ converges exponentially fast to zero.

Lemma 1. *Let $d \in \mathbb{N}$ and $\theta \in (0,1)$. Let*

$$(\boldsymbol{X},\boldsymbol{Y}) \sim \mathcal{N}^{\otimes d}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right), \tag{5.23}$$

and

$$\tilde{\boldsymbol{X}} = \frac{\boldsymbol{X}}{\|\boldsymbol{X}\|}, \quad \tilde{\boldsymbol{Y}} = \frac{\boldsymbol{Y}}{\|\boldsymbol{Y}\|}. \tag{5.24}$$

Then,

$$\mathbb{P}\left\{\tilde{\boldsymbol{X}}^{\mathsf{T}}\tilde{\boldsymbol{Y}} \geq \theta\right\} \leq \frac{1}{\theta\sqrt{d}} \cdot \left(1-\theta^2\right)^{\frac{d-1}{2}}. \tag{5.25}$$

In addition, one can prove that $P(d,\rho,\theta)$ converges to one as $d \to \infty$ for any $\theta \in (0,\rho)$. Since the proof of this result is quite long and relatively technical, we refrain from presenting it here. Going back to the bounds in Theorem 1, we conclude the following. Regarding the bound in (5.18), it can be further upper-bounded as

$$P_{\mathrm{FA}}(\phi_{\mathrm{Count}}) \leq \frac{n^2(n^2-1)Q(d,\theta)^2 + n^2Q(d,\theta)}{(\beta[nP(d,\rho,\theta) + n(n-1)Q(d,\theta)])^2}, \tag{5.26}$$

which vanishes exponentially fast for any fixed n as $d \to \infty$. Regarding the bound in (5.19), the situation is different, because when $d \to \infty$, we find that

$$P_{\mathrm{MD}}(\phi_{\mathrm{Count}}) \leq \exp\left\{-n \cdot \min\left\{(1-\beta)^2/2, (1-\beta)/12\right\}\right\}, \tag{5.27}$$

which is exponentially small but fixed in n.

Several remarks are now in order.

Remark 1. The threshold $\beta \in (0, 1)$ trades-off between the two error probabilities. When β is relatively low, then relatively few "dots" are required to reject H_0, hence the type I error probability is high and the type II error probability is low. When β grows towards 1, an increasing number of dots are needed to reject H_0; the type I error probability gradually decreases while the type II error probability increases.

Remark 2. The relatively interesting fact (at least to the writer of these lines) is the apparent dichotomy between two regions in the parameter space $\{(n, d)\}$. On the one hand, if $n^2Q(d, \theta)$, which is the expected number of dots under H_0, is smaller than 1, then the α-th moment in (5.15) is proportional to $n^2Q(d, \theta)$ for any $\alpha \in (0, 1] \cup \{2, 3, \ldots\}$. In this case, the type I error probability is relatively small, since the typical number of "correlated" pairs of vectors is zero. On the other hand, if (n, d) are such that $n^2Q(d, \theta) \geq 1$, then the α-th moment in (5.15) is proportional to $[n^2Q(d, \theta)]^\alpha$. In this case, any increment in $n^2Q(d, \theta)$ (due to an increasing n or a decreasing d) causes a relatively sharp increment in the type I error probability. Such phenomena, where moments of enumerators (i.e., sums of IID or weakly dependent indicator random variables) undergo phase transitions have already been encountered multiple times, e.g., in information theory [39, Appendix B, Lemma 3], [42, p. 168], [43, Appendix A, Lemma 1.1].

Remark 3. In this manuscript, we only consider a noiseless setup. Regarding noisy setups where the data are affected by additive white Gaussian noise[III], consider the following. Let

$$(\boldsymbol{X}, \boldsymbol{Y}) \sim \mathcal{N}^{\otimes d}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}\right), \tag{5.28}$$

and

$$\hat{\boldsymbol{X}} = \frac{\boldsymbol{X} + \boldsymbol{U}}{\sqrt{1 + \sigma^2}} \tag{5.29}$$

$$\hat{\boldsymbol{Y}} = \frac{\boldsymbol{Y} + \boldsymbol{V}}{\sqrt{1 + \sigma^2}}, \tag{5.30}$$

where $\boldsymbol{U}, \boldsymbol{V} \overset{\text{IID}}{\sim} \mathcal{N}(\boldsymbol{0}_d, \sigma^2\boldsymbol{I}_d)$ are independent of $(\boldsymbol{X}, \boldsymbol{Y})$. We can immediately see that $(\hat{\boldsymbol{X}}, \hat{\boldsymbol{Y}})$ are jointly Gaussian with a correlation coefficient of $\frac{\rho}{1+\sigma^2}$, and thus, the count test and its theoretical guarantees can be relied on also for this extended case.

Remark 4. Both the sum test and the count test assume perfect knowledge of ρ in order to choose desired trade-offs between the type I and type II error probabilities. In practice, the value of ρ may be unknown a-priori or only known to be in an interval of possible values. In the latter case, where $\rho \in [\rho_0, \rho_1]$, and this range is relatively narrow, the sum/count tests can still be applied with any ρ in this range, with some degraded performance due to the expected mismatch between the true ρ and the chosen one. Choosing ρ closer to ρ_0 has a preference towards lower type II error probabilities, and vice versa. The general case where ρ is unknown and may take any value in $(0, 1)$ seems to be more complicated and will not be discussed here in detail; we just mention that a possible solution may be in the spirit of the GLR test (as in (4.7)), where the likelihood under H_1 is now also maximized over $\rho \in (0, 1)$, in addition to the maximization over the permutations of the rows.

[III] Obfuscation, which refers to the deliberate addition of noise to the database entries, has been suggested as an additional measure to protect privacy [3].

5.4. On the type I probability bound

Let us begin by presenting an explicit expression for $P(d, \rho, \theta)$, which is the probability that the angle between a pair of correlated Gaussian vectors crosses a threshold. We first make a few definitions. For $\rho \in (0, 1)$ and $\theta \in (0, 1)$, define the constants

$$\alpha(\rho) = \frac{1-\rho^2}{\rho^2}, \tag{5.31}$$

$$\beta(\rho, \theta) = \frac{1-\rho^2}{\rho^2(1-\theta^2)}, \tag{5.32}$$

and the functions

$$F_1(u) = -\frac{\rho(1-\theta^2)}{\sqrt{1-\rho^2}}\sqrt{u} - \theta\sqrt{1 - \frac{\rho^2(1-\theta^2)}{1-\rho^2}u}, \tag{5.33}$$

$$F_2(u) = -\frac{\rho(1-\theta^2)}{\sqrt{1-\rho^2}}\sqrt{u} + \theta\sqrt{1 - \frac{\rho^2(1-\theta^2)}{1-\rho^2}u}. \tag{5.34}$$

Moreover, for any $d \in \mathbb{N}$, we define the two probability density functions

$$f(u) = \frac{1}{B\left(\frac{d}{2}, \frac{d}{2}\right)} u^{d/2-1} \left(1+u\right)^{-d}, \ u \geq 0, \tag{5.35}$$

and

$$g(s) = \frac{(1-s^2)^{\frac{d-3}{2}}}{B\left(\frac{d-1}{2}, \frac{1}{2}\right)}, \ s \in [-1, 1]. \tag{5.36}$$

The following lemma is proved in Appendix E.

Lemma 2. *Let $d \in \mathbb{N}$, $\rho \in (0, 1)$, and $\theta \in (0, 1)$. Let*

$$(\boldsymbol{X}, \boldsymbol{Y}) \sim \mathcal{N}^{\otimes d}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}\right), \tag{5.37}$$

and

$$\tilde{\boldsymbol{X}} = \frac{\boldsymbol{X}}{\|\boldsymbol{X}\|}, \quad \tilde{\boldsymbol{Y}} = \frac{\boldsymbol{Y}}{\|\boldsymbol{Y}\|}. \tag{5.38}$$

Then,

$$\begin{aligned} P(d, \rho, \theta) = \mathbb{P}\left\{\tilde{\boldsymbol{X}}^{\mathsf{T}}\tilde{\boldsymbol{Y}} \geq \theta\right\} = &\int_0^{\alpha(\rho)} \left[\int_{F_2(u)}^1 g(s)\mathrm{d}s\right] f(u)\mathrm{d}u \\ &+ \int_{\alpha(\rho)}^{\beta(\rho,\theta)} \left[\int_{-1}^{F_1(u)} g(s)\mathrm{d}s + \int_{F_2(u)}^1 g(s)\mathrm{d}s\right] f(u)\mathrm{d}u \\ &+ \int_{\beta(\rho,\theta)}^{\infty} f(u)\mathrm{d}u. \end{aligned} \tag{5.39}$$

Regarding the bound on the type I error probability, which appears in (5.18), and is given as a minimization between three competing terms, it is definitely not obvious that all of these terms are required to achieve the best performance. In order to show that all of them are indeed required, let us compare

$$E_{\text{frac}}(\beta) = -\log\left(\min_{\alpha\in[0,1]} \frac{(n^2Q)^\alpha \cdot \mathbb{1}\{n^2Q > 1\} + n^2Q \cdot \mathbb{1}\{n^2Q \leq 1\}}{(\beta[nP + n(n-1)Q])^\alpha}\right), \tag{5.40}$$

$$E_{\text{sec}}(\beta) = -\log\left(\frac{n^2(n^2-1)Q^2 + n^2Q}{(\beta[nP + n(n-1)Q])^2}\right), \tag{5.41}$$

and

$$E_{\text{int}}(\beta) = -\log\left(\inf_{k\geq 3} \frac{k(k+1)\mathsf{B}(k)\left[(n^2Q)^k \cdot \mathbb{1}\{n^2Q > 1\} + n^2Q \cdot \mathbb{1}\{n^2Q \leq 1\}\right]}{(\beta[nP + n(n-1)Q])^k}\right). \tag{5.42}$$

In Figure 4 below, we present plots of $E_{\text{frac}}(\beta)$ (solid line) and $E_{\text{sec}}(\beta)$ (dashed line) for the parameters $n = 200$, $d = 50$, $\rho = 0.4$, and $\theta = 0.6$. The quantity $Q(d, \theta)$ is calculated numerically using (5.20), and $P(d, \rho, \theta)$ is calculated using (5.39) in Lemma 2. As can be seen in Figure 4, at relatively low βs, $E_{\text{frac}}(\beta)$ is a constant and it strictly improves upon $E_{\text{sec}}(\beta)$. In this range, the optimizer is $\alpha^* = 0$. At higher β values, the optimizer is $\alpha^* = 1$, and $E_{\text{frac}}(\beta)$ is actually related to Markov's inequality. In the range of relatively high β values, $E_{\text{sec}}(\beta)$ outperforms $E_{\text{frac}}(\beta)$, which is not very surprising, since Chebyshev's inequality is usually tighter than Markov's inequality.

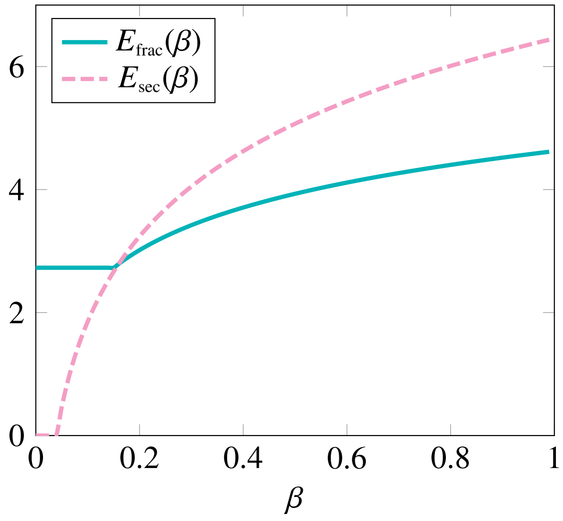
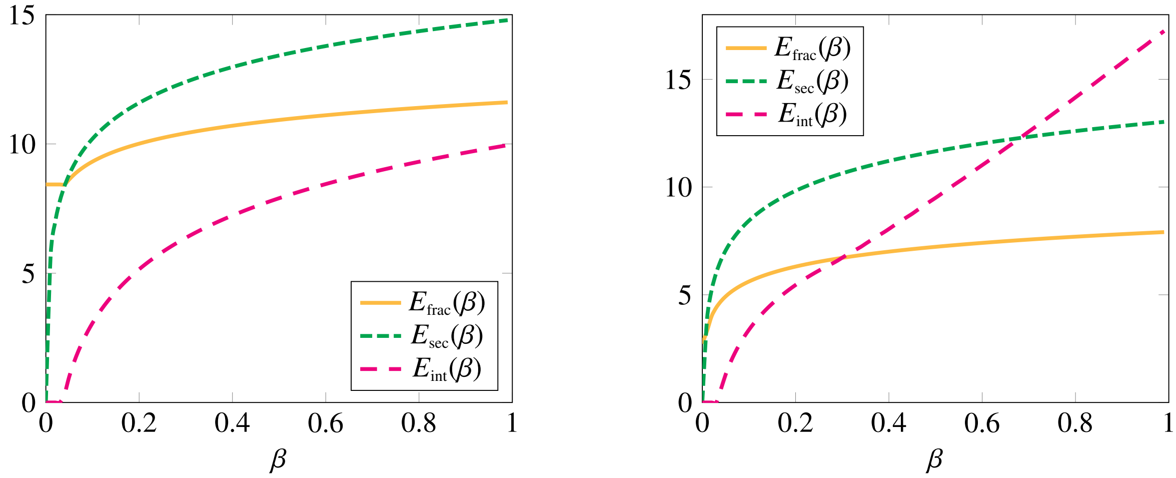


Figure 4. Comparison between $E_{\text{frac}}(\beta)$ and $E_{\text{sec}}(\beta)$ for $\rho = 0.4$ and $\theta = 0.6$.

With respect to the bound in $E_{\text{int}}(\beta)$, the situation is somewhat more complicated and depends on the model's parameters. For $E_{\text{int}}(\beta)$ to be higher than $E_{\text{sec}}(\beta)$, we have two contradictory conditions. On the one hand, we need $n^2Q(d, \theta) \leq 1$, which holds for relatively high θ values, but, on the other hand, we require $nP(d, \rho, \theta)$ relatively large, which holds for relatively low θ values. Thus, $E_{\text{int}}(\beta)$ improves on $E_{\text{sec}}(\beta)$ only in some cases but not generally. To capture this fact, we present plots of the three bounds in Figure 5 for the parameters $n = 200$ and $d = 50$, but for two different pairs of (ρ, θ) for which $n^2 Q(d, \theta) \leq 1$. Since the numbers $\mathsf{B}(k)$ in (5.17) grow quite rapidly with k, when calculating the bound in $E_{\text{int}}(\beta)$, we switched the infimum over $\{3, 4, \ldots\}$ with a minimization over $\{3, 4, \ldots, 30\}$. For $\rho = 0.6$ and $\theta = 0.7$, it turns out that $nP(d, \rho, \theta) \approx 24.2$; in this case, $E_{\text{sec}}(\beta)$ uniformly outperforms $E_{\text{int}}(\beta)$ for all $\beta \in [0, 1]$. For $\rho = 0.7$ and $\theta = 0.6$, we find that $nP(d, \rho, \theta) \approx 179.3$. In this alternative case, $E_{\text{int}}(\beta)$ is higher than $E_{\text{sec}}(\beta)$ for relatively high β values.



(a) $\rho = 0.6$ and $\theta = 0.7$ **(b)** $\rho = 0.7$ and $\theta = 0.6$

Figure 5. Comparison among $E_{\text{frac}}(\beta)$, $E_{\text{sec}}(\beta)$, and $E_{\text{int}}(\beta)$.

By comparing among $E_{\text{frac}}(\beta)$, $E_{\text{sec}}(\beta)$, and $E_{\text{int}}(\beta)$ for specific parameter values, and showing that none of them uniformly dominates the others, we can safely conclude that, generally, all three bounds are needed in order to achieve the best guarantee for the type I error probability.

6. Bounding the moments of $N(\theta)$

As we explained before in Section 5.3, a feasible way to upper-bound the type I probability of error is by using the generalized Chebyshev's inequality, which, in turn, requires to evaluate (or bound) the moments

$$\mathbb{E}_0\left[\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{1}\left\{\tilde{\boldsymbol{X}}_i^{\mathsf{T}}\tilde{\boldsymbol{Y}}_j \geq \theta\right\}\right)^{\alpha}\right], \tag{6.1}$$

where $\alpha > 0$. For $\alpha \in (0, 1]$ we relied on existing techniques from [41]. The case $\alpha > 1$ seems to be much more intricate; nonetheless, for k-th order integer moments, we were able to derive plausible bounds by using some graph-theoretic concepts. In what follows, we present a slightly more general result, which is the main theoretical contribution of this work.

