DISTRIBUTED DIFFERENTIALLY-PRIVATE CANONICAL CORRELATION ANALYSIS

Hafiz Imtiaz and Anand D. Sarwate

Rutgers University

Motivation

Goal: measure linear relationship among variables

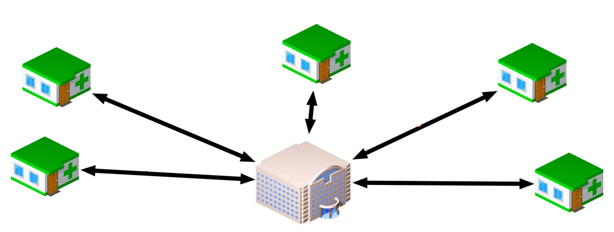
$\rightarrow$ can use correlation

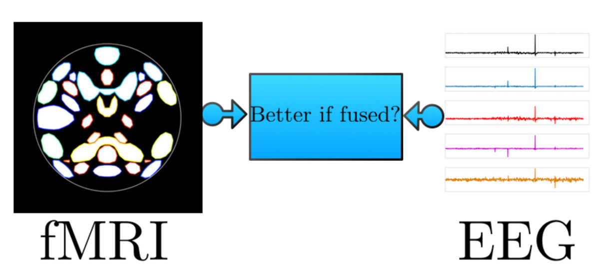
Challenges: data – privacy-sensitive and distributed

$\rightarrow$ how to guarantee privacy?

$\rightarrow$ how to measure the best correlation metric?

$\rightarrow$ how to do it in distributed setting?

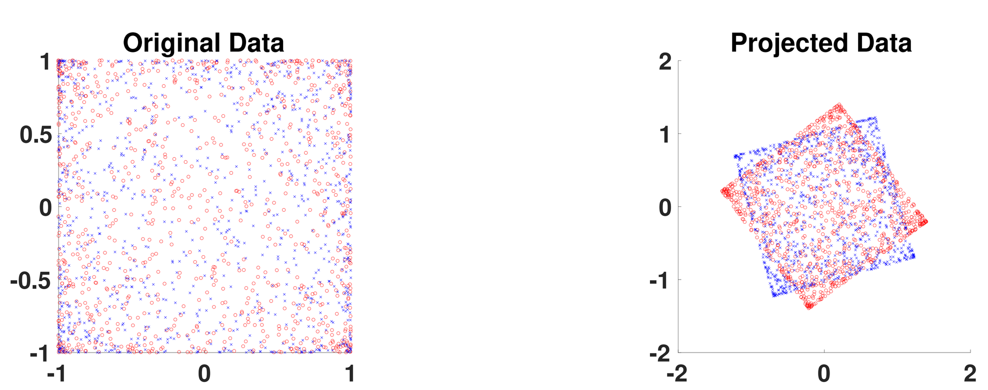




Canonical Correlation Analysis (CCA)

CCA finds subspaces for different "views" of data [1]

$\rightarrow$ "views" are maximally correlated after projection



Can we have a CCA algorithm that preserves privacy, provides good utility and operates in distributed-data setting?

Problem Formulation

$\rightarrow$ consider a system with S different sites

$\rightarrow$ site s contains views: $\mathbf{X}_s \in \mathbb{R}^{D_x \times N_s}$, $\mathbf{Y}_s \in \mathbb{R}^{D_y \times N_s}$

$\rightarrow$ pooled data scenario: $\mathbf{X} = [\mathbf{X}_1 \ldots \mathbf{X}_S] \in \mathbb{R}^{D_x \times N}$ and $\mathbf{Y} = [\mathbf{Y}_1 \ldots \mathbf{Y}_S] \in \mathbb{R}^{D_y \times N}$

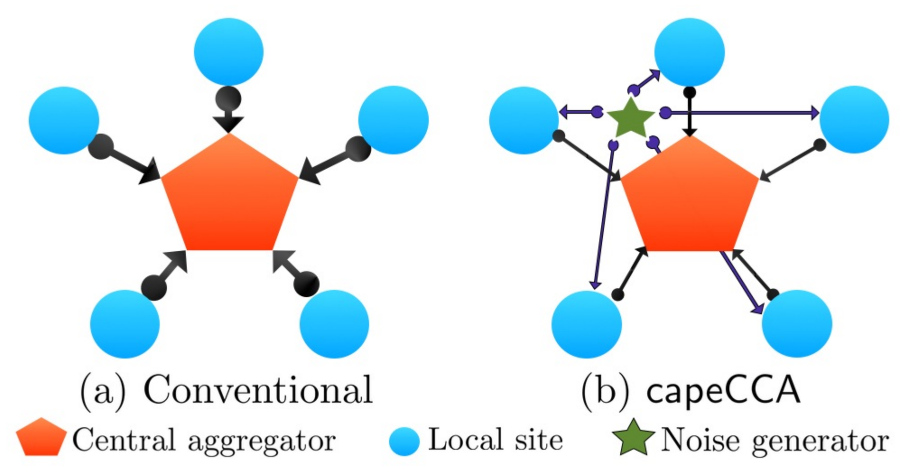
$\rightarrow$ **goal:** find subspaces $\mathbf{U} \in \mathbb{R}^{D_x \times K}$, $\mathbf{V} \in \mathbb{R}^{D_y \times K}$ [3]

