Latent Scale Prediction Model for Network Valued Covariates

Xin Ma[1], Suprateek Kundu[1]
[1]Department of Biostatistics and Bioinformatics, Emory University

EMORY ROLLINS SCHOOL OF PUBLIC HEALTH | CBIS Center for Biomedical Imaging Statistics

Background and Motivation

- A number of brain regions have been implicated in anxiety disorder [1]; however, none of the regions alone is fully responsible for anxiety disorder pathology.
- There is an increasing interest to identify dysfunctional brain circuits resulting from previous traumatic events as capturing the underlying biological susceptibility towards PTSD symptoms.
- Moreover, there is a lack of prediction and inferential tools for PTSD using a combination of both brain networks and trauma exposure, which are also able to identify whole brain network features that work in conjunction with trauma exposure to drive PTSD severity.
- Network valued data contain rich information, however there have been limited advances in regression approaches involving network valued covariates in literature. The high dimensionality of the networks often results in models with inflated number of parameters leading to computational burden and inaccurate estimation.
- Alternative approaches seek to reduce network dimension and then use the low-rank structure in prediction. This class of methods often lack interpretability and have reduced exploratory value.
- We wish to develop a novel two stage Bayesian framework to find a node-specific low-rank representation for the network covariates and then use a flexible regression framework for prediction. The approach results in a dramatic reduction in the number of regression parameters and is able to maintain interpretability at the node level.

Methods

Notation:

$\boldsymbol{Y} = (y_1, \cdots, y_n)^T$ denotes continuous scalar response vector for n subjects; e.g. log (PSS) score.

$\boldsymbol{G}_i, i = 1, \cdots, n$ denote undirected binary network predictors with p nodes; value 1 for element (l, s) represents connection between node l and node s, while value 0 represents no connection. This network may either correspond to the brain functional network computed from the rs-fMRI data or it may refer to the structural network obtained via probabilistic tractography using DTI data.

$\boldsymbol{Z}_i, i = 1, \cdots, n$ denote other exposure covariates of q dimension.

Stage I: Low-rank representation of network

Inspired by Hoff's latent space model [2], we assume that for subject i, each node k within the network can be accounted by a d dimensional latent scale $\boldsymbol{u}_{ik}$ such that the whole network obtains a representation $\boldsymbol{U}_i = (\boldsymbol{u}_{i1}, \cdots, \boldsymbol{u}_{ip})$ which satisfies the following model:

$$P(\boldsymbol{G}_{i,(l,s)}) = \left[1 + exp(-a_i - \boldsymbol{u}_{il}^T \boldsymbol{u}_{is})\right]^{-1}, \quad l \neq s$$

Typically, the dimension d would be much smaller compared to the node number p.

Methods continued

Stage II: Gaussian Process Regression

The regression model for the continuous scalar response is:

$$y_i = b_0 + \phi(\boldsymbol{U}_i) + \epsilon_i, \quad \epsilon_i \sim i.i.d. N(0, \tau^{-1})$$

We assume that function $\phi(\cdot)$ has a Gaussian process prior with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\kappa}$ whose element (i, i') associated with subject i and i' has the following structure:

$$\boldsymbol{\kappa}(i, i') = \psi_1 exp\Big(- \sum_{k=1}^{p} \psi_{uk} ||\boldsymbol{u}_{ik} - \boldsymbol{u}_{i'k}||_2^2 - \sum_{j=1}^{q} \psi_{zj} \big(\boldsymbol{z}_{ij} - \boldsymbol{z}_{i'j}\big)^2 \Big)$$

- As the likelihood in Stage I remains the same when the latent scales are rotated by any angle, we need to perform a Procrustes transformation with respect to rotations before calculating the kernel in Stage II.
- $\psi_{uk}, k = 1, \cdots, p$ and $\psi_{zj}, j = 1, \cdots, q$ are the characteristic length-scale parameters for network node and other exposure covariates. We can estimate them using maximum-likelihood method or using hierarchical Bayesian model. More specifically, if we place spike-and-slab prior on these length-scale parameters, we can obtain variable selection inference from posterior samples [3].

Data Augmentation Scheme for Stage I

An efficient EM algorithm for Stage I estimation is fascilitated by the following theorem from [4]:

Theorem 1. Let $p(\omega)$ denote the density of the random variable $\omega \sim PG(b, 0), b > 0$. Then the following integral identity holds for all $a \in \mathbb{R}$:

$$\frac{(e^{\psi})^a}{(1 + e^{\psi})^b} = 2^{-b} e^{\kappa\psi} \int_0^{\infty} e^{-\omega\psi^2/2} p(\omega) d\omega,$$

where $\kappa = a - b/2$. Moreover, the conditional distribution $p(\omega|\psi)$ is also in the Pólya-Gamma class: $(\omega|\psi) \sim PG(b, \psi)$.

References

1. Shin LM, Wright CI, Cannistraro PA, Wedig MM, McMullin K, Martis B, etal. A functional magnetic resonance imaging study of amygdala and medial prefrontal cortex responses to overtly presented fearful faces in posttraumatic stress disorder. Arch Gen Psychiatry (2005).
2. Hoff, Peter D., Adrian E. Raftery, and Mark S. Handcock. "Latent space approaches to social network analysis." Journal of the American Statistical Association (2002).
3. Zou, Fei, et al. "Nonparametric Bayesian variable selection with applications to multiple quantitative trait loci mapping with epistasis and gene-environment interaction." Genetics (2010).
4. Polson, Nicholas G., James G. Scott, and Jesse Windle. "Bayesian inference for logistic models using Pólya-Gamma latent variables." Journal of the American Statistical Association (2013).