Stochastic Convex Sparse Principal Component Analysis

Inci M. Baytas[1], Kaixiang Lin[1], Fei Wang[2], Anil K. Jain[1] and Jiayu Zhou[1]

[1]Michigan State University
[2]Cornell University

Principal Component Analysis (PCA)

A commonly used dimensionality reduction and data analysis tool

PROS:

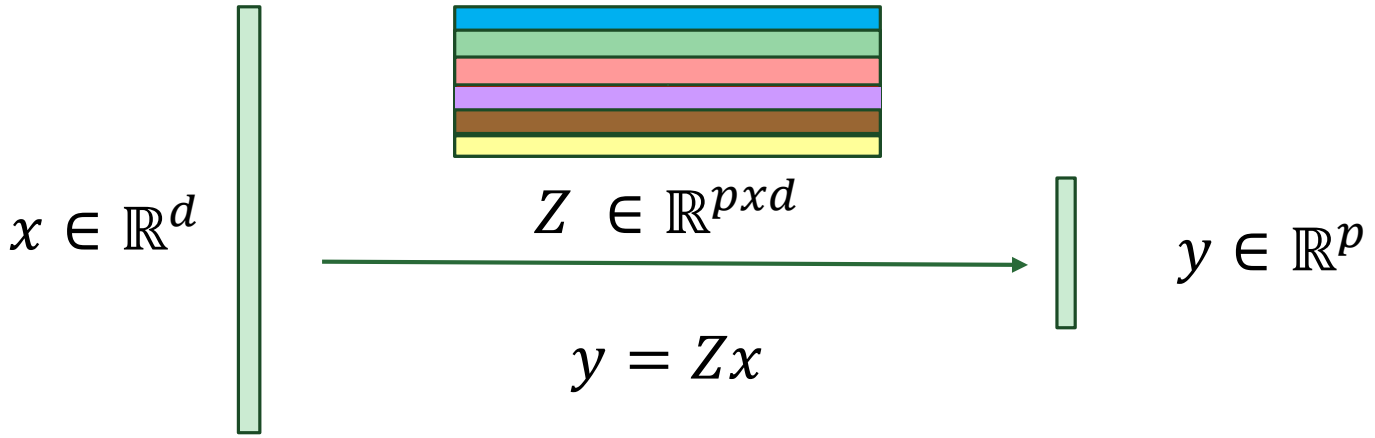
- A linear transformation **maximizing** the variance retained in the data.

$$\max_{\boldsymbol{Z}\in\mathbb{R}^{dxp}} \|\boldsymbol{XZ}\|_F^2, s.t. \boldsymbol{Z}^T\boldsymbol{Z} = \boldsymbol{I}$$

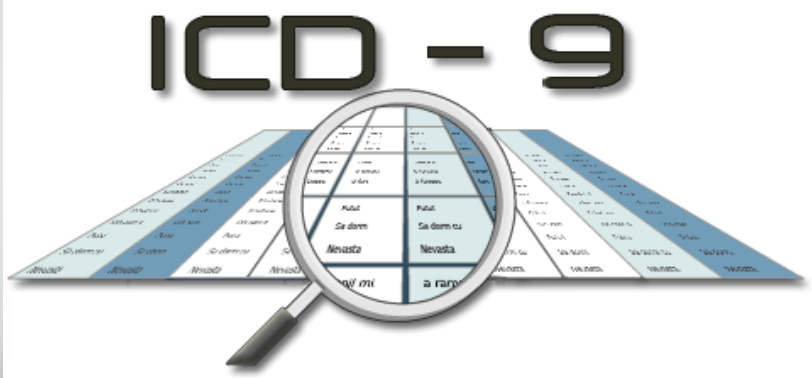
- Data compression with **minimum** information loss.

CONS:

- How to interpret the output dimensions?
  - Each principal component is linear combination of **all input dimensions**.



