Testing Dependency of Databases

Wasim Huleihel

Tel Aviv University
Department of Electrical Engineering - Systems
February 28, 2024

EnCORE Workshop on Computational vs Statistical Gaps in Learning and Optimization

Motivation: Data Alignment Problem

Correlated data structures

- Data collection (from many sources) is ubiquitous.

Motivation: Data Alignment Problem

Correlated data structures

- Data collection (from many sources) is ubiquitous.
- Different data structures/sources offer many great benefits for inference.

Motivation: Data Alignment Problem

Correlated data structures

- Data collection (from many sources) is ubiquitous.
- Different data structures/sources offer many great benefits for inference.
- Understanding and quantifying the correlation between data structures are among the most fundamental tasks in statistics!

Motivation: Data Alignment Problem

Correlated data structures

- Data collection (from many sources) is ubiquitous.
- Different data structures/sources offer many great benefits for inference.
- Understanding and quantifying the correlation between data structures are among the most fundamental tasks in statistics!
- Modern challenges: data structures are high-d, noisy, **unlabeled/scrambled**.

Motivation: Data Alignment Problem

Correlated data structures

- Data collection (from many sources) is ubiquitous.
- Different data structures/sources offer many great benefits for inference.
- Understanding and quantifying the correlation between data structures are among the most fundamental tasks in statistics!
- Modern challenges: data structures are high-d, noisy, **unlabeled/scrambled**.
- This precludes "direct" inference/data junction.

Motivation: Data Alignment Problem

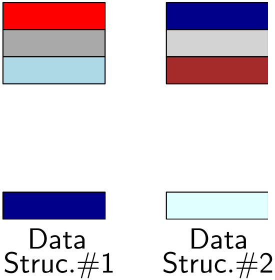
Correlated data structures

- Data collection (from many sources) is ubiquitous.
- Different data structures/sources offer many great benefits for inference.
- Understanding and quantifying the correlation between data structures are among the most fundamental tasks in statistics!
- Modern challenges: data structures are high-d, noisy, **unlabeled/scrambled**.
- This precludes "direct" inference/data junction.
- General goal: determine if $\exists$ a correspondence under which the sources are "correlated".

Motivation: Data Alignment Problem (Cont'd)

Pictorially...

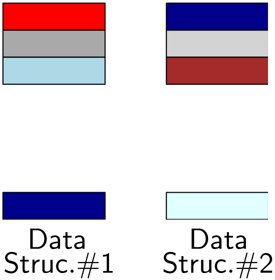
- Multiple data structures/sources are available.



Motivation: Data Alignment Problem (Cont'd)

Pictorially...

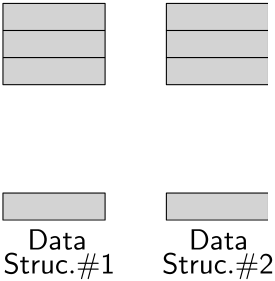
- Multiple data structures/sources are available.
- Each source provides information for entities (e.g., users).



Motivation: Data Alignment Problem (Cont'd)

Pictorially...

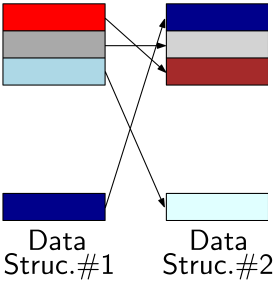
- Multiple data structures/sources are available.
- Each source provides information for entities (e.g., users).
- The correspondence between different sources is unknown/obfuscated.



Motivation: Data Alignment Problem (Cont'd)

Pictorially...

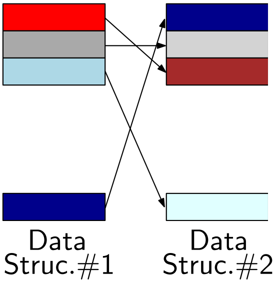
- Multiple data structures/sources are available.
- Each source provides information for entities (e.g., users).
- If "correlation" is sufficiently large maybe it is possible to glean something about the correspondence.



Motivation: Data Alignment Problem (Cont'd)

Pictorially...

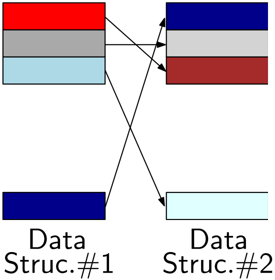
- Multiple data structures/sources are available.
- Each source provides information for entities (e.g., users).
- Valuable tool to recover missing information by labeling unlabeled features and allowing the junction of data coming from different sources.



Motivation: Data Alignment Problem (Cont'd)

Pictorially...

- Multiple data structures/sources are available.
- Each source provides information for entities (e.g., users).
- Crucial to understand limitations of data alignment so as to assess the feasibility and reliability of alignment procedures.



Motivation: Folklore Example

Netflix Prize

- Netflix prize: take dataset and come up with a better recommendation algorithm.
- Dataset: lists of features for a set of entities, say, users.

Motivation: Folklore Example

Netflix Prize

- Netflix prize: take dataset and come up with a better recommendation algorithm.
- Privacy concern: unique identifying sensitive information (e.g., names, user IDs) is deleted from a database while other features (e.g., movie ratings) are left unchanged.

Motivation: Folklore Example

Netflix Prize

- Netflix prize: take dataset and come up with a better recommendation algorithm.
- Privacy concern: unique identifying sensitive information (e.g., names, user IDs) is deleted from a database while other features (e.g., movie ratings) are left unchanged.
- No side information: could be effective for protecting user privacy (while providing access to data).

Motivation: Folklore Example

Netflix Prize

- Netflix prize: take dataset and come up with a better recommendation algorithm.
- Privacy concern: unique identifying sensitive information (e.g., names, user IDs) is deleted from a database while other features (e.g., movie ratings) are left unchanged.
- *Side information is abundant in the public domain!*

Motivation: Folklore Example

Netflix Prize

- Netflix prize: take dataset and come up with a better recommendation algorithm.
- Privacy concern: unique identifying sensitive information (e.g., names, user IDs) is deleted from a database while other features (e.g., movie ratings) are left unchanged.
- *Side information is abundant in the public domain!*
- [Narayanan&Shmatikov'08,09]: many Netflix user IDs can be matched with IMDb profiles.
- Netflix prize dataset (anonymized): User IDs, movie IDs, movie ratings.
- IMDb dataset (public): Usernames, movie names, movie ratings.

Motivation: Folklore Example

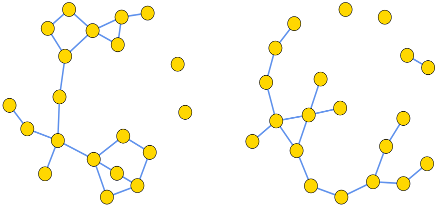
Netflix Prize

- Netflix prize: take dataset and come up with a better recommendation algorithm.
- Privacy concern: unique identifying sensitive information (e.g., names, user IDs) is deleted from a database while other features (e.g., movie ratings) are left unchanged.
- *Side information is abundant in the public domain!*
- Crucial to understand the conditions that allow/prevent privacy breaches, and vulnerability of de-anony. schemes.

Motivation: Graph Alignment/(Noisy) Graph Isomorphism

"Interactions among users"

- In many modern applications, observations appear as graphs.

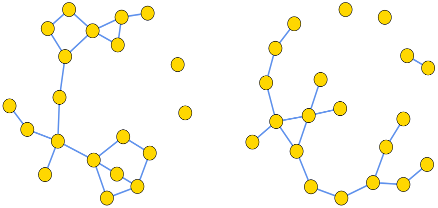


[Wu&Xu&Yu'21]

Motivation: Graph Alignment/(Noisy) Graph Isomorphism

"Interactions among users"

- In many modern applications, observations appear as graphs.
- Node labels may be absent or scrambled.

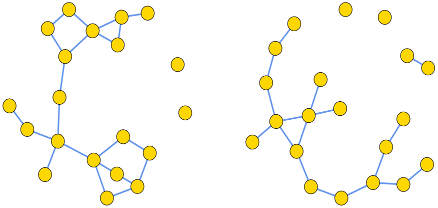


[Wu&Xu&Yu'21]

Motivation: Graph Alignment/(Noisy) Graph Isomorphism

"Interactions among users"

- In many modern applications, observations appear as graphs.
- **Goal: Find/detect node correspondence.**



Motivation: Graph Alignment/(Noisy) Graph Isomorphism

"Interactions among users"

- In many modern applications, observations appear as graphs.
- **Goal: Find/detect node correspondence.**
- Social network analysis: two friendship networks on different social platforms share structural similarities?
- Computational biology: assess the correlation of two biological networks in two different species.
- Natural language processing: uncovering the correlation between two knowledge graphs that are in either different languages.

Motivation: Graph Alignment/(Noisy) Graph Isomorphism

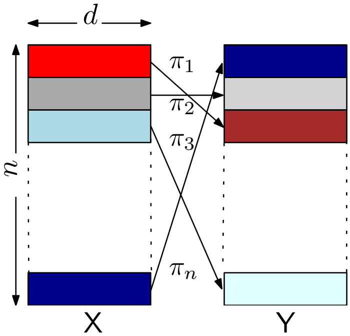
"Interactions among users"

- In many modern applications, observations appear as graphs.
- **Goal: Find/detect node correspondence.**
- Social network analysis: two friendship networks on different social platforms share structural similarities?
- Computational biology: assess the correlation of two biological networks in two different species.
- Natural language processing: uncovering the correlation between two knowledge graphs that are in either different languages.
- Significant attention and beautiful strong results, e.g., [Barak et. al.'19], [Cullina,Kiyavash'16,20], [Wu,Xu,Yu'21], [Ding, Ma, Wu, Xu'21], [Hall,Massoulié'21], [Ding,Li'22], [Ding,Du'23], and many references therein.

The Database Alignment Problem

Generative Correlation Model

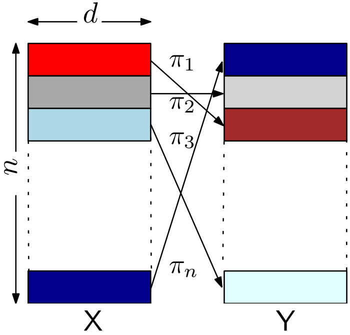
- Databases $\mathsf{X}, \mathsf{Y} \in \mathbb{R}^{n \times d}$: n "users" each with d "features".



The Database Alignment Problem

Generative Correlation Model

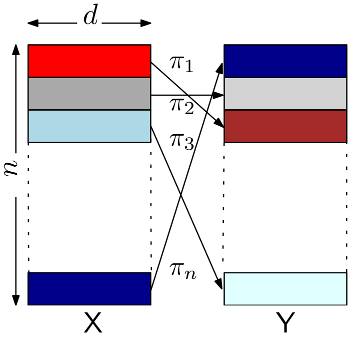
- Databases $\mathsf{X}, \mathsf{Y} \in \mathbb{R}^{n \times d}$: n "users" each with d "features".
- For now, databases include the same set of users.



The Database Alignment Problem

Generative Correlation Model

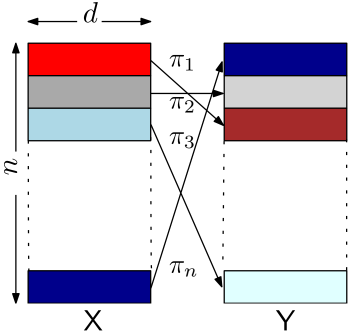
- Databases $X, Y \in \mathbb{R}^{n \times d}$: n "users" each with d "features".
- We will assume features are i.i.d.



The Database Alignment Problem

Generative Correlation Model

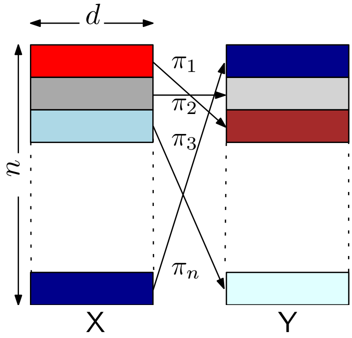
- Databases $\mathsf{X}, \mathsf{Y} \in \mathbb{R}^{n \times d}$: n "users" each with d "features".
- There is a latent (hidden, planted) correspondence (matching, permutation) $\pi \in \mathbb{S}_n$ between the rows of X and Y.



The Database Alignment Problem

Generative Correlation Model

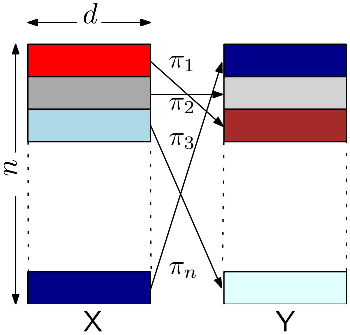
- Databases $\mathsf{X}, \mathsf{Y} \in \mathbb{R}^{n \times d}$: n "users" each with d "features".
- There is a latent (hidden, planted) correspondence (matching, permutation) $\pi \in \mathbb{S}_n$ between the rows of X and Y.
- Features $(\mathsf{X}_i, \mathsf{Y}_{\pi_i})$ associated with user i are dependent, while different pairs are independent.



The Database Alignment Problem

Generative Correlation Model

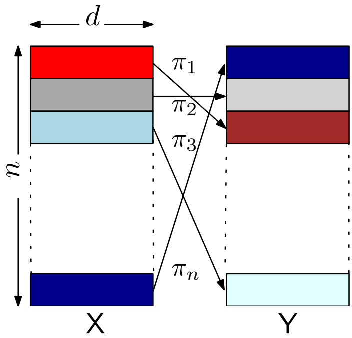
- Recovery/alignment problem: given X, Y recover π.



The Database Alignment Problem

Generative Correlation Model

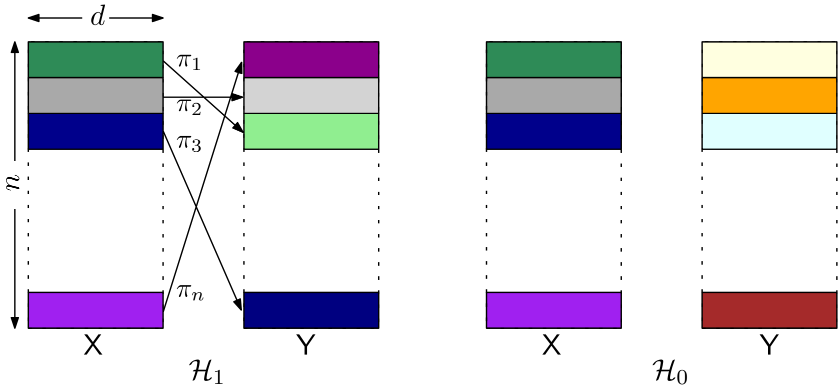
- Recovery/alignment problem: given X, Y recover π.
- Received significant attention, e.g., [Cullina,Mittal,Kiyavash'18],[Dai,Mittal,Kiyavash'19], [Wang,Wu,Xu,Yolou'22].



The Database Alignment Problem

Generative Correlation Model

- In this talk, we focus on the detection variant of this problem.

