Detecting Correlated Gaussian Databases

Zeynep K
Boston University
zeynepk@bu.edu

Bobak Nazer
Boston University
bobak@bu.edu

***Abstract*—This paper considers the problem of detecting whether two databases, each consisting of n users with d Gaussian features, are correlated. Under the null hypothesis, the databases are independent. Under the alternate hypothesis, the features are correlated across databases, under an unknown row permutation. A simple test is developed to show that detection is achievable above $\rho^2 \approx \frac{1}{d}$. For the converse, the truncated second moment method is used to establish that detection is impossible below roughly $\rho^2 \approx \frac{1}{d\sqrt{n}}$. These results are compared to the corresponding recovery problem, where the goal is to decode the row permutation, and a converse bound of roughly $\rho^2 \approx 1-n^{-4/d}$ has been previously shown. For certain choices of parameters, the detection achievability bound outperforms this recovery converse bound, demonstrating that detection can be easier than recovery in this scenario.**

I. Introduction

Consider the following naïve approach to database anonymization in order to protect the privacy of individuals: prior to public release, unique identifying information (e.g., names, user IDs) is deleted while other features are left unchanged. It is now well known that this approach is vulnerable to de-anonymization attacks, given another correlated database with identifying information intact, see examples [1]–[4].

The *database alignment* problem models the scenario above via two correlated databases. Specifically, the entries in the first database can be matched to the entries in the second data via a random permutation. For a pair of matched entries, the features are correlated (and are otherwise independent). The goal is to recover this unknown permutation using only the two correlated databases. Cullina et al. [5] proposed an information-theoretic model for the recovery problem, and derived achievability and converse bounds for the discrete memoryless version. Subsequent work extended this analysis to the case of Gaussian features [6]. More general distributions are studied using typicality in [7] and partial recovery for correlated Gaussian databases is studied in [8]. Other generalizations such as databases with random feature deletions [9] and repetitions [10] have also been recently studied.

The Gaussian database alignment problem is also equivalent to a certain idealized tracking problem studied by [11], [12], inspired by the application of particle tracking to infer the trajectories of objects from sequences of still images. Note that in this setting, the pairings are coupled whereas reconstruction of the planted matching in random bipartite graphs proposed by [13], [14] deals with independent random pairings.

The authors were supported by NSF Grant CCF-1955981.

We also note similarities between database alignment and the problem of graph alignment. In this setting, the edges of two random graphs are correlated and the node labels are shuffled. The recovery problem is to find the true node labeling, which is equivalent to a certain quadratic assignment optimization problem. Since randomness is associated with the edges that are simply pairs of nodes, neighbors create side information about a user during relabelling. Initial work [15] on this problem in the information theory literature proposed the correlated Erdös–Rényi graph model with dependent Bernoulli edge pairs. Subsequent work considered the recovery problem for the Gaussian setting [16], specifically correlated Gaussian Wigner matrices. Other recent papers have considered the corresponding detection problem for graph correlation [17], [18] and it is now known [19], [20] that detecting whether Gaussian graphs are correlated is as difficult as recovering the node labeling.

This paper considers the problem of detecting whether a pair of Gaussian databases is correlated. This problem can be phrased a binary hypothesis test where, under the alternate hypothesis, the databases are correlated as in the recovery problem, whereas under the null, the databases are independent independent. Similar to the previous work, the challenge is that the user identities are unknown and shuffled. Upper and lower bounds on the correlation needed for reliable detection are derived. Due to space limitations, full proofs will appear in the arXiv version.

II. Problem Statement

Notation. Random column vectors are denoted by capital letters such as X with corresponding transpose X^{T}. A collection of n random vectors is written as $X^n = (X_1, \ldots, X_n)$. The $n \times n$ identity matrix is denoted by $\mathbf{I}_n$, the $n \times n$ all-ones matrix by $\mathbf{1}_n$, and the length-n all-zeros column vector by 0_n. Let $[n] \triangleq \{1, \ldots, n\}$ and define S_n as the set of all permutations over $[n]$. For a given permutation $\sigma \in S_n$, let σ_i denote the value to which σ maps $i \in [n]$. The notation $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P_X$ means that the random vectors $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d.) according to P_X, or, equivalently, $X_1, \ldots, X_n \sim P_X^{\otimes n}$. $\mathbb{E}_\alpha X$ represents the expectation of X where $X \sim \mathbb{P}_\alpha$. We use $\mathcal{N}(\mu, \Sigma)$ to represent the multivariate normal distribution with mean vector μ and covariance matrix Σ. For two probability measures P and Q, with a slight abuse of notation, we write $\frac{P}{Q}$ to represent the Radon-Nikodym derivative $\frac{\mathrm{d}P}{\mathrm{d}Q}$. For two real-valued functions $f(x)$ and $g(x)$, the standard asymptotic

notations $f(x) = \mathcal{O}(g(x))$ and $g(x) = \Omega(f(x))$ imply that there exist C and x_0 such that $|f(x)| \leq C\,|g(x)|$ for all $x \geq x_0$; $f(x) = o(g(x))$ and $g(x) = \omega(f(x))$ mean that $f(x)/g(x) \to 0$ as $x \to \infty$.

A. Model

In our considerations, a *database* is a collection $X^n = (X_1, \ldots, X_n)$ of n i.i.d. random vectors in $\mathbb{R}^d$ where n is the number of users (or entries) and d is the number of features. For $i \in [n]$ and $k \in [d]$, let $X_i^{(k)}$ denote the k^{th} feature of the i^{th} user.

