Statistical and computational perspectives on latent variable models

Abstract: The growth in scope and complexity of modern data sets presents the field of statistics and data science with numerous inferential and computational challenges, among them how to deal with various forms of heterogeneity. Latent variable models provide a principled approach to modeling heterogeneous collections of data. However, due to the over-parameterization, it has been observed that parameter estimation and latent structures of these models have non-standard statistical and computational behaviors. In this talk, we provide new insights into these behaviors under mixture models, a building block of latent variable models.

From the statistical viewpoint, we propose a general framework for studying the convergence rates of parameter estimation in mixture models based on Wasserstein distance. Our study makes explicit the links between model singularities, parameter estimation convergence rates, and the algebraic geometry of the parameter space for mixtures of continuous distributions.

From the computational side, we study the non-asymptotic behavior of the EM algorithm under the over-specified settings of mixture models in which the likelihood need not be strongly concave, or, equivalently, the Fisher information matrix might be singular. Focusing on the simple setting of a two-component mixture fit with equal mixture weights to a multivariate Gaussian distribution, we demonstrate that EM updates converge to a fixed point at Euclidean distance $O((d/n)^{1/4})$ from the true parameter after $O((n/d)^{1/2})$ steps where $d$ is the dimension.

From the methodological standpoint, we develop computationally efficient optimization-based methods for the multilevel clustering problem based on Wasserstein distance. Experimental results with large-scale real-world datasets demonstrate the flexibility and scalability of our approaches. If time allows, we further discuss a novel post-processing procedure, named Merge-Truncate-Merge algorithm, to determine the true number of components in a wide class of latent variable models.

Bio: Nhat Ho is currently a postdoctoral fellow in the Electrical Engineering and Computer Science (EECS) Department where he is supervised by Professor Michael I. Jordan and Professor Martin J. Wainwright. Before going to Berkeley, he finished his Ph.D. degree in 2017 at the Department of Statistics, University of Michigan, Ann Arbor where he was advised by Professor Long Nguyen and Professor Ya’acov Ritov. His current research focuses on the interplay of four principles of statistics and data science: heterogeneity of data, interpretability of models, stability, and scalability of optimization and sampling algorithms.