Learning latent features in images with applications to brain imaging

Hafiz Imtiaz (advisor: Prof. Anand D. Sarwate)

Rutgers, The State University of New Jersey

Context

Big-picture project: develop a system that allows researchers studying brain disorders/conditions to collaboratively analyze their data without sharing "raw" data or violating patient/subject privacy.

Example task: discover regions in the brain whose combined activity "explains" measured activity.

Challenges: MRIs are big images, but we don't have too many scans – high dimension, low sample size.

→ need to use a simple mathematical model

→ model should be effective to "capture" the relevant parts of the brain

→ better privacy guarantees ⟹ encourages sharing ⟹ better sample size

Approach: start with decentralized/distributed algorithms and then incorporate more rigorous privacy guarantees such as *differential privacy*.

Benefit: promising testbed for understanding where to improve differentially private learning:

- closed systems with trusted parties
- sharing data derivatives may satisfy privacy concerns
- explore losses from more rigorous privacy models

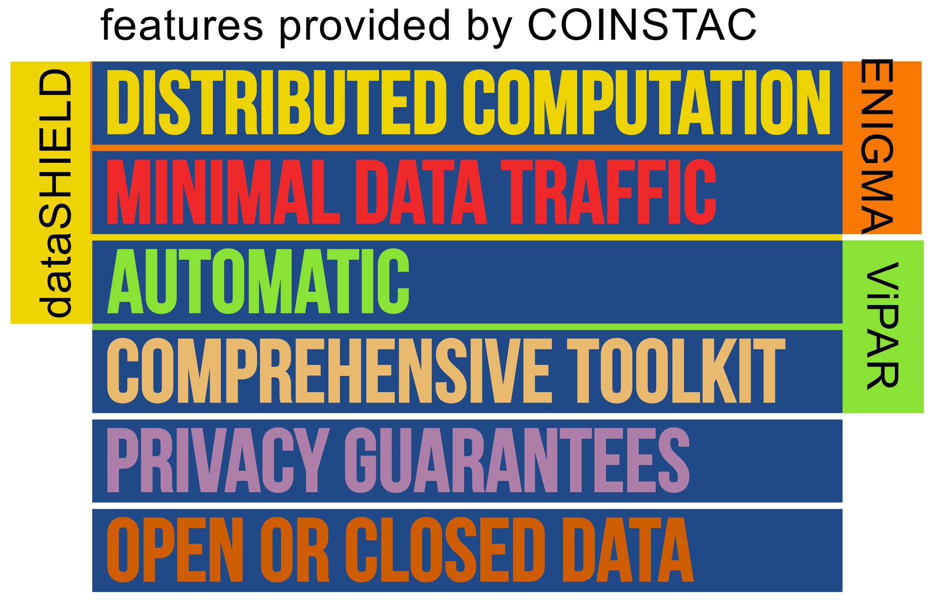
Algorithms we want to support

Many statistical/signal processing tasks can be useful in studying brain imaging:

1. Simple point estimators (means, standard deviations etc.): "what is the average volume of the hippocampus in people with a disease X?"
2. Regression and classification: "how well can we predict disease state from brain measurements?"
3. Unsupervised and supervised feature learning: "what regions in the brain are more active in patients with schizophrenia?"
4. Higher-order (tensor) analysis: "can we learn more by using the 3D structure of the brain?"
5. Data visualization: "if we cluster the patients by similarity, how many clusters do we get?"

Example goal: find structural differences that can allow classification of individuals into schizophrenic or healthy [2].

COINSTAC: a system for collaborative neuroscience



Improvements over previous systems (ViPAR, ENIGMA, dataSHIELD):

- Easier to develop and test new learning methods.
- More control over privacy and sharing policies.