We say that a pair of databases $X^n = (X_1, \ldots, X_n)$ and $Y^n = (Y_1, \ldots, Y_n)$ is *correlated with permutation* σ if, for some $\sigma \in S_n$, we have that $(X_1, Y_{\sigma_1}), \ldots, (X_n, Y_{\sigma_n}) \overset{\text{iid}}{\sim} P_{XY}$ for some joint distribution P_{XY} over $\mathbb{R}^d \times \mathbb{R}^d$. We will focus on the special case of *Gaussian correlated databases*. Define

$$\mathbf{\Sigma}_\rho \triangleq \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

for some correlation coefficient $-1 \leq \rho \leq 1$. Then, $P_{XY} = \mathcal{N}^{\otimes d}(0_2, \mathbf{\Sigma}_\rho)$, i.e., the features of a matched pair of entries are distributed as $(X_i^{(1)}, Y_{\sigma_i}^{(1)}), \ldots, (X_i^{(d)}, Y_{\sigma_i}^{(d)}) \overset{\text{iid}}{\sim} \mathcal{N}(0_2, \mathbf{\Sigma}_\rho)$.

B. Detection

Consider the following binary hypothesis testing problem. Under the *null hypothesis* H_0, the databases X^n and Y^n are generated independently with $X_1, \ldots, X_n, Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(0_d, \mathbf{I}_d)$. (This is equivalent to generating the databases as above with the covariance matrix $\mathbf{\Sigma}_0$.) Let $\mathbb{P}_0$ denote the resulting distribution over X^n, Y^n. Under the *alternate hypothesis* H_1, the databases X^n and Y^n are correlated with permutation σ for some (unknown) $\sigma \in S_n$ and some correlation coefficient $\rho \neq 0$. Let $\mathbb{P}_{1|\sigma}$ denote the resulting distribution over X^n, Y^n. Summarizing,

$$\begin{aligned} H_0 &: \ (X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{iid}}{\sim} \mathcal{N}^{\otimes d}(0_2, \mathbf{I}_2) \\ H_1 &: \ (X_1, Y_{\sigma_1}), \ldots, (X_n, Y_{\sigma_n}) \overset{\text{iid}}{\sim} \mathcal{N}^{\otimes d}(0_2, \mathbf{\Sigma}_\rho) \end{aligned} \tag{1}$$

for some $\sigma \in S_n$.

Given only the databases X^n and Y^n (and not the permutation σ), a *test* $\phi : \mathbb{R}^{d\times n} \times \mathbb{R}^{d\times n} \to \{0, 1\}$ guesses whether the null, $\phi(X^n, Y^n) = 0$, or alternate, $\phi(X^n, Y^n) = 1$ hypothesis occurred. For a given n and d, the *risk* of a test ϕ is

$$R(\phi) \triangleq \mathbb{P}_0\big\{\phi(X^n, Y^n) = 1\big\} + \max_{\sigma \in S_n} \mathbb{P}_{1|\sigma}\big\{\phi(X^n, Y^n) = 0\big\}$$

and the *minimax risk* is $R^* \triangleq \inf_\phi R(\phi)$ where the infimum is over all (measurable) decision rules. Note that this is the smallest possible sum of false alarm (FA) and missed detection (MD) probabilities.

C. Recovery

Consider the following recovery problem. Gaussian correlated databases X^n and Y^n are generated according to

$$(X_1, Y_{\sigma_1}), \ldots, (X_n, Y_{\sigma_n}) \overset{\text{iid}}{\sim} \mathcal{N}^{\otimes d}(0_2, \mathbf{\Sigma}_\rho) \quad \text{for some } \sigma \in S_n,$$

and the resulting distribution is denoted as $\mathbb{P}_{1|\sigma}$ to be consistent with our detection notation. The goal is to decode the permutation $\sigma \in S_n$ given X^n and Y^n. Specifically, the *decoder* is a function $\hat{\sigma} : \mathbb{R}^{d\times n} \times \mathbb{R}^{d\times n} \to S_n$. We define the probability of error

$$P_{\text{err}}(\hat{\sigma}) \triangleq \max_{\sigma \in S_n} \mathbb{P}_{1|\sigma}\big\{\hat{\sigma}(X^n, Y^n) \neq \sigma\big\}$$

and the *minimax probability of error* $P_{\text{err}}^* \triangleq \inf_{\hat{\sigma}} P_{\text{err}}(\hat{\sigma})$ where the infimum is over all (measurable) decoders.[1]

III. Main Results

Our achievability argument is based on a sum-of-inner-products statistic $T \triangleq \text{sign}(\rho) \sum_{i,j} X_i^\mathsf{T} Y_j$ with corresponding test $\phi_{T,t}$ that compares the resulting T value to a threshold t

$$\phi_{T,t}(X^n, Y^n) = \begin{cases} 0, & \text{sign}(\rho) \displaystyle\sum_{i,j} X_i^\mathsf{T} Y_j < t \\ 1, & \text{sign}(\rho) \displaystyle\sum_{i,j} X_i^\mathsf{T} Y_j \geq t \ . \end{cases} \tag{2}$$

Let ϕ_T be the optimal test over such functions, i.e., $R(\phi_T) = \inf_t R(\phi_{T,t})$ and the minimizer $t \in (\mathbb{E}_0\, T, \min_{\sigma\in S_n} \mathbb{E}_{1|\sigma}\, T) = (0, |\rho| nd)$ may depend on n, d, and ρ.

