THE UNIVERSITY OF CHICAGO | DEPARTMENT OF STATISTICS

PHD DISSERTATION PROPOSAL PRESENTATION

Huy Dang Tran

Department of Statistics
The University of Chicago

“NEW METHODOLOGIES FOR HIGH-DIMENSIONAL DATA UNDER STRUCTURED SPARSITY”

FRIDAY, May 19, 2023, at 11:00 AM
Jones 303, 5747 S. Ellis Avenue

ABSTRACT

The ubiquity of high-dimensional data, where the number of parameters to be estimated may far exceed the number of observations, necessitates new estimation procedures that take advantage of structural properties of the data in order to reduce the effective dimension of the parameter space. This dissertation explores three different statistical problems of this type, with an emphasis on proposing and analyzing statistically optimal procedures that are also computationally scalable. Applications of these methods are wide-ranging and include spatial statistics, topic modeling and neuroscience.

1. We will talk about the Generalized Elastic Net penalty, an extension of the $\ell 1 + \ell 2$ Elastic Net penalty that is applicable to regression problems where the true regression vector is believed to align with a given graph and the design matrix may be highly correlated. The estimation and prediction error bounds in the $p \gg n$ regime for some commonly encountered graphs, as well as an efficient coordinate descent algorithm to compute the proposed estimator, will be provided. Further empirical observations based on synthetic data as well as applications to COVID-19 prediction and Alzheimer’s disease detection will be discussed.

2. We will then discuss the Probabilistic Latent Semantic Indexing (PLSI) model, a commonly used model in topic modeling. Our main goal is to use singular value decomposition to estimate the p-by-K topic matrix A given the word frequency matrix computed based on word counts from n documents, with particular focus on the $p \gg N$ setting where p is the vocabulary size and N is the average document length. Prior works have demonstrated that the minimax-optimal rate under $\ell 1$ norm is $\surd\ pnN$. Furthermore, optimal statistical guarantees for SVD-based methods have only been provided under the highly unrealistic assumption that $p \leq N$, and these methods computationally do not scale well with p. We study the estimation of A under the Frobenius norm and show that the minimax-optimal error bound for this norm is $(N\ n)-1/4$ up to a logarithmic term. The corresponding estimation procedure involves a thresholding step to exploit the weakly sparse structure of A, which helps reduce the computational complexity significantly as a function of p. Beside from text analysis, our procedure can also be applicable to the analysis of microbiome data, where we hope to show that it is competitive relative to the more popular methods based on Latent Dirichlet Allocation.

3. Finally, we will discuss the generalized sparse additive models, where the true regression function is assumed to be the sum of p smooth univariate functions. Prior works have shown that using a penalty combining the empirical norms and structured norms (for example, RKHS norms) of the univariate functions can lead to minimax-optimal rates and can also be computable. We will discuss some issues with the theoretical analysis in prior papers and propose a re-analysis using the notion of localized sub-Gaussian (or sub-exponential) complexity. In particular, this may help us extend the analysis to generalized models where $\{Yi - E(Yi|X) : i = 1, \cdots, n\}$, are uniformly sub-exponential and not just sub-Gaussian (for example if the Yi's are Poisson-distributed). We will also discuss why group sparsity does not help improve the rate in this nonparametric setting, as well as an estimator that may work when each univariate function space is not associated with a structured norm (for example, the space of monotonic functions).