AWARDS & ANNOUNCEMENTS

Dr. Lee Dicker

Maximum likelihood and nonparametric empirical Bayes methods in high dimensions

The main objective of the project is to develop new methodologies, computational strategies, and theoretical results for the use of nonparametric maximum likelihood (NPML) techniques and empirical Bayes methods in high-dimensional data analysis. This work is fundamentally related to the analysis of nonparametric mixture models. Empirical Bayes methods have a long and rich history in statistics, and are particularly well-suited to high-dimensional problems. Moreover, recent computational results and convex approximations have greatly simplified the implementation of NPML-based methods. Leveraging these computational breakthroughs, the investigator is developing novel and scalable NPML-based methods for high-dimensional classification, high-dimensional regression, and other statistical problems. New still-faster algorithms for computing NPML estimators, which take advantage of certain types of sparsity in the estimated mixing-measure, are also being developed. The investigator is studying theoretical properties of the proposed methods in high-dimensional settings. Areas of emphasis for theoretical analysis include convergence rates and frequentist risk properties of the proposed empirical Bayes methods.

The investigator is combining classical and elegant ideas from statistics (empirical Bayes, mixture models, and nonparametric maximum likelihood), with important recent breakthroughs in computing to help develop a rigorous, practical framework for many

problems in modern data analysis. Applications in genomics and other areas of biology where high-throughput data are generated form an important part of the project. Beyond biology, the methods developed during the course of the project are expected to have applications in finance (e.g. fraud detection), machine learning (e.g. speech, text, and pattern recognition), and other fields where vast high-dimensional datasets are being rapidly generated and require accurate, incisive analysis. Another important aspect of the project addresses questions about reproducibility, which have come to the forefront in many applications involving high-dimensional data analysis. To address these questions, the investigator is studying fundamental properties of statistical risk and risk estimation in high dimensions. Algorithms and methods developed during the course of the project are being implemented in easy-to-use and freely available software packages. Project research is closely integrated with education, via graduate student training and newly developed courses for graduate and undergraduate students.