THE UNIVERSITY OF CHICAGO | DEPARTMENT OF STATISTICS

PHD DISSERTATION PROPOSAL PRESENTATION

JOONSUK KANG

Department of Statistics
The University of Chicago

Learning Meaningful Representations of Data with Empirical Bayes Methods

WEDNESDAY, November 16, 2022, at 1:30 PM
Jones 304, 5747 S. Ellis Avenue

ABSTRACT

Learning meaningful representations of data is one of the most fundamental statistical inference tasks. Meaningful representations can provide answers to scientific inquiries and improve the performance of downstream analyses. Learning representations involves uncovering multiple factors of variation, which are expected to have varying levels of structure, such as sparsity. Therefore, it is desirable to have factor-wise hyperparameters which control the degree of penalty in penalized likelihood methods or parameterize the prior in the Bayesian methods. However, properly tuning these factor-wise hyperparameters can be computationally challenging. This challenge can be overcome with empirical Bayes, which estimates prior hyperparameters from observed data. Hence, as a probabilistic model, the empirical Bayes approach can be flexible enough to model a wide range of structures, such as sparsity, non-negativity, and bimodality.

I propose to develop three empirical Bayes methods to learn meaningful representations of data, with primary motivations from genetics. The first part of the proposal will start with introducing the empirical Bayes covariance decomposition (EBCD) framework. EBCD decomposes a covariance matrix into multiple additive structured loadings under the assumption that the factors in the (possibly unknown) data matrix are orthogonal. Combining the EBCD with sparsity-inducing point-Laplace priors, we develop a sparse PCA method. The second part of the proposal will describe the overlapping clustering method that combines EBCD with generalized binary priors, and its application to identifying individuals' overlapping and potentially partial memberships in shared genetic drifts from a genetic relatedness matrix. The last part of the proposal will briefly discuss future work on a semi-nonnegative matrix factorization method for post-GWAS/PheWAS analysis that is invariant to an arbitrary selection of reference alleles. The change of a reference allele leads to a row-wise sign flip in the data matrix, which is absorbed by row-wise sign flips in the loading matrix that is restricted to be row-wise non-negative or non-positive.