Theorem 1 (Detection Achievability)**.** *Let $t = \sqrt{\gamma}\frac{nd}{2}$ with $\gamma \in (0, 4\rho^2)$. The risk of the sum-of-inner-products test ϕ_T for the binary hypothesis testing problem (1) is upper bounded by*

$$\begin{aligned} R(\phi_T) &\leq \min_{\gamma \in (0, 4\rho^2)} \exp\Big(-\frac{d}{2} g_{FA}(\gamma)\Big) + \exp\Big(-\frac{d}{2} g_{MD}(\gamma)\Big) \\ &\leq 2\exp\Big(-\frac{d\rho^2}{60}\Big) \\ g_{FA}(\gamma) &\triangleq \sqrt{1+\gamma} - 1 - \ln\left(\frac{1+\sqrt{1+\gamma}}{2}\right) \\ g_{MD}(\gamma) &\triangleq \frac{1}{1-\rho^2}\Big(\sqrt{\gamma + (1-\rho^2)^2} - \sqrt{\rho^2\gamma}\Big) \\ &\quad - \left(1 + \ln\left(\frac{\sqrt{\gamma + (1-\rho^2)^2} + 1 - \rho^2}{2}\right)\right) . \end{aligned}$$

Thus, if $\rho^2 = \omega\big(\frac{1}{d}\big)$, then $R(\phi_T) \to 0$ as $d \to \infty$. The proof can be found in Section IV.

Our converse exhibits a similar d–asymptotic behavior, up to a constant factor, but also dependent on n.

Theorem 2 (Detection Converse)**.** *If*

$$\rho^2 = o\left(\frac{e^{\sqrt{\ln n}}}{dn}\right) \ \text{ and } \ d = \Omega(e^{\sqrt{\ln n}}) \ ,$$

then the minimax risk for the binary hypothesis testing problem (1) goes to 1, i.e., $R^* \to 1$ as d, $n \to \infty$.

The proof uses a truncated second moment method and is outlined in Section V. (The full proof is available on arXiv.)

It follows immediately from our results that for fixed n, the minimax risk undergoes a *phase transition* at $\rho^2 \approx \frac{1}{d}$. That

[1] Note that prior work [6] considers the average-case probability of error with respect to a uniform distribution over permutations. However, owing to the symmetries of the distribution, these two models are essentially equivalent, and the results of [6] translate directly to minimax bounds.

is, for $\rho^2 = \omega\left(\frac{1}{d}\right)$, detection succeeds with high probability and, for $\rho^2 = o\left(\frac{1}{d}\right)$, the risk of any test is lower bounded by an absolute constant. Compared to the behavior of the recovery problem as determined in [6], and reproduced in our notation below, for sufficiently large d, our detection boundary is significantly lower. However, our achievability and converse bounds do not exhibit the same scaling with respect to n.

Theorem 3 (Recovery Achievability, [6, Theorem 1]). *The probability of error of the maximum likelihood (ML) decoder* $\hat{\sigma}_{ML} = \arg\max_{\sigma \in S_n} \mathbb{P}_{1|\sigma}\left(X^n, Y^n\right)$ *is upper bounded by*

$$P_{err}(\hat{\sigma}_{ML}) \leq n(1-\rho^2)^{\frac{d}{4}} \frac{1-\left(n(1-\rho^2)^{\frac{d}{4}}\right)^n}{1-n(1-\rho^2)^{\frac{d}{4}}}$$

Consequently, if $\rho^2 = 1 - o\left(n^{-\frac{4}{d}}\right)$, then the ML decoder returns the true permutation with high probability.

Theorem 4 (Recovery Converse, [6, Theorem 2]). *The minimax probability of error is lower bounded by*

$$P^*_{err} \geq 1 - \left(n(1-\rho^2)^{\frac{d}{4}(1+\epsilon(d))}\right)^{-2} - 4\left(n(1-\rho^2)^{\frac{d}{4}(1+\epsilon(d))}\right)^{-1}$$

where $\epsilon(d) \to 0$ *as* $d \to \infty$.

Consequently, if $\rho^2 = 1 - \omega(n^{-\frac{4}{d}})$, then the probability of error of any decoder is close to 1.

In Figures 1 and 2, we have the plotted upper and lower bounds on the squared correlation coefficient ρ^2 required to attain a specified risk R that follow from Theorems 1 and 2.[2] For comparison, we have also plotted the achievable and converse results from the recovery problem from Theorems 3 and 4. Specifically, any recovery scheme can be converted into a detection scheme by first estimating the permutation as $\hat{\sigma}$ and then thresholding the statistic $\operatorname{sign}(\rho)\sum_i X_i^T Y_{\hat{\sigma}_i}$. These plots show that detection is easier than recovery.

[2]For the plot of Theorem 2, we assume the constants in the bound are 1, i.e., we plot $\rho^2 = \frac{e^{\sqrt{\ln n}}}{dn}$ and $d = e^{\sqrt{\ln n}}$ This is primarily to show that the converse bound has the same scaling behavior as our achievability bound.

IV. Achievability

For our proposed sum-of-inner-products statistic T, the moment generating function (MGF) can be calculated explicitly under H_0 and H_1.

