Illinois Center for Wireless Systems

ICWS Seminar Series

Robustness in High-dimensional Statistics: PCA and Collaborative Filtering

Professor Constantine Caramanis
Dept. of Electrical & Computer Eng.
University of Texas @ Austin

Monday, March 7, 2011
4:00 – 5:00 p.m.
141 Coordinated Science Lab

Abstract:
The analysis of very high dimensional data - data sets where the dimensionality of each observation is comparable to or even larger than the number of observations - has drawn increasing attention in the last few decades due to a broad array of applications, from DNA microarrays to video processing, to consumer preference modeling and collaborative filtering, and beyond. As we discuss, many of our tried-and-true statistical techniques fail in this regime.

We revisit one of the perhaps most widely used statistical techniques for dimensionality reduction: Principal Component Analysis (PCA). In the standard setting, PCA is computationally efficient, and statistically consistent, i.e., as the number of samples goes to infinity, we are guaranteed to recover the optimal low-dimensional subspace. On the other hand, PCA is well-known to be exceptionally brittle -- even a single corrupted point can lead to arbitrarily bad PCA output.

We consider PCA in the high-dimensional regime, where a constant fraction of the observations in the data set are arbitrarily corrupted. We show that standard techniques fail in this setting, and discuss some of the unique challenges (and also opportunities) that the high-dimensional regime poses. For example, one of the (many) confounding features of the high-dimensional regime, is that the noise magnitude dwarfs the signal magnitude. While in the classical regime, dimensionality recovery would fail under these conditions, sharp concentration-of-measure phenomena in high dimensions provide a way forward.

Then, for the main part of the talk, we propose a High-dimensional Robust Principal Component Analysis (HR-PCA) algorithm that is computationally tractable, robust to contaminated points, and easily kernelizable. The resulting subspace has a bounded deviation from the desired one, for up to 50% corrupted points. No algorithm can possibly do better than that, and there is currently no known polynomial-time algorithm that can handle anything above 0%. Finally, unlike ordinary PCA algorithms, HR-PCA has perfect recovery in the limiting case where the proportion of corrupted points goes to zero.

Bio:
Constantine Caramanis is an assistant professor at The University of Texas at Austin in Electrical and Computer Engineering. He got his PhD in EECS at MIT, and his AB in Mathematics from Harvard. He received the NSF CAREER award in 2011. His current research interests include optimization, machine learning and statistics, and networks. Of particular interest are applications to wireless networks, energy and air traffic control.

ICWS Seminar series is supported by a grant from Rockwell Collins