Applications of PCA



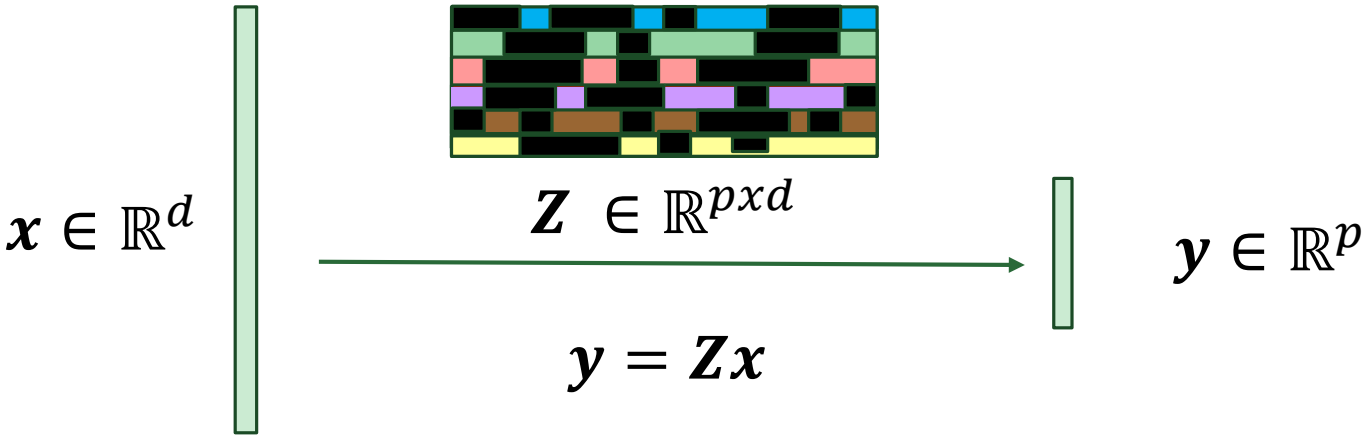
Healthcare

Biology

Data Analysis and Visualization

Sparse PCA

- Addresses the interpretability limitation of PCA
- Learns sparse loading vectors
- Easier interpretation when the data has physical meaning
- Obtaining “key” features instead of manually thresholding loading values



Optimization Approach

- Sparse PCA is posed as an ℓ_1 regularized optimization problem.

$$\min_{\boldsymbol{z}} -\boldsymbol{z}^T \boldsymbol{X}\boldsymbol{z} + \gamma \|\boldsymbol{z}\|_1$$

- **Possible Solution:** Proximal stochastic gradient (Prox-SGD) method
  - Only one gradient is calculated instead of full gradient at each iteration.
  - Stochastic gradient methods are more scalable!
  - **Downside:** Low convergence due to random sampling.
- **Proposed Solution:** Using proximal stochastic variance reduced gradient method (Prox-SVRG).

Convex Optimization for Sparse PCA

- **Properties of Prox-SVRG**
  - ✓ Variance of the gradient convergences to zero at the optimal point.
  - ✓ Two assumptions:
    1. Lipschitz continuity of the gradient
    2. Strong convexity ⟶ But $\min_{\boldsymbol{z}} -\boldsymbol{z}^T\boldsymbol{X}\boldsymbol{z} + \gamma\|\boldsymbol{z}\|_1$ is **Non-Convex!**
- Use convex formulation of calculating the first principal component:

$$\min_{\boldsymbol{z}} \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}} \left(\frac{1}{2}\boldsymbol{z}^T\left(\lambda \boldsymbol{I} - \boldsymbol{x_i}\boldsymbol{x_i}^{\boldsymbol{T}}\right)\boldsymbol{z} - \boldsymbol{w}^T\boldsymbol{z} \right) + \gamma\|\boldsymbol{z}\|_1$$

$\lambda > \lambda_1(\boldsymbol{X})$ ($\lambda_1(\boldsymbol{X})$ is the largest eigenvalue of the covariance matrix)