Lemma 5. *For* $-\frac{1}{n(1-|\rho|)} < \lambda < \frac{1}{n(1+|\rho|)}$, *the MGF under* H_1 *is*

$$\mathbb{E}_{1|\sigma}[e^{\lambda T}] = \left(1 - 2n\lambda|\rho| - n^2\lambda^2(1-\rho^2)\right)^{-d/2} .$$

For $|\lambda| < \frac{1}{n}$, *the MGF under* H_0 *is*

$$\mathbb{E}_0[e^{\lambda T}] = \left(\mathbb{E}_{1|\sigma}[e^{\lambda T}]\right)\big|_{\rho=0} = \left(1 - n^2\lambda^2\right)^{-d/2} .$$

Proof. Observe that under $\mathbb{P}_{1|\sigma}$, $X_i = \rho Y_{\sigma_i} + \sqrt{1-\rho^2} Z_i$ with $Z_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{I}_d)$ and independent of Y_{σ_i} for any $i \in [n]$. For some λ, we get

$$\begin{aligned}
&\mathbb{E}_{1|\sigma}[e^{\lambda T}] \\
&= \mathbb{E}_{1|\sigma}\Big[\exp\Big(\lambda \operatorname{sign}(\rho) \sum_{i,j}^{n} X_i^{\mathsf{T}} Y_j\Big)\Big] \\
&= \mathbb{E}_{Y^n, Z^n \overset{\text{iid}}{\sim}\mathcal{N}(0,\mathbf{I}_d)}\Big[\exp\Big(\lambda \sum_{i,j}^{n} (|\rho| Y_{\sigma_i} + \operatorname{sign}(\rho)\sqrt{1-\rho^2} Z_i)^{\mathsf{T}} Y_j\Big)\Big] \\
&\overset{(a)}{=} \mathbb{E}\Big[\exp\Big(\lambda|\rho| \sum_{i,j} Y_j^{\mathsf{T}} Y_{\sigma_i}\Big) \\
&\qquad \cdot \mathbb{E}\Big[\exp\Big(\lambda \operatorname{sign}(\rho)\sqrt{1-\rho^2} \sum_i Z_i^{\mathsf{T}}\Big(\sum_j Y_j\Big)\Big) \mid Y^n\Big]\Big]
\end{aligned}$$

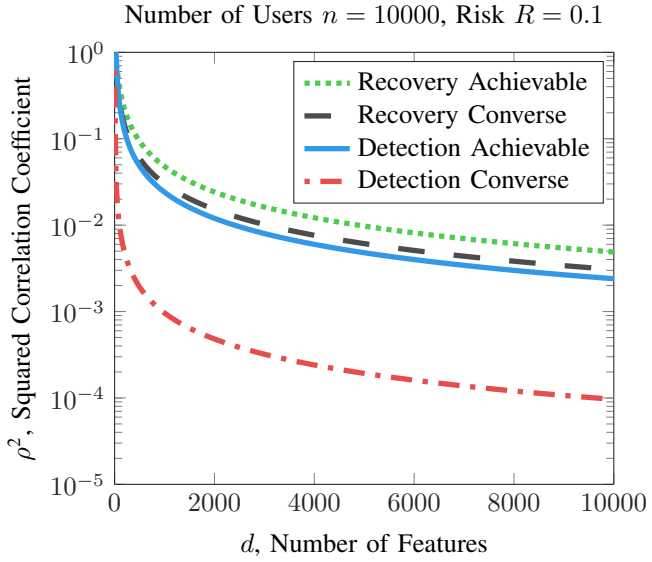

Fig. 1. Upper and lower bounds on the squared correlation coefficient ρ^2 needed to attain risk $R = 0.1$ for $n = 10000$ users, with respect to the number of features d.

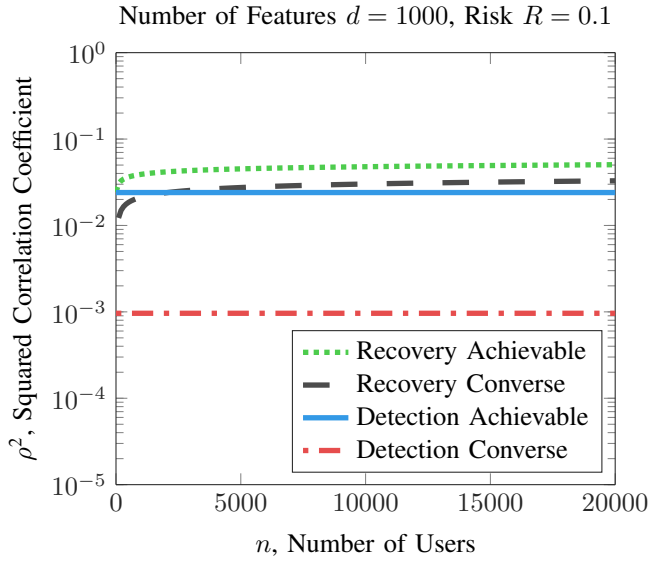

Fig. 2. Upper and lower bounds on the squared correlation coefficient ρ^2 needed to attain risk $R = 0.1$ for $d = 1000$ features, with respect to the number of users n.

$$\overset{(b)}{=} \mathbb{E}\Big[\exp\Big(\lambda|\rho|\sum_{i,j} Y_j^\mathsf{T} Y_{\sigma_i}\Big) \cdot \prod_i \mathbb{E}\Big[\exp\Big(\lambda\,\mathrm{sign}(\rho)\sqrt{1-\rho^2}\, Z_i^\mathsf{T}\Big(\sum_j Y_j\Big)\Big) \mid Y^n\Big]\Big]$$

$$\overset{(c)}{=} \mathbb{E}\Big[\exp\Big(\lambda|\rho|\sum_{i,j} Y_j^\mathsf{T} Y_{\sigma_i} + n\frac{\lambda^2(1-\rho^2)}{2}\Big\|\sum_j Y_j\Big\|^2\Big)\Big]$$

where (a) follows by the law of total expectation, (b) the conditional independence of Z^n given Y^n and (c) the moment generating function of a normal variable. Due to the bijectivity of a permutation, we observe that $\sum_{i,j} Y_j^\mathsf{T} Y_{\sigma_i} = \big\|\sum_j Y_j\big\|^2$. To simplify the notation, let $Y = \sum_j Y_j$ so that $\frac{Y}{\sqrt{n}} \sim \mathcal{N}(0, \mathbf{I}_d)$. Plugging back into the last equation and using the moment generating function of the χ^2_d distribution, we get