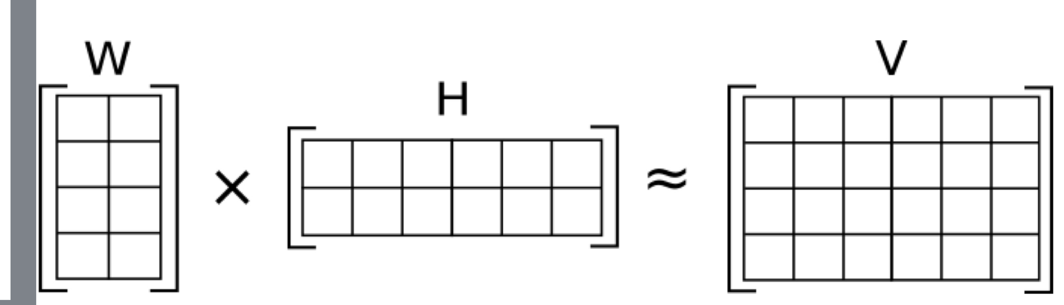
The COINSTAC system [1] extends the existing COINS (`coins.mrn.org`) system to allow automated analyses:

- Users can form ad-hoc research consortia.
- Algorithms will comply with local access policies.

Potential benefits:

- Easy deployment and testing of distributed learning methods.
- "Buy-in" to try out *privacy-sensitive* learning methods.

Nonnegative matrix factorization



Non-negative Matrix Factorization (NMF):

- Model: $\mathbf{V} = \mathbf{W}\mathbf{H}$
- Assumption: basis $\mathbf{W}$ and coefficients $\mathbf{H}$ are entry-wise non-negative
- Objective function to minimize

$$f(\mathbf{W}, \mathbf{H}) = \underset{\mathbf{W},\mathbf{H}}{\operatorname{argmin}} \|\mathbf{V} - \mathbf{W}\mathbf{H}\|_F^2$$

Independent Components Analysis (ICA):

- Model: $\mathbf{V} = \mathbf{A}\mathbf{S}$
- Assumption: sources in $\mathbf{S}_p$ are independent
- Objective function to minimize

$$I(\mathbf{A}^*) = \sum_{i=1}^{d} \sum_{p=1}^{P} h(\mathbf{s}_{p,d}) - \log|\det \mathbf{A}^*|.$$