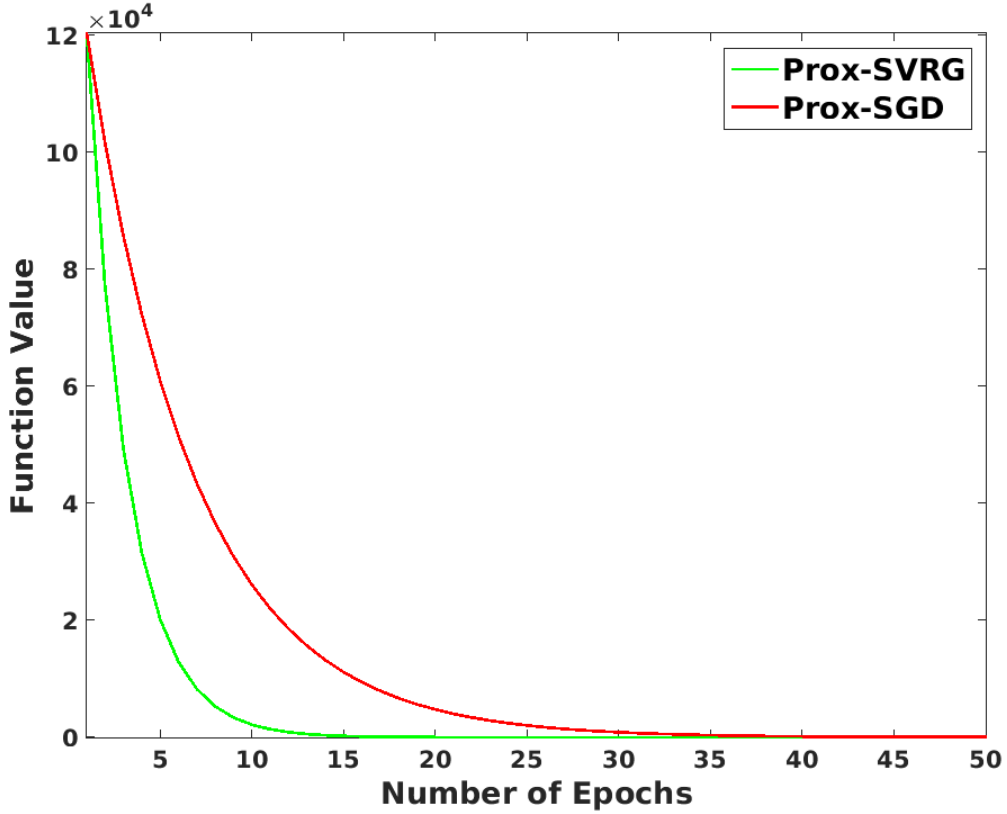
a random vector

Convergence Analysis

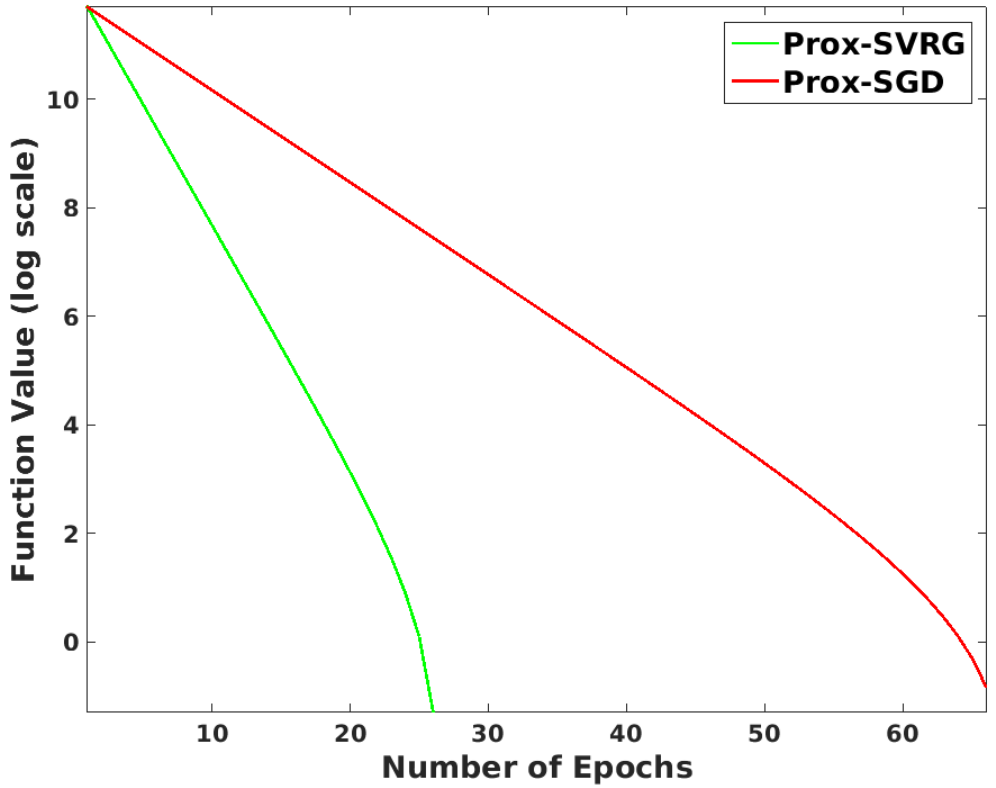
- Strong convexity provides faster convergence rates.
  - In many applications, objective functions are not necessarily strongly convex.
- We relax the strong convexity assumption by using convexity and the property of the proximal operator.
- Our convergence analysis for Prox-SVRG can be used for a broader class of objective functions.