$$\mathbb{E}_{1|\sigma}[e^{\lambda T}] = \mathbb{E}\Big[\exp\Big(\Big\|\frac{Y}{\sqrt{n}}\Big\|^2\Big(n\lambda|\rho| + n^2\frac{\lambda^2(1-\rho^2)}{2}\Big)\Big)\Big] = \big(1 - 2n\lambda|\rho| - n^2\lambda^2(1-\rho^2)\big)^{-d/2}$$

for $1-2n\lambda|\rho|-n^2\lambda^2(1-\rho^2) > 0$. By completing the square, we can rewrite the constraint as

$$\begin{aligned} &1 - 2n\lambda|\rho| - n^2\lambda^2(1-\rho^2) > 0 \\ &\Leftrightarrow \Big(\lambda + \frac{|\rho|}{n(1-\rho^2)}\Big)^2 < \frac{1}{n^2(1-\rho^2)} + \frac{\rho^2}{n^2(1-\rho^2)^2} \\ &\Leftrightarrow \frac{-1}{n(1-|\rho|)} < \lambda < \frac{1}{n(1+|\rho|)}\,. \end{aligned}$$

For $\mathbb{E}_0[e^{\lambda T}]$, we simply note that the null model is equal to the alternate model evaluated at zero correlation, $\mathbb{P}_0 = (\mathbb{P}_{1|\sigma})|_{\rho=0}$, and thus $\mathbb{E}_0[e^{\lambda T}] = \big(\mathbb{E}_{1|\sigma}[e^{\lambda T}]\big)\big|_{\rho=0}$. ∎

Proof of Theorem 1. From the definition of the threshold test (2), we have that

$$R(\phi_T) = \min_t\Big(\mathbb{P}_0\{T \ge t\} + \max_{\sigma\in S_n}\mathbb{P}_{1|\sigma}\{T < t\}\Big)\,.$$

Next, we apply the Chernoff bound. For the false alarm probability, we have

$$\begin{aligned} \mathbb{P}_0\{T \ge t\} &\le \min_{\lambda>0}\exp\Big(-\lambda t + \ln\big(\mathbb{E}_0[e^{\lambda T}]\big)\Big) \\ &= \min_{0<\lambda<\frac{1}{n}}\exp\Big(-\lambda t - \frac{d}{2}\ln(1-n^2\lambda^2)\Big) \\ &= \exp\Big(-\frac{d}{2}g_{\text{FA}}(\gamma)\Big) \end{aligned}$$

where the last step follows from plugging in the optimal value $\lambda^*_{\text{FA}} = -\frac{d}{2t} + \sqrt{\frac{1}{n^2} + (\frac{d}{2t})^2}$, defining $\gamma = (\frac{2t}{nd})^2 \in (0, 4\rho^2)$, and simplifying. Similarly, for the missed detection probability, we have

$$\begin{aligned} &\mathbb{P}_{1|\sigma}\{T < t\} \\ &\le \min_{\lambda>0}\exp\Big(\lambda t + \ln\big(\mathbb{E}_{1|\sigma}[e^{\lambda T}]\big)\Big) \\ &= \min_{0<\lambda<\frac{1}{n(1-|\rho|)}}\exp\Big(\lambda t - \frac{d}{2}\ln\big(1 + 2n\lambda|\rho| - n^2\lambda^2(1-\rho^2)\big)\Big) \\ &= \exp\Big(-\frac{d}{2}g_{\text{MD}}(\gamma)\Big) \end{aligned}$$

where the last step follows from plugging in the optimal value $\lambda^*_{\text{MD}} = \frac{|\rho|}{n(1-\rho^2)} + \frac{d}{2t} - \sqrt{(\frac{1}{n(1-\rho^2)})^2 + (\frac{d}{2t})^2}$, defining $\gamma = (\frac{2t}{nd})^2 \in (0, 4\rho^2)$, and simplifying.

By using the inequality $\ln x \le x - 1$ and the concavity of the square root function, we get

$$g_{\text{FA}}(\gamma) \ge \frac{\sqrt{2}-1}{2}\gamma\,.$$

Set $\gamma = \rho^2$. By using the inequalities for $x \in (0,1]$ $\ln x \le 2 - \frac{4}{1+x}$ [21] and $x^2(\sqrt{1+x}-1) \le \sqrt{1+x^3}-1 \le x\,(\sqrt{1+x}-1)$, one can show that

$$g_{\text{MD}}(\rho^2) \ge g_{\text{FA}}(\rho^2) - (\sqrt{2}-1)^2\rho^2.$$

Finally, the concise (but loose) upper bound $2\exp\big(-\frac{d\rho^2}{60}\big)$ follows by combining these results. ∎

V. Converse

We begin with the well-known fact that the minimax risk is lower bounded by the Bayes risk. For a prior distribution π over the set of permutations S_n, we define the corresponding Bayes risk as

$$R^*_\pi \triangleq \inf_\phi \mathbb{P}_0\big[\phi(X^n, Y^n) = 1\big] + \mathbb{E}_{\sigma\sim\pi}\mathbb{P}_{1|\sigma}\big[\phi(X^n, Y^n) = 0\big]\,.$$

Then, $R^* \ge R^*_\pi$ for any prior π. The Bayes risk stems from a binary hypothesis test between simple hypotheses, $\mathbb{P}_0$ for the null and the mixture distribution $\mathbb{P}_1 = \mathbb{E}_{\sigma\sim\pi}\,\mathbb{P}_{1|\sigma}$ for the alternate. We assume that $\mathbb{P}_0$ and $\mathbb{P}_1$ are mutually absolutely continuous and we define their likelihood ratio as $L \triangleq \mathbb{P}_1/\mathbb{P}_0$. Thus, we immediately have that

$$1 - R^* \le 1 - R^*_\pi = d_{\text{TV}}(\mathbb{P}_0, \mathbb{P}_1) = \mathbb{E}_0\,|L-1| \le \sqrt{\mathbb{E}_0 L^2 - 1} \quad (3)$$

where the last step is due to Cauchy-Schwarz inequality and the total variation distance is $d_{\text{TV}}(P,Q) = \frac{1}{2}\mathbb{E}_Q\big|\frac{P}{Q} - 1\big|$.