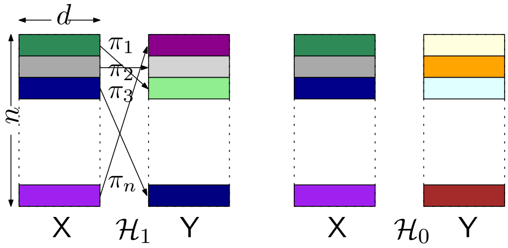


Detecting Correlated Databases

Detection/Hypothesis Testing

- **Null**: X and Y are Gaussian and independent, i.e.,

$$(\mathsf{X}_1, \mathsf{Y}_1), \ldots, (\mathsf{X}_n, \mathsf{Y}_n) \overset{\text{i.i.d.}}{\sim} \mathcal{N}^{\otimes d}(0_{2\times 1}, \mathbf{I}_{2\times 2})$$



Detecting Correlated Databases

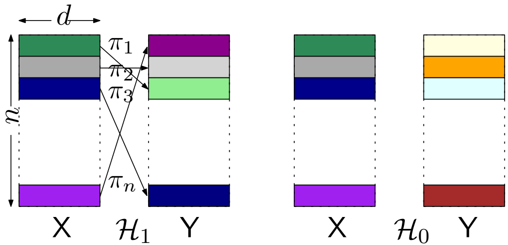
Detection/Hypothesis Testing

- **Null**: X and Y are Gaussian and independent, i.e.,

$$(\mathsf{X}_1,\mathsf{Y}_1),\ldots,(\mathsf{X}_n,\mathsf{Y}_n) \overset{\text{i.i.d.}}{\sim} \mathcal{N}^{\otimes d}(0_{2\times 1},\mathbf{I}_{2\times 2})$$

- **Alternative**: cond. on $\pi \sim \mathsf{Unif}(\mathbb{S}_n)$ (or $\exists \pi \in \mathbb{S}_n$)

$$(\mathsf{X}_1,\mathsf{Y}_{\pi_1}),\ldots,(\mathsf{X}_n,\mathsf{Y}_{\pi_n}) \overset{\text{i.i.d.}}{\sim} \mathcal{N}^{\otimes d}\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix} \triangleq \Sigma_\rho\right)$$



Detecting Correlated Databases

Detection/Hypothesis Testing

- **Null**: X and Y are Gaussian and independent, i.e.,

$$(\mathsf{X}_1,\mathsf{Y}_1),\ldots,(\mathsf{X}_n,\mathsf{Y}_n) \overset{\text{i.i.d.}}{\sim} \mathcal{N}^{\otimes d}(0_{2\times 1},\mathbf{I}_{2\times 2})$$

- **Alternative**: cond. on $\pi \sim \mathsf{Unif}(\mathbb{S}_n)$ (or $\exists\pi \in \mathbb{S}_n$)

$$(\mathsf{X}_1,\mathsf{Y}_{\pi_1}),\ldots,(\mathsf{X}_n,\mathsf{Y}_{\pi_n}) \overset{\text{i.i.d.}}{\sim} \mathcal{N}^{\otimes d}\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix} \triangleq \Sigma_\rho\right)$$

- For a test $\phi : \mathbb{R}^{n\times d}\times\mathbb{R}^{n\times d}\to\{0,1\}$, the "risk" is:

$$\mathsf{R}(\phi) \triangleq \mathbb{P}_{\mathcal{H}_0}[\phi(\mathsf{X},\mathsf{Y})=1] + \mathbb{E}_{\pi\sim\mathsf{Unif}(\mathbb{S}_n)}\mathbb{P}_{\mathcal{H}_1|\pi}[\phi(\mathsf{X},\mathsf{Y})=0].$$

Detecting Correlated Databases

Detection/Hypothesis Testing

- **Null**: X and Y are Gaussian and independent, i.e.,

$$(\mathsf{X}_1,\mathsf{Y}_1),\ldots,(\mathsf{X}_n,\mathsf{Y}_n) \overset{\text{i.i.d.}}{\sim} \mathcal{N}^{\otimes d}(0_{2\times 1},\mathbf{I}_{2\times 2})$$

- **Alternative**: cond. on $\pi \sim \mathsf{Unif}(\mathbb{S}_n)$ (or $\exists \pi \in \mathbb{S}_n$)

$$(\mathsf{X}_1,\mathsf{Y}_{\pi_1}),\ldots,(\mathsf{X}_n,\mathsf{Y}_{\pi_n}) \overset{\text{i.i.d.}}{\sim} \mathcal{N}^{\otimes d}\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix} \triangleq \Sigma_\rho\right)$$

- For a test $\phi : \mathbb{R}^{n\times d}\times\mathbb{R}^{n\times d}\to\{0,1\}$, the "risk" is:

$$\mathsf{R}(\phi) \triangleq \mathbb{P}_{\mathcal{H}_0}[\phi(\mathsf{X},\mathsf{Y})=1] + \mathbb{E}_{\pi\sim\mathsf{Unif}(\mathbb{S}_n)}\mathbb{P}_{\mathcal{H}_1|\pi}[\phi(\mathsf{X},\mathsf{Y})=0].$$

- Minimal (optimal) risk $\mathsf{R}^\star = \inf_\phi \mathsf{R}(\phi)$.

Detecting Correlated Databases

Detection/Hypothesis Testing

- **Null**: X and Y are Gaussian and independent, i.e.,

$$(\mathsf{X}_1, \mathsf{Y}_1), \ldots, (\mathsf{X}_n, \mathsf{Y}_n) \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}^{\otimes d}(0_{2\times 1}, \mathbf{I}_{2\times 2})$$

- **Alternative**: cond. on $\pi \sim \mathsf{Unif}(\mathbb{S}_n)$ (or $\exists \pi \in \mathbb{S}_n$)

$$(\mathsf{X}_1, \mathsf{Y}_{\pi_1}), \ldots, (\mathsf{X}_n, \mathsf{Y}_{\pi_n}) \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}^{\otimes d}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \triangleq \Sigma_\rho\right)$$

- For a test $\phi : \mathbb{R}^{n\times d} \times \mathbb{R}^{n\times d} \to \{0,1\}$, the "risk" is:

$$\mathsf{R}(\phi) \triangleq \mathbb{P}_{\mathcal{H}_0}[\phi(\mathsf{X},\mathsf{Y}) = 1] + \mathbb{E}_{\pi\sim\mathsf{Unif}(\mathbb{S}_n)}\mathbb{P}_{\mathcal{H}_1|\pi}[\phi(\mathsf{X},\mathsf{Y}) = 0].$$

- Possibility: *strong detection* if $\mathsf{R}(\phi) = o(1)$, and *weak detection* if $\lim \mathsf{R}(\phi) < 1$.

Detecting Correlated Databases

Detection/Hypothesis Testing

- **Null**: X and Y are Gaussian and independent, i.e.,

$$(\mathsf{X}_1, \mathsf{Y}_1), \ldots, (\mathsf{X}_n, \mathsf{Y}_n) \overset{\text{i.i.d.}}{\sim} \mathcal{N}^{\otimes d}(0_{2\times 1}, \mathbf{I}_{2\times 2})$$

- **Alternative**: cond. on $\pi \sim \mathsf{Unif}(\mathbb{S}_n)$ (or $\exists \pi \in \mathbb{S}_n$)

$$(\mathsf{X}_1, \mathsf{Y}_{\pi_1}), \ldots, (\mathsf{X}_n, \mathsf{Y}_{\pi_n}) \overset{\text{i.i.d.}}{\sim} \mathcal{N}^{\otimes d}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \triangleq \Sigma_\rho \right)$$

- For a test $\phi : \mathbb{R}^{n\times d} \times \mathbb{R}^{n\times d} \to \{0,1\}$, the "risk" is:

$$\mathsf{R}(\phi) \triangleq \mathbb{P}_{\mathcal{H}_0}[\phi(\mathsf{X},\mathsf{Y}) = 1] + \mathbb{E}_{\pi\sim\mathsf{Unif}(\mathbb{S}_n)}\mathbb{P}_{\mathcal{H}_1|\pi}[\phi(\mathsf{X},\mathsf{Y}) = 0].$$

- Possibility: *strong detection* if $\mathsf{R}(\phi) = o(1)$, and *weak detection* if $\lim \mathsf{R}(\phi) < 1$.
- Impossibility: *strong detection* if $\mathsf{R}^\star = \Omega(1)$, and *weak detection* if $\mathsf{R}^\star = 1 - o(1)$.

Detecting Correlated Databases

Detection/Hypothesis Testing

- **Null**: X and Y are Gaussian and independent, i.e.,

$$(\mathsf{X}_1,\mathsf{Y}_1),\ldots,(\mathsf{X}_n,\mathsf{Y}_n) \overset{\text{i.i.d.}}{\sim} \mathcal{N}^{\otimes d}(0_{2\times 1},\mathbf{I}_{2\times 2})$$

- **Alternative**: cond. on $\pi \sim \mathsf{Unif}(\mathbb{S}_n)$ (or $\exists \pi \in \mathbb{S}_n$)

$$(\mathsf{X}_1,\mathsf{Y}_{\pi_1}),\ldots,(\mathsf{X}_n,\mathsf{Y}_{\pi_n}) \overset{\text{i.i.d.}}{\sim} \mathcal{N}^{\otimes d}\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix} \triangleq \Sigma_\rho\right)$$

- For a test $\phi : \mathbb{R}^{n\times d}\times\mathbb{R}^{n\times d}\to\{0,1\}$, the "risk" is:

$$\mathsf{R}(\phi) \triangleq \mathbb{P}_{\mathcal{H}_0}[\phi(\mathsf{X},\mathsf{Y})=1] + \mathbb{E}_{\pi\sim\mathsf{Unif}(\mathbb{S}_n)}\mathbb{P}_{\mathcal{H}_1|\pi}[\phi(\mathsf{X},\mathsf{Y})=0].$$

- Possibility: *strong detection* if $\lim d_{\mathsf{TV}}(\mathbb{P}_{\mathcal{H}_0},\mathbb{P}_{\mathcal{H}_1})=1$, and *weak detection* if $\liminf d_{\mathsf{TV}}(\mathbb{P}_{\mathcal{H}_0},\mathbb{P}_{\mathcal{H}_1})>0$.
- Impossibility: *strong detection* if $d_{\mathsf{TV}}(\mathbb{P}_{\mathcal{H}_0},\mathbb{P}_{\mathcal{H}_1})\leq 1-\Omega(1)$, and *weak detection* if $d_{\mathsf{TV}}(\mathbb{P}_{\mathcal{H}_0},\mathbb{P}_{\mathcal{H}_1})=o(1)$.

Prior Work (Correlated Databases)

Known Results and Gaps

- [Dai,Cullina,Kiyavash'19]: Perfect recovery is *possible* if $\rho^2 = 1 - o(n^{-4/d})$ and *impossible* if $\rho^2 = 1 - \omega(n^{-4/d})$, assuming $1 \ll d = O(\log n)$.

  E.g., if $d = \omega(\log n)$ then rec. is possible if $\rho^2 = \omega\left(\frac{\log n}{d}\right)$.

Prior Work (Correlated Databases)

Known Results and Gaps

- [Dai,Cullina,Kiyavash'19]: Perfect recovery is *possible* if $\rho^2 = 1 - o(n^{-4/d})$ and *impossible* if $\rho^2 = 1 - \omega(n^{-4/d})$, assuming $1 \ll d = O(\log n)$.
- [Wang,Wu,Xu,Yolou'22]: Improved the above result by a factor of $\log d$, and hold for any $d \geq 1$.

Prior Work (Correlated Databases)

Known Results and Gaps

- [Dai,Cullina,Kiyavash'19]: Perfect recovery is *possible* if $\rho^2 = 1 - o(n^{-4/d})$ and *impossible* if $\rho^2 = 1 - \omega(n^{-4/d})$, assuming $1 \ll d = O(\log n)$.
- [Wang,Wu,Xu,Yolou'22]: Improved the above result by a factor of $\log d$, and hold for any $d \geq 1$.
- Almost perfect recovery [Dai,Cullina,Kiyavash'20], feature deletions and repetitions [Bakirtas,Erkip'20,21], etc.

Prior Work (Correlated Databases)

Known Results and Gaps

- [Dai,Cullina,Kiyavash'19]: Perfect recovery is *possible* if $\rho^2 = 1 - o(n^{-4/d})$ and *impossible* if $\rho^2 = 1 - \omega(n^{-4/d})$, assuming $1 \ll d = O(\log n)$.
- [Wang,Wu,Xu,Yolou'22]: Improved the above result by a factor of $\log d$, and hold for any $d \geq 1$.
- Almost perfect recovery [Dai,Cullina,Kiyavash'20], feature deletions and repetitions [Bakirtas,Erkip'20,21], etc.
- [Zeynep,Nazer'21,22]: (Efficient) strong detection *possible* if $\rho^2 d \to \infty$, and *impossible* if $\rho^2 d\sqrt{n} \to 0$ and $d = \Omega(\log n)$

Prior Work (Correlated Databases)

Known Results and Gaps

- [Dai,Cullina,Kiyavash'19]: Perfect recovery is *possible* if $\rho^2 = 1 - o(n^{-4/d})$ and *impossible* if $\rho^2 = 1 - \omega(n^{-4/d})$, assuming $1 \ll d = O(\log n)$.
- [Wang,Wu,Xu,Yolou'22]: Improved the above result by a factor of $\log d$, and hold for any $d \geq 1$.
- Almost perfect recovery [Dai,Cullina,Kiyavash'20], feature deletions and repetitions [Bakirtas,Erkip'20,21], etc.
- [Zeynep,Nazer'21,22]: (Efficient) strong detection *possible* if $\rho^2 d \to \infty$, and *impossible* if $\rho^2 d\sqrt{n} \to 0$ and $d = \Omega(\log n)$
- Most notably, there is a $\sqrt{n}$ gap, and upper bound is independent of n.

Prior Work (Correlated Databases)

Known Results and Gaps

- [Dai,Cullina,Kiyavash'19]: Perfect recovery is *possible* if $\rho^2 = 1 - o(n^{-4/d})$ and *impossible* if $\rho^2 = 1 - \omega(n^{-4/d})$, assuming $1 \ll d = O(\log n)$.
- [Wang,Wu,Xu,Yolou'22]: Improved the above result by a factor of $\log d$, and hold for any $d \geq 1$.
- Almost perfect recovery [Dai,Cullina,Kiyavash'20], feature deletions and repetitions [Bakirtas,Erkip'20,21], etc.
- [Zeynep,Nazer'21,22]: (Efficient) strong detection *possible* if $\rho^2 d \to \infty$, and *impossible* if $\rho^2 d\sqrt{n} \to 0$ and $d = \Omega(\log n)$
- Most notably, there is a $\sqrt{n}$ gap, and upper bound is independent of n.
- [Tamir'22,23]: Joint correlation detection and recovery.