Experiments : Prox-SVRG vs Prox-SGD

- 1000 dimensional synthetic data with 100,000 samples is randomly generated.



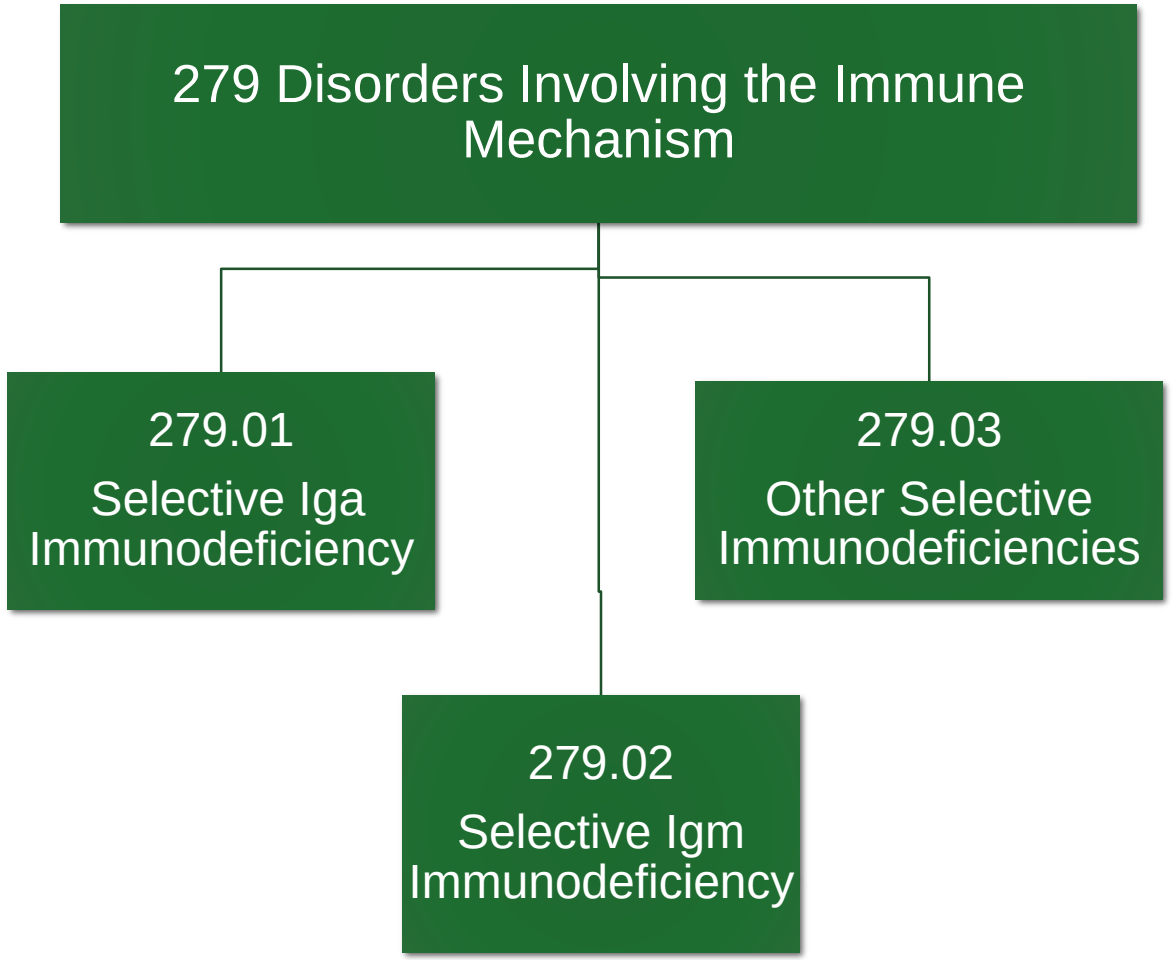
Convergence when maximum number of epochs is 50 .