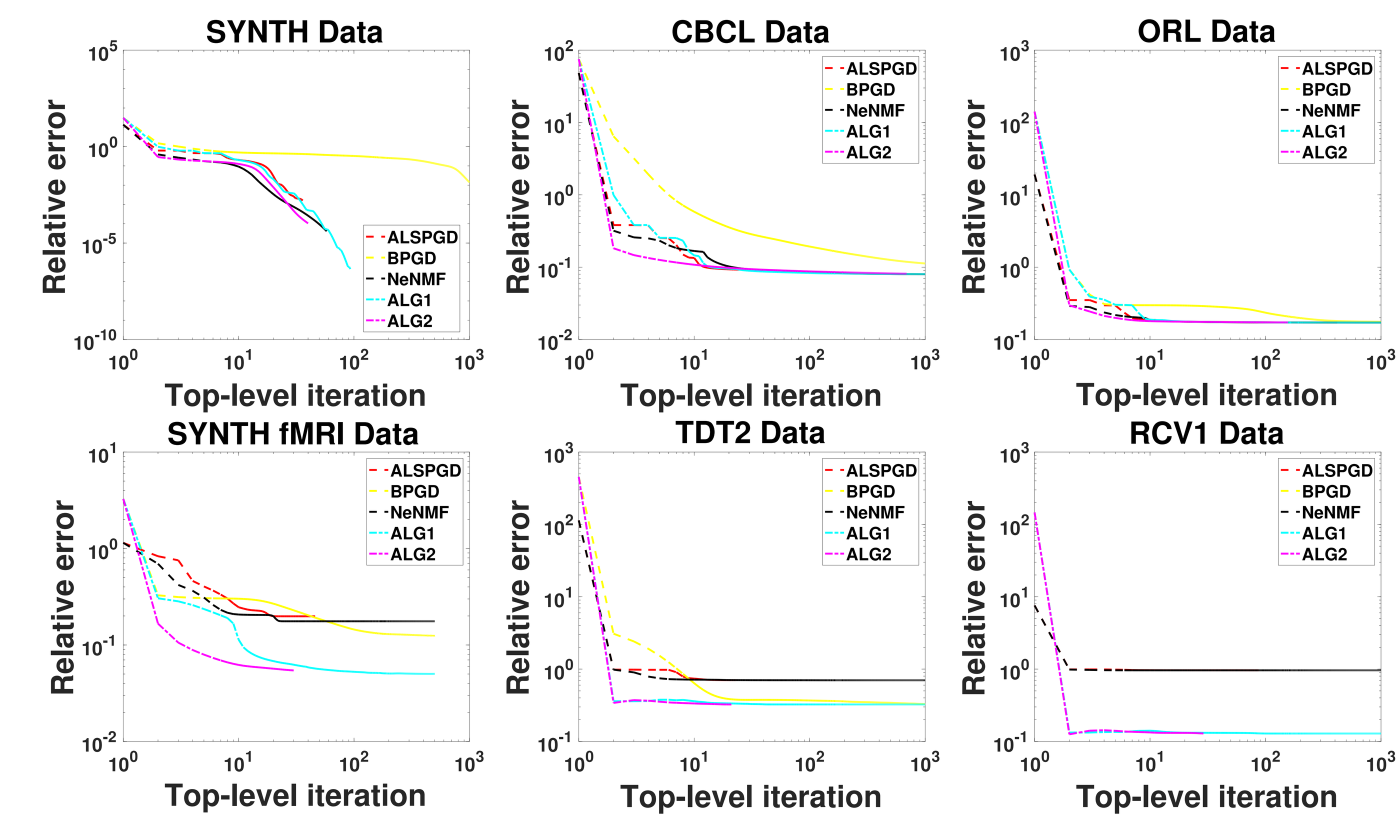
Main Idea: use iterative message exchange (e.g. gradients) simulate the centralized algorithm.

1. ICA - Pre-processing step to project data into lower dimension (e.g. PCA)
2. Both ICA & NMF - Iterative gradient descent procedure to minimize the loss
3. ICA - Incorporate differential privacy into PCA step and gradient descent.
4. NMF - Find and discard "outliers" before estimating the basis $\mathbf{W}$

Pros and Cons:

- ✓ Consortium participants may be satisfied with decentralized operation alone.
- ✓ ICA - easy use of differentially-private PCA and gradient descent step.
- ✓ Low computational burden on data holders.
- ✗ Requires a master node: not fully distributed.
- ✗ Privacy loss accumulates rapidly over iterations.
- ✗ Hard to find a bound on coefficients $\mathbf{H}$

Results: NMF



Proposed NMF with outliers:

- better relative error
- sharper decrease in objective value per iteration
- can be employed in a distributed setting
- distributed algorithm can achieve as low an error as the centralized version

Moving forward

Preliminary evidence shows what?

Some future directions in making things distributed:

- decentralized IVA and other feature learning methods
- decentralized tensor decomposition

Future directions in making things privacy-sensitive

- integrating *differential privacy* into the algorithms
- designing new models for measuring privacy loss in repeated analyses

References

[1] S. Plis et al., COINSTAC: A Privacy Enabled Model and Prototype for Leveraging and Processing Decentralized Brain Imaging Data, Frontiers in Neuroscience 10 (365), 2016.

[2] A.D. Sarwate et al., Sharing privacy-sensitive access to neuroimaging and genetics data: a review and preliminary validation, Frontiers in Neuroinformatics 8(35): 2014.

[3] H. Imtiaz, A. D. Sarwate, Non-negative Matrix Factorization with Outliers, manuscript under preparation.