Theorem 2. *Let $n, d \in \mathbb{N}$. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ and $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ be two sets of IID random variables taking values in $\mathbb{R}^d$. Let $J : \mathbb{R}^d \times \mathbb{R}^d \to \{0, 1\}$ and assume that*

$$\mathbb{P}(J(\boldsymbol{x}, \boldsymbol{Y}) = 1) \leq Q, \quad \forall \boldsymbol{x} \in \mathbb{R}^d, \tag{6.2}$$

$$\mathbb{P}(J(\boldsymbol{X}, \boldsymbol{y}) = 1) \leq Q, \quad \forall \boldsymbol{y} \in \mathbb{R}^d, \tag{6.3}$$

$$\mathbb{P}(J(\boldsymbol{X}, \boldsymbol{Y}) = 1) \leq Q, \tag{6.4}$$

where $Q \in [0, 1]$ is independent of $\boldsymbol{x}$ and $\boldsymbol{y}$. Let $k \in \mathbb{N}$. Then,

$$\mathbb{E}\left[\left(\sum_{\ell=1}^{n} \sum_{m=1}^{n} J(\boldsymbol{X}_\ell, \boldsymbol{Y}_m)\right)^k\right] \leq k(k+1)\mathsf{B}(k) \cdot \begin{cases} (n^2 Q)^k & \text{if } n^2 Q > 1, \\ n^2 Q & \text{if } n^2 Q \leq 1. \end{cases} \tag{6.5}$$

A proof of Theorem 2 is given in Appendix C. Due to the three technical uniformity conditions in (6.2)–(6.4), we can better understand the motivation for the initial normalization of the feature vectors in (5.4) and (5.5). Thanks to these normalizations and the spherical symmetry of the Gaussian distribution, we have the following for any $\boldsymbol{x} \in \mathbb{R}^d$ and $\boldsymbol{y} \in \mathbb{R}^d$:

$$\mathbb{P}\left\{\tilde{\boldsymbol{x}}^\mathsf{T} \tilde{\boldsymbol{Y}} \geq \theta\right\} = \mathbb{P}\left\{\tilde{\boldsymbol{X}}^\mathsf{T} \tilde{\boldsymbol{y}} \geq \theta\right\} = \mathbb{P}\left\{\tilde{\boldsymbol{X}}^\mathsf{T} \tilde{\boldsymbol{Y}} \geq \theta\right\} = Q(d, \theta). \tag{6.6}$$

Thus, we can use Theorem 2 when deriving theoretical guarantees on the type I error probability of the count test as defined with $N(\theta)$. On the other hand, referring to the statistic in (5.3), which does not involve data normalization, we see that the conditions in (6.2)–(6.4) are not satisfied. For example, the probability $\mathbb{P}\left\{\boldsymbol{x}^\mathsf{T} \boldsymbol{Y} \geq \theta\right\}$ is given by a function of θ and $\|\boldsymbol{x}\|$.

We now move to explaining the main proof concepts of Theorem 2 by referring to a slightly different probabilistic problem. The main benefit of switching to a close problem is that its proof techniques are very close to those required in the proof of Theorem 2, but they are still easy to understand from simple diagrams.

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_M$ be a set of IID random variables taking values in $\mathbb{R}^d$ and let $J : \mathbb{R}^d \times \mathbb{R}^d \to \{0, 1\}$. Let us denote the indicator random variables $\mathcal{I}(m, m') = J(\boldsymbol{X}_m, \boldsymbol{X}_{m'})$ and the enumerator

$$N = \sum_{(m,m') \in [M]^2} \mathcal{I}(m, m'), \tag{6.7}$$

where the set $[M]^2$ is an abbreviation for the set $\{(m, m') : m, m' \in \{1, 2, \ldots, M\}, m \neq m'\}$. The main objective is to derive an upper bound for $\mathbb{E}[N^k]$. In what follows, we provide and explain only the novel concepts in analyzing this k-th-order moment. We divide the analysis into three main steps: Graph counting, graph pruning, and graph reduction.

6.1. Step I: Graph counting

For any $k \in \mathbb{N}$, let $\mathsf{S}(k, d)$ be the number of ways to partition a set of k labeled objects into $d \in \{1, 2, \ldots, k\}$ nonempty unlabeled subsets, which is given by Stirling numbers of the second kind [45]:

$$\mathsf{S}(k, d) = \frac{1}{d!} \sum_{i=0}^{d} (-1)^i \binom{d}{i} (d-i)^k. \tag{6.8}$$

We begin as follows:

$$\mathbb{E}[N^k] = \mathbb{E}\left[\left(\sum_{(m,m') \in [M]^2} \mathcal{I}(m, m')\right)^k\right] \tag{6.9}$$

$$= \sum_{(m_1,m_1')\in[M]^2} \sum_{(m_2,m_2')\in[M]^2} \cdots \sum_{(m_k,m_k')\in[M]^2} \mathbb{E}\left[\mathcal{I}(m_1,m_1')\cdot\mathcal{I}(m_2,m_2')\cdot\ldots\cdot\mathcal{I}(m_k,m_k')\right] \tag{6.10}$$

$$= \sum_{d=1}^{k} \sum_{\left\{\substack{(m_i,m_i')\in[M]^2,\ 1\le i\le k,\\ \text{divided into } d \text{ subsets of identical pairs}}\right\}} \mathbb{E}\left[\mathcal{I}(m_1,m_1')\cdot\mathcal{I}(m_2,m_2')\cdot\ldots\cdot\mathcal{I}(m_k,m_k')\right] \tag{6.11}$$

$$= \sum_{d=1}^{k} \mathsf{S}(k,d) \sum_{\left\{\substack{(m_i,m_i')\in[M]^2,\ 1\le i\le d,\\ (m_i,m_i')\neq(m_j,m_j')\ \forall i\neq j}\right\}} \mathbb{E}\left[\mathcal{I}(m_1,m_1')\cdot\mathcal{I}(m_2,m_2')\cdot\ldots\cdot\mathcal{I}(m_d,m_d')\right], \tag{6.12}$$

where, in the inner summation in (6.11), we sum over all possible k pairs of vector indices, which are divided in any possible way into exactly d subsets, and all pairs in each subset are identical[IV]. In (6.12), we use the Stirling numbers of the second kind and sum over exactly d distinct pairs of vector indices, where all the identical pairs of indices in (6.11) have been merged together, using the trivial fact that multiplying any number of identical indicator random variables is equal to any one of them.

Let us handle the inner sum in (6.12). The idea is as follows. Instead of summing over the set $\{(m_i,m_i') \in [M]^2,\ 1 \le i \le d,\ (m_i,m_i') \neq (m_j,m_j')\ \forall i \neq j\}$ of d distinct pairs of vector indices, we represent each possible configuration of the indices in this set as a *graph* G, and sum over *all the different graphs* with exactly d distinct edges. In our graph representation, each vector index $m_i \in \{1,2,\ldots,M\}$ is denoted by a vertex, and each pair of indices (m_i,m_i'), $m_i \neq m_i'$, is connected by an edge[V]. Hence, the number of edges is fixed, but the numbers of vertices and subgraphs (i.e., disconnected parts of the graph) are variable.

Now, given $d \in \{1,2,\ldots,k\}$, we sum over the *range* of integers, which consists of all possible numbers of vertices needed to support a graph with d different edges. At one extreme, if all of the d edges are disconnected, then we must have $2d$ vertices. At the other extreme, let $\mathcal{V}(d)$ be the minimal number of vertices that can support a graph of d edges.

Next, given $d \in \{1,2,\ldots,k\}$ and $v \in [\mathcal{V}(d),2d]$, we sum over the range of possible number of subgraphs. Let $\mathcal{S}_{\min}(d,v)$ ($\mathcal{S}_{\max}(d,v)$) be the minimal (maximal) number of subgraphs that a graph with d edges and v vertices can have. For a given triplet (d,v,s), where $s \in [\mathcal{S}_{\min}(d,v),\mathcal{S}_{\max}(d,v)]$ is the number of subgraphs within G, note that one can create many different graphs (see Figure 6), and we have to take all of them into account. Hence, let $\mathsf{T}(d,v,s)$ be the number of distinct ways (or topologies) to connect a graph with d edges, v vertices, and s subgraphs. Finally, for any $1 \le i \le \mathsf{T}(d,v,s)$, let $\mathsf{G}_i(d,v,s)$ be the set of different graphs with d edges, v vertices, s subgraphs, and the topology i, which can be defined on the set of M vectors, as we explained before.

[IV] Two pairs (m_1,m_1') and (m_2,m_2') are called identical if and only if $m_1 = m_2$ and $m_1' = m_2'$, otherwise, they said to be distinct.

[V] Indices that belong to distinct pairs may be joined together, e.g., if (m_1,m_1') and (m_2,m_2') are distinct, then it may be that $m_1 = m_2$ or $m_1' = m_2'$, but not both.

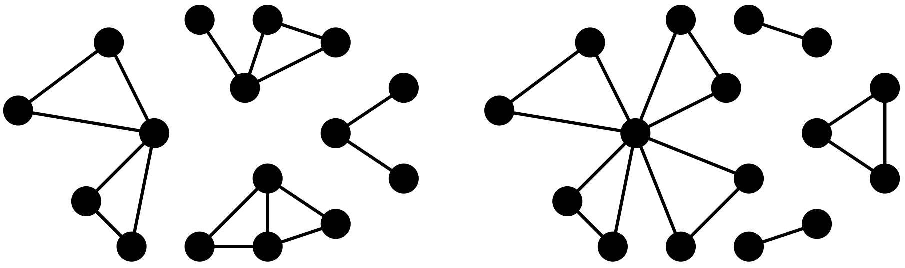


Figure 6. Examples of two different graphs with $d = 17$, $v = 16$, and $s = 4$.

Let $\Theta(G)$ be an indicator random variable that equals one if and only if all the indicator random variables related to the pairs of vectors that are linked by the edges of G equals one. With the definitions above, the inner sum in (6.12) can now be written as:

$$\sum_{\{(m_i,m_i')\in[M]^2,\ 1\le i\le d,\ (m_i,m_i')\neq(m_j,m_j')\ \forall i\neq j\}} \mathbb{E}\left[\mathcal{I}(m_1,m_1')\cdot\mathcal{I}(m_2,m_2')\cdot\ldots\cdot\mathcal{I}(m_d,m_d')\right]$$

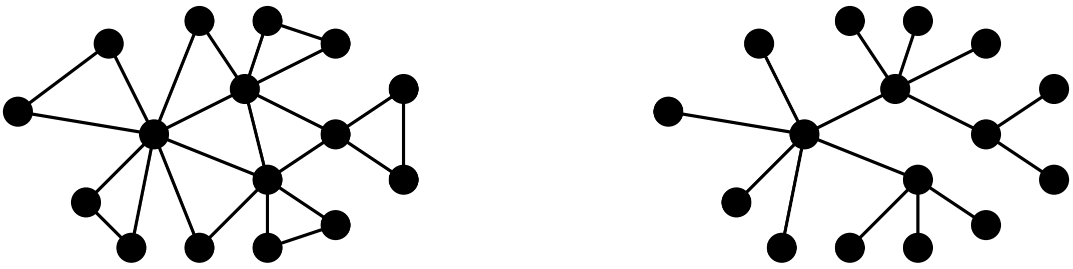
$$\equiv \sum_{\{(m_i,m_i')\in[M]^2,\ 1\le i\le d,\ (m_i,m_i')\neq(m_j,m_j')\ \forall i\neq j\}} \mathbb{E}\left[\prod_{i=1}^{d} J(\boldsymbol{X}_{m_i},\boldsymbol{X}_{m_i'})\right] \tag{6.13}$$

$$= \sum_{v=\mathcal{V}(d)}^{2d}\ \sum_{s=\mathcal{S}_{\min}(d,v)}^{\mathcal{S}_{\max}(d,v)}\ \sum_{i=1}^{\mathsf{T}(d,v,s)}\ \sum_{G\in\mathsf{G}_i(d,v,s)} \mathbb{E}\left[\Theta(G)\right], \tag{6.14}$$

i.e., for any $d \in \{1,2,\ldots,k\}$, we first sum over the number of vertices, then over the number of subgraphs. Later, for a fixed triplet (d, v, s), we sum over all possible $\mathsf{T}(d, v, s)$ topologies and finally over all specific graphs $G \in \mathsf{G}_i(d, v, s)$ with a given topology.

6.2. Step II: Graph pruning

It turns out that the expectation of $\Theta(G)$ can be easily evaluated if all subgraphs of G are trees (also known as a *forest* in the terminology of graph theory). If at least one subgraph of G contains loops, we apply a process of *graph pruning*, in which we cut out the minimal amount of edges[VI], while keeping all vertices intact, until we get a forest (for example, see Figure 7).



(a) The original graph

(b) The pruned graph

Figure 7. Example of the process of graph pruning.

Denote by $\mathcal{P}(G)$ the pruned graph of G. Notice that the expectation of $\Theta(G)$ is upper-bounded[VII]

[VI] In fact, this procedure is equivalent to upper-bounding some of the indicator functions in (6.13) by one.

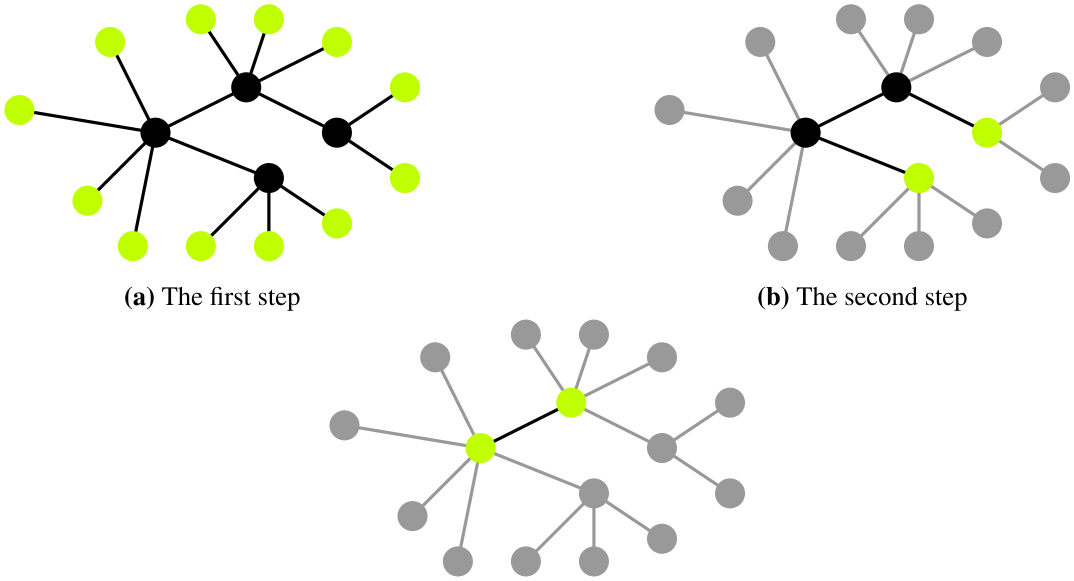
[VII] It follows from the fact that $\Theta(G) \le \Theta(\mathcal{P}(G))$ with probability one.

by the expectation of $\Theta(\mathcal{P}(G))$, and equality holds between them if G is a tree. Before we proceed, let us calculate the quantity $\mathcal{E}(G)$, which is the number of edges in $\mathcal{P}(G)$. Of course, if G is already a forest, then $\mathcal{E}(G) = d$. For any $G \in \mathsf{G}_i(d, v, s)$ which is not necessarily a forest, we find $\mathcal{E}(G)$ as follows. Assume that $v_1, v_2, \ldots, v_s$ are the number of vertices in each of the s subgraphs of G. Then, in the process of graph pruning, each of the $j \in \{1, 2, \ldots, s\}$ subgraphs of G will transform into a tree with exactly $v_j - 1$ edges. Hence,

$$\mathcal{E}(G) = \sum_{j=1}^{s} (v_j - 1) = v - s. \tag{6.15}$$

6.3. *Step III: Graph reduction*

The expectation of $\Theta(\mathcal{P}(G))$ can be upper-bounded in a simple iterative process of *graph reduction*. In the first step, we take the expectation with respect to all vectors that are labels of leaf vertices in $\mathcal{P}(G)$, while we condition on the realizations of all other vectors, namely those corresponding to inner vertices in $\mathcal{P}(G)$; afterwards, we erase the leaf vector vertices and the corresponding edges. The successive steps are identical to the first one on the remaining (unerased) graph, continuing until all vectors that are attributed to G have been considered (for example, see Figure 8).


(a) The first step

(b) The second step

(c) The last step

Figure 8. Example of the process of graph reduction.

In each step of graph reduction, the expectation with respect to each of the leaf vectors is upper-bounded as $\mathbb{P}\{J(\boldsymbol{x}, \boldsymbol{X}_{\text{leaf}}) = 1\} \leq Q$, regardless of $\boldsymbol{x}$, since we condition on the realizations of vectors that are attributed to the inner vertices. For any graph $G \in \mathsf{G}_i(d, v, s)$, we conclude that the expectation of $\Theta(\mathcal{P}(G))$ is upper-bounded by $Q^{\mathcal{E}(G)}$. Thus, for any $G \in \mathsf{G}_i(d, v, s)$, we prove the following upper bound on the expectation of $\Theta(G)$ as follows:

$$\mathbb{E}\left[\Theta(G)\right] \leq \mathbb{E}\left[\Theta(\mathcal{P}(G))\right] \leq Q^{\mathcal{E}(G)} = Q^{v-s}. \tag{6.16}$$

This bound involves, in fact, the main conceptual steps in the derivation of an upper bound on $\mathbb{E}[N^k]$; after upper-bounding (6.14) using (6.16), the remaining steps are much more technical and will not be presented here, as they resemble the second half of the proof of Theorem 2 (in Appendix C).

7. Comparison with the sum test

We now show that for some parameter choices, the new proposed statistical test in (5.12) achieves lower type I and type II error probabilities than the sum test in (4.2), which was proposed in [13]. The fact that we could prove that the new test has lower error probabilities is not trivial, since the analysis in [13] is based on Chernoff's bound, which is known to be relatively tight[VIII] in many cases, while the analysis in the current manuscript follows other methods that are able to accommodate the statistical dependencies between the indicator random variables in (5.6).