Cvx-SPCA with prox-SVRG took 49 epochs and with prox-SGD took 67 epochs to converge to a similar sparsity pattern.

Healthcare Dataset

- A private healthcare dataset containing electronic medical records of 177,856 patients over 4 years.
- Total of 11,982 features: ICD9 codes
  - Aggregating features belonging to the same group resulted in 918 features.
- Age and gender were obtained from ICD9 codes.



An example of ICD9 code hierarchy

Demographic Groups

- Determined by looking at the descriptions of ICD9 codes.
  - **Female:** patients with female specific diagnoses; pregnancy, delivery, etc.
  - **Male:** patients with male specific diagnoses; neoplasm of prostate, etc.
  - **Old patients**: assumed > 60 years
  - **Child patients:** assumed < 18

Demographic	Number of features	Number of patients
Female	1,268	130,035
Male	106	24,184
Old	66	2,060
Child	596	38,434

Experimental Results for General Patient Population

- Cvx-SPCA was applied to the whole patient population.
- Cvx-SPCA can help to investigate key medical features.
- Proposed approach provides insight about diagnoses of the patient population.

ICD9 Code	Description
7	Balantidiasis/Infectious
72	Mumps Orchitisn/Infectious
115	Infection by Histoplasma Capsulatum
266	Ariboflavinosis/Metabolic Disorder
507	Pnemonitis/Bacterial
695	Toxic Erythema/Dermatological
697	Lichen Planus/Dermatological
761	Incompetent Cervix Affecting Fetus or Newborn
795	Abnormal Glandular Papanicolaou Smear of Cervix
924	Contusion of Thigh/Injury

Features that will contribute the first principal component

Experimental Results for Demographic Groups

ICD9 Code	Description
281	Pernicious Anemia
392	Valvular and Rheumatic Heart Disease
614	Female Genital Disorders
778	Serious Perinatal Problem Affecting Newborn
905	Major Head Injury

Key features for female patients.

ICD9 Code	Description
153	Malignant Neoplasm of Colon
173	Other Malignant Neoplasm of Skin
337	Disorders of the Autonomic Nervous System
368	Visual Disturbance

Key features for old patients.

ICD9 Code	Description
185	Malignant Neoplasm of Prostate
298	Depressive Type Psychosis
719	Effusion of Joint
800	Closed Fracture of Vault of Skull
811	Closed Fracture of Scapula
860	Traumatic Pneumothorax

