The Broken Sample Problem revisited

Jiaming Xu

The Fuqua School of Business
Duke University

Joint work with
Simiao Jiao (Duke) and Yihong Wu (Yale)

Allerton Conference 2024

Matchmaking [DeGroot-Feder-Goel '71]

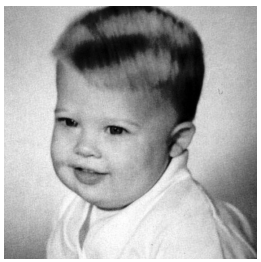
The Annals of Mathematical Statistics
1971, Vol. 42, No. 2, 578–593

MATCHMAKING[1]

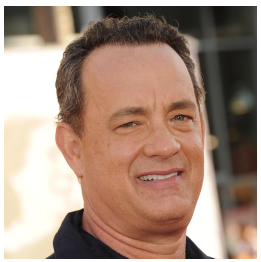
BY MORRIS H. DEGROOT, PAUL I. FEDER, AND PREM K. GOEL
Carnegie–Mellon University and Yale University

Matchmaking: Toy example

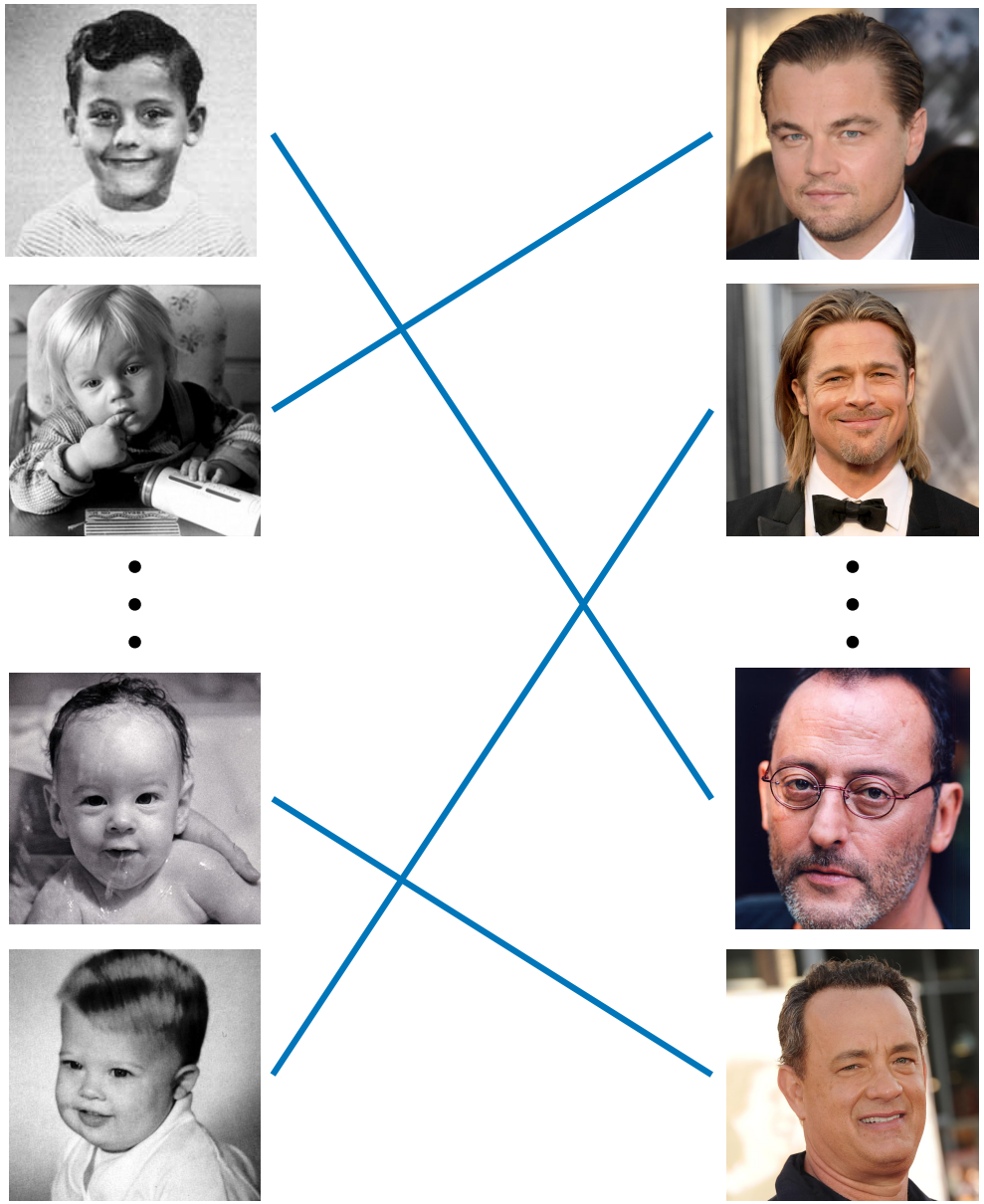




Matchmaking: Toy example



The broken sample problem

Probab. Theory Relat. Fields 131, 528–552 (2005)
Digital Object Identifier (DOI) 10.1007/s00440-004-0384-5

Zhidong Bai · Tailen Hsing

The broken sample problem

Dedicated to Professor Xiru Chen on His 70th Birthday

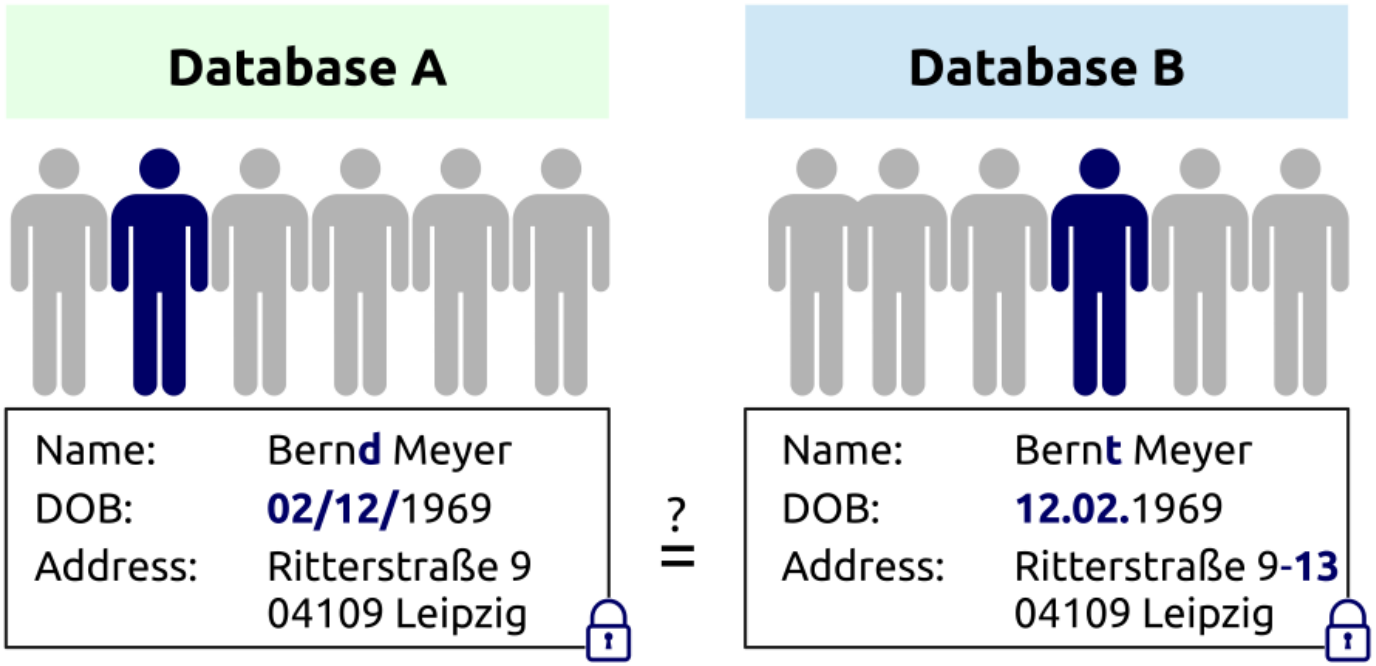
Received: 20 February 2002 / Revised version: 16 June 2004
Published online: 12 September 2004 –

Abstract. Suppose that $(X_i, Y_i), i = 1, 2, \ldots, n$, are iid. random vectors with uniform marginals and a certain joint distribution F_ρ, where ρ is a parameter with $\rho = \rho_o$ corresponds to the independence case. However, the X's and Y's are observed separately so that the pairing information is missing. Can ρ be consistently estimated? This is an extension of a problem considered in DeGroot and Goel (1980) which focused on the bivariate normal distribution with ρ being the correlation. In this paper we show that consistent discrimination between two distinct parameter values ρ_1 and ρ_2 is impossible if the density f_ρ of F_ρ is square integrable and the second largest singular value of the linear operator $h \to \int_0^1 f_\rho(x, \cdot)h(x)dx$, $h \in L^2[0, 1]$, is strictly less than 1 for $\rho = \rho_1$ and ρ_2. We also consider this result from the perspective of a bivariate empirical process which contains information equivalent to that of the broken sample.