Main Results (Correlated Databases)

We show in [Elimelech,Huleihel'23,24]

	Weak Detection		Strong Detection	
Asymptotics	**Possible**	**Impossible**	**Possible**	**Impossible**
$n, d \to \infty$	$\Omega(d^{-1})$	$o(d^{-1})$	$\omega(d^{-1})$	$(1-\varepsilon)d^{-1}$
$d \to \infty$, n constant	$\Omega(d^{-1})$	$o(d^{-1})$	$\omega(d^{-1})$	$O(d^{-1})$
$n \to \infty$, d constant	$\rho^2 = \Omega(1)$	$o(1)$	$1 - o(n^{-\frac{4}{d}})$	$\rho^\star(d)$

- If at least $d \to \infty$, then $\sqrt{n}$ is not needed, namely, upper bound from [Zeynep,Nazer'21,22] is the truth.

Main Results (Correlated Databases)

We show in [Elimelech,Huleihel'23,24]

	Weak Detection		Strong Detection	
Asymptotics	**Possible**	**Impossible**	**Possible**	**Impossible**
$n, d \to \infty$	$\Omega(d^{-1})$	$o(d^{-1})$	$\omega(d^{-1})$	$(1-\varepsilon)d^{-1}$
$d \to \infty$, n constant	$\Omega(d^{-1})$	$o(d^{-1})$	$\omega(d^{-1})$	$O(d^{-1})$
$n \to \infty$, d constant	$\rho^2 = \Omega(1)$	$o(1)$	$1 - o(n^{-\frac{4}{d}})$	$\rho^\star(d)$

- If at least $d \to \infty$, then $\sqrt{n}$ is not needed, namely, upper bound from [Zeynep,Nazer'21,22] is the truth.
- Fixed d is the interesting and more challenging regime.

Main Results (Correlated Databases)

We show in [Elimelech,Huleihel'23,24]

	Weak Detection		Strong Detection	
Asymptotics	**Possible**	**Impossible**	**Possible**	**Impossible**
$n, d \to \infty$	$\Omega(d^{-1})$	$o(d^{-1})$	$\omega(d^{-1})$	$(1-\varepsilon)d^{-1}$
$d \to \infty$, n constant	$\Omega(d^{-1})$	$o(d^{-1})$	$\omega(d^{-1})$	$O(d^{-1})$
$n \to \infty$, d constant	$\rho^2 = \Omega(1)$	$o(1)$	$1 - o(n^{-\frac{4}{d}})$	$\rho^\star(d)$

- If at least $d \to \infty$, then $\sqrt{n}$ is not needed, namely, upper bound from [Zeynep,Nazer'21,22] is the truth.
- Fixed d is the interesting and more challenging regime.
- We use: $d\rho^2 \to 0 \Leftrightarrow \rho^2 = o(d^{-1}) \Leftrightarrow d\rho^2 = o(1)$.

Upper Bounds (or, Algorithms)

Upper Bounds (or, Algorithms)

- **Total sum** [Zeynep,Nazer'21,22]: Threshold the sum of inner-products

$$\phi_{\mathsf{sum}}(\mathsf{X},\mathsf{Y}) \triangleq \mathbb{1}\left\{\sum_{i,j=1}^{n} \mathsf{X}_i^T \mathsf{Y}_j > \frac{dn\rho}{2}\right\}.$$

Upper Bounds (or, Algorithms)

- **Total sum** [Zeynep,Nazer'21,22]: Threshold the sum of inner-products

$$\phi_{\mathsf{sum}}(\mathsf{X},\mathsf{Y}) \triangleq \mathbb{1}\left\{\sum_{i,j=1}^{n} \mathsf{X}_i^T \mathsf{Y}_j > \frac{dn\rho}{2}\right\}.$$

Chernoff's bound gives:

$$\mathsf{R}(\phi_{\mathsf{sum}}) \leq 2 \cdot \exp\left(-\frac{d\rho^2}{60}\right).$$

Upper Bounds (or, Algorithms)

- **Total sum** [Zeynep,Nazer'21,22]: Threshold the sum of inner-products

$$\phi_{\mathsf{sum}}(\mathsf{X},\mathsf{Y}) \triangleq \mathbb{1}\left\{\sum_{i,j=1}^{n} \mathsf{X}_i^T \mathsf{Y}_j > \frac{dn\rho}{2}\right\}.$$

Chernoff's bound gives:

$$\mathsf{R}(\phi_{\mathsf{sum}}) \leq 2 \cdot \exp\left(-\frac{d\rho^2}{60}\right).$$

1. Strong detection if $d\rho^2 = \omega_d(1)$.

Upper Bounds (or, Algorithms)

- **Total sum** [Zeynep,Nazer'21,22]: Threshold the sum of inner-products

$$\phi_{\mathsf{sum}}(\mathsf{X},\mathsf{Y}) \triangleq \mathbb{1}\left\{\sum_{i,j=1}^{n} \mathsf{X}_i^T \mathsf{Y}_j > \frac{dn\rho}{2}\right\}.$$

Chernoff's bound gives:

$$\mathsf{R}(\phi_{\mathsf{sum}}) \leq 2 \cdot \exp\left(-\frac{d\rho^2}{60}\right).$$

1. Strong detection if $d\rho^2 = \omega_d(1)$.
2. Weak detection if $\rho^2 > \frac{60 \log 2}{d}$.

Upper Bounds (or, Algorithms)

- **Total sum** [Zeynep,Nazer'21,22]: Threshold the sum of inner-products

$$\phi_{\mathsf{sum}}(\mathsf{X},\mathsf{Y}) \triangleq \mathbb{1}\left\{\sum_{i,j=1}^{n} \mathsf{X}_i^T \mathsf{Y}_j > \frac{dn\rho}{2}\right\}.$$

Chernoff's bound gives:

$$\mathsf{R}(\phi_{\mathsf{sum}}) \leq 2 \cdot \exp\left(-\frac{d\rho^2}{60}\right).$$

1. Strong detection if $d\rho^2 = \omega_d(1)$.
2. Weak detection if $\rho^2 > \frac{60 \log 2}{d}$.
3. Completely independent of n.

Upper Bounds (or, Algorithms)

- **Total sum** [Zeynep,Nazer'21,22]: Threshold the sum of inner-products

$$\phi_{\mathsf{sum}}(\mathsf{X},\mathsf{Y}) \triangleq \mathbb{1}\left\{\sum_{i,j=1}^{n} \mathsf{X}_i^T \mathsf{Y}_j > \frac{dn\rho}{2}\right\}.$$

Chernoff's bound gives:

$$\mathsf{R}(\phi_{\mathsf{sum}}) \leq 2 \cdot \exp\left(-\frac{d\rho^2}{60}\right).$$

1. Strong detection if $d\rho^2 = \omega_d(1)$.
2. Weak detection if $\rho^2 > \frac{60 \log 2}{d}$.
3. Completely independent of n.
4. **If d is fixed, then strong detection using ϕ_{sum} is not possible.**

Upper Bounds (or, Algorithms)

- **Counting products** [Elimelech,Huleihel'24]: Consider

$$\phi_{\mathsf{count}}(\mathsf{X},\mathsf{Y}) \triangleq \mathbb{1}\left\{\sum_{i,j=1}^{n} \mathbb{1}\left\{\mathsf{L}(\mathsf{X}_i,\mathsf{Y}_j) \geq d\cdot\tau_{\mathsf{count}}\right\} \geq \frac{n\mathcal{P}_d}{2}\right\}$$

where

$$\mathsf{L}(\mathsf{X}_i,\mathsf{Y}_j) \triangleq -\frac{d}{2}\log(1-\rho^2) - \frac{d\rho^2}{2(1-\rho^2)} + \frac{\rho}{1-\rho^2}\mathsf{X}_i^T\mathsf{Y}_j$$

Upper Bounds (or, Algorithms)

- **Counting products** [Elimelech,Huleihel'24]: Consider

$$\phi_{\mathsf{count}}(\mathsf{X},\mathsf{Y}) \triangleq \mathbb{1}\left\{\sum_{i,j=1}^{n} \mathbb{1}\left\{\mathsf{L}(\mathsf{X}_i,\mathsf{Y}_j) \geq d\cdot\tau_{\mathsf{count}}\right\} \geq \frac{n\mathcal{P}_d}{2}\right\}$$

where

$$\mathsf{L}(\mathsf{X}_i,\mathsf{Y}_j) \triangleq -\frac{d}{2}\log(1-\rho^2) - \frac{d\rho^2}{2(1-\rho^2)} + \frac{\rho}{1-\rho^2}\mathsf{X}_i^T\mathsf{Y}_j$$
$$= \log\frac{P_{XY}^{\otimes d}(\mathsf{X}_i,\mathsf{Y}_j)}{Q_{XY}^{\otimes d}(\mathsf{X}_i,\mathsf{Y}_j)}.$$

Upper Bounds (or, Algorithms)

- **Counting products** [Elimelech,Huleihel'24]: Consider

$$\phi_{\mathsf{count}}(\mathsf{X},\mathsf{Y}) \triangleq \mathbb{1}\left\{\sum_{i,j=1}^{n} \mathbb{1}\left\{\mathsf{L}(\mathsf{X}_i,\mathsf{Y}_j) \geq d\cdot\tau_{\mathsf{count}}\right\} \geq \frac{n\mathcal{P}_d}{2}\right\}$$

where

$$\mathsf{L}(\mathsf{X}_i,\mathsf{Y}_j) \triangleq -\frac{d}{2}\log(1-\rho^2) - \frac{d\rho^2}{2(1-\rho^2)} + \frac{\rho}{1-\rho^2}\mathsf{X}_i^T\mathsf{Y}_j$$

Theorem (Count test strong detection)

Fix $d \in \mathbb{N}$. Then, $\mathsf{R}(\phi_{\mathsf{count}}) \to 0$, as $n \to \infty$, if $\rho^2 = 1 - o(n^{-4/d})$.

Upper Bounds (or, Algorithms)

- **Counting products** [Elimelech,Huleihel'24]: Consider

$$\phi_{\mathsf{count}}(\mathsf{X},\mathsf{Y}) \triangleq \mathbb{1}\left\{\sum_{i,j=1}^{n} \mathbb{1}\left\{\mathsf{L}(\mathsf{X}_i,\mathsf{Y}_j) \geq d\cdot\tau_{\mathsf{count}}\right\} \geq \frac{n\mathcal{P}_d}{2}\right\}$$

where

$$\mathsf{L}(\mathsf{X}_i,\mathsf{Y}_j) \triangleq -\frac{d}{2}\log(1-\rho^2) - \frac{d\rho^2}{2(1-\rho^2)} + \frac{\rho}{1-\rho^2}\mathsf{X}_i^T\mathsf{Y}_j$$

Theorem (Count test strong detection)

Fix $d \in \mathbb{N}$. Then, $\mathsf{R}(\phi_{\mathsf{count}}) \to 0$, as $n \to \infty$, if $\rho^2 = 1 - o(n^{-4/d})$.

1. Coincides with the recovery threshold (via ML).

Upper Bounds (or, Algorithms)

- **Counting products** [Elimelech,Huleihel'24]: Consider

$$\phi_{\mathsf{count}}(\mathsf{X},\mathsf{Y}) \triangleq \mathbb{1}\left\{\sum_{i,j=1}^{n} \mathbb{1}\left\{\mathsf{L}(\mathsf{X}_i,\mathsf{Y}_j) \geq d\cdot\tau_{\mathsf{count}}\right\} \geq \frac{n\mathcal{P}_d}{2}\right\}$$

where

$$\mathsf{L}(\mathsf{X}_i,\mathsf{Y}_j) \triangleq -\frac{d}{2}\log(1-\rho^2) - \frac{d\rho^2}{2(1-\rho^2)} + \frac{\rho}{1-\rho^2}\mathsf{X}_i^T\mathsf{Y}_j$$

Theorem (Count test strong detection)

Fix $d \in \mathbb{N}$. Then, $\mathsf{R}(\phi_{\mathsf{count}}) \to 0$, as $n \to \infty$, if $\rho^2 = 1 - o(n^{-4/d})$.

1. Coincides with the recovery threshold (via ML).
2. Decay rate is not optimal.

Upper Bounds (or, Algorithms)

- **Counting products** [Elimelech,Huleihel'24]: Consider

$$\phi_{\mathsf{count}}(\mathsf{X},\mathsf{Y}) \triangleq \mathbb{1}\left\{\sum_{i,j=1}^{n} \mathbb{1}\left\{\mathsf{L}(\mathsf{X}_i,\mathsf{Y}_j) \geq d\cdot\tau_{\mathsf{count}}\right\} \geq \frac{n\mathcal{P}_d}{2}\right\}$$

Proof sketch: first moment

$$\mathbb{P}_{\mathcal{H}_0}\left(\phi_{\mathsf{count}} = 1\right) = \mathbb{P}_{\mathcal{H}_0}\left(\sum_{i,j=1}^{n} \mathsf{G}_{ij} \geq \frac{n\mathcal{P}_d}{2}\right) \leq \frac{2n\mathcal{Q}_d}{\mathcal{P}_d},$$

where

$$\mathcal{Q}_d \triangleq \mathbb{P}_{\mathcal{N}^{\otimes d}(0,\mathbf{I})}[\mathsf{L}(\mathsf{A},\mathsf{B}) \geq d\cdot\tau_{\mathsf{count}}] \leq e^{-d\cdot E_Q(\tau_{\mathsf{count}})}$$

$$\mathcal{P}_d \triangleq \mathbb{P}_{\mathcal{N}^{\otimes d}(0,\Sigma_\rho)}[\mathsf{L}(\mathsf{A},\mathsf{B}) \geq d\cdot\tau_{\mathsf{count}}] \geq 1 - e^{-d\cdot E_P(\tau_{\mathsf{count}})}$$

Upper Bounds (or, Algorithms)

- **Counting products** [Elimelech,Huleihel'24]: Consider

$$\phi_{\text{count}}(\mathsf{X},\mathsf{Y}) \triangleq \mathbb{1}\left\{\sum_{i,j=1}^{n} \mathbb{1}\left\{\mathsf{L}(\mathsf{X}_i,\mathsf{Y}_j) \geq d\cdot\tau_{\text{count}}\right\} \geq \frac{n\mathcal{P}_d}{2}\right\}$$

Proof sketch: second moment (w.l.o.g. $\pi = \mathsf{Id}$),

$$\begin{aligned}
\mathbb{P}_{\mathcal{H}_1}(\phi_{\text{count}} = 0) &= \mathbb{P}_{\mathcal{H}_1}\left(\sum_{i,j=1}^{n} \mathsf{G}_{ij} < \frac{n\mathcal{P}_d}{2}\right) \\
&\leq \mathbb{P}_{\mathcal{H}_1}\left(\sum_{i=1}^{n} \mathsf{G}_{ii} < \frac{n\mathcal{P}_d}{2}\right) \\
&\leq \frac{4\cdot \mathsf{Var}_\rho\left(\sum_{i=1}^{n} \mathsf{G}_{ii}\right)}{n^2\mathcal{P}_\rho^2} = \frac{4(1-\mathcal{P}_d)}{n\mathcal{P}_d} \leq \frac{4}{n\mathcal{P}_d}.
\end{aligned}$$

Upper Bounds (or, Algorithms)

- **Comparison test** [Elimelech,Huleihel'24]: Define,

$$\phi_{\mathsf{comp}}(\mathsf{X},\mathsf{Y}) \triangleq \mathbb{1}\left\{\left|\sum_{i,j}(\mathsf{X}_{ij}-\mathsf{Y}_{ij})\right| \leq \theta\right\}$$

Upper Bounds (or, Algorithms)

- **Comparison test** [Elimelech,Huleihel'24]: Define,

$$\phi_{\mathsf{comp}}(\mathsf{X},\mathsf{Y}) \triangleq \mathbb{1}\left\{\left|\sum_{i,j}(\mathsf{X}_{ij}-\mathsf{Y}_{ij})\right| \leq \theta\right\}$$

Take θ as the value for which

$$d_{\mathsf{TV}}\left(\mathcal{N}(0,1),\mathcal{N}\left(0,1-|\rho|\right)\right) = \mathbb{P}\left(|\mathsf{G}| \geq \frac{\theta}{\sqrt{2nd}}\right) - \mathbb{P}\left(|\mathsf{G}'| \geq \frac{\theta}{\sqrt{2nd}}\right),$$

where $\mathsf{G} \sim \mathcal{N}(0,1)$ and $\mathsf{G}' \sim \mathcal{N}\left(0,1-|\rho|\right)$.