For each specific ρ, the trade-off between the two error probabilities varies significantly with the choices of the thresholds β and θ; hence, these parameters should be tuned accordingly to achieve the desired trade-off. Let us denote the right-hand sides of (5.18) and (5.19) as $P_{\mathrm{FA}}(\beta,\theta)$ and $P_{\mathrm{MD}}(\beta,\theta)$, respectively, where the dependence on (n,d,ρ) is made implicit. We define a unique performance curve for the count test by minimizing the type II error probability while constraining the type I error probability to a desired level. More specifically,

$$E_{\mathrm{MD}}^{\mathrm{Count}}(n,d,\rho,E_{\mathrm{FA}}) = \max_{\{(\beta,\theta)\in(0,1)^2:\ -\log P_{\mathrm{FA}}(\beta,\theta)\geq E_{\mathrm{FA}}\}} -\log P_{\mathrm{MD}}(\beta,\theta). \tag{7.1}$$

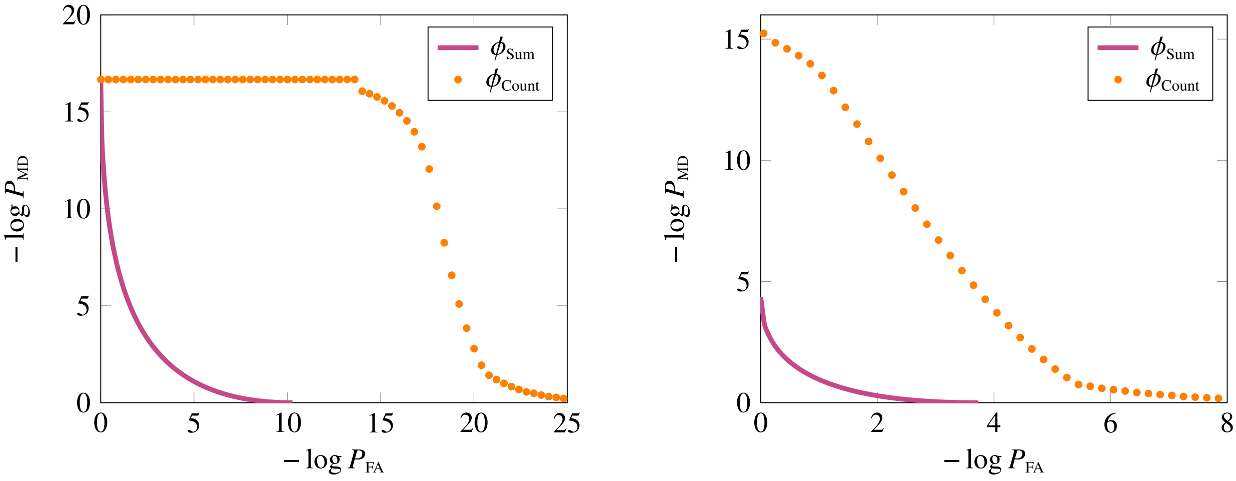
In Figure 9 below, we present a plot of the trade-off between $\frac{d}{2}G_{\mathrm{FA}}(\gamma)$ and $\frac{d}{2}G_{\mathrm{MD}}(\gamma)$ for $\gamma \in (0,4\rho^2)$ (solid line) together with a plot of $E_{\mathrm{MD}}^{\mathrm{Count}}(n,d,\rho,E_{\mathrm{FA}})$ as a function of E_{FA} (scattered points) for the parameters $n = 200$, $d = 50$, and for two values of ρ. The quantity $Q(d,\theta)$ is calculated numerically using (5.20) and $P(d,\rho,\theta)$ is calculated using (5.39) in Lemma 2. Since the numbers $\mathsf{B}(k)$ in (5.17) grow quite rapidly with k, when calculating the bound in (5.42), we switched the infimum over $\{3,4,5,\ldots\}$ with a minimization over $\{3,4,\ldots,30\}$. As can be seen in Figure 9, the new proposed statistical test achieves lower error probabilities for both $\rho = 0.7$ and $\rho = 0.4$.

In the comparison aboove, we see that the count test uniformly outperforms the sum test for specific choices of the parameter triplet n,d,ρ. We now show that the range of parameter choices for which the count test is better than the sum test is relatively wide. To compare between the two tests, let us define

$$E_{\mathrm{MD}}^{\mathrm{Sum}}(d,\rho,E_{\mathrm{FA}}) = \max_{\{\gamma\in(0,4\rho^2):\ \frac{d}{2}G_{\mathrm{FA}}(\gamma)\geq E_{\mathrm{FA}}\}} \tfrac{d}{2}G_{\mathrm{MD}}(\gamma). \tag{7.2}$$

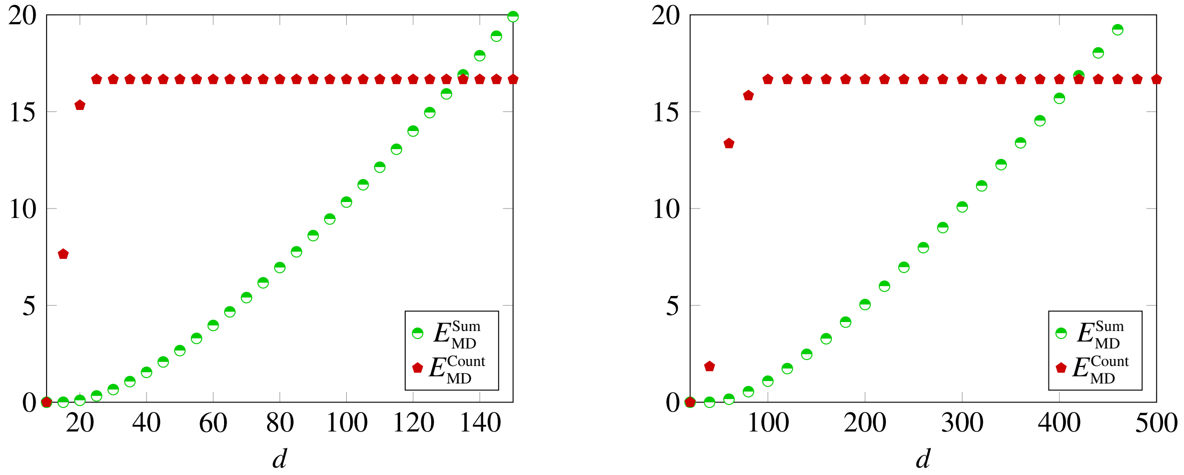
We let d vary while fixing the parameter values $n = 200$ and $E_{\mathrm{FA}} = 3$, such that the type I error probabilities are upper-bounded by approximately 0.05 for any value of d. In Figure 10 below, we present plots of $E_{\mathrm{MD}}^{\mathrm{Count}}(n,d,\rho,E_{\mathrm{FA}})$ and $E_{\mathrm{MD}}^{\mathrm{Sum}}(d,\rho,E_{\mathrm{FA}})$ as functions of d for two values of ρ. As can be seen in Figure 10, the count test outperforms the sum test for relatively low d values, while the sum test performs better otherwise. One should note that for lower values of ρ, the range of d in which the count test outperforms the sum test increases.

[VIII] In the sense that a lower bound with a matching exponent function can be derived.



(a) $\rho = 0.7$ **(b)** $\rho = 0.4$

Figure 9. Plots of the trade-offs between the exponential rates of decay of the two error probabilities for the sum test (4.2) (solid line) and for the count test (5.12) (scattered points).



(a) $\rho = 0.7$ **(b)** $\rho = 0.4$

Figure 10. Plots of the exponential rates of decay of the type II error probabilities versus d, maximized under the constraint that the type I error probabilities do not exceed 0.05.

It may be somewhat surprising to some readers that the sum test was found to perform worse than the count test in the Gaussian setup. Moreover, it is possible that the sum test may actually perform better in practice, but the (Chernoff-based) bounds for its error probabilities may not be tight enough to show this. In order to reject this hypothesis, we computed the type I and type II error probabilities of both tests via Monte–Carlo simulations. In Figure 11 below, we present points of the form $(-\log \hat{P}_{\text{FA}}, -\log \hat{P}_{\text{MD}})$ for some threshold values t in the sum test and for some threshold pairs (θ,β) in the count test for the parameters $n = 200$, $d = 50$, and $\rho = 0.3$. In order to facilitate the simulations for the count test, we fixed the local threshold as $\theta = 0.4$ and picked a few values for β in $(0, 1)$. The error probabilities of the sum test have been estimated on the basis of 20,000 samples at each point, while for the count test we increased the sample size to 200,000 at each point. As can be seen in Figure 11, the estimated error probabilities of the sum test are higher than those of the count test, thus the latter performs better also in practice, at least for the specific chosen parameter values.

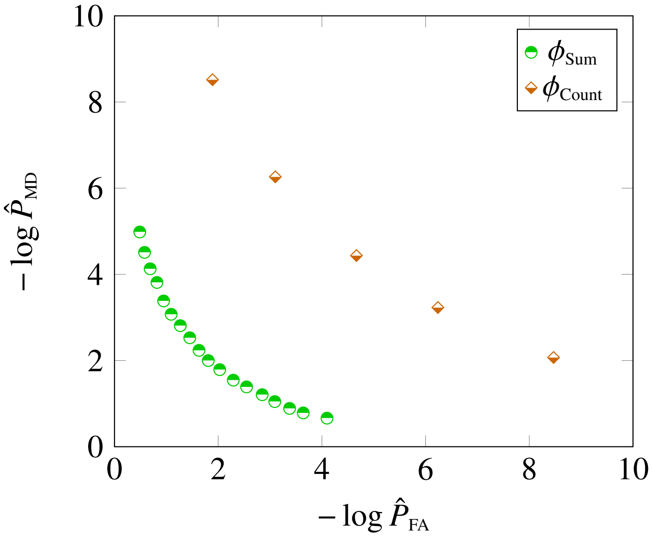


Figure 11. Plots of the trade-offs between the exponential rates of decay of the two error probabilities estimated via Monte–Carlo simulations for the sum test and the count test.

8. Summary and future research

In this manuscript, we propose the count test, which is a new statistical test that solves the correlation detection problem between two Gaussian databases. The proposed test has a reduced computational complexity in comparison with the GLR test, which makes it more desirable from a practical point of view. The count test relies on a statistic given by a sum of weakly dependent indicator random variables. Therefore, novel graph-theoretic techniques have been developed in order to derive theoretical guarantees on the type I probability of error. We compared the count test with the sum test, which was recently proposed as a valid solution for the same statistical problem. By comparing the analytical bounds of the two error probabilities, it is realized that the count test outperforms the sum test for a relatively wide range of system parameters. In principle, the probability bounds may not reflect the true performance of the tests, and thus we also present a comparison between the estimated error probabilities that result from Monte–Carlo simulations of both tests. In this comparison as well, the count test outperforms the sum test. Further possible directions for future research are suggested as follows.

(1) A goal for future research is to analyze the GLR test for the correlation detection problem and compare its performance with those of the sum test and the count test. Bounding the type II error probability of the GLR test should be relatively straightforward, but deriving a plausible bound for its type I error probability seems to be more challenging and may require new probabilistic tools. In addition, one can also propose a generalization of the simplified form of the GLR test in (4.10), which trades-off between computational complexity and reliability; we suggest a statistical test with a running time of $O(n^\delta)$ for some $\delta \in (1, 3)$. Specifically, we suggest the following test. For two given

databases $\boldsymbol{X}^n$ and $\boldsymbol{Y}^n$ and a predefined $\alpha \in [0, 1]$, we consider the first n^α rows in $\boldsymbol{X}^n$ and solve the partial assignment problem between $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n^\alpha}\}$ and $\boldsymbol{Y}^n$, i.e., we calculate the statistic

$$W(\alpha) = \max_{\{\mathcal{A}\subseteq[n], |\mathcal{A}|=n^\alpha\}} \max_{\sigma\in\mathcal{S}_{n^\alpha}} \sum_{i=1}^{n^\alpha} \boldsymbol{X}_i^\mathsf{T} \boldsymbol{Y}_{\sigma_i}^{\mathcal{A}}, \tag{8.1}$$

where $\boldsymbol{Y}^{\mathcal{A}}$ is a database that contains only the rows from $\boldsymbol{Y}^n$, with indices as given by the set $\mathcal{A} \subseteq [n]$. The proposed test then compares $W(\alpha)$ to a threshold as follows:

$$\phi_{\mathrm{GLR}}(\boldsymbol{X}^n, \boldsymbol{Y}^n) = \begin{cases} 0 & W(\alpha) < \xi \\ 1 & W(\alpha) \geq \xi, \end{cases} \tag{8.2}$$

where $\xi \in (\mathbb{E}_0[W(\alpha)], \mathbb{E}_1[W(\alpha)])$ is a global threshold.

Concerning the running time of the test $\phi_{\mathrm{GLR}}(\boldsymbol{X}^n, \boldsymbol{Y}^n)$, it is directly connected to the computational complexity of the partial assignment problem between J jobs and W workers ($J \leq W$). Such an assignment problem can be solved sequentially; one can add the j-th job and update the total cost in time $O(jW)$, yielding an overall time complexity of $O(\sum_{j=1}^{J} jW) = O(J^2W)$. In our case, the time complexity is given by $O(n^{1+2\alpha})$. It may be of special interest to compare this test at $\alpha = 0$ with the sum test (whose computational complexity scales like $O(n)$) and, additionally, to compare this test at $\alpha = 1/2$ with the count test (whose computational complexity scales like $O(n^2)$).

(2) In the current manuscript, we have considered a relatively simplified model, in which the components of every feature vector are IID and Gaussian. In practice, both of these assumptions may not be valid. In the first step towards more general settings, one should consists a model where each feature vector is still Gaussian but with a general covariance matrix. In a second step, a model with non-Gaussian feature vectors may be considered as well. It is of great interest to explore statistical tests for more general models, which are characterized by high reliability and low computational complexity.

Use of Generative-AI tools declaration

The author declares that he has not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The author gratefully acknowledges the timely and constructive reports of the anonymous referees, as well as editorial comments provided by the editor. Special thanks are extended to Edward Hall (EPFL) for his help in executing the Monte–Carlo simulations in Section 7. The author also thanks Argaman and Omri Tamir for sharing their artistic preferences when finalizing Figure 4.

Conflict of interest

The author declares no conflicts of interest.

References

1. P. Ohm, Broken promises of privacy: Responding to the surprising failure of anonymization, *UCLA Law Rev.*, **57** (2009), 1701.

2. J. Sedayao, R. Bhardwaj, N. Gorade, *Making big data, privacy, and anonymization work together in the enterprise: Experiences and issues,* In: 2014 IEEE International Congress on Big Data, USA: IEEE, 2014, 601–607. https://doi.org/10.1109/BigData.Congress.2014.92

3. L. Sweeney, Weaving technology and policy together to maintain confidentiality, *J. Law Med. Ethics*, **25** (1997), 98–110. https://doi.org/10.1111/j.1748-720X.1997.tb01885.x

4. F. M. Naini, J. Unnikrishnan, P. Thiran, M. Vetterli, Where you are is who you are: User identification by matching statistics, *IEEE T. Inf. Foren. Sec.*, **11** (2016), 358–372. https://doi.org/10.1109/TIFS.2015.2498131

5. A. Datta, D. Sharma, A. Sinha, *Provable de-anonymization of large datasets with sparse dimensions,* In: International Conference on Principles of Security and Trust, Berlin, Germany: Springer, 2012, 229–248. https://doi.org/10.1007/978-3-642-28641-4_13

6. A. Narayanan, V. Shmatikov, *Robust de-anonymization of large sparse datasets,* In: 2008 IEEE Symposium on Security and Privacy, USA: IEEE, 2008, 111–125. https://doi.org/10.1109/SP.2008.33

7. N. Takbiri, A. Houmansadr, D. L. Goeckel, H. Pishro-Nik, Matching anonymized and obfuscated time series to users' profiles, *IEEE T. Inform. Theory*, **65** (2019), 724–741. https://doi.org/10.1109/TIT.2018.2873134

8. G. Wondracek, T. Holz, E. Kirda, C. Kruegel, *A practical attack to de-anonymize social network users,* In: 2010 IEEE Symposium on Security and Privacy, USA: IEEE, 2010, 223–238. https://doi.org/10.1109/SP.2010.21

9. J. Su, A. Shukla, S. Goel, A. Narayanan, *De-anonymizing web browsing data with social networks,* In: Proc. 26th Int. Conf. World Wide Web, 2017, 1261–1269. https://doi.org/10.1145/3038912.3052714

10. L. Bilge, T. Strufe, D. Balzarotti, E. Kirda, *All your contacts are belong to us: Automated identity theft attacks on social networks,* In: Proc. 18th Int. Conf. World wide web, 2009, 551–560. https://doi.org/10.1145/1526709.1526784

11. M. Srivatsa M. Hicks, *Deanonymizing mobility traces: Using social network as a side-channel,* In: Proc. ACM Conf. Comput. Commun. Secur., 2012, 628–637. https://doi.org/10.1145/2382196.2382262

12. Z. Cheng, J. Caverlee, K. Lee, *You are where you tweet: A content based approach to geo-locating Twitter users,* In: Proc. 19th ACM Int. Conf. Inf. Knowl. Manage., 2010, 759–768. https://doi.org/10.1145/1871437.1871535

13. Z. K, B. Nazer, *Detecting correlated Gaussian databases,* In: IEEE International Symposium on Information Theory, Finland: IEEE, 2022, 2064–2069. https://doi.org/10.1109/ISIT50566.2022.9834731

14. D. Elimelech, W. Huleihel, *Phase transitions in the detection of correlated databases,* In: International Conference on Machine Learning (PMLR), 2023, 9246–9266.