Problem goes by many names

- Record linkage
- Data matching
- Feature matching
- Database alignment
- Data de-anonymization
- Identity Fragmentation/Identity resolution
- Shuffled/uncoupled regression
- ...

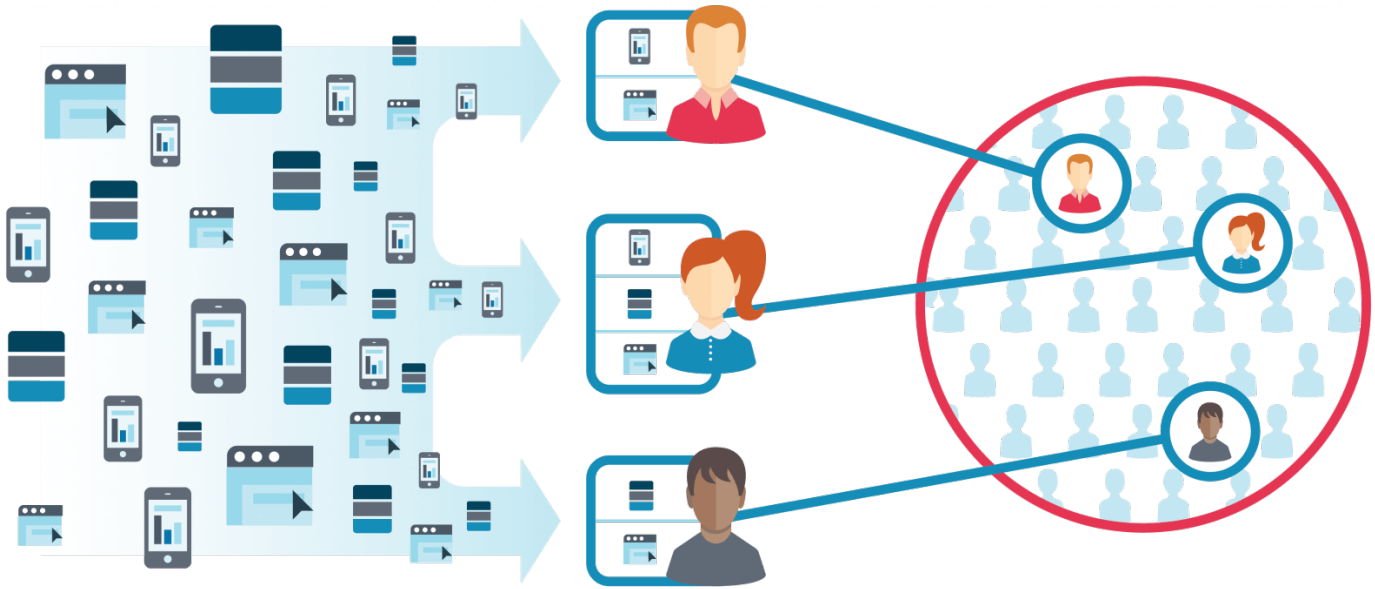
Applications: Record linkage/Data matching



Matching medical records or census records in two databases that refer to the same entity [Fellegi-Sunter '69]

Applications: Identity fragmentation

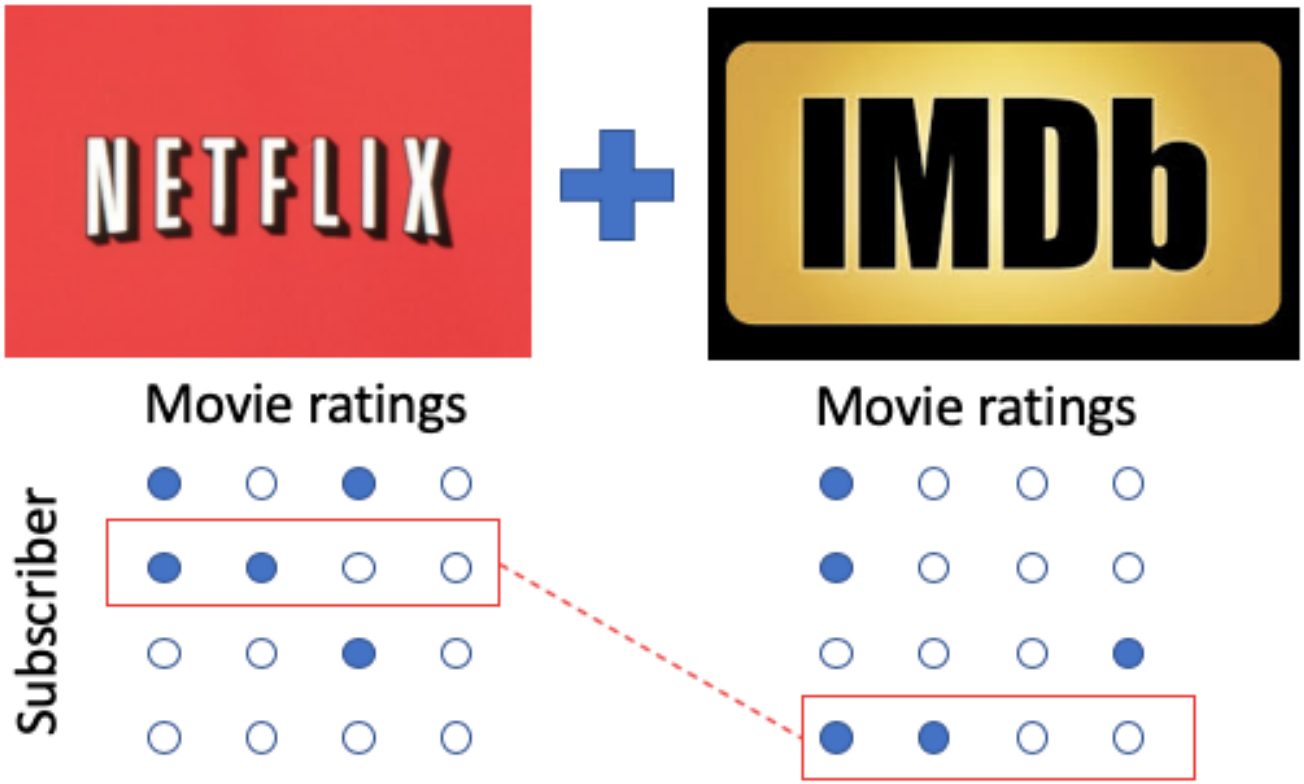
Consumers constantly explore Internet via multiple devices with different identifiers $\Rightarrow$ fragmented view of exposures and user behaviors



Detecting/linking same users based on their browsing logs on different devices $\Rightarrow$ key to the success of marketing and advertising [Lin-Misra Marketing Science '22]

Applications: Data de-anonymization

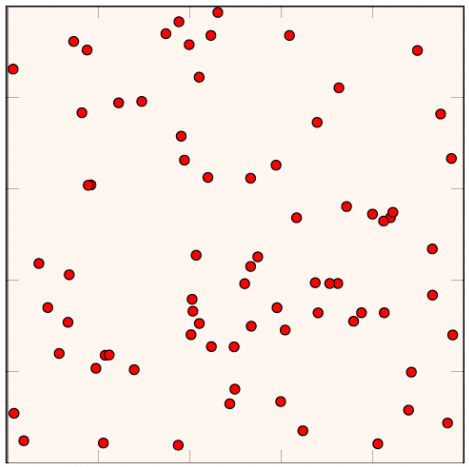
Collecting and disseminating datasets can expose customers to serious privacy breaches, even if datasets are anonymized and sanitized



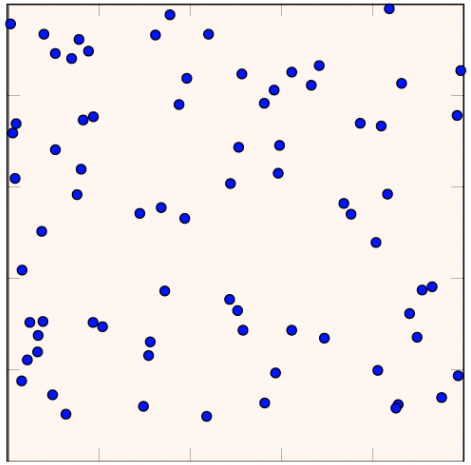
Successfully de-anonymize users by linking Netflix records and IMDB records [Narayanan-Shmatikov '08]

Applications: Particle tracking

Track mobile objects (birds in flocks, motile cells, or particles in fluid) from video frames taken at certain rate