Upper Bounds (or, Algorithms)

- **Comparison test [Elimelech,Huleihel'24]**: Define,

$$\phi_{\mathsf{comp}}(\mathsf{X},\mathsf{Y}) \triangleq \mathbb{1}\left\{\left|\sum_{i,j}(\mathsf{X}_{ij}-\mathsf{Y}_{ij})\right| \leq \theta\right\}$$

Take θ as the value for which

$$\begin{aligned} & d_{\mathsf{TV}}\left(\mathcal{N}(0,1),\mathcal{N}\left(0,1-|\rho|\right)\right) \\ & \quad = \mathbb{P}\left(|\mathsf{G}| \geq \frac{\theta}{\sqrt{2nd}}\right) - \mathbb{P}\left(|\mathsf{G}'| \geq \frac{\theta}{\sqrt{2nd}}\right), \end{aligned}$$

where $\mathsf{G} \sim \mathcal{N}(0,1)$ and $\mathsf{G}' \sim \mathcal{N}\left(0,1-|\rho|\right)$.

Theorem

Fix $d \in \mathbb{N}$. *If* $\rho^2 = \Omega(1)$ *then* $\lim_{n\to\infty} \mathsf{R}(\phi_{\mathsf{comp}}) < 1$.

Upper Bounds (or, Algorithms)

- **Comparison test** [Elimelech,Huleihel'24]: Define,

$$\phi_{\mathsf{comp}}(\mathsf{X},\mathsf{Y}) \triangleq \mathbb{1}\left\{\left|\sum_{i,j}(\mathsf{X}_{ij}-\mathsf{Y}_{ij})\right| \leq \theta\right\}$$

Proof sketch: Let $\mathsf{G}_1 \triangleq \sum_{ij} \mathsf{X}_{ij}$ and $\mathsf{G}_2 \triangleq \sum_{ij} \mathsf{Y}_{ij}$. Then, $\mathsf{G}_1 - \mathsf{G}_2 \overset{\mathcal{H}_0}{\sim} \mathcal{N}(0, 2nd)$ and $\mathsf{G}_1 - \mathsf{G}_2 \overset{\mathcal{H}_1}{\sim} \mathcal{N}(0, 2nd(1-\rho))$. Therefore,

$$\begin{aligned} 1 - \mathsf{R}(\phi_{\mathsf{comp}}) &= \mathbb{P}_{\mathcal{H}_0}(|\mathsf{G}_1 - \mathsf{G}_2| \geq \theta) - \mathbb{P}_{\mathcal{H}_1}(|\mathsf{G}_1 - \mathsf{G}_2| \geq \theta) \\ &= \mathbb{P}(|\mathcal{N}(0, 2nd)| \geq \theta) \\ &\quad - \mathbb{P}(|\mathcal{N}(0, 2n(1-\rho))| \geq \theta) \\ &= d_{\mathsf{TV}}\left(\mathcal{N}(0,1), \mathcal{N}\left(0, 1-\rho\right)\right) = \Omega(1). \end{aligned}$$

Lower Bound ($d \to \infty$)

We start with the regime where at least $d \to \infty$.

Lower Bound ($d \to \infty$)

We start with the regime where at least $d \to \infty$.

Second moment calculation: let $\mathsf{L}_n(\mathsf{X}, \mathsf{Y}) \triangleq \frac{\mathbb{P}_{\mathcal{H}_1}(\mathsf{X},\mathsf{Y})}{\mathbb{P}_{\mathcal{H}_0}(\mathsf{X},\mathsf{Y})}$, then

$$\boxed{\mathsf{R}^\star = 1 - d_{\mathsf{TV}}(\mathbb{P}_{\mathcal{H}_1}, \mathbb{P}_{\mathcal{H}_0})}$$

$$\mathbb{E}_{\mathcal{H}_0}\left[\mathsf{L}_n^2\right] = O(1) \implies d_{\mathsf{TV}}(\mathbb{P}_{\mathcal{H}_1}, \mathbb{P}_{\mathcal{H}_0}) \leq 1 - \Omega(1)$$

$$\mathbb{E}_{\mathcal{H}_0}\left[\mathsf{L}_n^2\right] = 1 + o(1) \implies d_{\mathsf{TV}}(\mathbb{P}_{\mathcal{H}_1}, \mathbb{P}_{\mathcal{H}_0}) \leq o(1)$$

Lower Bound ($d \to \infty$)

We start with the regime where at least $d \to \infty$.

Second moment calculation: let $\mathsf{L}_n(\mathsf{X}, \mathsf{Y}) \triangleq \frac{\mathbb{P}_{\mathcal{H}_1}(\mathsf{X},\mathsf{Y})}{\mathbb{P}_{\mathcal{H}_0}(\mathsf{X},\mathsf{Y})}$, then

$$\boxed{\mathsf{R}^\star = 1 - d_{\mathsf{TV}}(\mathbb{P}_{\mathcal{H}_1}, \mathbb{P}_{\mathcal{H}_0})}$$

$$\begin{aligned}
&\mathbb{E}_{\mathcal{H}_0}\left[\mathsf{L}_n^2\right] = O(1) \\
&\qquad\qquad \implies d_{\mathsf{TV}}(\mathbb{P}_{\mathcal{H}_1}, \mathbb{P}_{\mathcal{H}_0}) \leq 1 - \Omega(1) \\
&\mathbb{E}_{\mathcal{H}_0}\left[\mathsf{L}_n^2\right] = 1 + o(1) \\
&\qquad\qquad \implies d_{\mathsf{TV}}(\mathbb{P}_{\mathcal{H}_1}, \mathbb{P}_{\mathcal{H}_0}) \leq o(1)
\end{aligned}$$

Thus, it is suffice to analyze the second moment of the likelihood.

Lower Bound ($d \to \infty$)

Recall that

$$\mathsf{L}_n(\mathsf{X},\mathsf{Y}) = \frac{\mathbb{P}_{\mathcal{H}_1}(\mathsf{X},\mathsf{Y})}{\mathbb{P}_{\mathcal{H}_0}(\mathsf{X},\mathsf{Y})} = \frac{\mathbb{E}_\pi[\mathbb{P}_{\mathcal{H}_1|\pi}(\mathsf{X},\mathsf{Y})]}{\mathbb{P}_{\mathcal{H}_0}(\mathsf{X},\mathsf{Y})} = \mathbb{E}_\pi\left[\frac{\mathbb{P}_{\mathcal{H}_1|\pi}(\mathsf{X},\mathsf{Y})}{\mathbb{P}_{\mathcal{H}_0}(\mathsf{X},\mathsf{Y})}\right].$$

Lower Bound ($d \to \infty$)

Recall that

$$\begin{aligned} \mathsf{L}_n(\mathsf{X},\mathsf{Y}) &= \frac{\mathbb{P}_{\mathcal{H}_1}(\mathsf{X},\mathsf{Y})}{\mathbb{P}_{\mathcal{H}_0}(\mathsf{X},\mathsf{Y})} \\ &= \frac{\mathbb{E}_\pi[\mathbb{P}_{\mathcal{H}_1|\pi}(\mathsf{X},\mathsf{Y})]}{\mathbb{P}_{\mathcal{H}_0}(\mathsf{X},\mathsf{Y})} = \mathbb{E}_\pi\left[\frac{\mathbb{P}_{\mathcal{H}_1|\pi}(\mathsf{X},\mathsf{Y})}{\mathbb{P}_{\mathcal{H}_0}(\mathsf{X},\mathsf{Y})}\right]. \end{aligned}$$

Then,

$$[\mathsf{L}_n]^2 = \mathbb{E}_{\pi \perp\!\!\!\perp \pi'}\left[\frac{\mathbb{P}_{\mathcal{H}_1|\pi}}{\mathbb{P}_{\mathcal{H}_0}} \cdot \frac{\mathbb{P}_{\mathcal{H}_1|\pi'}}{\mathbb{P}_{\mathcal{H}_0}}\right].$$

Lower Bound ($d \to \infty$)

Recall that

$$\mathsf{L}_n(\mathsf{X}, \mathsf{Y}) = \frac{\mathbb{P}_{\mathcal{H}_1}(\mathsf{X}, \mathsf{Y})}{\mathbb{P}_{\mathcal{H}_0}(\mathsf{X}, \mathsf{Y})}$$
$$= \frac{\mathbb{E}_\pi[\mathbb{P}_{\mathcal{H}_1|\pi}(\mathsf{X}, \mathsf{Y})]}{\mathbb{P}_{\mathcal{H}_0}(\mathsf{X}, \mathsf{Y})} = \mathbb{E}_\pi \left[\frac{\mathbb{P}_{\mathcal{H}_1|\pi}(\mathsf{X}, \mathsf{Y})}{\mathbb{P}_{\mathcal{H}_0}(\mathsf{X}, \mathsf{Y})} \right].$$

Then,

$$[\mathsf{L}_n]^2 = \mathbb{E}_{\pi \perp\!\!\!\perp \pi'} \left[\frac{\mathbb{P}_{\mathcal{H}_1|\pi}}{\mathbb{P}_{\mathcal{H}_0}} \cdot \frac{\mathbb{P}_{\mathcal{H}_1|\pi'}}{\mathbb{P}_{\mathcal{H}_0}} \right].$$

Thus, Ingster-Suslina method (Fubini's theorem)

$$\mathbb{E}_{\mathcal{H}_0}[\mathsf{L}_n^2] = \mathbb{E}_{\pi \perp\!\!\!\perp \pi'} \left[\mathbb{E}_{\mathcal{H}_0} \left[\frac{\mathbb{P}_{\mathcal{H}_1|\pi}}{\mathbb{P}_{\mathcal{H}_0}} \cdot \frac{\mathbb{P}_{\mathcal{H}_1|\pi'}}{\mathbb{P}_{\mathcal{H}_0}} \right] \right].$$

Lower Bound $(d \to \infty)$

Invariance: fix $\pi' = \mathsf{Id}$,

$$\mathbb{E}_{\mathcal{H}_0}[\mathsf{L}_n^2] = \mathbb{E}_\pi \left[\mathbb{E}_{\mathcal{H}_0} \left[\frac{\mathbb{P}_{\mathcal{H}_1|\pi}}{\mathbb{P}_{\mathcal{H}_0}} \cdot \frac{\mathbb{P}_{\mathcal{H}_1|\mathsf{Id}}}{\mathbb{P}_{\mathcal{H}_0}} \right] \right].$$

Lower Bound ($d \to \infty$)

Invariance: fix $\pi' = \mathsf{Id}$,

$$\mathbb{E}_{\mathcal{H}_0}[\mathsf{L}_n^2] = \mathbb{E}_\pi \left[\mathbb{E}_{\mathcal{H}_0} \left[\frac{\mathbb{P}_{\mathcal{H}_1|\pi}}{\mathbb{P}_{\mathcal{H}_0}} \cdot \frac{\mathbb{P}_{\mathcal{H}_1|\mathsf{Id}}}{\mathbb{P}_{\mathcal{H}_0}} \right] \right].$$

Recall that pairs $\{(\mathsf{X}_i, \mathsf{Y}_{\pi_i})\}_{i \in [n]}$ are i.i.d.,

$$\frac{\mathbb{P}_{\mathcal{H}_1|\pi}(\mathsf{X}, \mathsf{Y})}{\mathbb{P}_{\mathcal{H}_0}(\mathsf{X}, \mathsf{Y})} = \prod_{i=1}^{n} \mathsf{L}(\mathsf{X}_i, \mathsf{Y}_{\pi_i})$$

$$\frac{\mathbb{P}_{\mathcal{H}_1|\mathsf{Id}}(\mathsf{X}, \mathsf{Y})}{\mathbb{P}_{\mathcal{H}_0}(\mathsf{X}, \mathsf{Y})} = \prod_{i=1}^{n} \mathsf{L}(\mathsf{X}_i, \mathsf{Y}_i),$$

where $\mathsf{L}(\mathsf{X}_i, \mathsf{Y}_i) \triangleq \frac{P_{XY}^{\otimes d}(\mathsf{X}_i, \mathsf{Y}_i)}{Q_{XY}^{\otimes d}(\mathsf{X}_i, \mathsf{Y}_i)}$.

Lower Bound ($d \to \infty$)

Invariance: fix $\pi' = \mathsf{Id}$,

$$\mathbb{E}_{\mathcal{H}_0}[\mathsf{L}_n^2] = \mathbb{E}_\pi \left[\mathbb{E}_{\mathcal{H}_0} \left[\frac{\mathbb{P}_{\mathcal{H}_1|\pi}}{\mathbb{P}_{\mathcal{H}_0}} \cdot \frac{\mathbb{P}_{\mathcal{H}_1|\mathsf{Id}}}{\mathbb{P}_{\mathcal{H}_0}} \right] \right].$$

Thus,

$$\frac{\mathbb{P}_{\mathcal{H}_1|\pi}}{\mathbb{P}_{\mathcal{H}_0}} \cdot \frac{\mathbb{P}_{\mathcal{H}_1|\mathsf{Id}}}{\mathbb{P}_{\mathcal{H}_0}} = \prod_{i=1}^n \mathsf{L}(\mathsf{X}_i, \mathsf{Y}_{\pi_i}) \mathsf{L}(\mathsf{X}_i, \mathsf{Y}_i) \triangleq \prod_{i=1}^n \mathsf{Z}_i$$

Lower Bound ($d \to \infty$)

Invariance: fix $\pi' = \mathsf{Id}$,

$$\mathbb{E}_{\mathcal{H}_0}[\mathsf{L}_n^2] = \mathbb{E}_\pi \left[\mathbb{E}_{\mathcal{H}_0} \left[\frac{\mathbb{P}_{\mathcal{H}_1|\pi}}{\mathbb{P}_{\mathcal{H}_0}} \cdot \frac{\mathbb{P}_{\mathcal{H}_1|\mathsf{Id}}}{\mathbb{P}_{\mathcal{H}_0}} \right] \right].$$

Accordingly,

$$\mathbb{E}_{\mathcal{H}_0}[\mathsf{L}_n^2] = \mathbb{E}_\pi \left[\mathbb{E}_{\mathcal{H}_0} \left(\prod_{i=1}^{n} \mathsf{Z}_i \right) \right].$$

Problem: $\{\mathsf{Z}_i\}_{i=1}^n$ are dependent random variables

Solution: cycle decomposition!

