Outline for Upcoming Projects

January 31, 2022

Outline

PCCA. Background

Simpler to begin with probabilistic PCA

$$\begin{aligned} X|Z &\sim N(WZ + \mu, \Sigma) \\ Z &\sim N(0, I_d) \end{aligned} \tag{1}$$

where $W \in \mathbb{R}^{p \times d}, Z \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{p \times p}, \mu \in \mathbb{R}^p$ and I_d is $d \times d$ identity.

In turn probabilistic CCA

$$\begin{aligned} X_1|Z &\sim N(W_1 Z + \mu_1, \Sigma_1) \\ X_2|Z &\sim N(W_2 Z + \mu_2, \Sigma_2) \\ Z &\sim N(0, I_d) \end{aligned} \tag{2}$$

where $Z \in \mathbb{R}^d, W_j \in \mathbb{R}^{p_j \times d}, \mu_j \in \mathbb{R}^{p_j}, \Sigma_j \in \mathbb{R}^{p_j \times p_j}$ for $j = 1, 2$.

PCCA. Background cont.

Issues with probabilistic PCA and in turn probabilistic CCA,

- Label switching / non-identifiability when running MCMC.
- Difficulties with interpretation.

Some positives from the probabilistic framing,

- Generative model.
- Latent variables Z can be used for visualization.
- EM algorithms (at least for PPCA) can be very fast.

PCCA. Objectives and next steps

What do we want out of a method?

- Visualization
- Variable selection

Group-sparse Regression (GSR). Background

Regression framework with,

$$y = X\beta + \epsilon \tag{3}$$

where $\beta = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$, $X \in \mathbb{R}^{n \times p}$, $y \in \mathbb{R}^n$ and ϵ is a noise term.

Define the groups, $G_k = \{G_{k_1}, \ldots, G_{k_p}\}$ for $k = 1, \ldots, M$ to be disjoint sets of indices, such that $\bigcup_{k=1}^{M} G_k = \{1, \ldots, p\}$.

Group-sparse Regression. Background cont.

Patterns of sparsity include:

- **Coordinate sparsity**: few coordinate of β are non-zero.
- **Group sparsity**: few vectors $\beta_{G_k} = (\beta_{G_{k_1}}, \ldots, \beta_{G_{k_p}})$ are non-zero.
- **Sparse-group sparsity**: few vectors $\beta_{G_k} = (\beta_{G_{k_1}}, \ldots, \beta_{G_{k_p}})$ are non-zero and the vectors themselves are sparse.

GSR. Proposal

Project proposal

Variational extension to the group-sparse setting.

Does this fit our objectives (or would we ideally seek an extension to the sparse group-sparse setting?)

GSR. Outline

1. Extension of linear model to the group-sparse setting
2. Simulations
3. Theoretical results?
4. Extend to logistic / survival

3D Breast cancer data (BCD), Background

- Eric's group have access to 3D imaging data and tumor class labels
- The project involves using the data to construct a classification model
- The raw data has been split into 1cm x 1cm x 1cm regions and passed through the lab's radiomics pipeline, giving $K_i = \sum$ regions realizations for each feature.

BCD, Background cont.

Sample	x_1	x_2	$\dots$	x_{p-1}	x_p
			$\uparrow$		
1			K_1		
			$\downarrow$		
			$\vdots$		
			$\uparrow$		
n			K_n		
			$\downarrow$		

BCD, Next steps

- Exploratory data analysis
- Literature review for methods dealing with multiple different realizations of the same features for each sample.
- Model proposals

Comments

- All project's involve high-dimensional data of some sort.
- Unclear how to proceed with data integration project.
- Can we combine GSR with the Breast cancer dataset?