$x(t)$

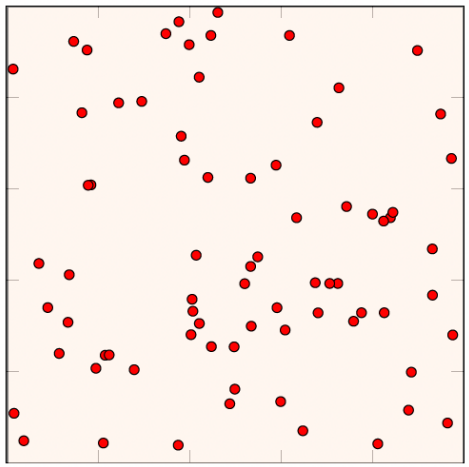


$x(t + \Delta t)$

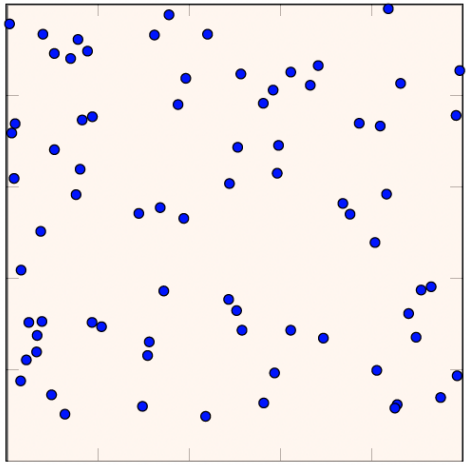
Picture courtesy of Gabriele Sicuro

Applications: Particle tracking

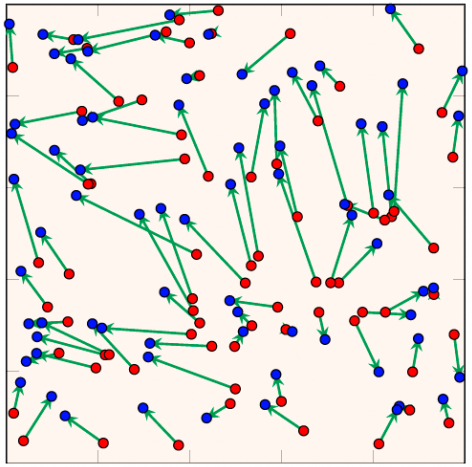
Track mobile objects (birds in flocks, motile cells, or particles in fluid) from video frames taken at certain rate



$x(t)$



$x(t + \Delta t)$

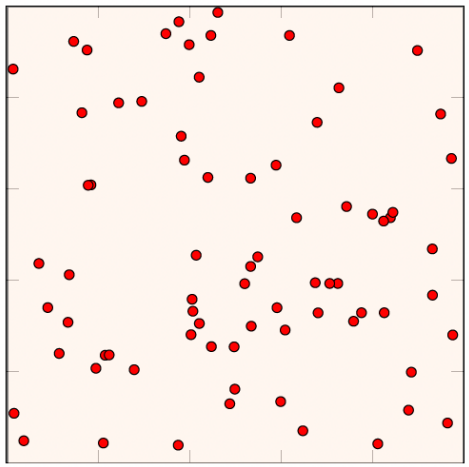


?

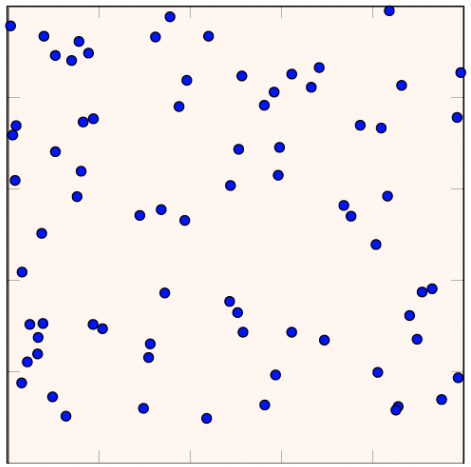
Picture courtesy of Gabriele Sicuro

Applications: Particle tracking

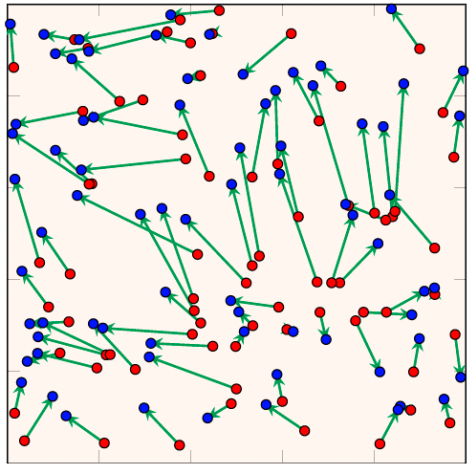
Track mobile objects (birds in flocks, motile cells, or particles in fluid) from video frames taken at certain rate



$x(t)$



$x(t + \Delta t)$



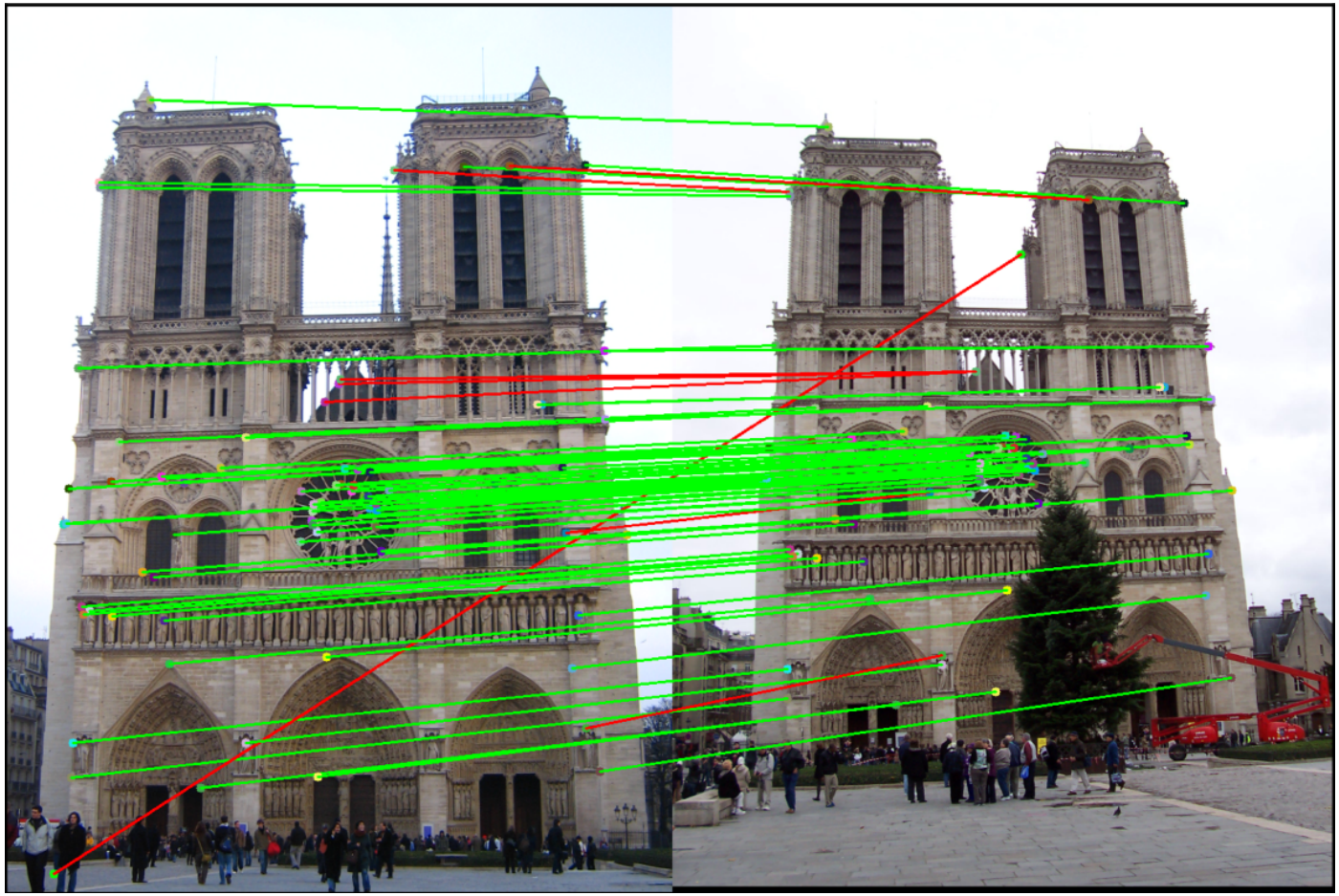
?

Picture courtesy of Gabriele Sicuro

Match objects in two consecutive frames based on their positional vectors; the noises are determined by the density and mobility of objects and the acquisition rate of frames [Chertkov-Kroc-Krzakala-Vergassola-Zdeborová '10, Semerjian-Sicuro-Zdeborová '20, Moharrami-Moore-X '21, Kunisky-Niels-Weed '21, Ding-Wu-X-Yang '23,...]