Lower Bound ($d \to \infty$)

Facts on cycles (orbits)

- For each element $a \in [n]$, its orbit is a cycle $(a_0, \ldots, a_{k-1})$, where $a_i = \pi^i(a)$, for $i = 0, \ldots, k-1$ and $\pi(a_{k-1}) = a$.

  For example: Consider $\pi \in \mathbb{S}_7$ that

  1. Keeps 1 in the same place
  2. Swaps 2 with 3
  3. Cyclically shifts 4567

  Then, π consists of <u>three</u> orbits in canonical notation

$$\pi = (1)(23)(4567)$$

Lower Bound ($d \to \infty$)

Facts on cycles (orbits)

- For each element $a \in [n]$, its orbit is a cycle $(a_0, \ldots, a_{k-1})$, where $a_i = \pi^i(a)$, for $i = 0, \ldots, k-1$ and $\pi(a_{k-1}) = a$.
- If $|O| = k$, we call O a k-orbit.
- Set of orbits of a permutation induce a partition of $[n]$

Let $\{O\}_{O \in \mathcal{O}}$ be the orbit/cycle decomposition of π. For $O \in \mathcal{O}$,

$$\mathsf{Z}_O \triangleq \prod_{i \in O} \mathsf{Z}_i \quad \Longrightarrow \quad \prod_{i=1}^{n} \mathsf{Z}_i = \prod_{O \in \mathcal{O}} \mathsf{Z}_O$$

The random variables $\{Z_O\}_O$ are independent (under $\mathbb{P}_{\mathcal{H}_0}$),

$$\mathbb{E}_{\mathcal{H}_0}[\mathsf{L}_n^2] = \mathbb{E}_\pi \mathbb{E}_{\mathcal{H}_0}\left[\prod_{i=1}^{n} \mathsf{Z}_i\right] = \mathbb{E}_\pi \mathbb{E}_{\mathcal{H}_0}\left[\prod_{O \in \mathcal{O}} \mathsf{Z}_O\right] = \mathbb{E}_\pi \prod_{O \in \mathcal{O}} \mathbb{E}_{\mathcal{H}_0}[\mathsf{Z}_O].$$

Lower Bound ($d \to \infty$)

For a fixed orbit O of a permutation π,

$$\mathbb{E}_{\mathcal{H}_0}[\mathsf{Z}_O] = \frac{1}{(1-\rho^{2|O|})^d}.$$

Lower Bound ($d \to \infty$)

For a fixed orbit O of a permutation π,

$$\mathbb{E}_{\mathcal{H}_0}[\mathsf{Z}_O] = \frac{1}{(1-\rho^{2|O|})^d}.$$

If $N_k(\pi)$ is the number of k-orbits of π, then

$$\mathbb{E}_0[\mathsf{L}_n^2] = \mathbb{E}_\pi\left[\prod_C \mathbb{E}_{\mathcal{H}_0}[\mathsf{Z}_O]\right] = \mathbb{E}_\pi\left[\prod_{k=1}^n \frac{1}{(1-\rho^{2k})^{d\cdot N_k}}\right].$$

Lower Bound ($d \to \infty$)

For a fixed orbit O of a permutation π,

$$\mathbb{E}_{\mathcal{H}_0}[\mathsf{Z}_O] = \frac{1}{(1-\rho^{2|O|})^d}.$$

If $N_k(\pi)$ is the number of k-orbits of π, then

$$\mathbb{E}_0[\mathsf{L}_n^2] = \mathbb{E}_\pi\left[\prod_C \mathbb{E}_{\mathcal{H}_0}[\mathsf{Z}_O]\right] = \mathbb{E}_\pi\left[\prod_{k=1}^n \frac{1}{(1-\rho^{2k})^{d\cdot N_k}}\right].$$

Use statistical properties of k-orbits of $\pi \sim \mathsf{Unif}(\mathbb{S}_n)$.

Lower Bound ($d \to \infty$)

For a fixed orbit O of a permutation π,

$$\mathbb{E}_{\mathcal{H}_0}[\mathsf{Z}_O] = \frac{1}{(1-\rho^{2|O|})^d}.$$

If $N_k(\pi)$ is the number of k-orbits of π, then

$$\mathbb{E}_0[\mathsf{L}_n^2] = \mathbb{E}_\pi\left[\prod_C \mathbb{E}_{\mathcal{H}_0}[\mathsf{Z}_O]\right] = \mathbb{E}_\pi\left[\prod_{k=1}^n \frac{1}{(1-\rho^{2k})^{d\cdot N_k}}\right].$$

In particular, [Arratia,Tavaré'92]

$$d_{\mathsf{TV}}\left(\mathcal{L}\left(N_1, N_2, \ldots, N_k\right), \mathcal{L}\left(P_1, P_2, \ldots, P_k\right)\right) \leq F\left(\frac{n}{k}\right),$$

for any $1 \leq k \leq n$, and $\{P_i\}_{i=1}^n$ independent sequence with $P_i \sim$ Poisson $\left(i^{-1}\right)$, and $\log F(x) = -x \log x(1+o(1))$ as $x \to \infty$

Lower Bound ($d \to \infty$)

In the Poisson world, for any m,

$$\mathbb{E}\left[\prod_{k=1}^{m} \frac{1}{(1-\rho^{2k})^{d \cdot P_k}}\right] \le \exp\left(\frac{d\rho^2}{1-\rho^2} + \frac{c(d,\rho^2)\rho^4}{1-\rho^4}\right)$$

Lower Bound ($d \to \infty$)

In the Poisson world, for any m,

$$\mathbb{E}\left[\prod_{k=1}^{m} \frac{1}{(1-\rho^{2k})^{d \cdot P_k}}\right] \leq \exp\left(\frac{d\rho^2}{1-\rho^2} + \frac{c(d,\rho^2)\rho^4}{1-\rho^4}\right)$$

Decompose,

$$\prod_{k=1}^{n} \frac{1}{(1-\rho^{2k})^{d \cdot N_k}} = \prod_{k=1}^{\lceil \log n \rceil} \frac{1}{(1-\rho^{2k})^{d \cdot N_k}} \prod_{k=\lceil \log n \rceil + 1}^{n} \frac{1}{(1-\rho^{2k})^{d \cdot N_k}}$$

Lower Bound ($d \to \infty$)

In the Poisson world, for any m,

$$\mathbb{E}\left[\prod_{k=1}^{m} \frac{1}{(1-\rho^{2k})^{d \cdot P_k}}\right] \leq \exp\left(\frac{d\rho^2}{1-\rho^2} + \frac{c(d,\rho^2)\rho^4}{1-\rho^4}\right)$$

Decompose,

$$\prod_{k=1}^{n} \frac{1}{(1-\rho^{2k})^{d \cdot N_k}} = \prod_{k=1}^{\lceil \log n \rceil} \frac{1}{(1-\rho^{2k})^{d \cdot N_k}} \prod_{k=\lceil \log n \rceil + 1}^{n} \frac{1}{(1-\rho^{2k})^{d \cdot N_k}}$$

For the tail ($m = \lceil \log n \rceil$),

$$\prod_{k=m+1}^{n} \frac{1}{(1-\rho^{2k})^{d \cdot N_k}} \leq \left(\frac{1}{1-\rho^{2m}}\right)^{d \sum_{k=m}^{n} N_k}$$

$$= \left(\frac{1}{1-\rho^{2m}}\right)^{dn} \leq \exp\left(\frac{dn\rho^{2m}}{1-\rho^{2m}}\right) = 1 + o(1),$$

for $d\rho^2 = o(1)$.

Lower Bound ($d \to \infty$)

In the Poisson world, for any m,

$$\mathbb{E}\left[\prod_{k=1}^{m} \frac{1}{(1-\rho^{2k})^{d \cdot P_k}}\right] \leq \exp\left(\frac{d\rho^2}{1-\rho^2} + \frac{c(d,\rho^2)\rho^4}{1-\rho^4}\right)$$

Decompose,

$$\prod_{k=1}^{n} \frac{1}{(1-\rho^{2k})^{d \cdot N_k}} = \prod_{k=1}^{\lceil \log n \rceil} \frac{1}{(1-\rho^{2k})^{d \cdot N_k}} \prod_{k=\lceil \log n \rceil+1}^{n} \frac{1}{(1-\rho^{2k})^{d \cdot N_k}}$$

Thus,

$$\prod_{k=1}^{n} \frac{1}{(1-\rho^{2k})^{d \cdot N_k}} = (1+o(1)) \cdot \prod_{k=1}^{\lceil \log n \rceil} \frac{1}{(1-\rho^{2k})^{d \cdot N_k}}$$

Lower Bound ($d \to \infty$)

In the Poisson world, for any m,

$$\mathbb{E}\left[\prod_{k=1}^{m} \frac{1}{(1-\rho^{2k})^{d \cdot P_k}}\right] \leq \exp\left(\frac{d\rho^2}{1-\rho^2} + \frac{c(d,\rho^2)\rho^4}{1-\rho^4}\right)$$

Now,

$$\mathbb{E}_\pi\left[\prod_{k=1}^{m}\left(\frac{1}{1-\rho^{2k}}\right)^{dN_k}\right] \leq \mathbb{E}_\pi\left[\prod_{k=1}^{m}\left(\frac{1}{1-\rho^{2k}}\right)^{dP_k}\right] + d_{\mathsf{TV}}\left(\mathcal{L}\left(N_1^m\right), \mathcal{L}\left(P_1^m\right)\right) \cdot \left(\frac{1}{1-\rho^2}\right)^{dn}$$

Lower Bound ($d \to \infty$)

In the Poisson world, for any m,

$$\mathbb{E}\left[\prod_{k=1}^{m} \frac{1}{(1-\rho^{2k})^{d \cdot P_k}}\right] \leq \exp\left(\frac{d\rho^2}{1-\rho^2} + \frac{c(d,\rho^2)\rho^4}{1-\rho^4}\right)$$

Now,

$$\begin{aligned}
\mathbb{E}_\pi\left[\prod_{k=1}^{m}\left(\frac{1}{1-\rho^{2k}}\right)^{dN_k}\right] &\leq \mathbb{E}_\pi\left[\prod_{k=1}^{m}\left(\frac{1}{1-\rho^{2k}}\right)^{dP_k}\right] \\
&\quad + d_{\mathsf{TV}}\left(\mathcal{L}\left(N_1^m\right), \mathcal{L}\left(P_1^m\right)\right) \cdot \left(\frac{1}{1-\rho^2}\right)^{dn} \\
&\leq \exp\left(\frac{d\rho^2}{1-\rho^2} + \frac{c(d,\rho^2)\rho^4}{1-\rho^4}\right) \\
&\quad + F\left(\frac{n}{\lceil \log n \rceil}\right)\left(\frac{1}{1-\rho^2}\right)^{dn} \\
&= 1 + o(1),
\end{aligned}$$

if $d\rho^2 = o(1)$.

Finite d: Detecting Correlated Vectors [Elimelech, H'24]

When d is fixed, and $n \to \infty$, the above technique gives

Theorem (Impossibility)

Strong detection is impossible if $d < \frac{\log(\rho^2)}{\log(1-\rho^2)}$.

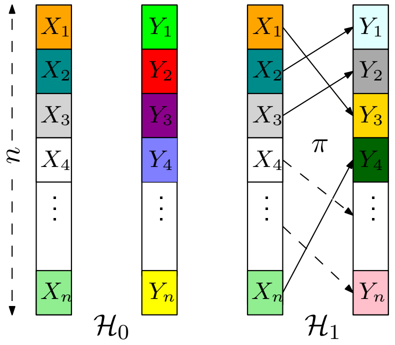
Finite d: Detecting Correlated Vectors [Elimelech, H'24]

When d is fixed, and $n \to \infty$, the above technique gives

Theorem (Impossibility)

Strong detection is impossible if $d < \frac{\log(\rho^2)}{\log(1-\rho^2)}$.

Consider the simple case of $d = 1$,



Finite d: Detecting Correlated Vectors [Elimelech, H'24]

When d is fixed, and $n \to \infty$, the above technique gives

Theorem (Impossibility)

Strong detection is impossible if $d < \frac{\log(\rho^2)}{\log(1-\rho^2)}$.

Lower bound: for $d = 1$, we get the condition $\rho^2 < 1/2$.

Finite d: Detecting Correlated Vectors [Elimelech, H'24]

When d is fixed, and $n \to \infty$, the above technique gives

Theorem (Impossibility)

Strong detection is impossible if $d < \frac{\log(\rho^2)}{\log(1-\rho^2)}$.

Lower bound: for $d = 1$, we get the condition $\rho^2 < 1/2$.

Upper bound is $\rho^2 = 1 - o(n^{-4})$.

Finite d: Detecting Correlated Vectors [Elimelech, H'24]

When d is fixed, and $n \to \infty$, the above technique gives

Theorem (Impossibility)

Strong detection is impossible if $d < \frac{\log(\rho^2)}{\log(1-\rho^2)}$.

Lower bound: for $d = 1$, we get the condition $\rho^2 < 1/2$.

Upper bound is $\rho^2 = 1 - o(n^{-4})$.

What is the source for this significant gap? Computational?

Finite d: Detecting Correlated Vectors [Elimelech, H'24]

When d is fixed, and $n \to \infty$, the above technique gives

Theorem (Impossibility)

Strong detection is impossible if $d < \frac{\log(\rho^2)}{\log(1-\rho^2)}$.

Lower bound: for $d = 1$, we get the condition $\rho^2 < 1/2$.

Upper bound is $\rho^2 = 1 - o(n^{-4})$.

What is the source for this significant gap? Computational?

Not clear yet! But, we can prove a better lower bound.

Finite d: Detecting Correlated Vectors [Elimelech, H'24]

We have the following result.

Theorem (Impossibility for $d = 1$)

Strong detection is impossible for any $\rho^2 < 1$.

Finite d: Detecting Correlated Vectors [Elimelech, H'24]

Proof sketch: We use polynomial decomposition:

Finite d: Detecting Correlated Vectors [Elimelech, H'24]

Proof sketch: We use polynomial decomposition:

- We want to analyze $\mathbb{E}_{\mathcal{H}_0}[\mathsf{L}_n^2]$.

Finite d: Detecting Correlated Vectors [Elimelech, H'24]

Proof sketch: We use polynomial decomposition:

- We want to analyze $\mathbb{E}_{\mathcal{H}_0}[\mathsf{L}_n^2]$.
- Idea: decompose L_n into its orthogonal components.