$$\underset{\mathbf{U},\mathbf{V}}{\text{minimize}} \quad \|\mathbf{U}^\top \mathbf{X} - \mathbf{V}^\top \mathbf{Y}\|_F^2$$

$$\text{subject to} \quad \frac{1}{N}\mathbf{U}^\top \mathbf{X}\mathbf{X}^\top \mathbf{U} = \mathbf{I}, \frac{1}{N}\mathbf{V}^\top \mathbf{Y}\mathbf{Y}^\top \mathbf{V} = \mathbf{I}, \quad \frac{1}{N}\mathbf{U}^\top \mathbf{X}\mathbf{Y}^\top \mathbf{V} = \mathbf{I}.$$

Want to estimate U and V in the distributed setting while preserving privacy

Differential Privacy (DP)



Definition: Algorithm $\mathcal{A}(\mathbb{D})$ taking values in a set $\mathbb{T}$ provides (ϵ, δ)-differential privacy [2] if $P(\mathcal{A}(\mathbb{D}) \in \mathbb{S}) \leq e^{\epsilon} P(\mathcal{A}(\mathbb{D}') \in \mathbb{S}) + \delta$ for all measurable $\mathbb{S} \subseteq \mathbb{T}$ and all *neighboring* data sets $\mathbb{D}$ and $\mathbb{D}'$ differing in a single entry.

A conventional scheme:

- Compute $\mathbf{Z}_s = \left[\mathbf{X}_s^\top \ \mathbf{Y}_s^\top\right]^\top$ and $\mathbf{C}_s = \frac{1}{N_s}\mathbf{Z}_s\mathbf{Z}_s^\top$
- Send $\hat{\mathbf{C}}_s = \mathbf{C}_s + \mathbf{E}_s$ to aggregator, where $\{[\mathbf{E}_s]_{ij} : i \in [D], j \leq i\}$ drawn i.i.d. from $\mathcal{N}(0, \tau_s^2)$
- Aggregator computes $\hat{\mathbf{C}} = \frac{1}{S}\sum_{s=1}^{S} \hat{\mathbf{C}}_s$
- Variance of the estimator: $\tau_{\text{ag}}^2 \triangleq \frac{\tau_s^2}{S}$

$\rightarrow$ In pooled-data setting: noise variance $\tau_c^2 = \frac{\tau_s^2}{S^2}$

How can we achieve the same noise variance in the distributed setting? $\rightarrow$ employ CAPE protocol [4]

Proposed Algorithm: capeCCA

Input: $\mathbf{0}$-centered samples $\mathbf{X}_s$ and $\mathbf{Y}_s$ as $\mathbf{Z}_s = \left[\mathbf{X}_s^\top \ \mathbf{Y}_s^\top\right]^\top$, $\|\mathbf{z}_{s,n}\|_2 \leq 1$ for $s \in [S]$; privacy parameters ϵ, δ

Stage 1: Generate $\mathbf{E}_s \in \mathbb{R}^{D \times D}$

Stage 1: Generate $D \times D$ symmetric $\mathbf{G}_s$

Stage 2: Compute and send $\hat{\mathbf{C}}_s \leftarrow \frac{1}{N_s}\mathbf{Z}_s\mathbf{Z}_s^\top + \mathbf{E}_s + \mathbf{F}_s + \mathbf{G}_s$

Stage 1: Generate $\mathbf{F}_s \in \mathbb{R}^{D \times D}$

Stage 3: Compute $\hat{\mathbf{C}} \leftarrow \frac{1}{S}\sum_{s=1}^{S}\left(\hat{\mathbf{C}}_s - \mathbf{F}_s\right)$

Stage 4: Extract sub-matrices from $\hat{\mathbf{C}}$

Output: Differentially-private approximates: $\hat{\mathbf{U}}^*$ and $\hat{\mathbf{V}}^*$