Applications: Feature matching

Align multi-views of objects to generate panoramic views or 3D models



Detect and match important features in two images [Szeliski '22]

The broken sample model

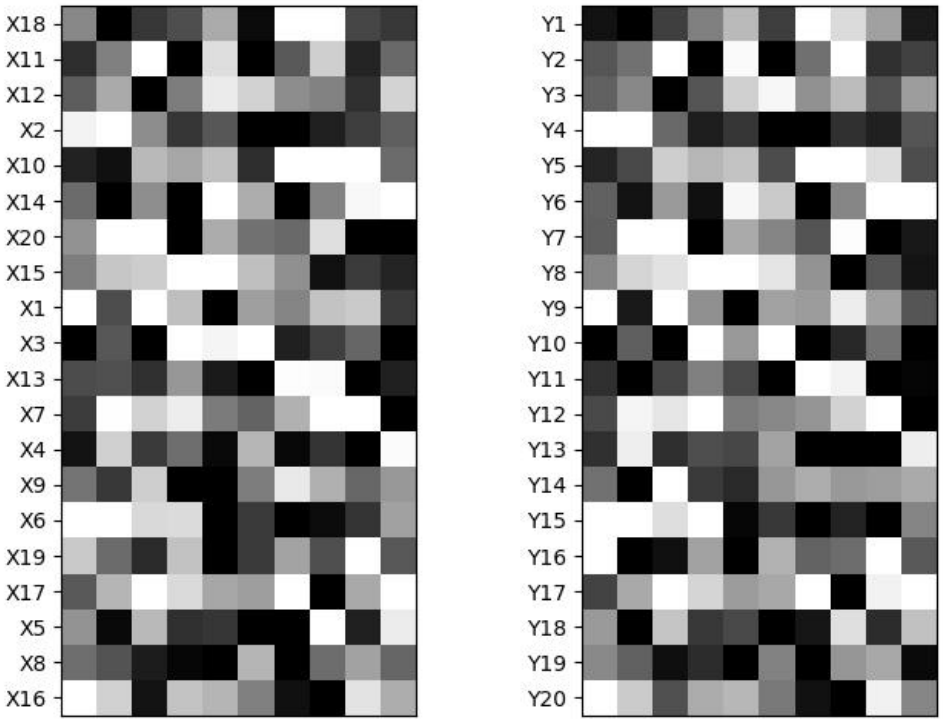
- π^* is a random uniform permutation on $[n]$

The broken sample model

- π^* is a random uniform permutation on $[n]$
- $(X_{\pi^*(i)}, Y_i) \overset{\text{i.i.d.}}{\sim} P_{X,Y}$ for $i \in [n]$

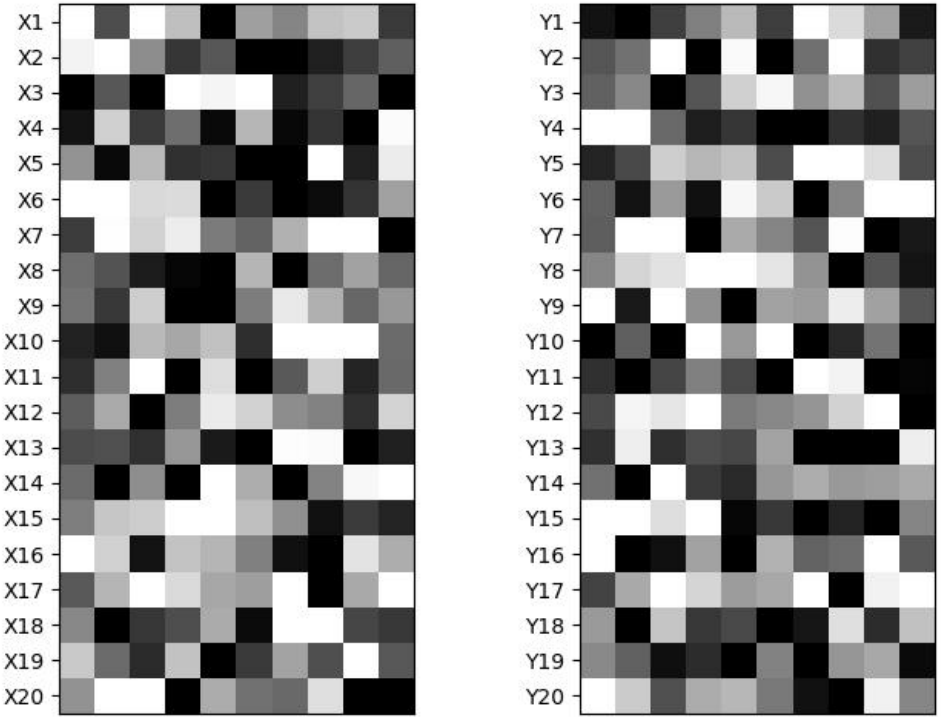
The broken sample model

- π^* is a random uniform permutation on $[n]$
- $(X_{\pi^*(i)}, Y_i) \overset{\text{i.i.d.}}{\sim} P_{X,Y}$ for $i \in [n]$



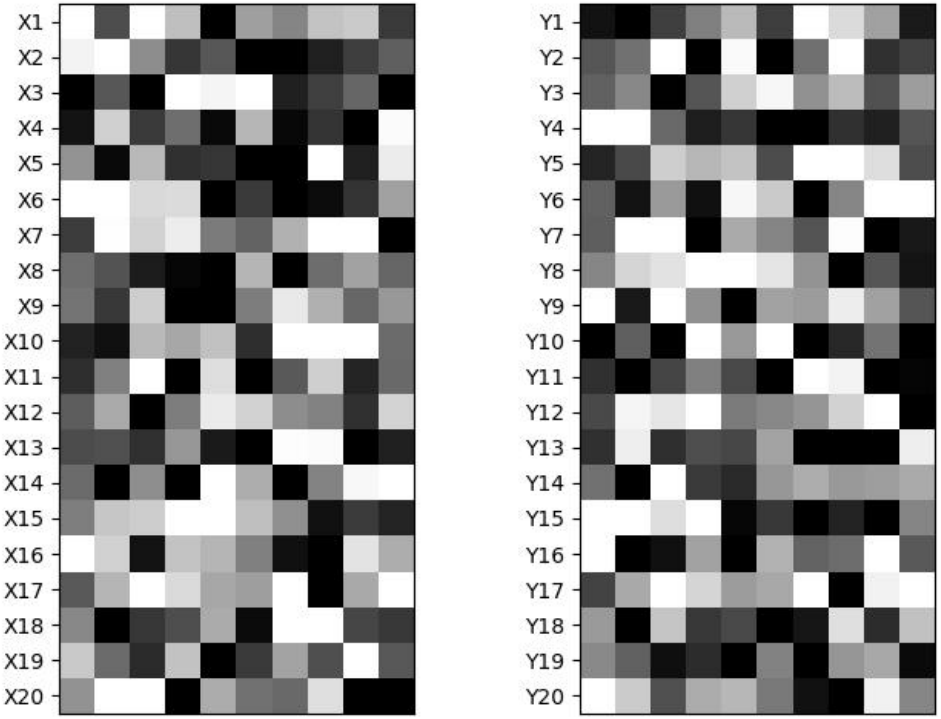
The broken sample model

- π^* is a random uniform permutation on $[n]$
- $(X_{\pi^*(i)}, Y_i) \overset{\text{i.i.d.}}{\sim} P_{X,Y}$ for $i \in [n]$
- Observe $(X_i)_{i=1}^n$ and $(Y_i)_{i=1}^n$



The broken sample model

- π^* is a random uniform permutation on $[n]$
- $(X_{\pi^*(i)}, Y_i) \overset{\text{i.i.d.}}{\sim} P_{X,Y}$ for $i \in [n]$
- Observe $(X_i)_{i=1}^n$ and $(Y_i)_{i=1}^n$



Goal:

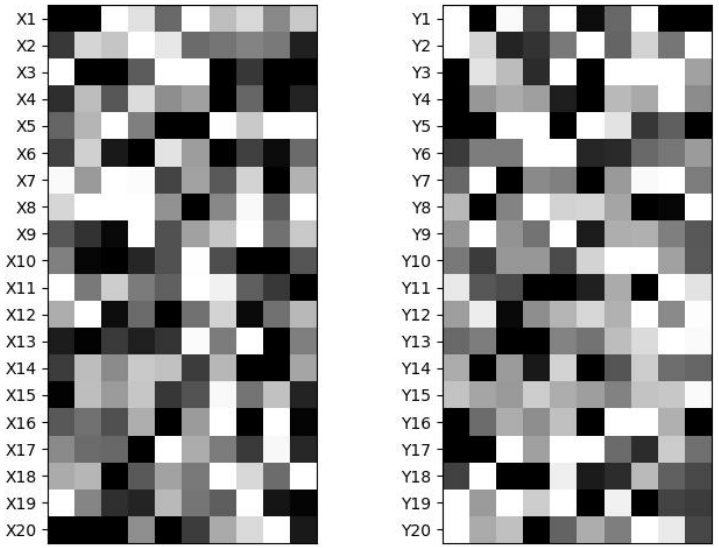
- Estimation: Recover π^* [Dai-Cullina-Kiyavash '19,'20, Kunisky-Niles-Weed '22, Wang-Wu-X-Yolou '22]
- Detection: Test correlated against independence model ($P_{X,Y}$ versus $P_X \otimes P_Y$)

Example: Detecting correlated Gaussian databases

Definition (DeGroot-Feder-Goel '71, Bai-Hsing '05)

$$H_0 : (X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{i.i.d.}}{\sim} N^{\otimes d}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1, 0\\0, 1\end{bmatrix}\right)$$

$$H_1 : (X_{\pi^*(1)}, Y_1), \ldots, (X_{\pi^*(n)}, Y_n) \overset{\text{i.i.d.}}{\sim} N^{\otimes d}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1, \rho\\\rho, 1\end{bmatrix}\right)$$



H_0 or H_1?