This second moment method (i.e., upper bounding $\mathbb{E}_0\,L^2$) is a common approach to find the lower bound for hypothesis testing problems (see, e.g., [17], [22], [23]). The composite law under H_1 makes it challenging to compute $\mathbb{E}_0\,L^2 = \mathbb{E}_0(\,\mathbb{E}_{\sigma\sim\pi}\,\mathbb{P}_{1|\sigma}\,/\,\mathbb{P}_0)^2$. The lemma below gives an alternative expression by using Fubini's theorem to exchange the order of integration. To the best of our knowledge, this approach was introduced by Ingster and Suslina in the context of a Gaussian location model (see [24, Equation (3.66)]).

Lemma 6 (Ingster-Suslina method)**.** *Let $\mathbb{P}_{1|\sigma}$ and $\mathbb{P}_0$ be distributions with σ from a parameter space. Given a prior on the space of σ, define the mixture distribution as $\mathbb{P}_1 = \mathbb{E}_{\sigma\sim\pi}\,\mathbb{P}_{1|\sigma}$. Then, with $\sigma, \tilde\sigma \overset{\text{iid}}{\sim} \pi$ and $G(\sigma,\tilde\sigma) = \int \frac{\mathbb{P}_{1|\sigma}\mathbb{P}_{1|\tilde\sigma}}{\mathbb{P}_0}$, we have*

$$\mathbb{E}_0\,L^2 = \mathbb{E}_0\Big(\frac{\mathbb{E}_{\sigma\sim\pi}\,\mathbb{P}_{1|\sigma}}{\mathbb{P}_0}\Big)^2 = \mathbb{E}_{\sigma,\tilde\sigma\sim\pi}\,G(\sigma,\tilde\sigma)\,.$$

This leads to our first converse result.

Lemma 7.

$$R^* \ge 1 - \sqrt{(1-\rho^2)^{-dn} - 1}$$

Consequently, if $\rho^2 = o\big(\frac{1}{dn}\big)$*, then* $R^* \to 1$ *as* $n \to \infty$ *or* $d \to \infty$.

Proof. The proof uses a version of [17, Proposition 1], Lemma 6 with $\pi = \mathrm{Unif}(S_n)$ and Equation (3). ∎

A. Truncation

To sharpen the converse, we truncate the likelihood ratio by conditioning on a carefully-chosen event that occurs with high probability under H_1, following the approach of [17], [22]. For an event Γ_σ, define the truncated likelihood ratio as

$$\tilde{L} = \mathbb{E}_{\sigma\sim\pi} L_\sigma \mathbb{1}_{\Gamma_\sigma} \text{ where } L_\sigma = \mathbb{P}_{1|\sigma}/\mathbb{P}_0 \ .$$

Then, by the triangle and Cauchy-Schwarz inequalities combined with the fact that $\tilde{L} \le L = \mathbb{E}_{\sigma\sim\pi} L_\sigma$, we get

$$\begin{aligned} 1 - R^* \le 1 - R^*_\pi = \mathbb{E}_0 |L - 1| &\le \mathbb{E}_0 |\tilde{L} - 1| + \mathbb{E}_0(L - \tilde{L}) \\ &= \sqrt{\mathbb{E}_0 \tilde{L}^2 - 1 + 2\big(1 - \mathbb{E}_0 \tilde{L}\big)} + \big(1 - \mathbb{E}_0 \tilde{L}\big) \ . \end{aligned} \tag{4}$$

Thus, if we simultaneously lower bound $\mathbb{E}_0 \tilde{L}$ and upper bound on $\mathbb{E}_0 \tilde{L}^2$, we can obtain a converse bound. Observe that by Fubini's theorem, the truncated first moment becomes

$$\mathbb{E}_0 \tilde{L} = \mathbb{E}_0 \mathbb{E}_{\sigma\sim\pi} L_\sigma \mathbb{1}_{\Gamma_\sigma} = \mathbb{E}_{\sigma\sim\pi} \mathbb{E}_0 L_\sigma \mathbb{1}_{\Gamma_\sigma} = \mathbb{E}_{\sigma\sim\pi} \mathbb{P}_{1|\sigma}(\Gamma_\sigma) \ .$$

Setting $\pi = \mathrm{Unif}(S_n)$, we will select an event Γ_σ that occurs with high probability under H_1 for any $\sigma \in S_n$, which in turns shows that $\mathbb{E}_0 \tilde{L} \to 1$. Moreover, Lemma 6 can be used to compute $\mathbb{E}_0 \tilde{L}^2$ as before.

