Statistical and Computational Properties of Some "User-Friendly" Methods for High-Dimensional Estimation

Abstract:
As high-dimensional estimation is ubiquitous these days, it is important for practitioners (as well as statisticians) to understand that statistical and computational properties are not the only considerations when choosing a method. Another important consideration is "user-friendliness"—a term we use to encapsulate the various properties that make a method easy to work with in practice, e.g., its interpretability, ease of implementation, and ability to easily interpolate between simple and complex fits. In this thesis, we present new results on user-friendly methods in various high-dimensional estimation settings.

From a statistical standpoint, we analyze four user-friendly methods for regression and graphical modeling. First, we show under very weak conditions that the generalized lasso estimate is unique, even in a high-dimensional setup, a helpful result from the point-of-view of interpretability. Second, we show that the estimates given by g-stagewise (a general framework for deriving easy-to-implement estimates, for a variety of regression problems) can be viewed as discretizations of a continuous-time dynamical system; as part of planned work, we intend to use this insight to obtain rates for the prediction error of the g-stagewise estimates. Third, as part of other planned work, we intend to derive rates for the prediction error of sparse additive trend filtering (a highly interpretable additive model for sparse regression, where the component functions are the univariate trend filtering fits along each dimension), showing that these rates are minimax optimal. Fourth, we present guarantees for the support recovery of a new pseudolikelihood-based approach (based on sparse quantile regression) to undirected graphical modeling—a helpful result, once again, from the point-of-view of interpreting the resulting estimates. On the computational side, we present specialized, scalable algorithms that are sometimes an order of magnitude faster than the state-of-the-art, for fitting the aforementioned additive model and pseudolikelihood-based graphical model to high-dimensional, potentially non-Gaussian data.