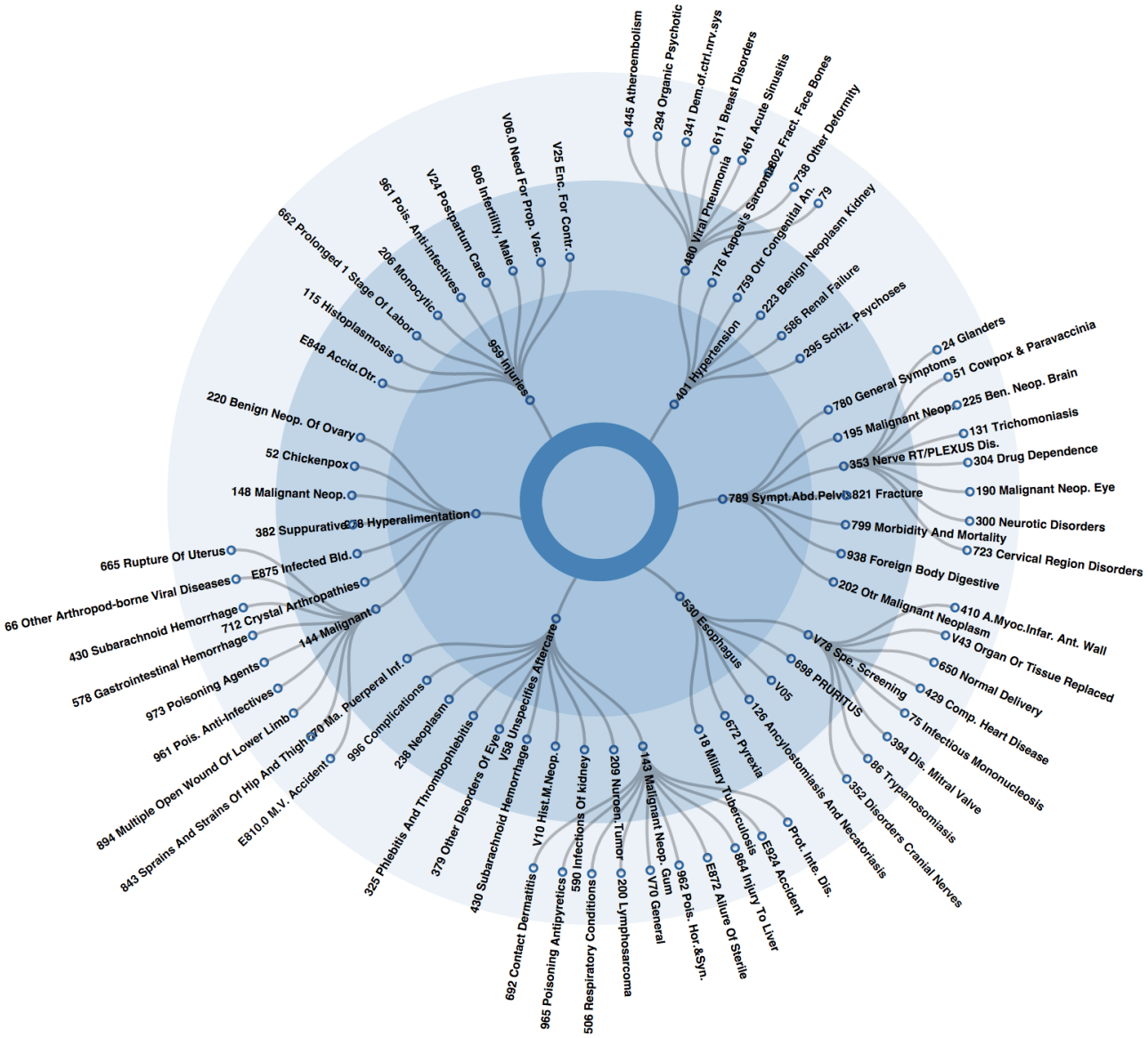
Key features for male patients.

ICD9 Code	Description
8	Intestinal Infection Due to Other Organisms
11	Pulmonary Tuberculosis
78	Other Diseases Due to Viruses and Chlamydiae
10	Primary Tuberculous Infection
204	Lymphoid Leukemia

Key features for child patients.

Future Work

- From key medical features to phenotypes
  - Hierarchical phenotypes
- From phenotyping to classifying /clustering patients
- Designing a clinical decision support system (CDSS)



This material is based in part upon work supported by the National Science Foundation under Grant **IIS-1565596**

CRII: III: Integrating Domain Knowledge via Interactive Multi-Task Learning

Principal Investigator: Jiayu Zhou

Michigan State University

Thank you, any questions?