For a permutation $\sigma \in S_n$, let $F_\sigma = \{i \in [n] : i = \sigma_i\}$ be the set of corresponding fixed points and let $N_1^\sigma = |F_\sigma|$. For any σ, define the truncation event Γ_σ as

$$\begin{aligned} \Gamma_\sigma \triangleq \bigcap_{k=k^*}^{N_1^\sigma} \bigcap_{\substack{T\subseteq F_\sigma \\ |T|=k}} \Bigg\{ \sum_{i\in T} X_i^\mathsf{T} X_i > w_k, \sum_{i\in T} Y_i^\mathsf{T} Y_i > w_k, \\ \operatorname{sign}(\rho) \sum_{i\in T} X_i^\mathsf{T} Y_{\sigma_i} < v_k \Bigg\} \end{aligned} \tag{5}$$

where $k^* \in [n]$, $w_k > 0$ and v_k for $k = k^*, \dots, N_1^\sigma$ will be chosen later.

In the following, we will bound the first and second moments of the truncated likelihood ratio $\tilde{L} = \mathbb{E}_{\sigma\sim\mathrm{Unif}(S_n)} \frac{\mathbb{P}_{1|\sigma}}{\mathbb{P}_0} \mathbb{1}_{\Gamma_\sigma}$ where $\mathbb{P}_0$ and $\mathbb{P}_{1|\sigma}$ are the likelihoods of the binary hypothesis testing problem (1) with prior distribution $\mathrm{Unif}(S_n)$ on S_n.

Lemma 8 (Truncated First Moment)**.** *If*

$$\begin{aligned} &w_k = dk - 2\sqrt{dk}\, r_k \quad \text{and} \quad v_k = |\rho|\, dk + 4\,|\rho|\, \sqrt{dk}\, s_k \\ &\text{with} \quad \frac{\sqrt{d}}{2} > r_k > \sqrt{\ln \frac{en}{k}} \quad \text{and} \\ &s_k > \sqrt{\frac{1-\rho^2}{\rho^2} \ln \frac{en}{k}} \max\left\{1, \frac{2}{\sqrt{d}} \sqrt{\ln \frac{en}{k}}\right\} \end{aligned}$$

for all $k = k^*, \dots, n$ *with* $k^* = \omega(1)$ *and* $d = \Omega\big(\ln \frac{n}{k^*}\big)$*, then* $\mathbb{E}_0 \tilde{L} \to 1$ *as* $n \to \infty$ *with* $\mathbb{E}_0 \tilde{L} \ge 1 - \mathcal{O}(1)$.

Proof outline. As discussed above, it suffices to upper bound $\mathbb{P}_{1|\sigma}(\Gamma_\sigma^C)$ for any σ where Γ_σ^C is the complement of Γ_σ. Upon applying the union bound, we get

$$\begin{aligned} \mathbb{P}_{1|\sigma}(\,\Gamma_\sigma^C\,) \le \sum_{k=k^*}^{N_1^\sigma} \sum_{\substack{T\subseteq F_\sigma \\ |T|=k}} 2\mathbb{P}_{1|\sigma}\Big(\sum_{i\in T} \|X_i\|^2 \le w_k\Big) \\ + \mathbb{P}_{1|\sigma}\Big(\operatorname{sign}(\rho) \sum_{i\in T} X_i^\mathsf{T} Y_{\sigma_i} \ge v_k\Big) \ . \end{aligned}$$

Then, the proof follows by applying two concentration inequalities: the Laurent-Massart lemma [25, Lemma 1] and Lemma 10 from the Appendix, applied to the first and second terms, respectively. ∎

Lemma 9 (Truncated Second Moment)**.** *If the conditions of Lemma 8 hold by choosing* $k^* = \frac{n}{e^{\sqrt{\ln n} - 1}}$ *and* $\rho^2 = o\Big(\frac{e^{\sqrt{\ln n}}}{dn}\Big)$*, then* $\mathbb{E}_0 \tilde{L}^2 \to 1$ *as* $n \to \infty$ *with* $\mathbb{E}_0 \tilde{L}^2 \le 1 + \mathcal{O}(1)$.

Proof Sketch. The proof follows by examining the behavior of $\mathbb{E}_0 \tilde{L}^2$ for small and large number of fixed points of a permutation separately as

$$\mathbb{E}_0 \tilde{L}^2 = \underbrace{\mathbb{E}_0 \tilde{L}^2 \mathbb{1}\{N_1^\sigma \le k^*\}}_{\triangleq \text{(I)}} + \underbrace{\mathbb{E}_0 \tilde{L}^2 \mathbb{1}\{N_1^\sigma > k^*\}}_{\triangleq \text{(II)}} \ .$$

Then, the aim is converted into finding two functions of k^*, call them $f_1(k^*)$ and $f_2(k^*)$, such that

$$\begin{aligned} &\text{for } \rho^2 < f_1(k^*), \quad \text{(I)} = 1 + \mathcal{O}(1) \text{ and} \\ &\text{for } \rho^2 < f_2(k^*), \quad \text{(II)} = \mathcal{O}(1) \end{aligned}$$

both hold. Combining them, we find the desired condition:

$$\text{For } \rho^2 < \min\{f_1(k^*),\, f_2(k^*)\},\ \mathbb{E}_0 \tilde{L}^2 = \text{(I)} + \text{(II)} = 1 + \mathcal{O}(1) \ .$$

In order to find a converse region of ρ^2 as large as possible – as tight as possible – we need to choose k^* as the maximizer of $\min\{f_1(k^*),\, f_2(k^*)\}$. ∎

Proof of Theorem 2. Apply Equation (4) along with Lemma 8 and Lemma 9 by setting the prior distribution as $\pi = \mathrm{Unif}(S_n)$, if $\rho^2 = o\big(\frac{e^{\sqrt{\ln n}}}{dn}\big)$ and $d = \Omega\big(\ln \frac{n}{k^*}\big) = \Omega(e^{\sqrt{\ln n}})$ for $k^* = \frac{n}{e^{\sqrt{\ln n} - 1}}$, then one gets $R^* \to 1$ as $n \to \infty$. ∎

Acknowledgments

The authors would like to thank Reviewer 2 for pointing out an error in the submitted version of Theorem 2.