Example: Detecting correlated Gaussian databases

Definition (DeGroot-Feder-Goel '71, Bai-Hsing '05)

$$H_0 : (X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{i.i.d.}}{\sim} N^{\otimes d}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1, 0\\0, 1\end{bmatrix}\right)$$

$$H_1 : (X_{\pi^*(1)}, Y_1), \ldots, (X_{\pi^*(n)}, Y_n) \overset{\text{i.i.d.}}{\sim} N^{\otimes d}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1, \rho\\\rho, 1\end{bmatrix}\right)$$

Example: Detecting correlated Gaussian databases

Definition (DeGroot-Feder-Goel '71, Bai-Hsing '05)

$$H_0 : (X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{i.i.d.}}{\sim} N^{\otimes d}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1, 0\\0, 1\end{bmatrix}\right)$$

$$H_1 : (X_{\pi^*(1)}, Y_1), \ldots, (X_{\pi^*(n)}, Y_n) \overset{\text{i.i.d.}}{\sim} N^{\otimes d}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1, \rho\\\rho, 1\end{bmatrix}\right)$$

- Marginally, both X_i's and Y_i's are i.i.d. standard Gaussian random vectors in $\mathbb{R}^d$

Example: Detecting correlated Gaussian databases

Definition (DeGroot-Feder-Goel '71, Bai-Hsing '05)

$$H_0 : (X_1, Y_1), \dots, (X_n, Y_n) \overset{\text{i.i.d.}}{\sim} N^{\otimes d}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1, 0\\0, 1\end{bmatrix}\right)$$

$$H_1 : (X_{\pi^*(1)}, Y_1), \dots, (X_{\pi^*(n)}, Y_n) \overset{\text{i.i.d.}}{\sim} N^{\otimes d}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1, \rho\\\rho, 1\end{bmatrix}\right)$$

- Marginally, both X_i's and Y_i's are i.i.d. standard Gaussian random vectors in $\mathbb{R}^d$
- The inherent correlation under H_1 is obscured by latent matching π^*

Example: Detecting correlated Gaussian databases

Definition (DeGroot-Feder-Goel '71, Bai-Hsing '05)

$$H_0 : (X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{i.i.d.}}{\sim} N^{\otimes d}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1, 0\\0, 1\end{bmatrix}\right)$$

$$H_1 : (X_{\pi^*(1)}, Y_1), \ldots, (X_{\pi^*(n)}, Y_n) \overset{\text{i.i.d.}}{\sim} N^{\otimes d}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1, \rho\\\rho, 1\end{bmatrix}\right)$$

- Marginally, both X_i's and Y_i's are i.i.d. standard Gaussian random vectors in $\mathbb{R}^d$
- The inherent correlation under H_1 is obscured by latent matching π^*
- Tests need to be permutation invariant

Example: Detecting correlated Gaussian databases

Definition (DeGroot-Feder-Goel '71, Bai-Hsing '05)

$$H_0 : (X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{i.i.d.}}{\sim} N^{\otimes d}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1, 0\\0, 1\end{bmatrix}\right)$$

$$H_1 : (X_{\pi^*(1)}, Y_1), \ldots, (X_{\pi^*(n)}, Y_n) \overset{\text{i.i.d.}}{\sim} N^{\otimes d}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1, \rho\\\rho, 1\end{bmatrix}\right)$$

- Marginally, both X_i's and Y_i's are i.i.d. standard Gaussian random vectors in $\mathbb{R}^d$
- The inherent correlation under H_1 is obscured by latent matching π^*
- Tests need to be permutation invariant
- What is the minimum ρ needed for achieving vanishing testing error as $n \to \infty$?

Example: Detecting correlated Gaussian databases

Definition (DeGroot-Feder-Goel '71, Bai-Hsing '05)

$$H_0 : (X_1, Y_1), \dots, (X_n, Y_n) \overset{\text{i.i.d.}}{\sim} N^{\otimes d}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1,0\\0,1\end{bmatrix}\right)$$

$$H_1 : (X_{\pi^*(1)}, Y_1), \dots, (X_{\pi^*(n)}, Y_n) \overset{\text{i.i.d.}}{\sim} N^{\otimes d}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1,\rho\\\rho,1\end{bmatrix}\right)$$

- Marginally, both X_i's and Y_i's are i.i.d. standard Gaussian random vectors in $\mathbb{R}^d$
- The inherent correlation under H_1 is obscured by latent matching π^*
- Tests need to be permutation invariant
- What is the minimum ρ needed for achieving vanishing testing error as $n \to \infty$?
- Recent works consider high dimensions [Dai-Cullina-Kiyavash '19,'20, Kunisky-Niles-Weed '22, Wang-Wu-X-Yolou '22, K-Nazer '22, Elimelech-Huleihel '23]

Optimal test

- Optimal Type-I+II error equals $1 - \mathrm{TV}(\mathbb{P}_0, \mathbb{P}_1)$, attained by LRT

$$L(\mathbf{X}, \mathbf{Y}) \triangleq \frac{\mathbb{P}_1(\mathbf{X}, \mathbf{Y})}{\mathbb{P}_0(\mathbf{X}, \mathbf{Y})} = \frac{1}{n!} \sum_{\pi \in S_n} \prod_{i=1}^{n} K(X_{\pi(i)}, Y_i),$$

where $K(x, y) = \frac{dP_{XY}}{dP_X \otimes dP_Y}(x, y)$.
This is hard to compute (equivalent to permanent) or to analyze.

Optimal test

- Optimal Type-I+II error equals $1 - \mathrm{TV}(\mathbb{P}_0, \mathbb{P}_1)$, attained by LRT

$$L(\mathbf{X}, \mathbf{Y}) \triangleq \frac{\mathbb{P}_1(\mathbf{X}, \mathbf{Y})}{\mathbb{P}_0(\mathbf{X}, \mathbf{Y})} = \frac{1}{n!} \sum_{\pi \in S_n} \prod_{i=1}^{n} K(X_{\pi(i)}, Y_i),$$

  where $K(x, y) = \frac{dP_{XY}}{dP_X \otimes dP_Y}(x, y)$.
  This is hard to compute (equivalent to permanent) or to analyze.

- Our goal: Strong detection

$$\mathrm{TV}(\mathbb{P}_0, \mathbb{P}_1) \to 1, \quad n \to \infty$$

Optimal test

- Optimal Type-I+II error equals $1 - \mathrm{TV}(\mathbb{P}_0, \mathbb{P}_1)$, attained by LRT

$$L(\mathbf{X}, \mathbf{Y}) \triangleq \frac{\mathbb{P}_1(\mathbf{X}, \mathbf{Y})}{\mathbb{P}_0(\mathbf{X}, \mathbf{Y})} = \frac{1}{n!} \sum_{\pi \in S_n} \prod_{i=1}^{n} K(X_{\pi(i)}, Y_i),$$

  where $K(x, y) = \frac{dP_{XY}}{dP_X \otimes dP_Y}(x, y)$.
  This is hard to compute (equivalent to permanent) or to analyze.

- Our goal: Strong detection

$$\mathrm{TV}(\mathbb{P}_0, \mathbb{P}_1) \to 1, \quad n \to \infty$$

- Key questions: What are the optimal detection thresholds? Can we achieve them in poly-time?

Two key quantities

- χ^2-information:

$$I_{\chi^2}(X;Y) \triangleq \chi^2(P_{XY} \| P_X \otimes P_Y) = \mathsf{var}_{P_X \otimes P_Y}[K(X,Y)],$$

where $K(x,y) = \frac{dP_{XY}}{dP_X \otimes dP_Y}(x,y)$.