Finite d: Detecting Correlated Vectors [Elimelech, H'24]

Proof sketch: We use polynomial decomposition:

- We want to analyze $\mathbb{E}_{\mathcal{H}_0}[\mathsf{L}_n^2]$.
- Idea: decompose L_n into its orthogonal components.
- Univariate Hermite polynomials: for $k \in \mathbb{N}$,

$$h_k(x) \triangleq (-1)^k e^{x^2/2} \frac{d^k}{dx^k} e^{-x^2/2},$$

  are orthonormal w.r.t. the standard Gaussian measure,

$$\mathbb{E}_{X \sim\sim N(0,1)} \left[h_k(X) h_\ell(X) \right] = \delta[k - \ell].$$

Finite d: Detecting Correlated Vectors [Elimelech, H'24]

Proof sketch: We use polynomial decomposition:

- We want to analyze $\mathbb{E}_{\mathcal{H}_0}[\mathsf{L}_n^2]$.
- Idea: decompose L_n into its orthogonal components.
- Multivariate Hermite polynomials:
  Let $H_\theta(x) = \prod_{i=1}^n h_{\theta_i}(x_i)$ for $\theta \in \mathbb{N}^n$, and it holds

$$\mathbb{E}_{\mathsf{X}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[H_\alpha(\mathsf{X})H_\gamma(\mathsf{X})\right] = \delta[\alpha - \gamma].$$

Finite d: Detecting Correlated Vectors [Elimelech, H'24]

Proof sketch: We use polynomial decomposition:

- We want to analyze $\mathbb{E}_{\mathcal{H}_0}[\mathsf{L}_n^2]$.
- Idea: decompose L_n into its orthogonal components.
- Multivariate Hermite polynomials:
  Let $H_\theta(x) = \prod_{i=1}^n h_{\theta_i}(x_i)$ for $\theta \in \mathbb{N}^n$, and it holds

$$\mathbb{E}_{\mathsf{X}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[H_\alpha(\mathsf{X})H_\gamma(\mathsf{X})\right] = \delta[\alpha-\gamma].$$

- Form a complete orthonormal system in $L^2(\mathcal{H}_0)$,

$$\mathsf{L}_n(\mathsf{X},\mathsf{Y}) = \sum_{\alpha,\beta\in\mathbb{N}^n} \left\langle H_{\alpha,\beta}(\mathsf{X},\mathsf{Y}), \mathsf{L}_n(\mathsf{X},\mathsf{Y})\right\rangle_{\mathcal{H}_0} H_{\alpha,\beta}(\mathsf{X},\mathsf{Y}),$$

  where $H_{\alpha,\beta}(\mathsf{X},\mathsf{Y}) \triangleq H_\alpha(\mathsf{X})H_\beta(\mathsf{Y})$, and

$$\langle\phi,\psi\rangle_{\mathcal{H}_0} \triangleq \mathbb{E}_{\mathcal{H}_0}\left[\psi(\mathsf{X},\mathsf{Y})\cdot\phi(\mathsf{X},\mathsf{Y})\right].$$

Finite d: Detecting Correlated Vectors [Elimelech, H'24]

Proof sketch: We use polynomial decomposition:

- We want to analyze $\mathbb{E}_{\mathcal{H}_0}[\mathsf{L}_n^2]$.
- Idea: decompose L_n into its orthogonal components.
- Parseval's identity,

$$\mathbb{E}_{\mathcal{H}_0}\left[\mathsf{L}_n^2\right] = \|\mathsf{L}_n\|_{\mathcal{H}_0}^2 = \sum_{\alpha,\beta\in\mathbb{N}^n} \langle H_{\alpha,\beta}(\mathsf{X},\mathsf{Y}), \mathsf{L}_n(\mathsf{X},\mathsf{Y})\rangle_{\mathcal{H}_0}^2$$

Finite d: Detecting Correlated Vectors [Elimelech, H'24]

Proof sketch: We use polynomial decomposition:

- We want to analyze $\mathbb{E}_{\mathcal{H}_0}[\mathsf{L}_n^2]$.
- Idea: decompose L_n into its orthogonal components.
- Parseval's identity,

$$\mathbb{E}_{\mathcal{H}_0}\left[\mathsf{L}_n^2\right] = \|\mathsf{L}_n\|_{\mathcal{H}_0}^2 = \sum_{\alpha,\beta\in\mathbb{N}^n} \langle H_{\alpha,\beta}(\mathsf{X},\mathsf{Y}), \mathsf{L}_n(\mathsf{X},\mathsf{Y})\rangle_{\mathcal{H}_0}^2$$

- It can be shown that

$$\langle H_{\alpha,\beta}(\mathsf{X},\mathsf{Y}), \mathsf{L}_n(\mathsf{X},\mathsf{Y})\rangle_{\mathcal{H}_0} = \rho^{|\alpha|} \cdot \mathbb{P}[\pi(\beta) = \alpha]$$

  where $\pi(\alpha) \in \mathbb{N}^n$ denotes the vector obtained by permuting the coordinates of α using π

Finite d: Detecting Correlated Vectors [Elimelech, H'24]

Proof sketch: We use polynomial decomposition:

- We want to analyze $\mathbb{E}_{\mathcal{H}_0}[\mathsf{L}_n^2]$.
- Idea: decompose L_n into its orthogonal components.
- Parseval's identity,

$$\mathbb{E}_{\mathcal{H}_0}\left[\mathsf{L}_n^2\right] = \|\mathsf{L}_n\|_{\mathcal{H}_0}^2 = \sum_{\alpha,\beta\in\mathbb{N}^n} \langle H_{\alpha,\beta}(\mathsf{X},\mathsf{Y}), \mathsf{L}_n(\mathsf{X},\mathsf{Y})\rangle_{\mathcal{H}_0}^2$$

- It can be shown that

$$\langle H_{\alpha,\beta}(\mathsf{X},\mathsf{Y}), \mathsf{L}_n(\mathsf{X},\mathsf{Y})\rangle_{\mathcal{H}_0} = \rho^{|\alpha|} \cdot \mathbb{P}[\pi(\beta) = \alpha]$$

where $\pi(\alpha) \in \mathbb{N}^n$ denotes the vector obtained by permuting the coordinates of α using π

Goal: find $\mathbb{P}[\pi(\beta) = \alpha]$.

Finite d: Detecting Correlated Vectors [Elimelech, H'24]

Proof sketch: We use polynomial decomposition:

- We want to analyze $\mathbb{E}_{\mathcal{H}_0}[\mathsf{L}_n^2]$.
- Integer distribution function: for $\alpha \in \mathbb{N}^n$,

$$p_\alpha(\ell) \triangleq |i \in [n] : \alpha_i = \ell|, \quad \ell \in \mathbb{N}.$$

Note that,

Finite d: Detecting Correlated Vectors [Elimelech, H'24]

Proof sketch: We use polynomial decomposition:

- We want to analyze $\mathbb{E}_{\mathcal{H}_0}[\mathsf{L}_n^2]$.
- Integer distribution function: for $\alpha \in \mathbb{N}^n$,

$$p_\alpha(\ell) \triangleq |i \in [n] : \alpha_i = \ell|, \quad \ell \in \mathbb{N}.$$

Note that,

  - We say $\alpha \equiv \beta$ iff there is $\pi \in \mathbb{S}_n$ s.t. $\pi(\beta) = \alpha$.

Finite d: Detecting Correlated Vectors [Elimelech, H'24]

Proof sketch: We use polynomial decomposition:

- We want to analyze $\mathbb{E}_{\mathcal{H}_0}[\mathsf{L}_n^2]$.
- Integer distribution function: for $\alpha \in \mathbb{N}^n$,

$$p_\alpha(\ell) \triangleq |i \in [n] : \alpha_i = \ell|, \quad \ell \in \mathbb{N}.$$

Note that,

- We say $\alpha \equiv \beta$ iff there is $\pi \in \mathbb{S}_n$ s.t. $\pi(\beta) = \alpha$.
- $\alpha \equiv \beta$ iff $p_\alpha = p_\beta$.

Finite d: Detecting Correlated Vectors [Elimelech, H'24]

Proof sketch: We use polynomial decomposition:

- We want to analyze $\mathbb{E}_{\mathcal{H}_0}[\mathsf{L}_n^2]$.
- Integer distribution function: for $\alpha \in \mathbb{N}^n$,

$$p_\alpha(\ell) \triangleq |i \in [n] : \alpha_i = \ell|, \quad \ell \in \mathbb{N}.$$

  Note that,

  - We say $\alpha \equiv \beta$ iff there is $\pi \in \mathbb{S}_n$ s.t. $\pi(\beta) = \alpha$.
  - $\alpha \equiv \beta$ iff $p_\alpha = p_\beta$.
  - Let $[\alpha]$ denote the equivalence class of α.

Finite d: Detecting Correlated Vectors [Elimelech, H'24]

Proof sketch: We use polynomial decomposition:

- We want to analyze $\mathbb{E}_{\mathcal{H}_0}[\mathsf{L}_n^2]$.
- Integer distribution function: for $\alpha \in \mathbb{N}^n$,

$$p_\alpha(\ell) \triangleq |i \in [n] : \alpha_i = \ell|, \quad \ell \in \mathbb{N}.$$

  Note that,

  - We say $\alpha \equiv \beta$ iff there is $\pi \in \mathbb{S}_n$ s.t. $\pi(\beta) = \alpha$.
  - $\alpha \equiv \beta$ iff $p_\alpha = p_\beta$.
  - Let $[\alpha]$ denote the equivalence class of α.

  Then,

$$\mathbb{P}[\pi(\beta) = \alpha] = \frac{1}{|[\alpha]|} \mathbb{1}_{\alpha \equiv \beta}.$$

Finite d: Detecting Correlated Vectors [Elimelech, H'24]

Proof sketch: We use polynomial decomposition:

- We want to analyze $\mathbb{E}_{\mathcal{H}_0}[\mathsf{L}_n^2]$.
- Thus,

$$\begin{aligned}
\mathbb{E}_{\mathcal{H}_0}\left[\mathsf{L}_n^2\right] &= \sum_{\alpha,\beta\in\mathbb{N}^n} \langle H_{\alpha,\beta}(\mathsf{X},\mathsf{Y}), \mathsf{L}_n(\mathsf{X},\mathsf{Y})\rangle_{\mathcal{H}_0}^2 \\
&= \sum_{\alpha,\beta\in\mathbb{N}^n} \rho^{2|\alpha|}\frac{1}{|[\alpha]|^2}\mathbb{1}_{\alpha\equiv\beta} \\
&= \sum_{m=0}^{\infty} |\{[\alpha] : |\alpha| = m\}| \cdot \rho^{2m} \\
&= \sum_{m=0}^{\infty} |\mathsf{Par}(m, \leq_n)| \cdot \rho^{2m}
\end{aligned}$$

where $\mathsf{Par}(m, \leq_n)$ is the set of integer partitions of m to at most n elements.

Finite d: Detecting Correlated Vectors [Elimelech, H'24]

Proof sketch: We use polynomial decomposition:

- We want to analyze $\mathbb{E}_{\mathcal{H}_0}[\mathsf{L}_n^2]$.
- Thus,

$$\begin{aligned}
\mathbb{E}_{\mathcal{H}_0}\left[\mathsf{L}_n^2\right] &= \sum_{\alpha,\beta\in\mathbb{N}^n} \langle H_{\alpha,\beta}(\mathsf{X},\mathsf{Y}), \mathsf{L}_n(\mathsf{X},\mathsf{Y})\rangle_{\mathcal{H}_0}^2 \\
&= \sum_{\alpha,\beta\in\mathbb{N}^n} \rho^{2|\alpha|}\frac{1}{|[\alpha]|^2}\mathbb{1}_{\alpha\equiv\beta} \\
&= \sum_{m=0}^{\infty} |\{[\alpha] : |\alpha| = m\}| \cdot \rho^{2m} \\
&\leq \sum_{m=0}^{\infty} |\mathsf{Par}(m, \leq_\infty)| \cdot \rho^{2m}
\end{aligned}$$

where $|\mathsf{Par}(m, \leq_\infty)|$ is the number of integer partitions of the number m.

Finite d: Detecting Correlated Vectors [Elimelech, H'24]

Proof sketch: We use polynomial decomposition:

- We want to analyze $\mathbb{E}_{\mathcal{H}_0}[\mathsf{L}_n^2]$.

$$\mathbb{E}_{\mathcal{H}_0}\left[\mathsf{L}_n^2\right] \leq \sum_{m=0}^{\infty} |\mathsf{Par}(m, \leq_\infty)| \cdot \rho^{2m} \quad (\star)$$

- Hardy-Ramanujan Formula: $\exists c > 0$, s.t.

$$|\mathsf{Par}(m, \leq_\infty)| \leq c \cdot \frac{1}{4\sqrt{3}m} \exp\left(\pi\sqrt{\frac{2m}{3}}\right).$$

Finite d: Detecting Correlated Vectors [Elimelech, H'24]

Proof sketch: We use polynomial decomposition:

- We want to analyze $\mathbb{E}_{\mathcal{H}_0}[\mathsf{L}_n^2]$.

$$\mathbb{E}_{\mathcal{H}_0}\left[\mathsf{L}_n^2\right] \leq \sum_{m=0}^{\infty} |\mathsf{Par}(m, \leq_\infty)| \cdot \rho^{2m} \quad (\star)$$

- Hardy-Ramanujan Formula: $\exists c > 0$, s.t.

$$|\mathsf{Par}(m, \leq_\infty)| \leq c \cdot \frac{1}{4\sqrt{3}m} \exp\left(\pi\sqrt{\frac{2m}{3}}\right).$$

- Thus, $|\mathsf{Par}(m, \leq_\infty)|$ is sub-exponential in m, and hence $(\star)$ converges to a finite number, for any $\rho^2 < 1$.

Finite d: Detecting Correlated Vectors [Elimelech, H'24]

Theorem (Impossibility for $d \in \mathbb{N}$)

Strong detection is impossible for any $d\rho^2 < 1$.

This is proved using the same techniques ending up with complicated high-dimensional distribution functions.

Detecting Dependent Databases [Paslev, H'23]

- What if the databases are not Gaussian?

Detecting Dependent Databases [Paslev, H'23]

- What if the databases are not Gaussian?
- Consider the following detection problem:

$$\mathcal{H}_0 : (\mathsf{X}_1, \mathsf{Y}_1), \ldots, (\mathsf{X}_n, \mathsf{Y}_n) \overset{\text{i.i.d}}{\sim} P_X^{\otimes d} \times P_Y^{\otimes d}$$
$$\mathcal{H}_1 : (\mathsf{X}_1, \mathsf{Y}_{\pi_1}), \ldots, (\mathsf{X}_n, \mathsf{Y}_{\pi_n}) \overset{\text{i.i.d}}{\sim} P_{XY}^{\otimes d},$$

with $P_X = P_Y$ and denote $Q_{XY} = P_X \times P_Y$.

Detecting Dependent Databases [Paslev, H'23]

- What if the databases are not Gaussian?
- Consider the following detection problem:

$$\mathcal{H}_0 : (\mathsf{X}_1, \mathsf{Y}_1), \ldots, (\mathsf{X}_n, \mathsf{Y}_n) \overset{\text{i.i.d}}{\sim} P_X^{\otimes d} \times P_Y^{\otimes d}$$
$$\mathcal{H}_1 : (\mathsf{X}_1, \mathsf{Y}_{\pi_1}), \ldots, (\mathsf{X}_n, \mathsf{Y}_{\pi_n}) \overset{\text{i.i.d}}{\sim} P_{XY}^{\otimes d},$$

with $P_X = P_Y$ and denote $Q_{XY} = P_X \times P_Y$.