15. V. Paslev, W. Huleihel, Testing dependency of unlabeled databases, *IEEE T. Inform. Theory*, **70** (2024), 7410–7431. https://doi.org/10.1109/TIT.2024.3442977

16. D. Cullina, P. Mittal, N. Kiyavash, *Fundamental limits of database alignment,* In: IEEE International Symposium on Information Theory, USA: IEEE, 2018, 651–655. https://doi.org/10.1109/ISIT.2018.8437908

17. O. E. Dai, D. Cullina, N. Kiyavash, *Database alignment with Gaussian features,* In: Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics, 2019, 3225–3233.

18. O. E. Dai, D. Cullina, N. Kiyavash, *Achievability of nearly-exact alignment for correlated Gaussian databases,* In: IEEE International Symposium on Information Theory, USA: IEEE, 2020, 1230–1235. https://doi.org/10.1109/ISIT44484.2020.9174507

19. F. Shirani, S. Garg, E. Erkip, *A concentration of measure approach to database de-anonymization,* In: IEEE International Symposium on Information Theory, France: IEEE, 2019, 2748–2752. https://doi.org/10.1109/ISIT.2019.8849392

20. S. Bakırtaş, E. Erkip, *Database matching under column deletions,* In: IEEE International Symposium on Information Theory, Australia: IEEE, 2021, 2720–2725. https://doi.org/10.1109/ISIT45174.2021.9518145

21. S. Bakırtaş, E. Erkip, *Database matching under adversarial column deletions,* In: IEEE Information Theory Workshop, France: IEEE, 2023, 181–185. https://doi.org/10.1109/ITW55543.2023.10161615

22. S. Bakırtaş, E. Erkip, *Distribution-agnostic database de-anonymization under synchronization errors,* In: IEEE International Workshop on Information Forensics and Security (WIFS), Germany: IEEE, 2023, 1–6. https://doi.org/10.1109/WIFS58808.2023.10374831

23. S. Bakırtaş, E. Erkip, Database matching under noisy synchronization errors, *IEEE T. Inform. Theory*, **70** (2024), 4335–4367. https://doi.org/10.1109/TIT.2024.3388990

24. M. Chertkov, L. Kroc, F. Krzakala, M. Vergassola, L. Zdeborová, *Inference in particle tracking experiments by passing messages between images,* In: Proceedings of the National Academy of Sciences, **107** (2010), 7663–7668. https://doi.org/10.1073/pnas.0910994107

25. D. Kunisky, J. Niles-Weed, *Strong recovery of geometric planted matchings,* In: ACM-SIAM Symposium on Discrete Algorithms, 2022, 834–876. https://doi.org/10.1137/1.9781611977073.36

26. P. Pedarsani, M. Grossglauser, *On the privacy of anonymized networks,* In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2011, 1235–1243. https://doi.org/10.1145/2020408.2020596

27. J. Ding, Z. Ma, Y. Wu, J. Xu, Efficient random graph matching via degree profiles, *Probab. Theory Rel.*, **179** (2021), 29–115. https://doi.org/10.1007/s00440-020-00997-4

28. Y. Wu, J. Xu, S. H. Yu, Testing correlation of unlabeled random graphs, *Ann. Appl. Probab.*, **33** (2023), 2519–2558. https://doi.org/10.1214/22-AAP1786

29. C. Mao, Y. Wu, J. Xu, S. H. Yu, Testing network correlation efficiently via counting trees, *Ann. Stat.*, **52** (2024), 2483–2505. https://doi.org/10.1214/23-AOS2261

30. Y. Wu, J. Xu, H. Y. Sophie, Settling the sharp reconstruction thresholds of random graph matching, *IEEE T. Inform. Theory*, **68** (2022), 5391–5417. https://doi.org/10.1109/TIT.2022.3169005

31. L. Ganassali, *Sharp threshold for alignment of graph databases with Gaussian weights*, In: Mathematical and Scientific Machine Learning, **2** (2022), 314–335.

32. S. M. Kay, *Fundamentals of statistical signal processing: Detection theory*, University of Rhode Island: Prentice Hall PTR, 1998.

33. M. Basseville, I. V. Nikiforov, *Detection of abrupt changes: theory and application*, Prentice-Hall, 1993.

34. H. W. Kuhn, The Hungarian Method for the assignment problem, *Nav. Res. Logist. Quart.,* **2** (1955), 83–97. https://doi.org/10.1002/nav.3800020109

35. H. W. Kuhn, Variants of the Hungarian method for assignment problems, *Nav. Res. Logist. Quart.,* **3** (1956), 253–258. https://doi.org/10.1002/nav.3800030404

36. J. Munkres, Algorithms for the assignment and transportation problems, *J. Soc. Ind. Appl. Math.,* **5** (1957), 32–38. https://doi.org/10.1137/0105003

37. J. Edmonds, R. M. Karp, Theoretical improvements in algorithmic efficiency for network flow problems, *JACM,* **19** (1972), 248–264. https://doi.org/10.1145/321694.321699

38. S. Janson, New versions of Suen's correlation inequality, *Random Struct. Algor.*, **13** (1998), 467–483. https://doi.org/10.1002/(SICI)1098-2418(199810/12)13:3/4¡467::AID-RSA15¿3.0.CO;2-W

39. R. Tamir, N. Merhav, N. Weinberger, A. G. i Fàbregas, Large deviations behavior of the logarithmic error probability of random codes, *IEEE T. Inform. Theory*, **66** (2020), 6635–6659. https://doi.org/10.1109/TIT.2020.2995136

40. L. V. Truong, G. Cocco, J. Font-Segura, A. G. i Fàbregas, Concentration properties of random codes, *IEEE T. Inform. Theory*, **69** (2023), 7499–7537. https://doi.org/10.1109/TIT.2023.3312326

41. N. Merhav, Error exponents of typical random codes, *IEEE T. Inform. Theory*, **64** (2018), 6223–6235. https://doi.org/10.1109/TIT.2018.2834503

42. N. Merhav, Statistical physics and information theory, *Found. Trends Commun.,* **6** (2009), 1–212. https://doi.org/10.1561/0100000052

43. C. Bunte, A. Lapidoth, On the listsize capacity with feedback, *IEEE T. Inform. Theory*, **60** (2014), 6733–6748. https://doi.org/10.1109/TIT.2014.2355815

44. S. Li, Concise formulas for the area and volume of a hyperspherical cap, *Asian J. Math. Stat.*, **4** (2011), 66–70. https://doi.org/10.3923/ajms.2011.66.70

45. J. Riordan, *An introduction to combinatorial analysis*, New York: Wiley, 1980.

Appendix A

Preliminaries

The main purpose of this appendix is to provide the general setting and the main result borrowed from [38].

Let $\{U_{\boldsymbol{k}}\}_{\boldsymbol{k}\in\mathcal{K}}$ be a family of Bernoulli random variables, where $\mathcal{K}$ is a set of multidimensional indexes. Let $\mathcal{G}$ be a dependency graph for $\{U_{\boldsymbol{k}}\}_{\boldsymbol{k}\in\mathcal{K}}$, i.e., a graph with a vertex set $\mathcal{K}$ such that if $\mathcal{A}$ and $\mathcal{B}$ are two disjoint subsets of $\mathcal{K}$, and $\mathcal{G}$ contains no edge between $\mathcal{A}$ and $\mathcal{B}$, then the families $\{U_{\boldsymbol{k}}\}_{\boldsymbol{k}\in\mathcal{A}}$ and $\{U_{\boldsymbol{k}}\}_{\boldsymbol{k}\in\mathcal{B}}$ are independent. Let $S = \sum_{\boldsymbol{k}\in\mathcal{K}} U_{\boldsymbol{k}}$ and $\Delta = \mathbb{E}[S]$. Moreover, we write $\boldsymbol{i} \sim \boldsymbol{j}$ if $(\boldsymbol{i}, \boldsymbol{j})$ is an edge in the dependency graph G. Let

$$\Omega = \max_{\boldsymbol{i}\in\mathcal{K}} \sum_{\boldsymbol{j}\in\mathcal{K},\boldsymbol{j}\sim\boldsymbol{i}} \mathbb{E}[U_{\boldsymbol{j}}] \tag{A.1}$$

and

$$\Theta = \frac{1}{2}\sum_{\boldsymbol{i}\in\mathcal{K}} \sum_{\boldsymbol{j}\in\mathcal{K},\boldsymbol{j}\sim\boldsymbol{i}} \mathbb{E}[U_{\boldsymbol{i}}U_{\boldsymbol{j}}]. \tag{A.2}$$

The following result will be used in the proof of Theorem 1 in Appendix B:

Theorem 3. *([38], Theorem 10) With the notation as given above, then for any* $0 \le \beta \le 1$,

$$\mathbb{P}\{S \le \beta\Delta\} \le \exp\left\{-\min\left((1-\beta)^2\frac{\Delta^2}{8\Theta+2\Delta}, (1-\beta)\frac{\Delta}{6\Omega}\right)\right\}. \tag{A.3}$$

Appendix B - Proof of Theorem 1

Analysis of false alarms

For $\alpha \ge 0$, consider the following:

$$\begin{aligned}
P_{\text{FA}}(\phi_{\text{Count}}) &= \mathbb{P}_0\left\{N(\theta) \ge \beta[nP + n(n-1)Q]\right\} && \text{(B.1)}\\
&= \mathbb{P}_0\left\{\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{1}\left\{\tilde{\boldsymbol{X}}_i^\mathsf{T}\tilde{\boldsymbol{Y}}_j \ge \theta\right\} \ge \beta[nP + n(n-1)Q]\right\} && \text{(B.2)}\\
&= \mathbb{P}_0\left\{\left(\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{1}\left\{\tilde{\boldsymbol{X}}_i^\mathsf{T}\tilde{\boldsymbol{Y}}_j \ge \theta\right\}\right)^\alpha \ge (\beta[nP + n(n-1)Q])^\alpha\right\} && \text{(B.3)}\\
&\le \frac{\mathbb{E}_0\left[\left(\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{1}\left\{\tilde{\boldsymbol{X}}_i^\mathsf{T}\tilde{\boldsymbol{Y}}_j \ge \theta\right\}\right)^\alpha\right]}{(\beta[nP + n(n-1)Q])^\alpha}, && \text{(B.4)}
\end{aligned}$$

where (B.4) is due to Markov's inequality. Since $\alpha \ge 0$ is arbitrary, it holds that

$$P_{\text{FA}}(\phi_{\text{Count}}) \le \inf_{\alpha\ge0} \frac{\mathbb{E}_0\left[\left(\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{1}\left\{\tilde{\boldsymbol{X}}_i^\mathsf{T}\tilde{\boldsymbol{Y}}_j \ge \theta\right\}\right)^\alpha\right]}{(\beta[nP + n(n-1)Q])^\alpha} \tag{B.5}$$

$$= \min\left\{\min_{\alpha\in[0,1]} \frac{\mathbb{E}_0\left[\left(\sum_{i=1}^n \sum_{j=1}^n \mathbb{1}\left\{\tilde{\boldsymbol{X}}_i^\mathsf{T}\tilde{\boldsymbol{Y}}_j \geq \theta\right\}\right)^\alpha\right]}{(\beta[nP+n(n-1)Q])^\alpha},\right.$$

$$\left.\inf_{\alpha\geq 1} \frac{\mathbb{E}_0\left[\left(\sum_{i=1}^n \sum_{j=1}^n \mathbb{1}\left\{\tilde{\boldsymbol{X}}_i^\mathsf{T}\tilde{\boldsymbol{Y}}_j \geq \theta\right\}\right)^\alpha\right]}{(\beta[nP+n(n-1)Q])^\alpha}\right\} \tag{B.6}$$

$$\leq \min\left\{\min_{\alpha\in[0,1]} \frac{\mathbb{E}_0\left[\left(\sum_{i=1}^n \sum_{j=1}^n \mathbb{1}\left\{\tilde{\boldsymbol{X}}_i^\mathsf{T}\tilde{\boldsymbol{Y}}_j \geq \theta\right\}\right)^\alpha\right]}{(\beta[nP+n(n-1)Q])^\alpha},\right.$$

$$\left.\inf_{k\in\mathbb{N}} \frac{\mathbb{E}_0\left[\left(\sum_{i=1}^n \sum_{j=1}^n \mathbb{1}\left\{\tilde{\boldsymbol{X}}_i^\mathsf{T}\tilde{\boldsymbol{Y}}_j \geq \theta\right\}\right)^k\right]}{(\beta[nP+n(n-1)Q])^k}\right\}, \tag{B.7}$$

so the main tasks are to evaluate the fractional α-th moment and the integer k-th moment of $N(\theta)$. We start with the fractional α-th moment of $N(\theta)$ and proceed as follows. For a given $\alpha \in [0,1]$, let $\xi \in [\alpha, 1]$. In this case,

$$\mathbb{E}_0\left[\left(\sum_{i=1}^n \sum_{j=1}^n \mathbb{1}\left\{\tilde{\boldsymbol{X}}_i^\mathsf{T}\tilde{\boldsymbol{Y}}_j \geq \theta\right\}\right)^\alpha\right]$$

$$= \mathbb{E}_0\left\{\left[\left(\sum_{i=1}^n \sum_{j=1}^n \mathbb{1}\left\{\tilde{\boldsymbol{X}}_i^\mathsf{T}\tilde{\boldsymbol{Y}}_j \geq \theta\right\}\right)^\xi\right]^{\alpha/\xi}\right\} \tag{B.8}$$

$$\leq \mathbb{E}_0\left\{\left[\sum_{i=1}^n \left(\sum_{j=1}^n \mathbb{1}\left\{\tilde{\boldsymbol{X}}_i^\mathsf{T}\tilde{\boldsymbol{Y}}_j \geq \theta\right\}\right)^\xi\right]^{\alpha/\xi}\right\} \tag{B.9}$$

$$\leq \left\{\mathbb{E}_0\left[\sum_{i=1}^n \left(\sum_{j=1}^n \mathbb{1}\left\{\tilde{\boldsymbol{X}}_i^\mathsf{T}\tilde{\boldsymbol{Y}}_j \geq \theta\right\}\right)^\xi\right]\right\}^{\alpha/\xi} \tag{B.10}$$

$$= \left\{\sum_{i=1}^n \mathbb{E}_0\left[\left(\sum_{j=1}^n \mathbb{1}\left\{\tilde{\boldsymbol{X}}_i^\mathsf{T}\tilde{\boldsymbol{Y}}_j \geq \theta\right\}\right)^\xi\right]\right\}^{\alpha/\xi}, \tag{B.11}$$

where the first inequality is based on the fact that $(\sum_i a_i)^t \leq \sum_i a_i^t$ whenever $\{a_i\}$ are non-negative and $t \in [0,1]$, and the second inequality follows from Jensen's inequality and the concavity of the function $f(u) = u^{\alpha/\xi}$ when $u \geq 0$ and $0 < \alpha/\xi \leq 1$. In order to proceed, we use the following modification of [43, Appendix A, Lemma 1.1].

Lemma 3. *Let $I_1, \ldots, I_n$ be IID Bernoulli random variables with success probability bounded as:*

$$p = \mathbb{P}(I_i = 1) = 1 - \mathbb{P}(I_i = 0) \leq Q, \tag{B.12}$$

where $n \in \mathbb{N}$ and $Q \in [0,1]$. Let $\xi \in (0,1)$. In this case,

$$\mathbb{E}\left[\left(\sum_{i=1}^{n} I_i\right)^{\xi}\right] \leq \begin{cases} nQ & \text{if } nQ \leq 1, \\ (nQ)^{\xi} & \text{if } nQ > 1. \end{cases} \tag{B.13}$$

Proof. We distinguish between two cases. If $nQ \leq 1$, then we rely on the result in [43, Equation (149)] with $(\lceil \xi \rceil !)^2 \lceil \xi \rceil = 1$ to conclude that

$$\mathbb{E}\left[\left(\sum_{i=1}^{n} I_i\right)^{\xi}\right] \leq nQ. \tag{B.14}$$

Conversely, if $nQ > 1$, Jensen's inequality and the concavity of the function $f(u) = u^{\xi}$ when $u \geq 0$ and $0 < \xi < 1$ yields

$$\mathbb{E}\left[\left(\sum_{i=1}^{n} I_i\right)^{\xi}\right] \leq \left(\mathbb{E}\left[\sum_{i=1}^{n} I_i\right]\right)^{\xi} \leq (nQ)^{\xi}. \tag{B.15}$$

■

Note that, conditioned on $\boldsymbol{X}_i = \boldsymbol{x}$, the random variables $\left\{\mathbb{1}\{\tilde{\boldsymbol{x}}^{\mathsf{T}} \tilde{\boldsymbol{Y}}_j \geq \theta\}\right\}_{j=1}^{n}$ are IID Bernoulli, and thus

$$\mathbb{E}_0\left[\left(\sum_{j=1}^{n} \mathbb{1}\left\{\tilde{\boldsymbol{X}}_i^{\mathsf{T}} \tilde{\boldsymbol{Y}}_j \geq \theta\right\}\right)^{\xi} \middle| \boldsymbol{X}_i\right] \leq \begin{cases} nQ & \text{if } nQ \leq 1, \\ (nQ)^{\xi} & \text{if } nQ > 1. \end{cases} \tag{B.16}$$

Since the right-hand side of (B.16) does not depend on $\boldsymbol{X}_i$, it also holds that

$$\mathbb{E}_0\left[\left(\sum_{j=1}^{n} \mathbb{1}\left\{\tilde{\boldsymbol{X}}_i^{\mathsf{T}} \tilde{\boldsymbol{Y}}_j \geq \theta\right\}\right)^{\xi}\right] \leq \begin{cases} nQ & \text{if } nQ \leq 1, \\ (nQ)^{\xi} & \text{if } nQ > 1. \end{cases} \tag{B.17}$$