Two key quantities

- χ^2-information:

$$I_{\chi^2}(X;Y) \triangleq \chi^2(P_{XY} \| P_X \otimes P_Y) = \mathsf{var}_{P_X \otimes P_Y}[K(X,Y)],$$

  where $K(x,y) = \frac{dP_{XY}}{dP_X \otimes dP_Y}(x,y)$.

- Maximal correlation [Hirschfeld, Gebelein, Samanov, Rényi]:

$$\rho(X;Y) \triangleq \sup_{f,g} \{\mathsf{corr}(f(X), g(Y))\}$$

Two key quantities

- χ^2-information:

$$I_{\chi^2}(X;Y) \triangleq \chi^2(P_{XY} \| P_X \otimes P_Y) = \mathsf{var}_{P_X \otimes P_Y}[K(X,Y)],$$

where $K(x,y) = \frac{dP_{XY}}{dP_X \otimes dP_Y}(x,y)$.

- Maximal correlation [Hirschfeld, Gebelein, Samanov, Rényi]:

$$\rho(X;Y) \triangleq \sup_{f,g} \{\mathsf{corr}(f(X), g(Y))\}$$

- Interpretation: Define a linear operator K

$$(Kf)(x) \triangleq \mathbb{E}[f(Y)|X=x] = \int K(x,y) f(y) dP_Y(y).$$

Then $I_{\chi^2}(X;Y)$ is the squared HS norm of K and $\rho(X;Y)$ is the second largest singular value of K

Main result: fixed distributions

Theorem

Fix $P_{X,Y}$. Strong detection is possible iff $I_{\chi^2}(X;Y) = \infty$ or $\rho(X;Y) = 1$.

Main result: fixed distributions

Theorem

Fix $P_{X,Y}$. Strong detection is possible iff $I_{\chi^2}(X;Y) = \infty$ or $\rho(X;Y) = 1$.

Remarks

- Negative result shown by [Bai-Hsing '05], who also conjectured the positive result.
- Achievable, in theory, by computationally efficient tests in the sense that, for any ϵ, there exists an algorithm with run time $O_\epsilon(n)$ with test error $\leq \epsilon$.

Proof of impossibility

Goal: Assuming $I_{\chi^2}(X;Y) < \infty$ and $\rho(X;Y) < 1$, show that

$$\chi^2(\mathbb{P}_1 \| \mathbb{P}_0) + 1 = \mathbb{E}_0 \left[L^2(\mathbf{X}, \mathbf{Y}) \right] = O(1).$$

Proof of impossibility

Goal: Assuming $I_{\chi^2}(X;Y) < \infty$ and $\rho(X;Y) < 1$, show that

$$\chi^2(\mathbb{P}_1 \| \mathbb{P}_0) + 1 = \mathbb{E}_0\left[L^2(\mathbf{X}, \mathbf{Y})\right] = O(1).$$

- The operator $(Kf)(x) = \mathbb{E}[f(Y)|X = x]$ is Hilbert-Schmidt and admits a spectral decomposition (SVD):

$$\boxed{K(x,y) = \sum_{k=0}^{\infty} \lambda_k \psi_k(x) \phi_k(y)}$$

  - $1 = \lambda_0 > \lambda_1 \geq \cdots \geq 0$,
  - $I_{\chi^2}(X;Y) = \sum_{k \geq 1} \lambda_k^2$, $\rho(X;Y) = \lambda_1$
  - $\{\psi_k\}$ is an orthonormal basis for $L_2(P_X)$
  - $\{\phi_k\}$ is an orthonormal basis for $L_2(P_Y)$.

Proof of impossibility

Goal: Assuming $I_{\chi^2}(X;Y) < \infty$ and $\rho(X;Y) < 1$, show that

$$\chi^2(\mathbb{P}_1 \| \mathbb{P}_0) + 1 = \mathbb{E}_0\left[L^2(\mathbf{X}, \mathbf{Y})\right] = O(1).$$

- The operator $(Kf)(x) = \mathbb{E}[f(Y)|X = x]$ is Hilbert-Schmidt and admits a spectral decomposition (SVD):

$$\boxed{K(x,y) = \sum_{k=0}^{\infty} \lambda_k \psi_k(x) \phi_k(y)}$$

  - $1 = \lambda_0 > \lambda_1 \geq \cdots \geq 0$,
  - $I_{\chi^2}(X;Y) = \sum_{k \geq 1} \lambda_k^2$, $\rho(X;Y) = \lambda_1$
  - $\{\psi_k\}$ is an orthonormal basis for $L_2(P_X)$
  - $\{\phi_k\}$ is an orthonormal basis for $L_2(P_Y)$.
- A beautiful argument of [Bai-Hsing '05] shows

$$\chi^2(\mathbb{P}_1 \| \mathbb{P}_0) + 1 \to \sum_{k \geq 1} \frac{1}{1 - \lambda_k^2}, \quad n \to \infty.$$

Positive result [Jiao-Wu-X '24]

Goal: $\rho(X;Y) = 1$ or $I_{\chi^2}(X;Y) = \infty \implies$ strong detection. Instead of analyzing LRT, we construct explicit tests.

Positive result [Jiao-Wu-X '24]

Goal: $\rho(X;Y) = 1$ or $I_{\chi^2}(X;Y) = \infty \implies$ strong detection. Instead of analyzing LRT, we construct explicit tests.

- Suppose $\rho(X;Y) = 1$. Then

$$T = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_1(X_i) - \phi_1(Y_i) \to \begin{cases} N(0,2) & \text{under } H_0 \\ \delta_0 & \text{under } H_1 \end{cases}$$

Positive result [Jiao-Wu-X '24]

Goal: $\rho(X;Y) = 1$ or $I_{\chi^2}(X;Y) = \infty \implies$ strong detection. Instead of analyzing LRT, we construct explicit tests.

- Suppose $\rho(X;Y) = 1$. Then

$$T = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_1(X_i) - \phi_1(Y_i) \to \begin{cases} N(0,2) & \text{under } H_0 \\ \delta_0 & \text{under } H_1 \end{cases}$$

- Next, assume $I_{\chi^2}(X;Y) = \infty$.

Sufficient statistics

- Observation: empirical distributions

$$\hat{P}_X = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}, \quad \hat{P}_Y = \frac{1}{n}\sum_{i=1}^{n} \delta_{Y_i}$$

are sufficient statistics

Sufficient statistics

- Observation: empirical distributions

$$\hat{P}_X = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}, \quad \hat{P}_Y = \frac{1}{n}\sum_{i=1}^{n} \delta_{Y_i}$$

  are sufficient statistics

- For real-valued data, the empirical CDFs have Gaussian fluctuations:

$$\hat{F}_X(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i \le t\} \approx F_X(t) + \frac{1}{\sqrt{n}} B_X(t)$$

$$\hat{F}_Y(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{Y_i \le t\} \approx F_Y(t) + \frac{1}{\sqrt{n}} B_Y(t),$$

  where (B_X, B_Y) are independent Brownian bridges under H_0 and correlated Brownian bridges under H_1

Sufficient statistics

- Observation: empirical distributions

$$\hat{P}_X = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}, \quad \hat{P}_Y = \frac{1}{n}\sum_{i=1}^{n} \delta_{Y_i}$$

  are sufficient statistics

- For real-valued data, the empirical CDFs have Gaussian fluctuations:

$$\hat{F}_X(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i \le t\} \approx F_X(t) + \frac{1}{\sqrt{n}} B_X(t)$$

$$\hat{F}_Y(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{Y_i \le t\} \approx F_Y(t) + \frac{1}{\sqrt{n}} B_Y(t),$$

  where (B_X, B_Y) are independent Brownian bridges under H_0 and correlated Brownian bridges under H_1

- This motivates tests based on histograms [Ding-Ma-Wu-X '18]

Test based on histograms

- Variational representation of divergence [Gelfand-Yaglom-Perez '56]:

$$I_{\chi^2}(X;Y) = \sup I_{\chi^2}(X_{\mathcal{P}};Y_{\mathcal{P}'})$$

  with sup taken over all finite partitions $\mathcal{P} = (A_1, \ldots, A_m)$ and $\mathcal{P}' = (B_1, \ldots, B_m)$ of X and Y spaces respectively, and $X_{\mathcal{P}}, Y_{\mathcal{P}'}$ are the quantized version.
- Since $I_{\chi^2}(X;Y) = \infty$, we fix a partition s.t. $I_{\chi^2}(X_{\mathcal{P}};Y_{\mathcal{P}'}) \gg 1$.