Theorem (Impossibility of weak detection)

Weak detection is impossible if

$$d \cdot \chi^2(P_{XY} || Q_{XY}) = o(1).$$

where $\chi^2(\mathbb{P}||\mathbb{Q}) = \int \frac{\mathrm{d}\mathbb{P}^2}{\mathrm{d}\mathbb{Q}} - 1.$

Detecting Dependent Databases [Paslev, H'23]

- What if the databases are not Gaussian?
- Consider the following detection problem:

$$\mathcal{H}_0 : (\mathsf{X}_1, \mathsf{Y}_1), \ldots, (\mathsf{X}_n, \mathsf{Y}_n) \overset{\text{i.i.d}}{\sim} P_X^{\otimes d} \times P_Y^{\otimes d}$$
$$\mathcal{H}_1 : (\mathsf{X}_1, \mathsf{Y}_{\pi_1}), \ldots, (\mathsf{X}_n, \mathsf{Y}_{\pi_n}) \overset{\text{i.i.d}}{\sim} P_{XY}^{\otimes d},$$

with $P_X = P_Y$ and denote $Q_{XY} = P_X \times P_Y$.

Theorem (Possibility of strong detection)

If

$$d \cdot \frac{d^2_{\mathsf{SKL}}(P_{XY} || Q_{XY})}{\mathsf{Var}_{Q_{XY}}(\mathcal{K}(A, B))} = \omega(1)$$

then, $\mathsf{R}(\phi_{\mathsf{sum}}) \to 0$, *as* $d \to \infty$.

Detecting Dependent Databases [Paslev, H'23]

Proof sketch:

- For any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, we let $\mathcal{L}(x, y) \triangleq \frac{P_{XY}(x,y)}{Q_{XY}(x,y)}$.

Detecting Dependent Databases [Paslev, H'23]

Proof sketch:

- For any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, we let $\mathcal{L}(x,y) \triangleq \frac{P_{XY}(x,y)}{Q_{XY}(x,y)}$.
- For any f s.t. $\mathbb{E}_Q f^2 < \infty$, consider the induced operator defined by the projection $(\mathcal{L}f)(x) \triangleq \mathbb{E}_{Y \sim Q_Y}\left[\mathcal{L}(x,Y)f(Y)\right]$.

Detecting Dependent Databases [Paslev, H'23]

Proof sketch:

- For any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, we let $\mathcal{L}(x,y) \triangleq \frac{P_{XY}(x,y)}{Q_{XY}(x,y)}$.
- For any f s.t. $\mathbb{E}_Q f^2 < \infty$, consider the induced operator defined by the projection $(\mathcal{L}f)(x) \triangleq \mathbb{E}_{Y \sim Q_Y}\left[\mathcal{L}(x,Y)f(Y)\right]$.
- We assume that $\mathcal{L}(x,y) = \mathcal{L}(y,x)$, and hence self-adjoint and Hilbert-Schmidt, diagonazable, with eigenvalues $\{\lambda_i\}_{i \geq 0}$.

Detecting Dependent Databases [Paslev, H'23]

Proof sketch:

- For any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, we let $\mathcal{L}(x,y) \triangleq \frac{P_{XY}(x,y)}{Q_{XY}(x,y)}$.
- For any f s.t. $\mathbb{E}_Q f^2 < \infty$, consider the induced operator defined by the projection $(\mathcal{L}f)(x) \triangleq \mathbb{E}_{Y \sim Q_Y}\left[\mathcal{L}(x,Y)f(Y)\right]$.
- We assume that $\mathcal{L}(x,y) = \mathcal{L}(y,x)$, and hence self-adjoint and Hilbert-Schmidt, diagonazable, with eigenvalues $\{\lambda_i\}_{i \geq 0}$.
- Recall that

$$\mathbb{E}_{\mathcal{H}_0}[\mathsf{L}_n^2] = \mathbb{E}_\pi\left[\prod_{O \in \mathcal{O}} \mathbb{E}_{\mathcal{H}_0}[\mathsf{Z}_O]\right].$$

Detecting Dependent Databases [Paslev, H'23]

Proof sketch:

- For any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, we let $\mathcal{L}(x,y) \triangleq \frac{P_{XY}(x,y)}{Q_{XY}(x,y)}$.
- For any f s.t. $\mathbb{E}_Q f^2 < \infty$, consider the induced operator defined by the projection $(\mathcal{L}f)(x) \triangleq \mathbb{E}_{Y\sim Q_Y}\left[\mathcal{L}(x,Y)f(Y)\right]$.
- We assume that $\mathcal{L}(x,y) = \mathcal{L}(y,x)$, and hence self-adjoint and Hilbert-Schmidt, diagonazable, with eigenvalues $\{\lambda_i\}_{i\geq 0}$.
- Recall that

$$\mathbb{E}_{\mathcal{H}_0}[\mathsf{L}_n^2] = \mathbb{E}_\pi\left[\prod_{O\in\mathcal{O}} \mathbb{E}_{\mathcal{H}_0}[\mathsf{Z}_O]\right].$$

- Then, with the notation above, it can be shown that,

$$\mathbb{E}_{\mathcal{H}_0}[\mathsf{Z}_C] = \left(\sum_{i\in\mathbb{N}} \lambda_i^{2|C|}\right)^d.$$

Detecting Dependent Databases [Paslev, H'23]

Proof sketch:

- For any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, we let $\mathcal{L}(x,y) \triangleq \frac{P_{XY}(x,y)}{Q_{XY}(x,y)}$.
- For any f s.t. $\mathbb{E}_Q f^2 < \infty$, consider the induced operator defined by the projection $(\mathcal{L}f)(x) \triangleq \mathbb{E}_{Y \sim Q_Y}\left[\mathcal{L}(x,Y)f(Y)\right]$.
- We assume that $\mathcal{L}(x,y) = \mathcal{L}(y,x)$, and hence self-adjoint and Hilbert-Schmidt, diagonazable, with eigenvalues $\{\lambda_i\}_{i\geq 0}$.
- Recall that

$$\mathbb{E}_{\mathcal{H}_0}[\mathsf{L}_n^2] = \mathbb{E}_\pi\left[\prod_{O\in\mathcal{O}} \mathbb{E}_{\mathcal{H}_0}[\mathsf{Z}_O]\right].$$

- Substituting, massaging, it can be shown that weak detection is impossible if

$$d \cdot \sum_{i\geq 1} \frac{\lambda_i^2}{1-\lambda_i^2} = o(1).$$

Partial Correlation

- What if the databases are only partially correlated?

Partial Correlation

- What if the databases are only partially correlated?
- Consider the following detection problem:

$$\mathcal{H}_0 : (\mathsf{X}_1, \mathsf{Y}_1), \ldots, (\mathsf{X}_n, \mathsf{Y}_n) \overset{\text{i.i.d}}{\sim} \mathcal{N}^{\otimes d}(0, I_{2\times 2})$$

$$\mathcal{H}_1 : \begin{cases} \{(\mathsf{X}_i, \mathsf{Y}_{\pi_i})\}_{i\in\mathcal{K}} \overset{\text{i.i.d}}{\sim} \mathcal{N}^{\otimes d}(\mathbf{0}, \Sigma_\rho) \\ \{(\mathsf{X}_i, \mathsf{Y}_{\pi_i})\}_{i\notin\mathcal{K}} \overset{\text{i.i.d}}{\sim} \mathcal{N}^{\otimes d}(\mathbf{0}, \mathbf{I}_{2\times 2}) \\ \{(\mathsf{X}_i, \mathsf{Y}_{\pi_i})\}_{i\notin\mathcal{K}} \perp\!\!\!\perp \{(\mathsf{X}_i, \mathsf{Y}_{\pi_i})\}_{i\in\mathcal{K}} \end{cases}$$

where $\pi \sim \mathsf{Unif}(\mathbb{S}_n)$ and $\mathcal{K} \sim \mathsf{Unif}\binom{[n]}{k}$.

Partial Correlation

- What if the databases are only partially correlated?
- Consider the following detection problem:

$$\mathcal{H}_0 : (\mathsf{X}_1, \mathsf{Y}_1), \ldots, (\mathsf{X}_n, \mathsf{Y}_n) \overset{\text{i.i.d}}{\sim} \mathcal{N}^{\otimes d}(0, I_{2\times 2})$$

$$\mathcal{H}_1 : \begin{cases} \{(\mathsf{X}_i, \mathsf{Y}_{\pi_i})\}_{i\in\mathcal{K}} \overset{\text{i.i.d}}{\sim} \mathcal{N}^{\otimes d}(\mathbf{0}, \Sigma_\rho) \\ \{(\mathsf{X}_i, \mathsf{Y}_{\pi_i})\}_{i\notin\mathcal{K}} \overset{\text{i.i.d}}{\sim} \mathcal{N}^{\otimes d}(\mathbf{0}, \mathbf{I}_{2\times 2}) \\ \{(\mathsf{X}_i, \mathsf{Y}_{\pi_i})\}_{i\notin\mathcal{K}} \perp\!\!\!\perp \{(\mathsf{X}_i, \mathsf{Y}_{\pi_i})\}_{i\in\mathcal{K}} \end{cases}$$

  where $\pi \sim \mathsf{Unif}(\mathbb{S}_n)$ and $\mathcal{K} \sim \mathsf{Unif}\binom{[n]}{k}$.
- So, only a planted set $\mathcal{K}$ of $k \leq n$ "users" is common to the two databases.

Partial Correlation

Theorem (Impossibility weak detection)

If,

$$\left(\frac{k}{n}\right)^2 \left(\prod_{i=1}^{k} \frac{1}{1-(d\rho^2)^i} - 1\right) = o(1),$$

then weak detection is impossible.

Partial Correlation

Theorem (Impossibility weak detection)

If,

$$\left(\frac{k}{n}\right)^2 \left(\prod_{i=1}^{k} \frac{1}{1-(d\rho^2)^i} - 1\right) = o(1),$$

then weak detection is impossible.

For example, if $k = O(\log n)$, then we get

$$\rho^2 < \frac{1}{d}\left[1 - \left(\mathsf{C}\frac{k}{n}\right)^{\frac{2}{k}}\right],$$

and we note that $(k/n)^{\frac{2}{k}} = o(1)$.

Testing Dependency of Random Graphs [Oren,Paslev,H'24]

- In many modern applications, the observations may be in the form of graphs.

Testing Dependency of Random Graphs [Oren,Paslev,H'24]

- Consider the following detection problem:

$$\mathcal{H}_0 : (\mathsf{A}_{ij}, \mathsf{B}_{ij}) \overset{\text{i.i.d.}}{\sim} \mathcal{Q}_{\mathsf{AB}} = \mathcal{P}_{\mathsf{A}} \times \mathcal{P}_{\mathsf{B}}$$

$$\mathcal{H}_1 : (\mathsf{A}_{ij}, \mathsf{B}_{\pi_i \pi_j}) \overset{\text{i.i.d.}}{\sim} \mathcal{P}_{\mathsf{AB}} | \pi \sim \mathsf{Unif}(\mathbb{S}_n),$$

where $\mathcal{P}_{\mathsf{A}} = \mathcal{P}_{\mathsf{B}}$.

Testing Dependency of Random Graphs [Oren,Paslev,H'24]

- Consider the following detection problem:

$$\mathcal{H}_0 : (\mathsf{A}_{ij}, \mathsf{B}_{ij}) \overset{\text{i.i.d.}}{\sim} \mathcal{Q}_{\mathsf{AB}} = \mathcal{P}_{\mathsf{A}} \times \mathcal{P}_{\mathsf{B}}$$
$$\mathcal{H}_1 : (\mathsf{A}_{ij}, \mathsf{B}_{\pi_i\pi_j}) \overset{\text{i.i.d.}}{\sim} \mathcal{P}_{\mathsf{AB}} | \pi \sim \mathsf{Unif}(\mathbb{S}_n),$$

  where $\mathcal{P}_{\mathsf{A}} = \mathcal{P}_{\mathsf{B}}$.

- The Bernoulli case was analyzes thoroughly in the literature[a], both from the statistical and computational point of views! Here, $\mathcal{P}_{\mathsf{A}} = \mathcal{P}_{\mathsf{B}} = \mathsf{Bernoulli}(\tau p)$, for some $p \in (0,1)$ and $\tau \in [0,1]$. Under $\mathcal{P}_{\mathsf{AB}}$, we have $A \sim \mathsf{Bernoulli}(\tau p)$, and

$$\mathsf{B}|\mathsf{A} \sim \begin{cases} \mathsf{Bernoulli}(\tau), & \text{if } X = 1 \\ \mathsf{Bernoulli}\left(\frac{\tau p(1-\tau)}{1-\tau p}\right), & \text{if } X = 0. \end{cases}$$

[a] E.g., [Wu,Xu,Yu'21], [Ding,Du,'23], [Ding,Du,Li'23].

Testing Dependency of Random Graphs [Oren,Paslev,H'24]

- Consider the following detection problem:

$$\mathcal{H}_0 : (\mathsf{A}_{ij}, \mathsf{B}_{ij}) \overset{\text{i.i.d.}}{\sim} \mathcal{Q}_{\mathsf{AB}} = \mathcal{P}_{\mathsf{A}} \times \mathcal{P}_{\mathsf{B}}$$
$$\mathcal{H}_1 : (\mathsf{A}_{ij}, \mathsf{B}_{\pi_i \pi_j}) \overset{\text{i.i.d.}}{\sim} \mathcal{P}_{\mathsf{AB}} | \pi \sim \mathsf{Unif}(\mathbb{S}_n),$$

  where $\mathcal{P}_{\mathsf{A}} = \mathcal{P}_{\mathsf{B}}$.

- The Gaussian case was studied from the statistical point of view [Wu,Xu,Yu'21].

Testing Dependency of Random Graphs [Oren,Paslev,H'24]

- Consider the following detection problem:

$$\mathcal{H}_0 : (\mathsf{A}_{ij}, \mathsf{B}_{ij}) \overset{\text{i.i.d.}}{\sim} \mathcal{Q}_{\mathsf{AB}} = \mathcal{P}_{\mathsf{A}} \times \mathcal{P}_{\mathsf{B}}$$

$$\mathcal{H}_1 : (\mathsf{A}_{ij}, \mathsf{B}_{\pi_i \pi_j}) \overset{\text{i.i.d.}}{\sim} \mathcal{P}_{\mathsf{AB}} | \pi \sim \mathsf{Unif}(\mathbb{S}_n),$$

where $\mathcal{P}_{\mathsf{A}} = \mathcal{P}_{\mathsf{B}}$.

Theorem (Impossibility of weak detection)

Weak detection is statistically impossible if

$$\chi^2 (\mathcal{P} || \mathcal{Q}) \leq \frac{(2-\epsilon)\log n}{\alpha n}, \quad \text{and}$$

$$d_{\mathsf{KL}} (\mathcal{P} || \mathcal{Q}) + \delta_n \cdot \mathsf{Var}_{\mathcal{P}} (\log \mathcal{L}) \leq \frac{(2-\epsilon)\log n}{n},$$

for any $\omega(1) = \delta_n = o(\log n)$, *and any constant* $\epsilon > 0$.