Upper-bounding (B.11) using (B.17) yields

$$\mathbb{E}_0\left[\left(\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{1}\left\{\tilde{\boldsymbol{X}}_i^{\mathsf{T}} \tilde{\boldsymbol{Y}}_j \geq \theta\right\}\right)^{\alpha}\right] \leq \left(\sum_{i=1}^{n} \begin{cases} nQ & \text{if } nQ \leq 1 \\ (nQ)^{\xi} & \text{if } nQ > 1 \end{cases}\right)^{\alpha/\xi} \tag{B.18}$$

$$= \left(\begin{cases} n^2 Q & \text{if } nQ \leq 1 \\ n^{1+\xi} Q^{\xi} & \text{if } nQ > 1 \end{cases}\right)^{\alpha/\xi} \tag{B.19}$$

$$= \begin{cases} (n^2 Q)^{\alpha/\xi} & \text{if } nQ \leq 1 \\ n^{(1/\xi+1)\alpha} Q^{\alpha} & \text{if } nQ > 1 \end{cases}. \tag{B.20}$$

Finally, minimizing over $\xi \in [\alpha, 1]$ yields

$$\mathbb{E}_0\left[\left(\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{1}\left\{\tilde{\boldsymbol{X}}_i^{\mathsf{T}} \tilde{\boldsymbol{Y}}_j \geq \theta\right\}\right)^{\alpha}\right] \leq \min_{\xi \in [\alpha,1]} \begin{cases} (n^2 Q)^{\alpha/\xi} & \text{if } nQ \leq 1 \\ n^{(1/\xi+1)\alpha} Q^{\alpha} & \text{if } nQ > 1 \end{cases} \tag{B.21}$$

$$= \min_{\xi\in[\alpha,1]} \begin{cases} (n^2Q)^{\alpha/\xi} & \text{if } nQ \le 1, n^2Q \le 1 \\ (n^2Q)^{\alpha/\xi} & \text{if } nQ \le 1, n^2Q > 1 \\ n^{(1/\xi+1)\alpha}Q^{\alpha} & \text{if } nQ > 1, n^2Q > 1 \end{cases} \tag{B.22}$$

$$= \begin{cases} (n^2Q)^{\alpha/\alpha} & \text{if } nQ \le 1, n^2Q \le 1 \\ (n^2Q)^{\alpha/1} & \text{if } nQ \le 1, n^2Q > 1 \\ n^{(1/1+1)\alpha}Q^{\alpha} & \text{if } nQ > 1, n^2Q > 1 \end{cases} \tag{B.23}$$

$$= \begin{cases} n^2Q & \text{if } nQ \le 1, n^2Q \le 1 \\ (n^2Q)^{\alpha} & \text{if } nQ \le 1, n^2Q > 1 \\ (n^2Q)^{\alpha} & \text{if } nQ > 1, n^2Q > 1 \end{cases} \tag{B.24}$$

$$= \begin{cases} n^2Q & \text{if } n^2Q \le 1 \\ (n^2Q)^{\alpha} & \text{if } n^2Q > 1 \end{cases}. \tag{B.25}$$

It then follows that the first term inside the minimum in (B.7) is upper-bounded by

$$\min_{\alpha\in[0,1]} \frac{(n^2Q)^{\alpha}\cdot \mathbb{1}\{n^2Q > 1\} + n^2Q\cdot\mathbb{1}\{n^2Q \le 1\}}{(\beta[nP + n(n-1)Q])^{\alpha}}. \tag{B.26}$$

Moving to upper-bounding the k-th integer moments of $N(\theta)$, let us start with the case of $k = 2$.

$$\begin{aligned} &\mathbb{E}_0\left[\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{1}\left\{\tilde{\boldsymbol{X}}_i^{\mathsf{T}}\tilde{\boldsymbol{Y}}_j \ge \theta\right\}\right)^2\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{\ell=1}^{n}\mathbb{P}\left\{\tilde{\boldsymbol{X}}_i^{\mathsf{T}}\tilde{\boldsymbol{Y}}_j \ge \theta, \tilde{\boldsymbol{X}}_k^{\mathsf{T}}\tilde{\boldsymbol{Y}}_\ell \ge \theta\right\} && \text{(B.27)} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k\neq i}\sum_{\ell\neq j}\mathbb{P}\left\{\tilde{\boldsymbol{X}}_i^{\mathsf{T}}\tilde{\boldsymbol{Y}}_j \ge \theta, \tilde{\boldsymbol{X}}_k^{\mathsf{T}}\tilde{\boldsymbol{Y}}_\ell \ge \theta\right\} \\ &\quad + \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k\neq i}\mathbb{P}\left\{\tilde{\boldsymbol{X}}_i^{\mathsf{T}}\tilde{\boldsymbol{Y}}_j \ge \theta, \tilde{\boldsymbol{X}}_k^{\mathsf{T}}\tilde{\boldsymbol{Y}}_j \ge \theta\right\} \\ &\quad + \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{\ell\neq j}\mathbb{P}\left\{\tilde{\boldsymbol{X}}_i^{\mathsf{T}}\tilde{\boldsymbol{Y}}_j \ge \theta, \tilde{\boldsymbol{X}}_i^{\mathsf{T}}\tilde{\boldsymbol{Y}}_\ell \ge \theta\right\} \\ &\quad + \sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{P}\left\{\tilde{\boldsymbol{X}}_i^{\mathsf{T}}\tilde{\boldsymbol{Y}}_j \ge \theta\right\} && \text{(B.28)} \\ &= n^2(n-1)^2Q^2 + n^2(n-1)Q^2 + n^2(n-1)Q^2 + n^2Q && \text{(B.29)} \\ &= n^2(n^2-1)Q^2 + n^2Q, && \text{(B.30)} \end{aligned}$$

where (B.29) is due to

$$\begin{aligned} &\mathbb{P}\left\{\tilde{\boldsymbol{X}}_i^{\mathsf{T}}\tilde{\boldsymbol{Y}}_j \ge \theta, \tilde{\boldsymbol{X}}_i^{\mathsf{T}}\tilde{\boldsymbol{Y}}_\ell \ge \theta\right\} \\ &\quad = \int \mathbb{P}\left\{\tilde{\boldsymbol{x}}^{\mathsf{T}}\tilde{\boldsymbol{Y}}_j \ge \theta, \tilde{\boldsymbol{x}}^{\mathsf{T}}\tilde{\boldsymbol{Y}}_\ell \ge \theta\right\}\mathrm{d}F_{X_i}(\boldsymbol{x}) && \text{(B.31)} \\ &\quad = \int \mathbb{P}\left\{\tilde{\boldsymbol{x}}^{\mathsf{T}}\tilde{\boldsymbol{Y}}_j \ge \theta\right\}\cdot\mathbb{P}\left\{\tilde{\boldsymbol{x}}^{\mathsf{T}}\tilde{\boldsymbol{Y}}_\ell \ge \theta\right\}\mathrm{d}F_{X_i}(\boldsymbol{x}) && \text{(B.32)} \end{aligned}$$

$$= \int Q^2 \mathrm{d}F_{\boldsymbol{X}_i}(\boldsymbol{x}) \tag{B.33}$$

$$= Q^2. \tag{B.34}$$

The resulting second-order bound is given by

$$\frac{n^2(n^2-1)Q^2 + n^2 Q}{(\beta[nP + n(n-1)Q])^2}. \tag{B.35}$$

For the general case of $k \geq 3$, we use the result of Theorem 2 (which appears in Section 6). It follows that the second term inside the minimum in (B.7) is upper-bounded by

$$\inf_{k\geq 3} \frac{k(k+1)\mathsf{B}(k)\left[(n^2Q)^k \cdot \mathbb{1}\{n^2Q > 1\} + n^2Q \cdot \mathbb{1}\{n^2Q \leq 1\}\right]}{(\beta[nP + n(n-1)Q])^k}. \tag{B.36}$$

Upper-bounding (B.7) with (B.26), (B.35), and (B.36) proves (5.18) in Theorem 1.

Analysis of missed detection

Consider the following:

$$P_{\mathrm{MD}}(\phi_{\mathrm{Count}}) = \mathbb{P}_{1|\sigma}\left\{N(\theta) \leq \beta\left[nP + n(n-1)Q\right]\right\} \tag{B.37}$$

$$= \mathbb{P}_{1|\sigma}\left\{\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{1}\left\{\tilde{\boldsymbol{X}}_i^{\mathsf{T}}\tilde{\boldsymbol{Y}}_j \geq \theta\right\} \leq \beta\Delta_{n,d}\right\}. \tag{B.38}$$

Let us abbreviate the indicator random variables as $\mathcal{I}(i,j) = \mathbb{1}\left\{\tilde{\boldsymbol{X}}_i^{\mathsf{T}}\tilde{\boldsymbol{Y}}_j \geq \theta\right\}$. In order to upper-bound (B.38), we use the result of Theorem 3, which appears in Appendix A. In our case, we have $\Delta = \Delta_{n,d}$, and it only remains to assess the quantities Θ and Ω. One can easily check that the indicator random variables $\mathcal{I}(i,j)$ and $\mathcal{I}(k,\ell)$ are independent as long as $i \neq k$ and $j \neq \ell$. Thus, we define our dependency graph so that each vertex (i,j) is connected exactly to $2n-2$ vertices of the form (i,ℓ), $\ell \neq j$ or (k,j), $k \neq i$. If the vertices (i,j) and (k,ℓ) are connected, we write $(i,j) \sim (k,\ell)$.

Let us denote the set $\Psi_n = \{(m,m') : m,m' \in \{1,2,\ldots,n\}\}$. Concerning Θ and Ω, we get

$$\Theta = \frac{1}{2}\sum_{(i,j)\in\Psi_n}\sum_{\substack{(k,\ell)\in\Psi_n\\(k,\ell)\sim(i,j)}} \mathbb{E}[\mathcal{I}(i,j)\mathcal{I}(k,\ell)] \tag{B.39}$$

$$= \frac{1}{2}\sum_{i=1}^{n}\sum_{\substack{(k,\ell)\in\Psi_n\\(k,\ell)\sim(i,i)}} \mathbb{E}[\mathcal{I}(i,i)\mathcal{I}(k,\ell)] + \frac{1}{2}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{\substack{(k,\ell)\in\Psi_n\\(k,\ell)\sim(i,j)}} \mathbb{E}[\mathcal{I}(i,j)\mathcal{I}(k,\ell)]. \tag{B.40}$$

Regarding the summands in the left-hand term of (B.40), we derive them as follows:

$$\mathbb{E}[\mathcal{I}(i,i)\mathcal{I}(k,\ell)] = \mathbb{P}\left\{\tilde{\boldsymbol{X}}_i^{\mathsf{T}}\tilde{\boldsymbol{Y}}_i \geq \theta, \tilde{\boldsymbol{X}}_i^{\mathsf{T}}\tilde{\boldsymbol{Y}}_\ell \geq \theta\right\} \tag{B.41}$$

$$= \int \mathbb{P}\left\{\tilde{\boldsymbol{x}}^{\mathsf{T}}\tilde{\boldsymbol{Y}}_i \geq \theta, \tilde{\boldsymbol{x}}^{\mathsf{T}}\tilde{\boldsymbol{Y}}_\ell \geq \theta\right\} \mathrm{d}F_{\boldsymbol{X}_i}(\boldsymbol{x}) \tag{B.42}$$

$$= \int \mathbb{P}\left\{\tilde{\boldsymbol{x}}^{\mathsf{T}}\tilde{\boldsymbol{Y}}_i \geq \theta\right\} \cdot \mathbb{P}\left\{\tilde{\boldsymbol{x}}^{\mathsf{T}}\tilde{\boldsymbol{Y}}_\ell \geq \theta\right\} \mathrm{d}F_{\boldsymbol{X}_i}(\boldsymbol{x}) \tag{B.43}$$

$$= \int PQ\mathrm{d}F_{\boldsymbol{X}_i}(\boldsymbol{x}) \tag{B.44}$$

$$= PQ. \tag{B.45}$$

Deriving the summands in the right-hand term of (B.40) in a similar way, we eventually arrive at

$$\Theta \leq \frac{1}{2}PQ(2n-2)n + \frac{1}{2}[2Q^2(n-2) + 2PQ]n(n-1) \tag{B.46}$$

$$= PQ(n-1)n + [Q^2(n-2) + PQ]n(n-1) \tag{B.47}$$

$$= 2PQn(n-1) + Q^2 n(n-1)(n-2). \tag{B.48}$$

In addition

$$\Omega = \max_{(i,j)\in\Psi_n} \sum_{\substack{(k,\ell)\in\Psi_n \\ (k,\ell)\sim(i,j)}} \mathbb{E}[\mathcal{I}(k,\ell)] \tag{B.49}$$

$$= \max\left\{ \sum_{\substack{(k,\ell)\in\Psi_n \\ (k,\ell)\sim(i,i)}} \mathbb{E}[\mathcal{I}(k,\ell)], \sum_{\substack{(k,\ell)\in\Psi_n \\ (k,\ell)\sim(i,j),\ i\neq j}} \mathbb{E}[\mathcal{I}(k,\ell)] \right\} \tag{B.50}$$

$$= \max\{(2n-2)Q, 2P + (2n-4)Q\} \tag{B.51}$$

$$= 2P + (2n-4)Q. \tag{B.52}$$

Then

$$\frac{\Delta_{n,d}}{6\Omega} = \frac{nP + n(n-1)Q}{6[2P + (2n-4)Q]} \tag{B.53}$$

$$= \frac{n[P + (n-1)Q]}{12[P + (n-2)Q]} \tag{B.54}$$

$$\geq \frac{n[P + (n-2)Q]}{12[P + (n-2)Q]} \tag{B.55}$$

$$= \frac{n}{12}, \tag{B.56}$$

and

$$\frac{\Delta_{n,d}^2}{8\Theta + 2\Delta_{n,d}} \geq \frac{[nP + n(n-1)Q]^2}{8[2PQn(n-1) + Q^2 n(n-1)(n-2)] + 2[nP + n(n-1)Q]} \tag{B.57}$$

$$\geq \frac{[nP + n(n-1)Q]^2}{8[2PQn^2 + 2Q^2 n^2(n-1)] + 2[nP + n(n-1)Q]} \tag{B.58}$$

$$= \frac{[nP + n(n-1)Q]^2}{16Qn[Pn + Qn(n-1)] + 2[nP + n(n-1)Q]} \tag{B.59}$$

$$= \frac{nP + n(n-1)Q}{16Qn + 2}. \tag{B.60}$$

Hence

$$P_{\mathrm{MD}}(\phi_{\mathrm{Count}}) = \mathbb{P}_{1|\sigma}\left\{\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{1}\left\{\tilde{\boldsymbol{X}}_i^{\mathsf{T}}\tilde{\boldsymbol{Y}}_j \geq \theta\right\} \leq \beta\Delta_{n,d}\right\} \tag{B.61}$$

$$\leq \exp\left\{-\min\left((1-\beta)^2\frac{\Delta_{n,d}^2}{8\Theta+2\Delta_{n,d}}, (1-\beta)\frac{\Delta_{n,d}}{6\Omega}\right)\right\} \tag{B.62}$$

$$\leq \exp\left\{-\min\left((1-\beta)^2\frac{nP+n(n-1)Q}{16Qn+2}, (1-\beta)\frac{n}{12}\right)\right\}, \tag{B.63}$$

which completes the proof of Theorem 1.

Appendix C - Proof of Theorem 2

For any $k \in \mathbb{N}$, let $\mathsf{S}(k,d)$ be the number of ways to partition a set of k labeled objects into $d \in \{1,2,\ldots,k\}$ nonempty unlabeled subsets, which is given by Stirling numbers of the second kind [45], i.e.,

$$\mathsf{S}(k,d) = \frac{1}{d!}\sum_{i=0}^{d}(-1)^i\binom{d}{i}(d-i)^k. \tag{C.1}$$

Obviously, $\mathsf{S}(k,1) = \mathsf{S}(k,k) = 1$. As before, we denote $\Psi_n = \{(m,m') :\ m,m' \in \{1,2,\ldots,n\}\}$. In this case,

$$\mathbb{E}\left[N^k\right] = \mathbb{E}\left[\left(\sum_{(m,m')\in\Psi_n} J(\boldsymbol{X}_m,\boldsymbol{Y}_{m'})\right)^k\right] \tag{C.2}$$

$$= \sum_{(m_1,m_1')\in\Psi_n}\cdots\sum_{(m_k,m_k')\in\Psi_n}\mathbb{E}\left[J(\boldsymbol{X}_{m_1},\boldsymbol{Y}_{m_1'})\cdot\ldots\cdot J(\boldsymbol{X}_{m_k},\boldsymbol{Y}_{m_k'})\right] \tag{C.3}$$

$$= \sum_{d=1}^{k}\sum_{\left\{\substack{(m_i,m_i')\in\Psi_n,\ 1\leq i\leq k,\\ \text{divided into } d \text{ subsets of identical pairs}}\right\}}\mathbb{E}\left[J(\boldsymbol{X}_{m_1},\boldsymbol{Y}_{m_1'})\cdot\ldots\cdot J(\boldsymbol{X}_{m_k},\boldsymbol{Y}_{m_k'})\right] \tag{C.4}$$

$$= \sum_{d=1}^{k}\mathsf{S}(k,d)\sum_{\left\{\substack{(m_i,m_i')\in\Psi_n,\ 1\leq i\leq d,\\ (m_i,m_i')\neq(m_j,m_j')\ \forall i\neq j}\right\}}\mathbb{E}\left[J(\boldsymbol{X}_{m_1},\boldsymbol{Y}_{m_1'})\cdot\ldots\cdot J(\boldsymbol{X}_{m_d},\boldsymbol{Y}_{m_d'})\right], \tag{C.5}$$

where, in the inner summation in (C.4), we sum over all possible k pairs of vectors indices, which are divided in any possible way into exactly d subsets, and all pairs in each subset are identical. In (C.5), we use the Stirling numbers of the second kind and sum over exactly d distinct pairs of vector indices, where all the identical pairs of indices in (C.4) have been merged together, using the trivial fact that multiplying any number of identical indicator random variables is equal to any one of them.