Test based on histograms

- Centered histograms:

$$U = \left(\sqrt{\frac{1}{n}} \sum_{i=1}^{n} (\mathbf{1}\{X_i \in A_j\} - P_X(A_j)) \right)_{j=1,\ldots,m},$$

$$V = \left(\sqrt{\frac{1}{n}} \sum_{i=1}^{n} (\mathbf{1}\{Y_i \in B_j\} - P_Y(B_j)) \right)_{j=1,\ldots,m}$$

Test based on histograms

- Centered histograms:

$$U = \left(\sqrt{\frac{1}{n}} \sum_{i=1}^{n} (\mathbf{1}\{X_i \in A_j\} - P_X(A_j)) \right)_{j=1,\ldots,m},$$

$$V = \left(\sqrt{\frac{1}{n}} \sum_{i=1}^{n} (\mathbf{1}\{Y_i \in B_j\} - P_Y(B_j)) \right)_{j=1,\ldots,m}$$

- Gaussian limits: $(U, V) \xrightarrow{n\to\infty} N(0, \Sigma_i)$ under H_i, $i = 0, 1$, where

$$\Sigma_0 = \begin{bmatrix} * & 0 \\ 0 & * \end{bmatrix} \quad \Sigma_1 = \begin{bmatrix} * & ** \\ ** & * \end{bmatrix}$$

are simultaneously diagonalizable, with eigenvalues determined by singular values $\lambda_{k,m}$ of the discretized operator

Reduce to testing two Gaussians

Test $N(0, \Sigma_0)$ vs $N(0, \Sigma_1)$:

- Optimal test: quadratic classifier (QDA) $T = (u^\top, v^\top)(\Sigma_0^\dagger - \Sigma_1^\dagger) \binom{u}{v}$

Reduce to testing two Gaussians

Test $N(0, \Sigma_0)$ vs $N(0, \Sigma_1)$:

- Optimal test: quadratic classifier (QDA) $T = (u^\top, v^\top)(\Sigma_0^\dagger - \Sigma_1^\dagger) \binom{u}{v}$
- Optimal error satisfies a dimension-free bound [Hajék,1958]:

$$1 - \mathrm{TV} \lesssim \mathrm{SKL}^{-1/4}$$

The symmetric KL divergence can be found to be

$$\mathrm{SKL}(N(0, \Sigma_0), N(0, \Sigma_1)) = \sum_k \frac{\lambda_{k,m}^2}{1 - \lambda_{k,m}^2}$$

Reduce to testing two Gaussians

Test $N(0, \Sigma_0)$ vs $N(0, \Sigma_1)$:

- Optimal test: quadratic classifier (QDA) $T = (u^\top, v^\top)(\Sigma_0^\dagger - \Sigma_1^\dagger)\binom{u}{v}$
- Optimal error satisfies a dimension-free bound [Hajék,1958]:

$$1 - \mathrm{TV} \lesssim \mathrm{SKL}^{-1/4}$$

The symmetric KL divergence can be found to be

$$\mathrm{SKL}(N(0, \Sigma_0), N(0, \Sigma_1)) = \sum_k \frac{\lambda_{k,m}^2}{1 - \lambda_{k,m}^2}$$

- By construction,
$I_{\chi^2}(X_{\mathcal{P}}; Y_{\mathcal{P}'}) = \sum \lambda_{k,m}^2 \gg 1 \implies \mathrm{SKL} \gg 1 \implies 1 - \mathrm{TV} \ll 1.$

Main result: varying distributions

Theorem

Let P_{XY} be a general distribution which may depend on n. Under extra condition on $\mathbb{E}_0[K^3(X,Y)]$, strong detection is possible iff $I_{\chi^2}(X;Y) \to \infty$ or $\rho(X;Y) \to 1$.

Remarks

- Negative result applies the same argument of [Bai-Hsing '05]
- Positive results by analyzing non-asymptotical tests based on histograms or eigenfunctions.

Example: detecting correlated Gaussians

Consider $P_{X,Y} = N(\binom{0}{0}, \binom{1\ \rho}{\rho\ 1})^{\otimes d}$ and $K(x,y) = \frac{dP_{XY}}{dP_X \otimes dP_Y}(x,y)$

- K is Mehler kernel, diagonalized by Hermite polynomials:

$$K(x,y) = \prod_{j=1}^{d} \sum_{k=0}^{\infty} \rho^{2k} H_k(x_j) H_k(y_j)$$

- $I_{\chi^2}(X;Y) = (1-\rho^2)^{-d}$ and $\rho(X;Y) = \rho$.

Example: detecting correlated Gaussians

Consider $P_{X,Y} = N(\binom{0}{0}, \binom{1\ \rho}{\rho\ 1})^{\otimes d}$ and $K(x,y) = \frac{dP_{XY}}{dP_X \otimes dP_Y}(x,y)$

- K is Mehler kernel, diagonalized by Hermite polynomials:

$$K(x,y) = \prod_{j=1}^{d} \sum_{k=0}^{\infty} \rho^{2k} H_k(x_j) H_k(y_j)$$

- $I_{\chi^2}(X;Y) = (1-\rho^2)^{-d}$ and $\rho(X;Y) = \rho$.
- Strong detection is possible iff $(1-\rho^2)^{-d} \to \infty$:
  - Low dimensions $d = O(1)$: $\rho^2 \to 1$ (near perfect correlation)
  - High dimensions $d \to \infty$: $\rho^2 d \to \infty$ (vanishing correlation works)

Example: detecting correlated Gaussians

Consider $P_{X,Y} = N(\binom{0}{0}, \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right))^{\otimes d}$ and $K(x,y) = \frac{dP_{XY}}{dP_X \otimes dP_Y}(x,y)$

- K is Mehler kernel, diagonalized by Hermite polynomials:

$$K(x,y) = \prod_{j=1}^{d} \sum_{k=0}^{\infty} \rho^{2k} H_k(x_j) H_k(y_j)$$

- $I_{\chi^2}(X;Y) = (1-\rho^2)^{-d}$ and $\rho(X;Y) = \rho$.
- Strong detection is possible iff $(1-\rho^2)^{-d} \to \infty$:
  - Low dimensions $d = O(1)$: $\rho^2 \to 1$ (near perfect correlation)
  - High dimensions $d \to \infty$: $\rho^2 d \to \infty$ (vanishing correlation works)
- This resolves the detection limit and improves over SOTA [K-Nazer '22, Elimelech-Huleihel '23]

	Possible	Impossible
$d \to \infty$	$\rho^2 = \omega(1/d)$	$\rho^2 < 1/d$
$d = O(1)$	$\rho^2 = 1 - o(n^{-\frac{2}{d-1}})$	$\rho^2 < \rho^*(d)$

Comparison with recovery

Gaussian	Threshold
Detection	$(1-\rho^2)^{-d} \to \infty$

Comparison with recovery

Gaussian	Threshold
Detection	$(1-\rho^2)^{-d} \to \infty$
Almost exact recovery	
Exact recovery	

Comparison with recovery

Gaussian	Threshold
Detection	$(1-\rho^2)^{-d} \to \infty$
Almost exact recovery	$(1-\rho^2)^{-d} \geq n^2$
Exact recovery	$(1-\rho^2)^{-d} \geq n^4$

Comparison with recovery

Gaussian	Threshold
Detection	$(1-\rho^2)^{-d} \to \infty$
Almost exact recovery	$(1-\rho^2)^{-d} \geq n^2$
Exact recovery	$(1-\rho^2)^{-d} \geq n^4$

- Exact and almost exact recovery of π^* are achieved by maximum likelihood [Dai-Cullina-Kiyavash '19,'20, Kunisky-Niles-Weed '22, Wang-Wu-X-Yolou '22]:

$$\min_{\pi} \sum_{i=1}^{n} \|X_{\pi(i)} - Y_i\|^2$$

Comparison with recovery

Gaussian	Threshold
Detection	$(1-\rho^2)^{-d} \to \infty$
Almost exact recovery	$(1-\rho^2)^{-d} \geq n^2$
Exact recovery	$(1-\rho^2)^{-d} \geq n^4$

- Exact and almost exact recovery of π^* are achieved by maximum likelihood [Dai-Cullina-Kiyavash '19,'20, Kunisky-Niles-Weed '22, Wang-Wu-X-Yolou '22]:

$$\min_{\pi} \sum_{i=1}^{n} \|X_{\pi(i)} - Y_i\|^2$$

  - This is a linear assignment problem and can be solved in poly-time!
  - Optimal objective $=$ squared Wasserstein-2 distance between empirical distributions of X_i's and Y_i's

Comparison with recovery

Gaussian	Threshold
Detection	$(1-\rho^2)^{-d} \to \infty$
Almost exact recovery	$(1-\rho^2)^{-d} \geq n^2$
Exact recovery	$(1-\rho^2)^{-d} \geq n^4$

- Exact and almost exact recovery of π^* are achieved by maximum likelihood [Dai-Cullina-Kiyavash '19,'20, Kunisky-Niles-Weed '22, Wang-Wu-X-Yolou '22]:

$$\min_{\pi} \sum_{i=1}^{n} \|X_{\pi(i)} - Y_i\|^2$$

  - This is a linear assignment problem and can be solved in poly-time!
  - Optimal objective $=$ squared Wasserstein-2 distance between empirical distributions of X_i's and Y_i's
- There is no theory for recovery for general distributions.