Testing Dependency of Random Graphs [Oren,Paslev,H'24]

- Consider the following detection problem:

$$\mathcal{H}_0 : (\mathsf{A}_{ij}, \mathsf{B}_{ij}) \overset{\text{i.i.d.}}{\sim} \mathcal{Q}_{\mathsf{AB}} = \mathcal{P}_{\mathsf{A}} \times \mathcal{P}_{\mathsf{B}}$$

$$\mathcal{H}_1 : (\mathsf{A}_{ij}, \mathsf{B}_{\pi_i \pi_j}) \overset{\text{i.i.d.}}{\sim} \mathcal{P}_{\mathsf{AB}} | \pi \sim \mathsf{Unif}(\mathbb{S}_n),$$

where $\mathcal{P}_{\mathsf{A}} = \mathcal{P}_{\mathsf{B}}$.

- For the class of distributions for which there is a constant $\mathsf{C} > 1$ such that $\chi^2 (\mathcal{P}||\mathcal{Q}) \leq \mathsf{C} \cdot d_{\mathsf{KL}} (\mathcal{P}||\mathcal{Q})$, weak detection is impossible if

$$d_{\mathsf{KL}} (\mathcal{P}||\mathcal{Q}) \leq \frac{(2-\epsilon)\log n}{n}.$$

Testing Dependency of Random Graphs [Oren,Paslev,H'24]

- Consider the following detection problem:

$$\mathcal{H}_0 : (\mathsf{A}_{ij}, \mathsf{B}_{ij}) \overset{\text{i.i.d.}}{\sim} \mathcal{Q}_{\mathsf{AB}} = \mathcal{P}_{\mathsf{A}} \times \mathcal{P}_{\mathsf{B}}$$

$$\mathcal{H}_1 : (\mathsf{A}_{ij}, \mathsf{B}_{\pi_i \pi_j}) \overset{\text{i.i.d.}}{\sim} \mathcal{P}_{\mathsf{AB}} | \pi \sim \mathsf{Unif}(\mathbb{S}_n),$$

  where $\mathcal{P}_{\mathsf{A}} = \mathcal{P}_{\mathsf{B}}$.

- For the class of distributions for which there is a constant $\mathsf{C} > 1$ such that $\chi^2 (\mathcal{P} || \mathcal{Q}) \leq \mathsf{C} \cdot d_{\mathsf{KL}} (\mathcal{P} || \mathcal{Q})$, weak detection is impossible if

$$d_{\mathsf{KL}} (\mathcal{P} || \mathcal{Q}) \leq \frac{(2 - \epsilon) \log n}{n}.$$

- Coincides with [Wu,Xu,Yu'21] for Bernoulli and Gaussian.

Testing Dependency of Random Graphs [Oren,Paslev,H'24]

- Consider the following detection problem:

$$\mathcal{H}_0 : (\mathsf{A}_{ij}, \mathsf{B}_{ij}) \overset{\text{i.i.d.}}{\sim} \mathcal{Q}_{\mathsf{AB}} = \mathcal{P}_{\mathsf{A}} \times \mathcal{P}_{\mathsf{B}}$$

$$\mathcal{H}_1 : (\mathsf{A}_{ij}, \mathsf{B}_{\pi_i \pi_j}) \overset{\text{i.i.d.}}{\sim} \mathcal{P}_{\mathsf{AB}} | \pi \sim \mathsf{Unif}(\mathbb{S}_n),$$

where $\mathcal{P}_{\mathsf{A}} = \mathcal{P}_{\mathsf{B}}$.

Theorem (Strong detection upper bound)

Suppose there is a $\bar{\theta} \in (-d_{\mathsf{KL}}(\mathcal{Q}||\mathcal{P}), d_{\mathsf{KL}}(\mathcal{P}||\mathcal{Q}))$ *with*

$$E_{\mathcal{Q}}(\bar{\theta}) \geq \frac{2\log(n/e)}{n-1} + O(n^{-2}\log n),$$

$$E_{\mathcal{P}}(\bar{\theta}) = \omega(n^{-2}).$$

Then, $\mathsf{R}_n(\phi_{\mathsf{GLRT}}) \to 0$*, as* $n \to \infty$*.*

Testing Dependency of Random Graphs [Oren,Paslev,H'24]

- Consider the following detection problem:

$$\mathcal{H}_0 : (\mathsf{A}_{ij}, \mathsf{B}_{ij}) \overset{\text{i.i.d.}}{\sim} \mathcal{Q}_{\mathsf{AB}} = \mathcal{P}_{\mathsf{A}} \times \mathcal{P}_{\mathsf{B}}$$

$$\mathcal{H}_1 : (\mathsf{A}_{ij}, \mathsf{B}_{\pi_i \pi_j}) \overset{\text{i.i.d.}}{\sim} \mathcal{P}_{\mathsf{AB}} | \pi \sim \mathsf{Unif}(\mathbb{S}_n),$$

where $\mathcal{P}_{\mathsf{A}} = \mathcal{P}_{\mathsf{B}}$.

- For pairs of distributions $(\mathcal{P}, \mathcal{Q})$ with sub-exponential likelihood function, strong detection is possible if

$$d_{\mathsf{KL}}\left(\mathcal{P} || \mathcal{Q}\right) \geq \frac{2 \log n}{n - 1}.$$

Testing Dependency of Random Graphs [Oren,Paslev,H'24]

- Consider the following detection problem:

$$\mathcal{H}_0 : (\mathsf{A}_{ij}, \mathsf{B}_{ij}) \overset{\text{i.i.d.}}{\sim} \mathcal{Q}_{\mathsf{AB}} = \mathcal{P}_{\mathsf{A}} \times \mathcal{P}_{\mathsf{B}}$$

$$\mathcal{H}_1 : (\mathsf{A}_{ij}, \mathsf{B}_{\pi_i \pi_j}) \overset{\text{i.i.d.}}{\sim} \mathcal{P}_{\mathsf{AB}} | \pi \sim \mathsf{Unif}(\mathbb{S}_n),$$

  where $\mathcal{P}_{\mathsf{A}} = \mathcal{P}_{\mathsf{B}}$.

- For pairs of distributions $(\mathcal{P}, \mathcal{Q})$ with sub-exponential likelihood function, strong detection is possible if

$$d_{\mathsf{KL}} (\mathcal{P} || \mathcal{Q}) \geq \frac{2 \log n}{n-1}.$$

- Complements lower bound.

Testing Dependency of Random Graphs [Oren,Paslev,H'24]

- Consider the following detection problem:

$$\mathcal{H}_0 : (\mathsf{A}_{ij}, \mathsf{B}_{ij}) \overset{\text{i.i.d.}}{\sim} \mathcal{Q}_{\mathsf{AB}} = \mathcal{P}_{\mathsf{A}} \times \mathcal{P}_{\mathsf{B}}$$
$$\mathcal{H}_1 : (\mathsf{A}_{ij}, \mathsf{B}_{\pi_i \pi_j}) \overset{\text{i.i.d.}}{\sim} \mathcal{P}_{\mathsf{AB}} | \pi \sim \mathsf{Unif}(\mathbb{S}_n),$$

  where $\mathcal{P}_{\mathsf{A}} = \mathcal{P}_{\mathsf{B}}$.

- For pairs of distributions $(\mathcal{P}, \mathcal{Q})$ with sub-exponential likelihood function, strong detection is possible if

$$d_{\mathsf{KL}}(\mathcal{P} || \mathcal{Q}) \geq \frac{2 \log n}{n - 1}.$$

- Complements lower bound.
- GLRT is exhibits exponential computational complexity. What about poly-time algorithms?

Testing Dependency of Random Graphs [Oren,Paslev,H'24]

- Consider the following detection problem:

$$\mathcal{H}_0 : (\mathsf{A}_{ij}, \mathsf{B}_{ij}) \overset{\text{i.i.d.}}{\sim} \mathcal{Q}_{\mathsf{AB}} = \mathcal{P}_{\mathsf{A}} \times \mathcal{P}_{\mathsf{B}}$$

$$\mathcal{H}_1 : (\mathsf{A}_{ij}, \mathsf{B}_{\pi_i \pi_j}) \overset{\text{i.i.d.}}{\sim} \mathcal{P}_{\mathsf{AB}} | \pi \sim \mathsf{Unif}(\mathbb{S}_n),$$

where $\mathcal{P}_{\mathsf{A}} = \mathcal{P}_{\mathsf{B}}$.

Theorem (Weak detection upper bound)

If $|\mathsf{corr}(\mathcal{Q}, \mathcal{P})| \triangleq \frac{|\mathsf{cov}_{\mathcal{P}}(A,B)|}{\mathsf{Var}_{\mathcal{Q}}(A)} = \Omega(1)$, *and*

$$\frac{\mathbb{E}_{\mathcal{Q}}|A-B|^3}{\mathsf{Var}_{\mathcal{Q}}^{3/2}(A)}, \frac{\mathbb{E}_{\mathcal{P}}|A-B|^3}{\mathsf{Var}_{\mathcal{Q}}^{3/2}(A)(1-|\mathsf{corr}(\mathcal{Q}, \mathcal{P})|)^{3/2}} = o(n),$$

then $\lim_{n\to\infty} \mathsf{R}_n(\phi_{\mathsf{sum}}) < 1$.

Testing Dependency of Random Graphs [Oren,Paslev,H'24]

- Consider the following detection problem:

$$\mathcal{H}_0 : (\mathsf{A}_{ij}, \mathsf{B}_{ij}) \overset{\text{i.i.d.}}{\sim} \mathcal{Q}_{\mathsf{AB}} = \mathcal{P}_{\mathsf{A}} \times \mathcal{P}_{\mathsf{B}}$$
$$\mathcal{H}_1 : (\mathsf{A}_{ij}, \mathsf{B}_{\pi_i \pi_j}) \overset{\text{i.i.d.}}{\sim} \mathcal{P}_{\mathsf{AB}} | \pi \sim \mathsf{Unif}(\mathbb{S}_n),$$

  where $\mathcal{P}_{\mathsf{A}} = \mathcal{P}_{\mathsf{B}}$.

- In the Gaussian and Bernoulli cases this boils down to $\rho^2 = \Omega(1)$, while GLRT allows for a vanishing correlation.

Testing Dependency of Random Graphs [Oren,Paslev,H'24]

- Consider the following detection problem:

$$\mathcal{H}_0 : (\mathsf{A}_{ij}, \mathsf{B}_{ij}) \overset{\text{i.i.d.}}{\sim} \mathcal{Q}_{\mathsf{AB}} = \mathcal{P}_{\mathsf{A}} \times \mathcal{P}_{\mathsf{B}}$$
$$\mathcal{H}_1 : (\mathsf{A}_{ij}, \mathsf{B}_{\pi_i \pi_j}) \overset{\text{i.i.d.}}{\sim} \mathcal{P}_{\mathsf{AB}} | \pi \sim \mathsf{Unif}(\mathbb{S}_n),$$

  where $\mathcal{P}_{\mathsf{A}} = \mathcal{P}_{\mathsf{B}}$.

- In the Gaussian and Bernoulli cases this boils down to $\rho^2 = \Omega(1)$, while GLRT allows for a vanishing correlation.
- **Conjecture**: this is fundamental in the sense that this is a barrier for what can be achieved using polynomial-time algorithms.

Testing Dependency of Random Graphs [Oren,Paslev,H'24]

- Consider the following detection problem:

$$\mathcal{H}_0 : (\mathsf{A}_{ij}, \mathsf{B}_{ij}) \overset{\text{i.i.d.}}{\sim} \mathcal{Q}_{\mathsf{AB}} = \mathcal{P}_{\mathsf{A}} \times \mathcal{P}_{\mathsf{B}}$$
$$\mathcal{H}_1 : (\mathsf{A}_{ij}, \mathsf{B}_{\pi_i \pi_j}) \overset{\text{i.i.d.}}{\sim} \mathcal{P}_{\mathsf{AB}} | \pi \sim \mathsf{Unif}(\mathbb{S}_n),$$

  where $\mathcal{P}_{\mathsf{A}} = \mathcal{P}_{\mathsf{B}}$.

- In the Gaussian and Bernoulli cases this boils down to $\rho^2 = \Omega(1)$, while GLRT allows for a vanishing correlation.
- **Conjecture**: this is fundamental in the sense that this is a barrier for what can be achieved using polynomial-time algorithms.
- In the Bernoulli case [Ding,Du,Li'23] prove computational lower bound based on the low-degree polynomial conjecture.

Testing Dependency of Random Graphs [Oren,Paslev,H'24]

- Consider the following detection problem:

$$\mathcal{H}_0 : (\mathsf{A}_{ij}, \mathsf{B}_{ij}) \overset{\text{i.i.d.}}{\sim} \mathcal{Q}_{\mathsf{AB}} = \mathcal{P}_{\mathsf{A}} \times \mathcal{P}_{\mathsf{B}}$$

$$\mathcal{H}_1 : (\mathsf{A}_{ij}, \mathsf{B}_{\pi_i \pi_j}) \overset{\text{i.i.d.}}{\sim} \mathcal{P}_{\mathsf{AB}} | \pi \sim \mathsf{Unif}(\mathbb{S}_n),$$

where $\mathcal{P}_{\mathsf{A}} = \mathcal{P}_{\mathsf{B}}$.

Theorem (LDP computational lower bound)

For a certain class of pairs of distributions $(\mathcal{P}, \mathcal{Q})$, if $|\mathsf{corr}(\mathcal{Q}, \mathcal{P})| = o(1)$, then $\|\mathsf{L}_{n,\leq \mathsf{D}}\|_{\mathcal{H}_0} \leq O(1)$, for any $\mathsf{D} = O(|\mathsf{corr}(\mathcal{Q}, \mathcal{P})|^{-1})$.

Here, $\mathsf{L}_{n,\leq \mathsf{D}}$ is the projection of L_n to the linear subspace of polynomials of degree at most $\mathsf{D} \in \mathbb{N}$.

Concluding Remarks

- We considered the problem of testing correlated/dependent databases and characterize the statistical limits (in some asymptotic regimes).

Concluding Remarks

- We considered the problem of testing correlated/dependent databases and characterize the statistical limits (in some asymptotic regimes).
- The impossibility proofs are based on delicate analysis of the second moment using properties of random permutation cycles and integer partition function via polynomial decomposition.

Concluding Remarks

- We considered the problem of testing correlated/dependent databases and characterize the statistical limits (in some asymptotic regimes).
- The impossibility proofs are based on delicate analysis of the second moment using properties of random permutation cycles and integer partition function via polynomial decomposition.
- There is a gap between lower and upper bound when d is fixed.

Concluding Remarks

- We considered the problem of testing correlated/dependent databases and characterize the statistical limits (in some asymptotic regimes).
- The impossibility proofs are based on delicate analysis of the second moment using properties of random permutation cycles and integer partition function via polynomial decomposition.
- There is a gap between lower and upper bound when d is fixed.

Open Problems:

- Close the gap, and obtain sharp bounds.
- Prove existence of/close the computational gaps.

Thank You!