DIFFERENTIALLY-PRIVATE CANONICAL CORRELATION ANALYSIS

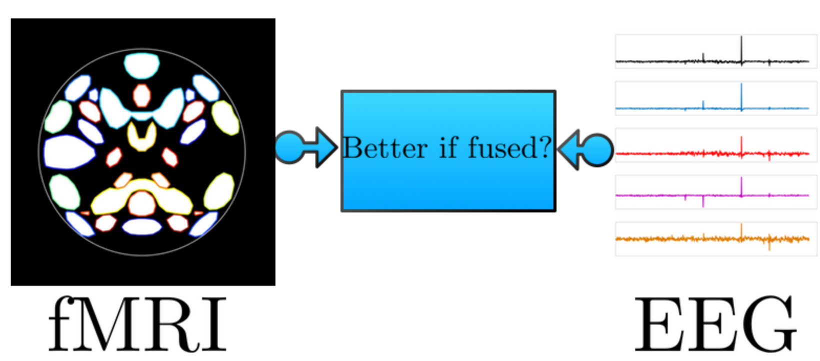
Hafiz Imtiaz and Anand D. Sarwate

Rutgers University

Motivation

Goal: measure linear relationship among variables
→ can use correlation

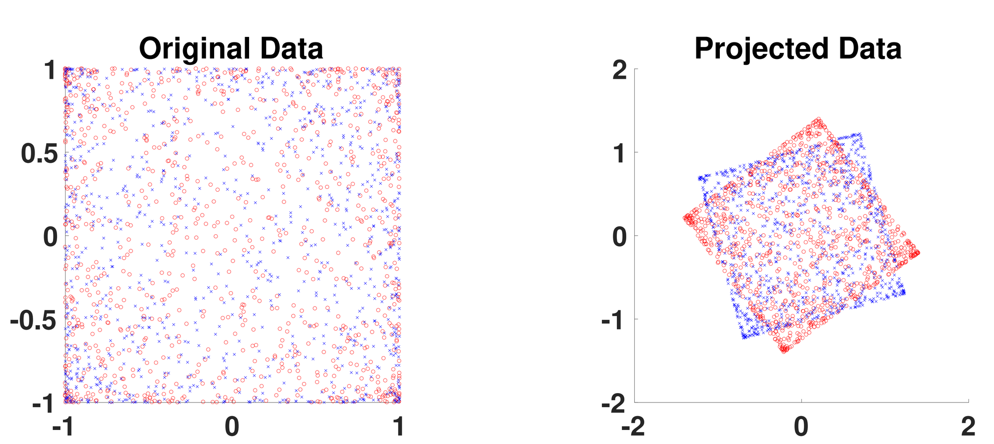
Challenges: data can be privacy-sensitive
→ how to guarantee privacy?
→ what is the best correlation metric?
→ how to measure it?



Is there a privacy-preserving way to compute the best correlation index?

Canonical Correlation Analysis (CCA)

CCA finds subspaces for different "views" of data
→ "views" are maximally correlated after projection



Can we have a CCA algorithm that preserves privacy and also provides good utility?

Problem Formulation

→ variables or views: $\mathbf{X} \in \mathbb{R}^{D_x \times N}$ and $\mathbf{Y} \in \mathbb{R}^{D_y \times N}$

→ **goal:** find subspaces $\mathbf{U} \in \mathbb{R}^{D_x \times K}$ and $\mathbf{V} \in \mathbb{R}^{D_y \times K}$

→ **how?:** solve the following optimization problem [1]

$$\begin{aligned} \underset{\mathbf{U},\mathbf{V}}{\text{minimize}} \quad & \|\mathbf{U}^\top \mathbf{X} - \mathbf{V}^\top \mathbf{Y}\|_F^2 \\ \text{subject to} \quad & \frac{1}{N}\mathbf{U}^\top \mathbf{X}\mathbf{X}^\top \mathbf{U} = \mathbf{I}, \frac{1}{N}\mathbf{V}^\top \mathbf{Y}\mathbf{Y}^\top \mathbf{V} = \mathbf{I}, \\ & \frac{1}{N}\mathbf{U}^\top \mathbf{X}\mathbf{Y}^\top \mathbf{V} = \mathbf{I}. \end{aligned}$$

Closed-form solution exists: [3]

- $\mathbf{U} \leftarrow$ the top-K eigenvectors of $\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}$
- $\mathbf{V} \leftarrow$ the top-K eigenvectors of $\mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}$

Differential Privacy (DP)