Concluding remarks

- Broken sample problems provide rich venues for theoretical study of statistical vs computational limits with many open problems
- They are deeply connected to assignment problems and optimal transport theory
- Many interesting variants (e.g., partially shuffled data) and connections: geometric matching, database alignment, particle tracking, uncoupled isotonic regression, ranking, data seriation
- Numerous applications in diverse fields

Concluding remarks

- Broken sample problems provide rich venues for theoretical study of statistical vs computational limits with many open problems
- They are deeply connected to assignment problems and optimal transport theory
- Many interesting variants (e.g., partially shuffled data) and connections: geometric matching, database alignment, particle tracking, uncoupled isotonic regression, ranking, data seriation
- Numerous applications in diverse fields

References

- Simiao Jiao, Yihong Wu, & Jiaming Xu. *The broken sample problem revisited: Proof of a conjecture by Bai-Hsing and high-dimensional extensions, Draft.*

Example: detecting correlated Gaussians

- Applying QDA to sample means is optimal:

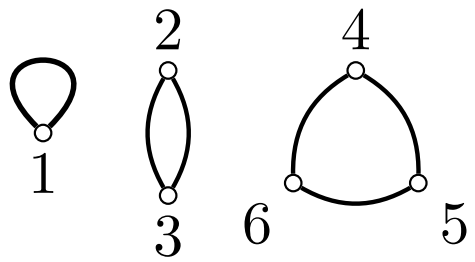
$$T = 2(1-\rho)\langle \overline{X}, \overline{Y} \rangle - \rho \|\overline{X} - \overline{Y}\|_2^2$$

  where $\overline{X} = \frac{1}{\sqrt{n}}(X_1 + \ldots + X_n)$ and $\overline{Y} = \frac{1}{\sqrt{n}}(Y_1 + \ldots + Y_n)$.
- Further simplification:
  - Low dimensions: $T = \|\overline{X} - \overline{Y}\|_2^2$.
  - High dimensions: $T = \langle \overline{X}, \overline{Y} \rangle$, previously considered by [K-Nazer '22]

Bai-Hsing's argument

Let N_ℓ denote the number of ℓ-cycles in the cycle decomposition of a random permutation π.

Example: $n = 6$ and $\pi = (1)(23)(456)$:

π: 

$N_1 = 1$, $N_2 = 1$, and $N_3 = 1$

Bai-Hsing's argument

$$\mathbb{E}_0 L^2(\mathbf{X}, \mathbf{Y})$$

$$= \mathbb{E}_{\pi, \mathbf{X}, \mathbf{Y}} \prod_{i=1}^{n} K(X_i, Y_i) K(X_i, Y_{\pi(i)})$$

Bai-Hsing's argument

$$
\begin{aligned}
&\mathbb{E}_0 L^2(\mathbf{X}, \mathbf{Y}) \\
&= \mathbb{E}_{\pi,\mathbf{X},\mathbf{Y}} \prod_{i=1}^{n} K(X_i, Y_i) K(X_i, Y_{\pi(i)}) \\
&= \mathbb{E}_{\pi,\mathbf{X},\mathbf{Y}} \prod_{i=1}^{n} \sum \lambda_k \psi_k(X_i) \phi_k(Y_i) \sum \lambda_k \psi_k(X_i) \phi_k(Y_{\pi(i)})
\end{aligned}
$$

Bai-Hsing's argument

$$
\begin{aligned}
& \mathbb{E}_0 L^2(\mathbf{X}, \mathbf{Y}) \\
= & \mathbb{E}_{\pi, \mathbf{X}, \mathbf{Y}} \prod_{i=1}^{n} K(X_i, Y_i) K(X_i, Y_{\pi(i)}) \\
= & \mathbb{E}_{\pi, \mathbf{X}, \mathbf{Y}} \prod_{i=1}^{n} \sum \lambda_k \psi_k(X_i) \phi_k(Y_i) \sum \lambda_k \psi_k(X_i) \phi_k(Y_{\pi(i)}) \\
= & \mathbb{E}_{\pi, \mathbf{Y}} \prod_{i=1}^{n} \sum \lambda_k^2 \phi_k(Y_i) \phi_k(Y_{\pi(i)}) \qquad \{\psi_k\} \text{ ON}
\end{aligned}
$$

Bai-Hsing's argument

$$
\begin{aligned}
& \mathbb{E}_0 L^2(\mathbf{X}, \mathbf{Y}) \\
= \; & \mathbb{E}_{\pi, \mathbf{X}, \mathbf{Y}} \prod_{i=1}^{n} K(X_i, Y_i) K(X_i, Y_{\pi(i)}) \\
= \; & \mathbb{E}_{\pi, \mathbf{X}, \mathbf{Y}} \prod_{i=1}^{n} \sum \lambda_k \psi_k(X_i) \phi_k(Y_i) \sum \lambda_k \psi_k(X_i) \phi_k(Y_{\pi(i)}) \\
= \; & \mathbb{E}_{\pi, \mathbf{Y}} \prod_{i=1}^{n} \sum \lambda_k^2 \phi_k(Y_i) \phi_k(Y_{\pi(i)}) && \{\psi_k\} \text{ ON} \\
= \; & \mathbb{E}_{\pi} \prod_{\ell=1}^{n} \left(\sum \lambda_k^{2\ell} \right)^{N_\ell} && \{\phi_k\} \text{ ON}
\end{aligned}
$$

Bai-Hsing's argument

$$
\begin{aligned}
&\mathbb{E}_0 L^2(\mathbf{X}, \mathbf{Y}) \\
&= \mathbb{E}_{\pi,\mathbf{X},\mathbf{Y}} \prod_{i=1}^{n} K(X_i, Y_i) K(X_i, Y_{\pi(i)}) \\
&= \mathbb{E}_{\pi,\mathbf{X},\mathbf{Y}} \prod_{i=1}^{n} \sum \lambda_k \psi_k(X_i) \phi_k(Y_i) \sum \lambda_k \psi_k(X_i) \phi_k(Y_{\pi(i)}) \\
&= \mathbb{E}_{\pi,\mathbf{Y}} \prod_{i=1}^{n} \sum \lambda_k^2 \phi_k(Y_i) \phi_k(Y_{\pi(i)}) && \{\psi_k\} \text{ ON} \\
&= \mathbb{E}_{\pi} \prod_{\ell=1}^{n} \left(\sum \lambda_k^{2\ell} \right)^{N_\ell} && \{\phi_k\} \text{ ON}
\end{aligned}
$$

To bound this we can apply Poisson approximation: $N_\ell \approx \text{Pois}(1/\ell)$

Bai-Hsing's argument

Let $f(z) = \prod_{k=0}^{\infty}(1 - z\lambda_k^2)^{-1}$ and $1 < r < 1/\lambda_1^2$:

Bai-Hsing's argument

Let $f(z) = \prod_{k=0}^{\infty}(1 - z\lambda_k^2)^{-1}$ and $1 < r < 1/\lambda_1^2$:

Bai-Hsing's argument

Let $f(z) = \prod_{k=0}^{\infty}(1 - z\lambda_k^2)^{-1}$ and $1 < r < 1/\lambda_1^2$:

$$\mathbb{E}\left[\prod_{\ell=1}^{n}\left(\sum_{k=0}^{\infty}\lambda_k^{2\ell}\right)^{N_\ell}\right] = [z^n]f(z)$$ (n-th coefficient in power series)

Bai-Hsing's argument

Let $f(z) = \prod_{k=0}^{\infty}(1 - z\lambda_k^2)^{-1}$ and $1 < r < 1/\lambda_1^2$:

$$\mathbb{E}\left[\prod_{\ell=1}^{n}\left(\sum_{k=0}^{\infty}\lambda_k^{2\ell}\right)^{N_\ell}\right] = [z^n]f(z) \text{ (n-th coefficient in power series)}$$

$$\text{(Cauchy's integral thm)} = \frac{1}{2\pi\mathbf{i}}\oint_{|z|=r}\frac{1}{z^{n+1}}f(z)\mathrm{d}z - \lim_{z\to 1}(z-1)f(z)$$

Bai-Hsing's argument

Let $f(z) = \prod_{k=0}^{\infty}(1 - z\lambda_k^2)^{-1}$ and $1 < r < 1/\lambda_1^2$:

$$\mathbb{E}\left[\prod_{\ell=1}^{n}\left(\sum_{k=0}^{\infty}\lambda_k^{2\ell}\right)^{N_\ell}\right] = [z^n]f(z) \text{ (n-th coefficient in power series)}$$

$$\text{(Cauchy's integral thm)} = \frac{1}{2\pi\mathbf{i}}\oint_{|z|=r}\frac{1}{z^{n+1}}f(z)\mathrm{d}z - \lim_{z\to 1}(z-1)f(z)$$

$$= o(1) + \prod_{k=1}^{\infty}(1-\lambda_k^2)^{-1}, \quad \text{as } n \to \infty$$