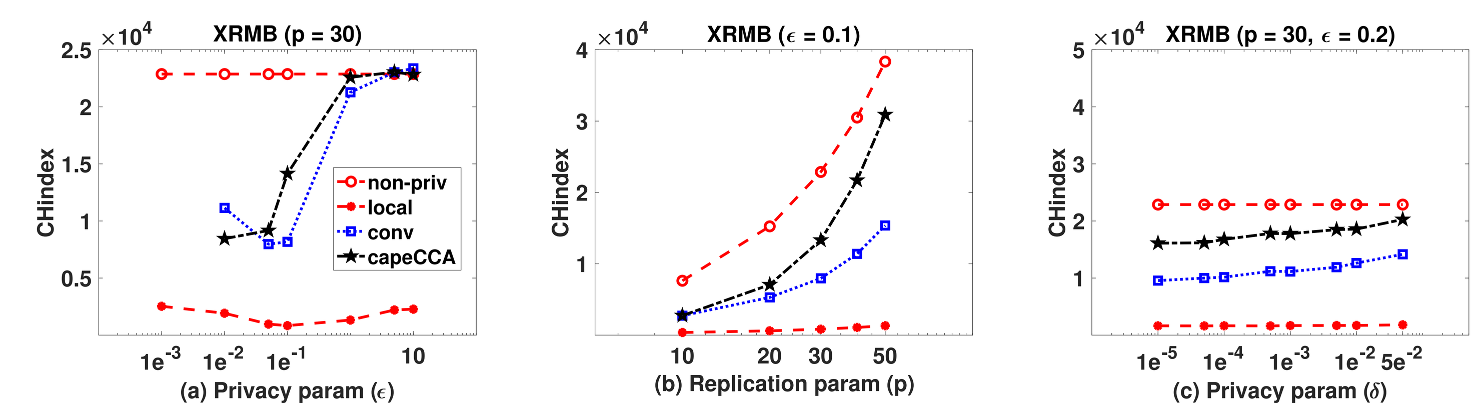
Privacy Analysis

- **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, δ)-DP
- Projection/clustering do not satisfy DP $\Rightarrow$ can be modified at the cost of utility

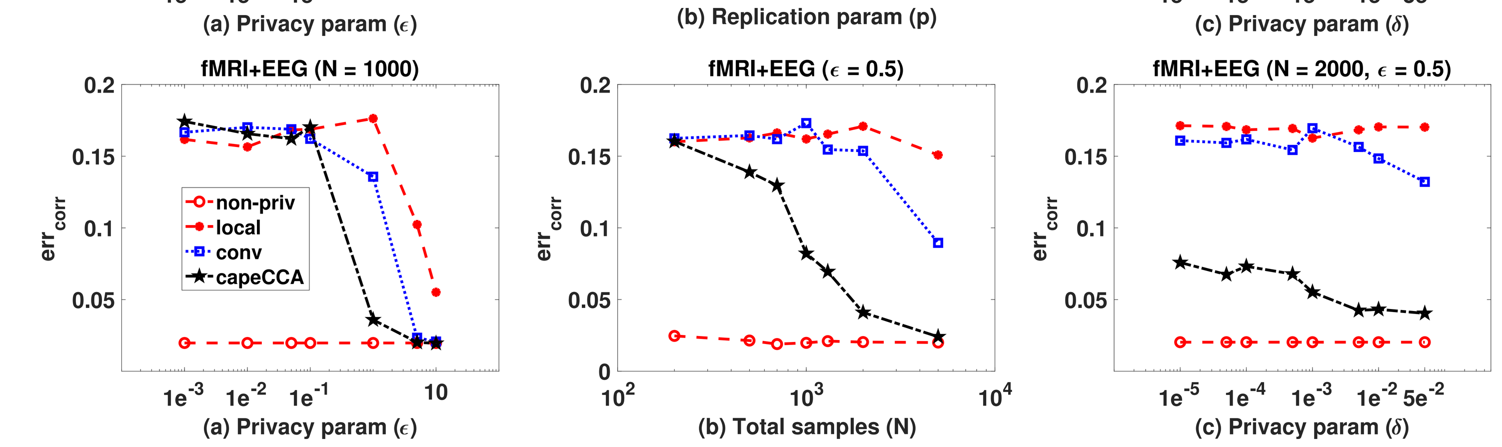
Simulation Results

- U of Wisc. X-ray Microbeam (XRMB) Dataset $\rightarrow$ view 1: speech; view 2: jaw movement
- fMRI+EEG Dataset $\rightarrow$ view 1: fMRI; view 2: EEG
- Clustering performance on XRMB $\rightarrow$ CHIndex
- Estimation of correlation on fMRI+EEG $\rightarrow$ $\mathrm{err}_{\mathrm{corr}}$

Performance Variation on XRMB Dataset



Performance Variation on fMRI + EEG Dataset



Conclusion and Future Works

capeCCA achieves the same utility as pooled-data scenario in the honest-but-curious setting

Takeaway:

- capeCCA has better utility than local and conv for the same privacy level
- capeCCA can reach non – priv in some regimes
- for fixed ϵ: more samples $\rightarrow$ better performance
- for fixed N and S: higher ϵ $\rightarrow$ better performance

Future directions:

- can we scrap the "trusted" noise generator? [4]
- can we achieve the same in an asymmetric network? [4]
- can we achieve adapt our approach to $\delta = 0$?

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. doi: 10.1145/2591796.2591883
[3] Hardoon, D. R. et al. (2004). Canonical Correlation Analysis: An Overview with Application to Learning Methods. doi: 10.1162/0899766042321814
[4] Imtiaz, H. et al. (2019). Distributed Differentially Private Computation of Functions with Correlated Noise. arXiv e-print: http://arxiv.org/abs/1904.10059