Differential privacy is *formal* and *quantifiable*

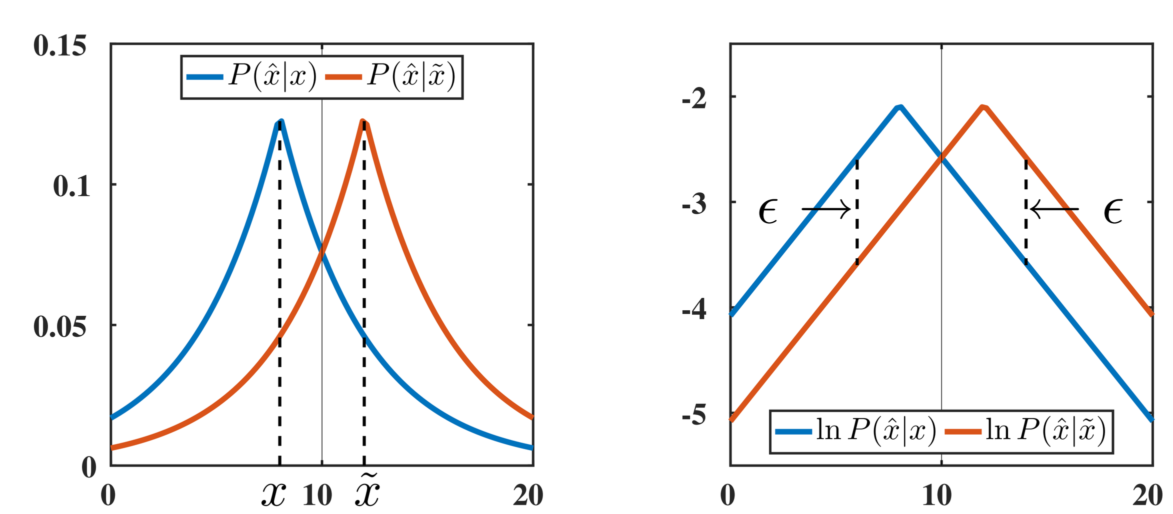
Definition: Algorithm $\mathcal{A}(\mathbb{D})$ taking values in a set $\mathbb{T}$ provides (ϵ, δ)-differential privacy if [2]

$$P(\mathcal{A}(\mathbb{D}) \in \mathbb{S}) \leq e^{\epsilon} P(\mathcal{A}(\mathbb{D}') \in \mathbb{S}) + \delta,$$

for all measurable $\mathbb{S} \subseteq \mathbb{T}$ and all *neighboring* data sets $\mathbb{D}$ and $\mathbb{D}'$ differing in a single entry.

Interpretation:

- $(\epsilon, \delta) \downarrow \Rightarrow$ privacy level $\uparrow \equiv$ noise level $\uparrow \Rightarrow$ utility $\downarrow$



Algorithm: Differentially-private CCA

Input:

- 0-centered samples $\mathbf{X}$ and $\mathbf{Y}$ as $\mathbf{Z} = [\mathbf{X};\ \mathbf{Y}]$; $\|\mathbf{z}_n\|_2 \leq 1$
- privacy parameters ϵ_o, δ

1. Compute $\mathbf{C} \leftarrow \frac{1}{N}\mathbf{Z}\mathbf{Z}^\top$
2. Generate $D \times D$ symmetric matrix $\mathbf{E}$ [2]:
   - $\{E_{ij} : i \in [D], j \leq i\}$ drawn i.i.d. $\sim \mathcal{N}(0, \tau^2)$
   - $\tau = \frac{\sqrt{2}}{N\epsilon_o}\sqrt{2\log\left(\frac{1.25}{\delta}\right)}$
   - $E_{ij} = E_{ji}$
3. Compute $\hat{\mathbf{C}} \leftarrow \mathbf{C} + \mathbf{E}$
4. Extract sub-matrices from $\hat{\mathbf{C}}$ according to:

$$\hat{\mathbf{C}} = \begin{bmatrix} \hat{\mathbf{C}}_{xx} & \hat{\mathbf{C}}_{xy} \\ \hat{\mathbf{C}}_{xy}^\top & \hat{\mathbf{C}}_{yy} \end{bmatrix}.$$

Output:

- Differentially-private approximates: $\hat{\mathbf{C}}_{xx}$, $\hat{\mathbf{C}}_{yy}$ and $\hat{\mathbf{C}}_{xy}$

Using $\hat{\mathbf{C}}_{xx}$, $\hat{\mathbf{C}}_{yy}$ and $\hat{\mathbf{C}}_{xy}$, we can compute the subspaces $\mathbf{U}$ and $\mathbf{V}$