Let us handle the inner sum of (C.5). The idea is as follows. Instead of summing over the set $\{(m_i,m_i') \in \Psi_n,\ 1 \leq i \leq d,\ (m_i,m_i') \neq (m_j,m_j')\ \forall i \neq j\}$ of d distinct pairs of indices of vectors, we represent each possible configuration of indices in this set as a *graph G* and sum over *all different* *graphs* with exactly d distinct edges. In our graph representation, each vector index $m_i \in \{1, 2, \ldots, n\}$ and $m'_i \in \{1, 2, \ldots, n\}$ is denoted by a vertex, and each pair of indices (m_i, m'_i), $m_i \neq m'_i$, is connected by an edge. Hence, the number of edges is fixed, but the numbers of vertices and subgraphs (i.e., disconnected parts of the graph) are variable.

Given $d \in \{1, 2, \ldots, k\}$, we sum over the *set* $\mathcal{V}(d) = \{(v_x, v_y)\}$ of pairs of integers, which consists of all possible pairs of vertices needed in order to support a graph with d different edges. Next, given $d \in \{1, 2, \ldots, k\}$ and $(v_x, v_y) \in \mathcal{V}(d)$, we sum over the range of the possible number of subgraphs. Let $\mathcal{S}_{\min}(d, v_x, v_y)$ ($\mathcal{S}_{\max}(d, v_x, v_y)$) be the minimal (maximal) number of subgraphs that a graph with d edges and (v_x, v_y) vertices can have. For a given quadruplet (d, v_x, v_y, s), where $s \in [\mathcal{S}_{\min}(d, v_x, v_y), \mathcal{S}_{\max}(d, v_x, v_y)]$ is the number of subgraphs within G, note that one can create many different graphs (see Figure 12), and we have to take all of them into account. Hence, let $\mathsf{T}(d, v_x, v_y)$ ($\mathsf{T}(d, v_x, v_y, s)$) be the number of distinct ways to connect a graph with d edges and (v_x, v_y) vertices (and s subgraphs). Finally, for any $1 \leq i \leq \mathsf{T}(d, v_x, v_y, s)$, let $\mathsf{G}_i(d, v_x, v_y, s)$ be the set of different graphs with d edges, (v_x, v_y) vertices, and s subgraph, that can be defined on the set Ψ_n of pairs of vectors, as we explained before.

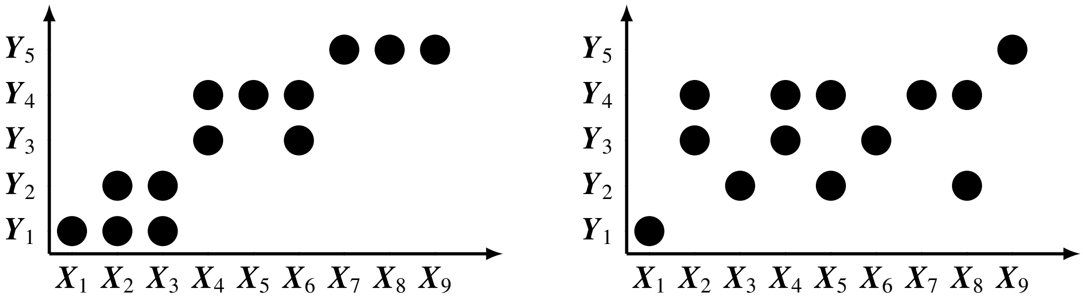

Y_5
Y_4
Y_3
Y_2
Y_1
X_1 X_2 X_3 X_4 X_5 X_6 X_7 X_8 X_9
Y_5
Y_4
Y_3
Y_2
Y_1
X_1 X_2 X_3 X_4 X_5 X_6 X_7 X_8 X_9

Figure 12. Examples of two different graphs with $d = 13$, $v_x = 9$, $v_y = 5$, and $s = 3$.

We now prove that the cardinality of the set $\mathsf{G}_i(d, v_x, v_y, s)$ is upper-bounded by an expression that depends only on the numbers of vertices (v_x, v_y). First, let $N(v_x)$ be the number of options to choose v_x different vectors from a set of cardinality n. In this case,

$$N(v_x) = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot (n - v_x + 1) \leq n^{v_x}. \tag{C.6}$$

Thus

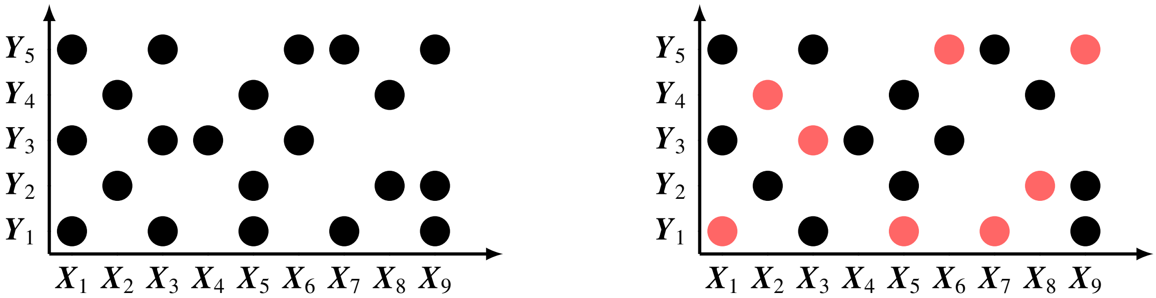
$$\begin{aligned} |\mathsf{G}_i(d, v_x, v_y, s)| &= N(v_x) \cdot N(v_y) && \text{(C.7)} \\ &\leq n^{v_x} \cdot n^{v_y}. && \text{(C.8)} \end{aligned}$$

Let $\Theta(G)$ be an indicator random variable that equals one if and only if all pairs of feature vectors that are linked by the edges of G have $J(\boldsymbol{X}, \boldsymbol{Y}) = 1$. With the definitions above, the inner sum of (C.5) can now be written as

$$\begin{aligned} &\sum_{\{(m_i, m'_i) \in \Psi_n,\ 1 \leq i \leq d,\ (m_i, m'_i) \neq (m_j, m'_j)\ \forall i \neq j\}} \mathbb{E}\left[J(\boldsymbol{X}_{m_1}, \boldsymbol{Y}_{m'_1}) \cdot \ldots \cdot J(\boldsymbol{X}_{m_d}, \boldsymbol{Y}_{m'_d})\right] \\ &\equiv \sum_{\{(m_i, m'_i) \in \Psi_n,\ 1 \leq i \leq d,\ (m_i, m'_i) \neq (m_j, m'_j)\ \forall i \neq j\}} \mathbb{E}\left[\prod_{i=1}^{d} J(\boldsymbol{X}_{m_i}, \boldsymbol{Y}_{m'_i})\right] \qquad \text{(C.9)} \end{aligned}$$

$$= \sum_{(v_x,v_y)\in\mathcal{V}(d)} \sum_{s=\mathcal{S}_{\min}(d,v_x,v_y)}^{\mathcal{S}_{\max}(d,v_x,v_y)} \sum_{i=1}^{\mathsf{T}(d,v_x,v_y,s)} \sum_{G\in\mathsf{G}_i(d,v_x,v_y,s)} \mathbb{E}\left[\Theta(G)\right], \tag{C.10}$$

i.e., for any $d \in \{1, 2, \dots, k\}$, we first sum over the numbers of vertices, then over the number of subgraphs, then, for a fixed quadruplet (d, v_x, v_y, s), over all possible $\mathsf{T}(d, v_x, v_y, s)$ topologies, and, finally, over all specific graphs $G \in \mathsf{G}_i(d, v_x, v_y, s)$ with a given topology. One should note that all limits of the three outer sums in (C.10) depend only on k, while $|\mathsf{G}_i(d, v_x, v_y, s)|$ is the only one that depends also on n. Similarly to the procedure described in Section 6.2, also here, the expectation of $\Theta(G)$ can be easily evaluated if all subgraphs of G are trees (i.e., G is a forest). If at least one subgraph of G contains loops, we apply a process of *graph pruning*, in which we cut out the minimal amount of edges[IX], while keeping all vertices intact, until we get a forest (for example, see Figure 13).


Y_5
Y_4
Y_3
Y_2
Y_1
X_1 X_2 X_3 X_4 X_5 X_6 X_7 X_8 X_9
Y_5
Y_4
Y_3
Y_2
Y_1
X_1 X_2 X_3 X_4 X_5 X_6 X_7 X_8 X_9

(a) The original graph

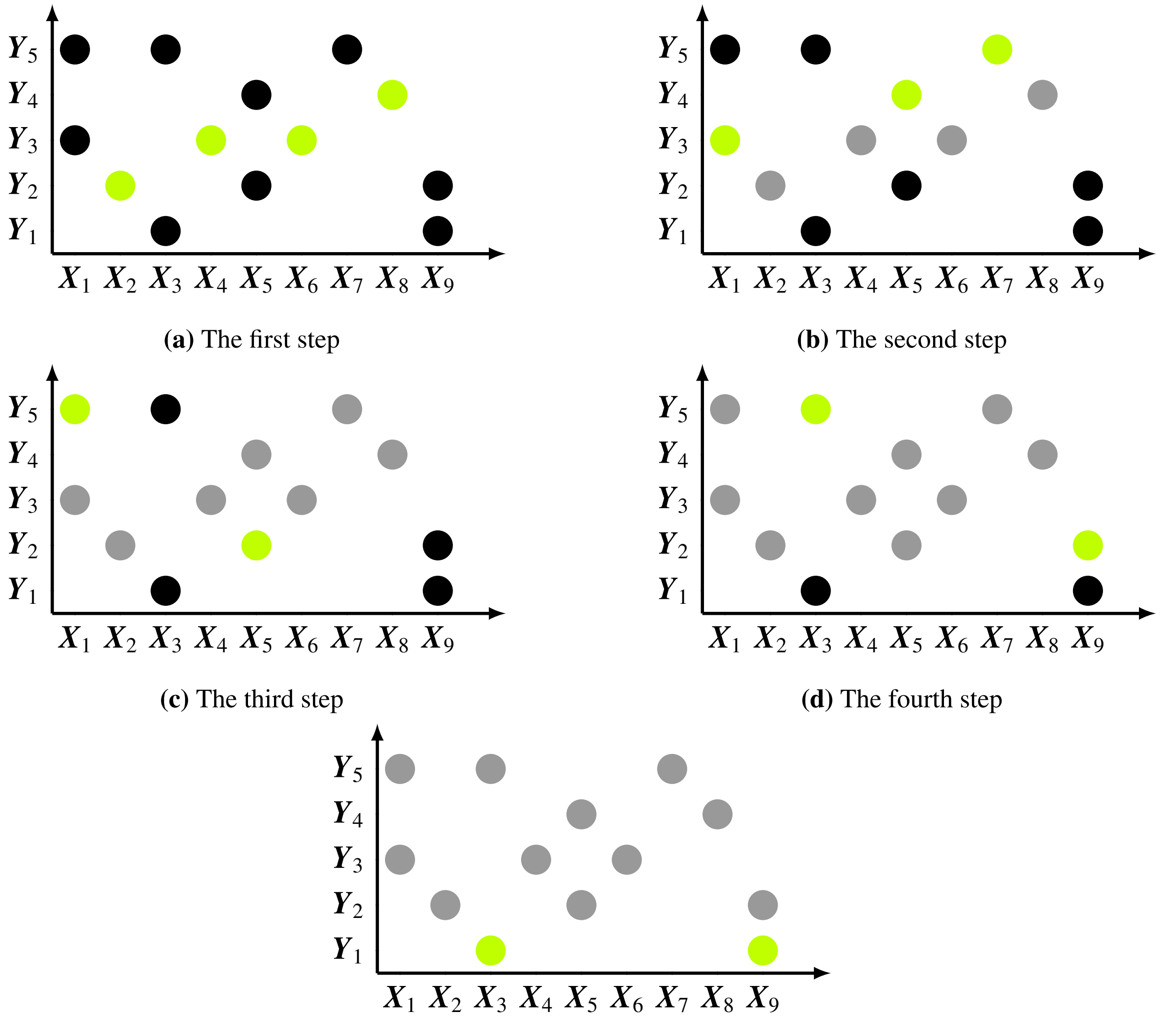
(b) The pruned graph

Figure 13. Example of the process of graph pruning.

We use $\mathcal{P}(G)$ to denote the pruned graph of G. Notice that the expectation of $\Theta(G)$ is upper-bounded[X] by the expectation of $\Theta(\mathcal{P}(G))$, which can be evaluated in a simple iterative process of *graph reduction*, similarly to the procedure described in Section 6.3. In the first step, we take the expectation with respect to all vectors that are the labels of leaf vertices in $\mathcal{P}(G)$, while we condition on the realizations of all other vectors, namely those corresponding to inner vertices in $\mathcal{P}(G)$; afterwards, we erase leaf vector vertices and corresponding edges. The successive steps are identical to the first one on the remaining (unerased) graph, continuing until all vectors that are attributed to G have been considered (for example, see Figure 14).

[IX] In fact, this procedure is equivalent to upper-bounding some of the indicator functions in (C.9) by one.

[X] It follows from the fact that $\Theta(G) \le \Theta(\mathcal{P}(G))$ with probability one.



(a) The first step

(b) The second step

(c) The third step

(d) The fourth step

(e) The last step

Figure 14. Example of the process of graph reduction.

In each step of graph reduction, the expectation with respect to each of the leaf vectors is given by $\mathbb{P}\{J(\boldsymbol{x}, \boldsymbol{Y}_{\text{leaf}}) = 1\} \leq Q$, since we condition on the realizations of vectors that are attributed to the inner vertices. For any graph $G \in \mathsf{G}_i(d, v_x, v_y, s)$, we conclude that the expectation of $\Theta(\mathcal{P}(G))$ is upper-bounded by $Q^{\mathcal{E}(G)}$, where $\mathcal{E}(G)$ is the number of edges in $\mathcal{P}(G)$. Of course, if G is already a forest, then $\mathcal{E}(G) = d$. For any $G \in \mathsf{G}_i(d, v_x, v_y, s)$ which is not a forest, we find $\mathcal{E}(G)$ as follows. Assume that $v_x(1), v_x(2), \ldots, v_x(s)$ and $v_y(1), v_y(2), \ldots, v_y(s)$ are the number of vertices in each of the s subgraphs of G. In the process of graph pruning, each of the $j \in \{1, 2, \ldots, s\}$ subgraphs of G will transform into a tree with exactly $v_x(j) + v_y(j) - 1$ edges. Hence

$$\mathcal{E}(G) = \sum_{j=1}^{s} (v_x(j) + v_y(j) - 1) = v_x + v_y - s. \tag{C.11}$$

The innermost sum of (C.10) can be treated as follows:

$$\sum_{G \in \mathsf{G}_i(d, v_x, v_y, s)} \mathbb{E}\left[\Theta(G)\right] \leq \sum_{G \in \mathsf{G}_i(d, v_x, v_y, s)} \mathbb{E}\left[\Theta(\mathcal{P}(G))\right] \tag{C.12}$$

$$\leq \sum_{G \in \mathsf{G}_i(d,v_x,v_y,s)} Q^{\mathcal{E}(G)} \tag{C.13}$$

$$= \sum_{G \in \mathsf{G}_i(d,v_x,v_y,s)} Q^{v_x+v_y-s} \tag{C.14}$$

$$= |\mathsf{G}_i(d,v_x,v_y,s)| \cdot Q^{v_x+v_y-s} \tag{C.15}$$

$$\leq n^{v_x} \cdot n^{v_y} \cdot Q^{v_x+v_y-s}, \tag{C.16}$$

where (C.14) follows from (C.11), and (C.16) is due to (C.8). Next, we substitute (C.16) back into (C.10) and get

$$\sum_{s=\mathcal{S}_{\min}(d,v_x,v_y)}^{\mathcal{S}_{\max}(d,v_x,v_y)} \sum_{i=1}^{\mathsf{T}(d,v_x,v_y,s)} n^{v_x} \cdot n^{v_y} \cdot Q^{v_x+v_y-s}$$

$$= \sum_{s=\mathcal{S}_{\min}(d,v_x,v_y)}^{\mathcal{S}_{\max}(d,v_x,v_y)} \mathsf{T}(d,v_x,v_y,s) \cdot n^{v_x} \cdot n^{v_y} \cdot Q^{v_x+v_y-s} \tag{C.17}$$

$$\leq \left(\sum_{s=\mathcal{S}_{\min}(d,v_x,v_y)}^{\mathcal{S}_{\max}(d,v_x,v_y)} \mathsf{T}(d,v_x,v_y,s) \right) \cdot n^{v_x} \cdot n^{v_y} \cdot Q^{v_x+v_y-\mathcal{S}_{\max}(d,v_x,v_y)} \tag{C.18}$$

$$= \mathsf{T}(d,v_x,v_y) \cdot n^{v_x} \cdot n^{v_y} \cdot Q^{v_x+v_y-\mathcal{S}_{\max}(d,v_x,v_y)}, \tag{C.19}$$

where (C.18) is true since $Q \in [0,1]$, such that replacing s by its maximal value $\mathcal{S}_{\max}(d, v_x, v_y)$ provides an upper bound. The passage (C.19) follows from the definitions of $\mathsf{T}(d, v_x, v_y)$ and $\mathsf{T}(d, v_x, v_y, s)$. Before we substitute it back into (C.10) and then into (C.5), we summarize some minor results that will be needed in the sequel. Let us define $d^* = \max\{v_x, v_y\}$.