Appendix

Lemma 10 (Gaussian chaos)**.** *Let* $X \sim \mathcal{N}(0, \mathbf{I}_d)$ *and* $A = O + D$ *be a symmetric matrix where* $D = \mathrm{diag}(\alpha)$ *with* $\alpha_i = A_{ii}$ *for* $i \in [d]$ *and* O *is the off-diagonal part of* A *with* $\lambda = (\lambda_1, \dots, \lambda_d)$ *eigenvalues of* O*. For* $t > 0$*,*

$$\begin{aligned} \mathbb{P}\left(X^\mathsf{T} A X - \mathbb{E} X^\mathsf{T} A X \ge t\right) \le 2 \exp\Bigg(-\frac{t}{16} \\ \min\left\{\frac{t}{2\,\|\alpha\|_2^2}, \frac{1}{\|\alpha\|_\infty}, \frac{t}{\|2\lambda\|_2^2}, \frac{1}{\|\lambda\|_\infty}\right\}\Bigg) \ . \end{aligned}$$

REFERENCES

[1] M. Barbaro, T. Zeller, and S. Hansell, "A face is exposed for AOL searcher no. 4417749," *New York Times*, vol. 9, no. 2008, p. 8, 2006.

[2] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pp. 111–125, IEEE, 2008.

[3] A. Narayanan and V. Shmatikov, "De-anonymizing social networks," in *2009 30th IEEE symposium on security and privacy*, pp. 173–187, IEEE, 2009.

[4] L. Sweeney, "Matching known patients to health records in Washington State data," *Available at SSRN 2289850*, 2013.

[5] D. Cullina, P. Mittal, and N. Kiyavash, "Fundamental limits of database alignment," in *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 651–655, IEEE, 2018.

[6] O. E. Dai, D. Cullina, and N. Kiyavash, "Database alignment with gaussian features," in *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, vol. 89 of *Proceedings of Machine Learning Research*, pp. 3225–3233, 2019.

[7] F. Shirani, S. Garg, and E. Erkip, "A concentration of measure approach to database de-anonymization," in *2019 IEEE International Symposium on Information Theory (ISIT)*, pp. 2748–2752, IEEE, 2019.

[8] O. E. Dai, D. Cullina, and N. Kiyavash, "Achievability of nearly-exact alignment for correlated gaussian databases," in *2020 IEEE International Symposium on Information Theory (ISIT)*, pp. 1230–1235, IEEE, 2020.

[9] S. Bakırtaş and E. Erkip, "Database matching under column deletions," in *2021 IEEE International Symposium on Information Theory (ISIT)*, pp. 2720–2725, IEEE, 2021.

[10] S. Bakirtas and E. Erkip, "Database matching under column repetitions," *arXiv preprint arXiv:2202.01730*, 2022.

[11] M. Chertkov, L. Kroc, F. Krzakala, M. Vergassola, and L. Zdeborová, "Inference in particle tracking experiments by passing messages between images," *Proceedings of the National Academy of Sciences*, vol. 107, no. 17, pp. 7663–7668, 2010.

[12] D. Kunisky and J. Niles-Weed, "Strong recovery of geometric planted matchings," in *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 834–876, SIAM, 2022.

[13] M. Moharrami, C. Moore, and J. Xu, "The planted matching problem: Phase transitions and exact results," *The Annals of Applied Probability*, vol. 31, no. 6, pp. 2663–2720, 2021.

[14] J. Ding, Y. Wu, J. Xu, and D. Yang, "The planted matching problem: Sharp threshold and infinite-order phase transition," *arXiv preprint arXiv:2103.09383*, 2021.

[15] P. Pedarsani and M. Grossglauser, "On the privacy of anonymized networks," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1235–1243, 2011.

[16] J. Ding, Z. Ma, Y. Wu, and J. Xu, "Efficient random graph matching via degree profiles," *Probability Theory and Related Fields*, vol. 179, no. 1, pp. 29–115, 2021.

[17] Y. Wu, J. Xu, and S. H. Yu, "Testing correlation of unlabeled random graphs," *arXiv preprint arXiv:2008.10097*, 2020.

[18] C. Mao, Y. Wu, J. Xu, and S. H. Yu, "Testing network correlation efficiently via counting trees," *arXiv preprint arXiv:2110.11816*, 2021.

[19] Y. Wu, J. Xu, and S. H. Yu, "Settling the sharp reconstruction thresholds of random graph matching," *arXiv preprint arXiv:2102.00082*, 2021.

[20] L. Ganassali, "Sharp threshold for alignment of graph databases with gaussian weights," *arXiv preprint arXiv:2010.16295*, 2020.

[21] F. Topsøe, "Some bounds for the logarithmic function," *RGMIA Res. Rep. Collection*, vol. 7, no. 2, pp. 1–20, 2004.

[22] E. Arias-Castro and N. Verzelen, "Community detection in dense random networks," *The Annals of Statistics*, vol. 42, no. 3, pp. 940 – 969, 2014.

[23] Y. Wu and J. Xu, "Statistical problems with planted structures: Information-theoretical and computational limits," *Information-Theoretic Methods in Data Science*, vol. 383, 13.

[24] Y. Ingster, J. I. Ingster, and I. Suslina, *Nonparametric goodness-of-fit testing under Gaussian models*, vol. 169. Springer Science & Business Media, 2003.

[25] B. Laurent and P. Massart, "Adaptive estimation of a quadratic functional by model selection," *The Annals of Statistics*, vol. 28, no. 5, pp. 1302 – 1338, 2000.