Some Remarks

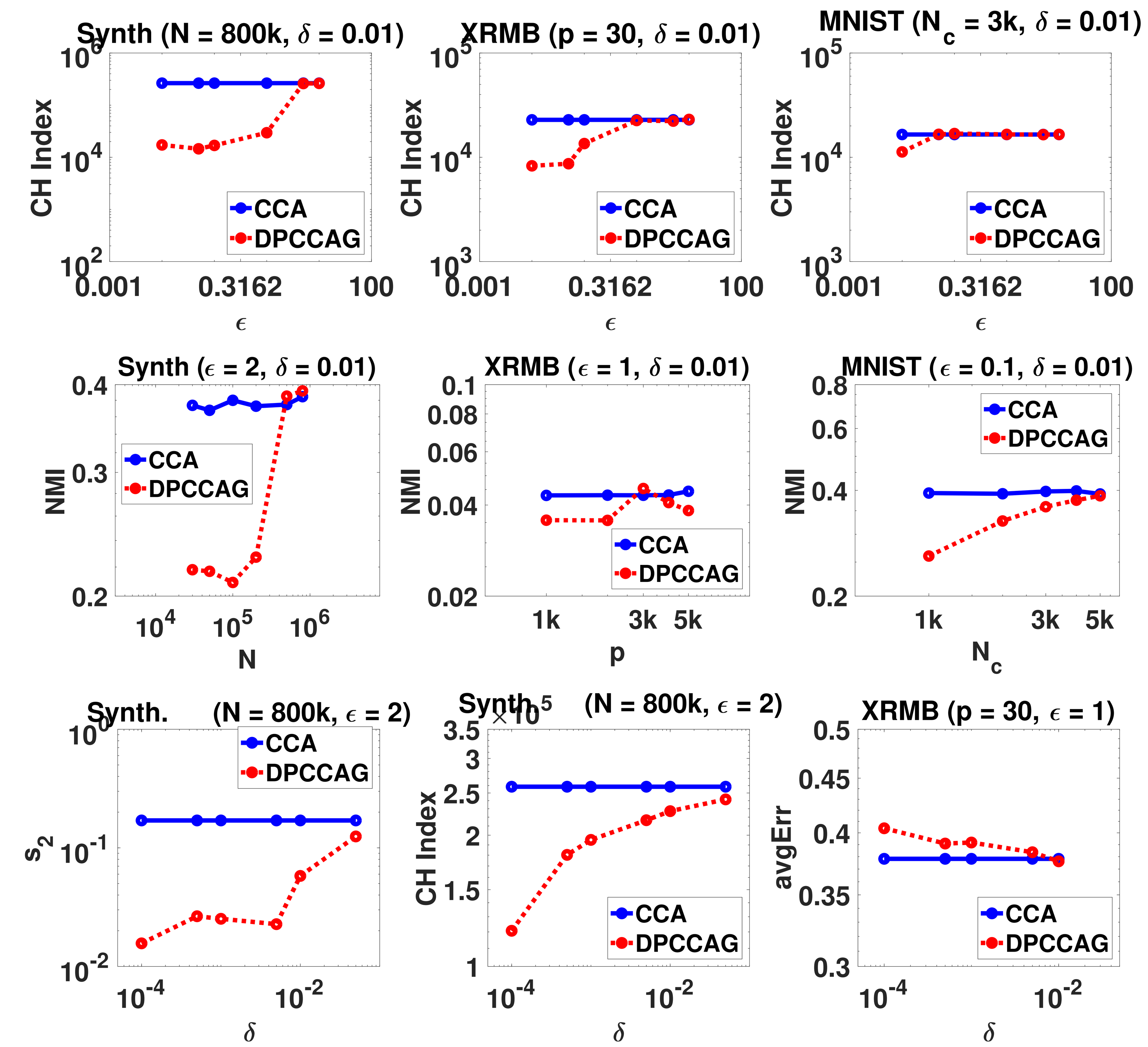
- **Analyze Gauss (AG)** algorithm: input perturbation on 2nd-moment matrix [2]
- DP is post-processing invariant $\Rightarrow$ computation of $\mathbf{U}$ and $\mathbf{V}$ is (ϵ_o, δ)-DP
- However, projection/clustering do not satisfy DP $\Rightarrow$ can be modified at the cost of utility

Simulation Results

Performance Variation with ϵ

Performance Variation with N

Performance Variation with δ



Conclusion and Future Works

Remarks:

- for fixed ϵ (privacy level): more samples → better performance
- for fixed N (sample size): higher ϵ → better performance
- **observation**: the proposed algorithm can achieved meaningful utility even with strict privacy

Future directions:

- novel utility bounds
- validation on high-dimensional real data (multi-modal location data)
- multi-view learning in neuroimaging (fMRI/EEG)

References

[1] Hotelling, H. (1936). Relations Between Two Sets of Variates. Biometrika, 28(3/4), 321-377. doi:10.2307/2333955

[2] Dwork, C. et al. (2014). Analyze Gauss: Optimal Bounds for Privacy-preserving Principal Component Analysis. 46-th Annual ACM Symposium on Theory of Computing. doi: doi.org/10.1145/2591796.2591883

[3] Hardoon, D. R. et al. (2004). Canonical Correlation Analysis: An Overview with Application to Learning Methods. Neural Comput. doi: doi.org/10.1162/0899766042321814