Lemma 4. *We have the following:*

1. *For fixed (v_x, v_y), $\mathcal{S}_{max}(d, v_x, v_y)$ is a non-increasing sequence of d.*
2. *For any (v_x, v_y), we have $\mathcal{S}_{max}(d^*, v_x, v_y) = \min\{v_x, v_y\}$.*

Proof. We prove each of the two arguments separately.

1. For fixed (v_x, v_y), if we add to the graph an edge, we have only two options. On the one hand, we may connect vertices that belong to the same subgraph so the number of subgraphs remains the same. On the other hand, we can connect vertices that belong to different subgraphs, in which case, the number of subgraphs decreases.
2. Without loss of generality, assume that $v_x \leq v_y$. Then $d^* = v_y$, and hence it follows by definition that each column contains exactly one edge, such that the set of edges in each row provides a subgraph. Thus, the number of subgraphs equals exactly v_x.

■

Now we have

$$\mathbb{E}\left[N^k\right] \leq \sum_{d=1}^{k} \mathsf{S}(k,d) \sum_{(v_x,v_y)\in\mathcal{V}(d)} \mathsf{T}(d,v_x,v_y) \cdot n^{v_x} \cdot n^{v_y} \cdot Q^{v_x+v_y-\mathcal{S}_{\max}(d,v_x,v_y)} \tag{C.20}$$

$$= \sum_{d=1}^{k} \sum_{(v_x,v_y)\in\mathcal{V}(d)} \mathsf{S}(k,d)\cdot\mathsf{T}(d,v_x,v_y)\cdot n^{v_x}\cdot n^{v_y}\cdot Q^{v_x+v_y-\mathcal{S}_{\max}(d,v_x,v_y)} \tag{C.21}$$

$$= \sum_{v_x=1}^{k}\sum_{v_y=1}^{k}\sum_{d=\max\{v_x,v_y\}}^{\min\{k,v_x\cdot v_y\}} \mathsf{S}(k,d)\cdot\mathsf{T}(d,v_x,v_y)\cdot n^{v_x}\cdot n^{v_y}\cdot Q^{v_x+v_y-\mathcal{S}_{\max}(d,v_x,v_y)} \tag{C.22}$$

$$\leq \sum_{v_x=1}^{k}\sum_{v_y=1}^{k}\left(\sum_{d=\max\{v_x,v_y\}}^{\min\{k,v_x\cdot v_y\}} \mathsf{S}(k,d)\cdot\mathsf{T}(d,v_x,v_y)\right)\cdot n^{v_x}\cdot n^{v_y}\cdot Q^{v_x+v_y-\mathcal{S}_{\max}(d^*,v_x,v_y)}. \tag{C.23}$$

In (C.22) we changed the order of summation, and (C.23) is true since $Q \in [0,1]$, and thus, according to the first point in Lemma 4, we reach an upper bound by substituting $d = d^*$. Moving forward, we use the trivial bound

$$\mathsf{T}(d,v_x,v_y) \leq \binom{v_xv_y}{d} \leq \frac{(v_xv_y)^d}{d!}, \tag{C.24}$$

and arrive at

$$\mathbb{E}\left[N^k\right] \leq \sum_{v_x=1}^{k}\sum_{v_y=1}^{k}\left(\sum_{d=\max\{v_x,v_y\}}^{\min\{k,v_x\cdot v_y\}} \mathsf{S}(k,d)\cdot\frac{(v_xv_y)^d}{d!}\right)\cdot n^{v_x}\cdot n^{v_y}\cdot Q^{v_x+v_y-\min\{v_x,v_y\}} \tag{C.25}$$

$$\leq \sum_{v_x=1}^{k}\sum_{v_y=1}^{k}\left(\sum_{d=1}^{k} \mathsf{S}(k,d)\cdot\frac{(v_xv_y)^d}{d!}\right)\cdot n^{v_x}\cdot n^{v_y}\cdot Q^{\max\{v_x,v_y\}} \tag{C.26}$$

$$\leq \sum_{v_x=1}^{k}\sum_{v_y=1}^{k}\left(\sum_{d=1}^{k} \mathsf{S}(k,d)\cdot\frac{k^{2d}}{d!}\right)\cdot n^{v_x}\cdot n^{v_y}\cdot Q^{\max\{v_x,v_y\}} \tag{C.27}$$

$$= \sum_{v_x=1}^{k}\sum_{v_y=1}^{k} \mathsf{B}(k)\cdot n^{v_x}\cdot n^{v_y}\cdot Q^{\max\{v_x,v_y\}}, \tag{C.28}$$

where (C.25) follows from the second point in Lemma 4 and (C.28) follows from the definition in (5.17). Finally,

$$\mathbb{E}\left[N^k\right] \leq \sum_{v_x=1}^{k}\sum_{v_y=1}^{v_x} \mathsf{B}(k)\cdot n^{v_x}\cdot n^{v_y}\cdot Q^{\max\{v_x,v_y\}}$$

$$+ \sum_{v_y=1}^{k}\sum_{v_x=1}^{v_y} \mathsf{B}(k)\cdot n^{v_x}\cdot n^{v_y}\cdot Q^{\max\{v_x,v_y\}} \tag{C.29}$$

$$= \sum_{v_x=1}^{k}\sum_{v_y=1}^{v_x} \mathsf{B}(k)\cdot n^{v_x}\cdot n^{v_y}\cdot Q^{v_x}$$

$$+ \sum_{v_y=1}^{k}\sum_{v_x=1}^{v_y} \mathsf{B}(k)\cdot n^{v_x}\cdot n^{v_y}\cdot Q^{v_y} \tag{C.30}$$

$$\leq \sum_{v_x=1}^{k} v_x\cdot\mathsf{B}(k)\cdot(n^2Q)^{v_x} + \sum_{v_y=1}^{k} v_y\cdot\mathsf{B}(k)\cdot(n^2Q)^{v_y} \tag{C.31}$$

$$= 2\mathsf{B}(k)\sum_{\ell=1}^{k}\ell\cdot(n^2Q)^{\ell}. \tag{C.32}$$

Now, if $n^2Q > 1$, then

$$\mathbb{E}\left[N^k\right] \le \left(2\mathsf{B}(k)\sum_{\ell=1}^{k}\ell\right)\cdot(n^2Q)^k \tag{C.33}$$

$$\le k(k+1)\mathsf{B}(k)\cdot(n^2Q)^k; \tag{C.34}$$

otherwise, if $n^2Q \le 1$, then

$$\mathbb{E}\left[N^k\right] \le k(k+1)\mathsf{B}(k)\cdot n^2Q, \tag{C.35}$$

which completes the proof of Theorem 2.

Appendix D - Proof of Lemma 1

We upper-bound the expression in (5.20), which is given by

$$Q(d,\theta) = \frac{1}{2}\frac{B\left(1-\theta^2;\frac{d-1}{2},\frac{1}{2}\right)}{B\left(\frac{d-1}{2},\frac{1}{2}\right)}. \tag{D.1}$$

As for the numerator, we have

$$B\left(1-\theta^2;\frac{d-1}{2},\frac{1}{2}\right) = \int_0^{1-\theta^2}\frac{t^{\frac{d-3}{2}}}{\sqrt{1-t}}\mathrm{d}t \tag{D.2}$$

$$\le \int_0^{1-\theta^2}\frac{t^{\frac{d-3}{2}}}{\sqrt{1-(1-\theta^2)}}\mathrm{d}t \tag{D.3}$$

$$= \frac{1}{\theta}\int_0^{1-\theta^2} t^{\frac{d-3}{2}}\mathrm{d}t \tag{D.4}$$

$$= \frac{2}{\theta(d-1)}\cdot\left(1-\theta^2\right)^{\frac{d-1}{2}} \tag{D.5}$$

$$\le \frac{4}{\theta d}\cdot\left(1-\theta^2\right)^{\frac{d-1}{2}}. \tag{D.6}$$

In order to bound the beta function, we use the following identity:

$$B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}, \tag{D.7}$$

where $\Gamma(\cdot)$ is the gamma function:

$$\Gamma(s) = \int_0^{\infty} x^{s-1}e^{-x}\mathrm{d}x. \tag{D.8}$$

In general

$$B\left(\frac{d-1}{2},\frac{1}{2}\right)=\frac{\Gamma\left(\frac{d-1}{2}\right)\Gamma\left(\frac{1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} \tag{D.9}$$

$$=\frac{\sqrt{\pi}\Gamma\left(\frac{d-1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}. \tag{D.10}$$

In order to bound the ratio of gamma functions, let us invoke Gautschi's inequality.

Lemma 5. *Let x be a positive real number and let $s \in (0,1)$. In this case,*

$$x^{1-s} \leq \frac{\Gamma(x+1)}{\Gamma(x+s)} \leq (x+1)^{1-s}. \tag{D.11}$$

For a lower bound

$$B\left(\frac{d-1}{2},\frac{1}{2}\right)=\frac{\sqrt{\pi}\Gamma\left(\frac{d-1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} \tag{D.12}$$

$$=\frac{\sqrt{\pi}}{\frac{\Gamma\left(\frac{d}{2}-1+1\right)}{\Gamma\left(\frac{d}{2}-1+\frac{1}{2}\right)}} \tag{D.13}$$

$$\geq \frac{\sqrt{\pi}}{(\frac{d}{2}-1+1)^{1-\frac{1}{2}}} \tag{D.14}$$

$$=\sqrt{\frac{2\pi}{d}}, \tag{D.15}$$

where (D.14) is due to the upper bound in Lemma 5. Now, combining (D.6) and (D.15) yields the upper bound

$$Q(d,\theta) \leq \frac{1}{2}\frac{\frac{4}{\theta d}\cdot\left(1-\theta^2\right)^{\frac{d-1}{2}}}{\sqrt{\frac{2\pi}{d}}} \tag{D.16}$$

$$=\sqrt{\frac{2}{\pi}}\frac{1}{\theta\sqrt{d}}\cdot\left(1-\theta^2\right)^{\frac{d-1}{2}} \tag{D.17}$$

$$\leq \frac{1}{\theta\sqrt{d}}\cdot\left(1-\theta^2\right)^{\frac{d-1}{2}}, \tag{D.18}$$

which completes the proof of Lemma 1.

Appendix E - Proof of Lemma 2

Observe that under the assumption of

$$(\boldsymbol{X},\boldsymbol{Y}) \sim \mathcal{N}^{\otimes d}\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}\right), \tag{E.1}$$

it holds that $\boldsymbol{Y} = \rho\boldsymbol{X} + \sqrt{1-\rho^2}\boldsymbol{Z}$ with $\boldsymbol{Z} \sim \mathcal{N}(\mathbf{0}_d, \boldsymbol{I}_d)$ and independent of $\boldsymbol{X}$. Consider the following for some $\theta \in (0,1)$:

$$\mathbb{P}\left\{\tilde{\boldsymbol{X}}^{\mathsf{T}}\tilde{\boldsymbol{Y}} \geq \theta\right\} = \mathbb{P}\left\{\frac{\boldsymbol{X}^{\mathsf{T}}\boldsymbol{Y}}{\|\boldsymbol{X}\|\cdot\|\boldsymbol{Y}\|} \geq \theta\right\} \tag{E.2}$$

$$= \mathbb{P}\left\{\frac{\boldsymbol{X}^{\mathsf{T}}\left(\rho\boldsymbol{X} + \sqrt{1-\rho^2}\boldsymbol{Z}\right)}{\|\boldsymbol{X}\|\left\|\rho\boldsymbol{X} + \sqrt{1-\rho^2}\boldsymbol{Z}\right\|} \geq \theta\right\} \tag{E.3}$$

$$= \mathbb{P}\left\{\frac{\rho\|\boldsymbol{X}\|^2 + \sqrt{1-\rho^2}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{Z}}{\|\boldsymbol{X}\|\sqrt{\rho^2\|\boldsymbol{X}\|^2 + 2\rho\sqrt{1-\rho^2}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{Z} + (1-\rho^2)\|\boldsymbol{Z}\|^2}} \geq \theta\right\}. \tag{E.4}$$

Since $\theta > 0$, the event in (E.4) can occur only if $\rho\|\boldsymbol{X}\|^2 + \sqrt{1-\rho^2}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{Z} \geq 0$. Let us denote the event

$$\mathcal{F}_\rho(\boldsymbol{X},\boldsymbol{Z}) = \left\{\rho\|\boldsymbol{X}\|^2 + \sqrt{1-\rho^2}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{Z} \geq 0\right\} \tag{E.5}$$

and the angle between $\boldsymbol{X}$ and $\boldsymbol{Z}$ as Φ_{XZ}. In this case,

$$\mathbb{P}\left\{\tilde{\boldsymbol{X}}^{\mathsf{T}}\tilde{\boldsymbol{Y}} \geq \theta\right\}$$

$$= \mathbb{P}\left\{\frac{\rho\|\boldsymbol{X}\|^2 + \sqrt{1-\rho^2}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{Z}}{\|\boldsymbol{X}\|\sqrt{\rho^2\|\boldsymbol{X}\|^2 + 2\rho\sqrt{1-\rho^2}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{Z} + (1-\rho^2)\|\boldsymbol{Z}\|^2}} \geq \theta, \mathcal{F}_\rho(\boldsymbol{X},\boldsymbol{Z})\right\} \tag{E.6}$$

$$= \mathbb{P}\left\{\frac{\rho^2\|\boldsymbol{X}\|^4 + 2\rho\sqrt{1-\rho^2}\|\boldsymbol{X}\|^2\boldsymbol{X}^{\mathsf{T}}\boldsymbol{Z} + (1-\rho^2)\left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{Z}\right)^2}{\rho^2\|\boldsymbol{X}\|^4 + 2\rho\sqrt{1-\rho^2}\|\boldsymbol{X}\|^2\boldsymbol{X}^{\mathsf{T}}\boldsymbol{Z} + (1-\rho^2)\|\boldsymbol{X}\|^2\|\boldsymbol{Z}\|^2} \geq \theta^2, \mathcal{F}_\rho(\boldsymbol{X},\boldsymbol{Z})\right\} \tag{E.7}$$

$$= \mathbb{P}\left\{\frac{\rho^2\|\boldsymbol{X}\|^4 + 2\rho\sqrt{1-\rho^2}\|\boldsymbol{X}\|^3\|\boldsymbol{Z}\|\cdot\cos(\Phi_{\mathrm{XZ}}) + (1-\rho^2)\|\boldsymbol{X}\|^2\|\boldsymbol{Z}\|^2\cdot\cos^2(\Phi_{\mathrm{XZ}})}{\rho^2\|\boldsymbol{X}\|^4 + 2\rho\sqrt{1-\rho^2}\|\boldsymbol{X}\|^3\|\boldsymbol{Z}\|\cdot\cos(\Phi_{\mathrm{XZ}}) + (1-\rho^2)\|\boldsymbol{X}\|^2\|\boldsymbol{Z}\|^2} \geq \theta^2, \mathcal{F}_\rho(\boldsymbol{X},\boldsymbol{Z})\right\} \tag{E.8}$$

$$= \mathbb{P}\left\{\frac{\rho^2\|\boldsymbol{X}\|^2 + 2\rho\sqrt{1-\rho^2}\|\boldsymbol{X}\|\|\boldsymbol{Z}\|\cdot\cos(\Phi_{\mathrm{XZ}}) + (1-\rho^2)\|\boldsymbol{Z}\|^2\cdot\cos^2(\Phi_{\mathrm{XZ}})}{\rho^2\|\boldsymbol{X}\|^2 + 2\rho\sqrt{1-\rho^2}\|\boldsymbol{X}\|\|\boldsymbol{Z}\|\cdot\cos(\Phi_{\mathrm{XZ}}) + (1-\rho^2)\|\boldsymbol{Z}\|^2} \geq \theta^2, \mathcal{F}_\rho(\boldsymbol{X},\boldsymbol{Z})\right\}. \tag{E.9}$$

Rearranging, we get

$$\begin{aligned}
&\mathbb{P}\left\{\tilde{\boldsymbol{X}}^{\mathsf{T}}\tilde{\boldsymbol{Y}} \geq \theta\right\} \\
&= \mathbb{P}\Big\{(1-\rho^2)\|\boldsymbol{Z}\|^2\cdot\cos^2(\Phi_{\mathrm{XZ}}) + 2\rho\sqrt{1-\rho^2}(1-\theta^2)\|\boldsymbol{X}\|\|\boldsymbol{Z}\|\cdot\cos(\Phi_{\mathrm{XZ}}) \\
&\qquad + \rho^2(1-\theta^2)\|\boldsymbol{X}\|^2 - (1-\rho^2)\theta^2\|\boldsymbol{Z}\|^2 \geq 0, \mathcal{F}_\rho(\boldsymbol{X},\boldsymbol{Z})\Big\} \qquad \text{(E.10)} \\
&= \mathbb{P}\Big\{(1-\rho^2)\cdot\cos^2(\Phi_{\mathrm{XZ}}) + 2\rho\sqrt{1-\rho^2}(1-\theta^2)\frac{\|\boldsymbol{X}\|}{\|\boldsymbol{Z}\|}\cdot\cos(\Phi_{\mathrm{XZ}}) \\
&\qquad + \rho^2(1-\theta^2)\frac{\|\boldsymbol{X}\|^2}{\|\boldsymbol{Z}\|^2} - (1-\rho^2)\theta^2 \geq 0, \mathcal{F}_\rho(\boldsymbol{X},\boldsymbol{Z})\Big\} \qquad \text{(E.11)} \\
&\triangleq \mathbb{P}\left\{\cos^2(\Phi_{\mathrm{XZ}}) + \frac{2\rho(1-\theta^2)}{\sqrt{1-\rho^2}}\sqrt{U}\cdot\cos(\Phi_{\mathrm{XZ}}) + \frac{\rho^2(1-\theta^2)}{1-\rho^2}U - \theta^2 \geq 0, \mathcal{F}_\rho(\boldsymbol{X},\boldsymbol{Z})\right\} \qquad \text{(E.12)}
\end{aligned}$$

$$= \mathbb{P}\left\{\left(\cos(\Phi_{\mathrm{XZ}}) + \frac{\rho(1-\theta^2)}{\sqrt{1-\rho^2}}\sqrt{U}\right)^2 + \frac{\rho^2\theta^2(1-\theta^2)}{1-\rho^2}U - \theta^2 \geq 0, \mathcal{F}_\rho(\boldsymbol{X},\boldsymbol{Z})\right\} \tag{E.13}$$

$$= \mathbb{P}\left\{\left(\cos(\Phi_{\mathrm{XZ}}) + \frac{\rho(1-\theta^2)}{\sqrt{1-\rho^2}}\sqrt{U}\right)^2 \geq \theta^2\left[1 - \frac{\rho^2(1-\theta^2)}{1-\rho^2}U\right], \mathcal{F}_\rho(\boldsymbol{X},\boldsymbol{Z})\right\}. \tag{E.14}$$

Note that $\rho\|\boldsymbol{X}\|^2 + \sqrt{1-\rho^2}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{Z} \geq 0$ is equivalent to

$$\cos(\Phi_{\mathrm{XZ}}) \geq -\frac{\rho}{\sqrt{1-\rho^2}}\sqrt{U}, \tag{E.15}$$

and therefore

$$\begin{aligned}&\mathbb{P}\left\{\tilde{\boldsymbol{X}}^{\mathsf{T}}\tilde{\boldsymbol{Y}} \geq \theta\right\}\\ &= \mathbb{P}\left\{\left(\cos(\Phi_{\mathrm{XZ}}) + \frac{\rho(1-\theta^2)}{\sqrt{1-\rho^2}}\sqrt{U}\right)^2 \geq \theta^2\left[1 - \frac{\rho^2(1-\theta^2)}{1-\rho^2}U\right], \cos(\Phi_{\mathrm{XZ}}) \geq -\frac{\rho}{\sqrt{1-\rho^2}}\sqrt{U}\right\}.\end{aligned} \tag{E.16}$$

Now, the left-hand side event in (E.16) becomes trivial if and only if

$$1 - \frac{\rho^2(1-\theta^2)}{1-\rho^2}U \leq 0 \iff U \geq \frac{1-\rho^2}{\rho^2(1-\theta^2)} \triangleq \beta(\rho,\theta), \tag{E.17}$$

while the right-hand side event in (E.16) becomes trivial if and only if

$$\frac{\rho}{\sqrt{1-\rho^2}}\sqrt{U} \geq 1 \iff U \geq \frac{1-\rho^2}{\rho^2} \triangleq \alpha(\rho). \tag{E.18}$$

In addition, note that for any $\theta \in (0,1)$, it holds that $\beta(\rho,\theta) \geq \alpha(\rho)$. Define the functions

$$F_1(u) \equiv F_1(u,\rho,\theta) = -\frac{\rho(1-\theta^2)}{\sqrt{1-\rho^2}}\sqrt{u} - \theta\sqrt{1 - \frac{\rho^2(1-\theta^2)}{1-\rho^2}u}, \tag{E.19}$$

$$F_2(u) \equiv F_2(u,\rho,\theta) = -\frac{\rho(1-\theta^2)}{\sqrt{1-\rho^2}}\sqrt{u} + \theta\sqrt{1 - \frac{\rho^2(1-\theta^2)}{1-\rho^2}u}, \tag{E.20}$$

$$F_3(u) \equiv F_3(u,\rho) = -\frac{\rho}{\sqrt{1-\rho^2}}\sqrt{u}, \tag{E.21}$$

such that

$$\mathbb{P}\left\{\tilde{\boldsymbol{X}}^{\mathsf{T}}\tilde{\boldsymbol{Y}} \geq \theta\right\} = \mathbb{P}\left\{\cos(\Phi_{\mathrm{XZ}}) \in [-1, F_1(U,\rho,\theta)] \cup [F_2(U,\rho,\theta), 1], \cos(\Phi_{\mathrm{XZ}}) \in [F_3(U,\rho), 1]\right\}. \tag{E.22}$$

Note that both $\|\boldsymbol{X}\|^2$ and $\|\boldsymbol{Z}\|^2$ follow a χ^2 distribution with d degrees of freedom. Hence, the random variable defined by the ratio

$$U = \frac{\|\boldsymbol{X}\|^2}{\|\boldsymbol{Z}\|^2} \tag{E.23}$$

follows the F-distribution with (d, d) degrees of freedom. The probability density function of a general F-distributed random variable with (d_1, d_2) degrees of freedom is given by

$$f_U(u; d_1, d_2) = \frac{1}{B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} \left(\frac{d_1}{d_2}\right)^{d_1/2} u^{d_1/2-1} \left(1 + \frac{d_1}{d_2} u\right)^{-(d_1+d_2)/2}, \quad u \geq 0, \tag{E.24}$$

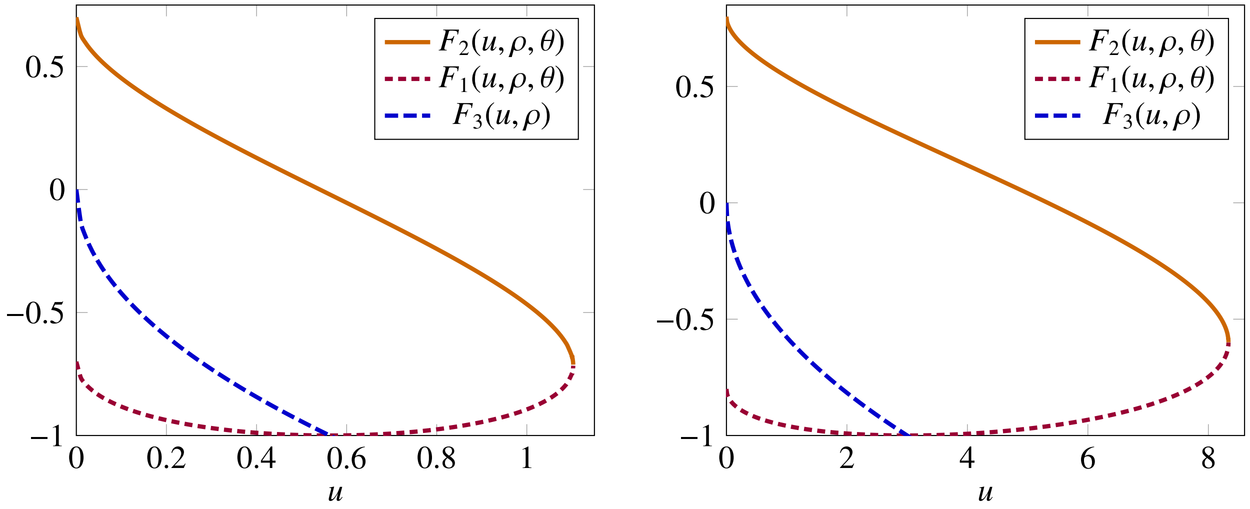
and for $d_1 = d_2 = d$, we denote it as

$$f_U(u) = \frac{1}{B\left(\frac{d}{2}, \frac{d}{2}\right)} u^{d/2-1} (1+u)^{-d}, \quad u \geq 0. \tag{E.25}$$

We now have

$$\begin{aligned}
&\mathbb{P}\left\{\tilde{\boldsymbol{X}}^{\mathsf{T}} \tilde{\boldsymbol{Y}} \geq \theta\right\} \\
&= \int_0^{\infty} \mathbb{P}\left\{\cos(\Phi_{\mathrm{XZ}}) \in [-1, F_1(u)] \cup [F_2(u), 1], \cos(\Phi_{\mathrm{XZ}}) \in [F_3(u), 1]\right\} f_U(u) \mathrm{d}u && \text{(E.26)} \\
&= \int_0^{\alpha(\rho)} \mathbb{P}\left\{\cos(\Phi_{\mathrm{XZ}}) \in [-1, F_1(u)] \cup [F_2(u), 1], \cos(\Phi_{\mathrm{XZ}}) \in [F_3(u), 1]\right\} f_U(u) \mathrm{d}u \\
&\quad + \int_{\alpha(\rho)}^{\beta(\rho,\theta)} \mathbb{P}\left\{\cos(\Phi_{\mathrm{XZ}}) \in [-1, F_1(u)] \cup [F_2(u), 1], \cos(\Phi_{\mathrm{XZ}}) \in [F_3(u), 1]\right\} f_U(u) \mathrm{d}u \\
&\quad + \int_{\beta(\rho,\theta)}^{\infty} \mathbb{P}\left\{\cos(\Phi_{\mathrm{XZ}}) \in [-1, F_1(u)] \cup [F_2(u), 1], \cos(\Phi_{\mathrm{XZ}}) \in [F_3(u), 1]\right\} f_U(u) \mathrm{d}u && \text{(E.27)} \\
&= \int_0^{\alpha(\rho)} \mathbb{P}\left\{\cos(\Phi_{\mathrm{XZ}}) \in [-1, F_1(u)] \cup [F_2(u), 1], \cos(\Phi_{\mathrm{XZ}}) \in [F_3(u), 1]\right\} f_U(u) \mathrm{d}u \\
&\quad + \int_{\alpha(\rho)}^{\beta(\rho,\theta)} \mathbb{P}\left\{\cos(\Phi_{\mathrm{XZ}}) \in [-1, F_1(u)] \cup [F_2(u), 1]\right\} f_U(u) \mathrm{d}u \\
&\quad + \int_{\beta(\rho,\theta)}^{\infty} f_U(u) \mathrm{d}u, && \text{(E.28)}
\end{aligned}$$

where (E.28) follows from the following considerations. For the second integral, we use the fact that for any $u \geq \alpha(\rho)$, the second event is trivial, and for the last integral, we use the fact that for any $u \geq \beta(\rho, \theta)$, both events become trivial. In order to facilitate the first integral in (E.28), we prove that for any $u \in [0, \alpha(\rho)]$, it holds that $F_1(u) \leq F_3(u) \leq F_2(u)$ (as can be seen in Figure 15 for two specific choices of the pair (ρ, θ)).



(a) $\rho = 0.8$ and $\theta = 0.7$

(b) $\rho = 0.5$ and $\theta = 0.8$

Figure 15. Plots of $F_1(u,\rho,\theta)$, $F_2(u,\rho,\theta)$, and $F_3(u,\rho)$.

It immediately follows from the definitions of $F_1(u)$ and $F_2(u)$ in (E.19) and (E.20) that for any $u \in [0,\beta(\rho,\theta)]$, $F_1(u) \le F_2(u)$. We start by proving that $F_1(u) \le F_3(u)$. Note that $F_3(u)$ is monotonically decreasing and $F_3(0) = 0$. In addition, $F_1(0) = -\theta < F_3(0)$. Furthermore, note that $F_1(\alpha(\rho)) = F_3(\alpha(\rho)) = -1$, so it only remains to show that $F_1(u)$ is also monotonically decreasing for any $u \in [0,\alpha(\rho)]$. To this end, we show that $F_1'(u) = 0$ has a unique solution at $u = \alpha(\rho)$. The derivative $F_1'(u)$ is given by

$$F_1'(u) = -\frac{\rho(1-\theta^2)}{\sqrt{1-\rho^2}} \cdot \frac{1}{2\sqrt{u}} + \theta\frac{\frac{\rho^2(1-\theta^2)}{1-\rho^2}}{2\sqrt{1-\frac{\rho^2(1-\theta^2)}{1-\rho^2}u}}. \tag{E.29}$$

$F_1'(u) = 0$ is then equivalent to

$$\sqrt{1-\frac{\rho^2(1-\theta^2)}{1-\rho^2}u} = \frac{\rho\theta}{\sqrt{1-\rho^2}} \cdot \sqrt{u}, \tag{E.30}$$

or, in turn, to the linear equation

$$1-\frac{\rho^2(1-\theta^2)}{1-\rho^2}u = \frac{\rho^2\theta^2}{1-\rho^2}u, \tag{E.31}$$

whose unique solution is given by

$$u^* = \frac{1-\rho^2}{\rho^2} = \alpha(\rho). \tag{E.32}$$

Since $F_1(0) = -\theta > -1 = F_1(\alpha(\rho))$, this proves that $F_1(u)$ is monotonically decreasing for any $u \in [0,\alpha(\rho)]$. This completes the proof of $F_1(u) \le F_3(u)$.

We next prove that $F_3(u) \leq F_2(u)$. The derivative of $F_2(u)$ is given by

$$F_2'(u) = -\frac{\rho(1-\theta^2)}{\sqrt{1-\rho^2}} \cdot \frac{1}{2\sqrt{u}} - \theta \frac{\frac{\rho^2(1-\theta^2)}{1-\rho^2}}{2\sqrt{1-\frac{\rho^2(1-\theta^2)}{1-\rho^2}u}}, \tag{E.33}$$

which is negative for any $u \in [0, \beta(\rho, \theta)]$, and thus $F_2(u)$ is a monotonically decreasing function on the same interval. Note that $F_2(0) = \theta > 0 = F_3(0)$ and, in addition

$$F_2(\beta(\rho,\theta)) = -\frac{\rho(1-\theta^2)}{\sqrt{1-\rho^2}}\sqrt{\frac{1-\rho^2}{\rho^2(1-\theta^2)}} + \theta\sqrt{1-\frac{\rho^2(1-\theta^2)}{1-\rho^2}\cdot\frac{1-\rho^2}{\rho^2(1-\theta^2)}} \tag{E.34}$$

$$= -\sqrt{1-\theta^2} \tag{E.35}$$

$$\geq -1. \tag{E.36}$$

Since $F_2(u)$ is monotonically decreasing and$\beta(\rho, \theta) \geq \alpha(\rho)$, we get

$$F_2(\alpha(\rho)) \geq F_2(\beta(\rho,\theta)) \geq -1 = F_3(\alpha(\rho)), \tag{E.37}$$

which completes the proof of $F_3(u) \leq F_2(u)$, since $F_3(u)$ is monotonically decreasing for any $u \geq 0$.

Using the double inequality $F_1(u) \leq F_3(u) \leq F_2(u)$, which holds for any $u \in [0, \alpha(\rho)]$, we arrive at

$$\begin{aligned}\mathbb{P}\left\{\tilde{\boldsymbol{X}}^{\mathsf{T}}\tilde{\boldsymbol{Y}} \geq \theta\right\} = &\int_0^{\alpha(\rho)} \mathbb{P}\left\{\cos(\Phi_{\mathrm{xz}}) \in [F_2(u), 1]\right\} f_U(u)\mathrm{d}u \\ &+ \int_{\alpha(\rho)}^{\beta(\rho,\theta)} \mathbb{P}\left\{\cos(\Phi_{\mathrm{xz}}) \in [-1, F_1(u)] \cup [F_2(u), 1]\right\} f_U(u)\mathrm{d}u \\ &+ \int_{\beta(\rho,\theta)}^{\infty} f_U(u)\mathrm{d}u.\end{aligned} \tag{E.38}$$

Regarding the probabilities inside the first two integrals in (E.38), consider the following result.

Lemma 6. *Let $d \in \mathbb{N}$ and $\boldsymbol{X}, \boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{0}_d, \boldsymbol{I}_d)$ be two independent random vectors. Let $\tilde{\boldsymbol{X}} = \frac{\boldsymbol{X}}{\|\boldsymbol{X}\|}$, $\tilde{\boldsymbol{Y}} = \frac{\boldsymbol{Y}}{\|\boldsymbol{Y}\|}$, and define $S = \cos(\Phi) = \tilde{\boldsymbol{X}}^{\mathsf{T}}\tilde{\boldsymbol{Y}}$. The probability density function of S is then given by*

$$g_S(s) = \frac{(1-s^2)^{\frac{d-3}{2}}}{B\left(\frac{d-1}{2}, \frac{1}{2}\right)}, \quad s \in [-1, 1], \tag{E.39}$$

where the complete beta function is

$$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}\mathrm{d}t, \quad a, b > 0. \tag{E.40}$$

Finally,

$$\mathbb{P}\left\{\tilde{\boldsymbol{X}}^{\mathsf{T}}\tilde{\boldsymbol{Y}} \geq \theta\right\} = \int_0^{\alpha(\rho)} \left[\int_{F_2(u)}^1 g_S(s)\mathrm{d}s\right] f_U(u)\mathrm{d}u$$

$$+\int_{\alpha(\rho)}^{\beta(\rho,\theta)}\left[\int_{-1}^{F_1(u)} g_S(s)\mathrm{d}s+\int_{F_2(u)}^{1} g_S(s)\mathrm{d}s\right] f_U(u)\mathrm{d}u$$
$$+\int_{\beta(\rho,\theta)}^{\infty} f_U(u)\mathrm{d}u, \tag{E.41}$$

which completes the proof of Lemma 2.