Compressive Information Extraction

A white paper submitted to the Pervasive Computing at Scale (PeCS) Workshop

Mario Sznaier, Electrical and Computer Engineering, Northeastern University

(i) ***Proposer's background:*** Dr. Mario Sznaier is currently the Dennis Picard Chaired Professor at the Electrical and Computer Engineering Department, Northeastern University. Prior to joining Northeastern University, Dr. Sznaier was a Professor of Electrical Engineering at the Pennsylvania State University and also held visiting positions at the California Institute of Technology. His research expertise includes actionable information extraction from very large data sets, robust identification and control of hybrid systems, robust optimization, and dynamical vision. He was a plenary speaker at the 2009 and 2010 International Conferences on the Dynamics of Information Systems, and will deliver a plenary lecture at the 2012 IFAC Systems Identification Symposium. Dr. Sznaier is currently serving as an associate editor for the journal Automatica, Executive Director of the IEEE Control Systems Society, and member of its Executive Committee and its Board of Governors. He served as program vice-chair of the 2008 IEEE Conf. on Decision and Control and as Program Chair of the 2009 IFAC Symposium on Robust Control Design. A list of publications and currently funded publications can be found at http://robustsystems.ece.neu.edu

(ii) ***Short vision statement:*** This white paper describes a new vision for substantially enhancing our ability to robustly extract and use information sparsely encoded in extremely high volume, disparate data streams. At its core is a novel, unified vision, centered on the use of dynamical models as information encapsulators, and emphasizing robustness and computational complexity issues. The central theme of our approach is the realization that actionable information can be often represented with a small number of invariants associated with an underlying dynamical system. Thus, in this context, the problem of actionable information extraction can be reformulated as identifying these invariants from (high dimensional) noisy data, and thought of as a generalization of sparse signal recovery problems to a dynamical systems framework. While in principle this approach leads to generically nonconvex, hard to solve problems, computationally tractable relaxations (and in some cases exact solutions) can be obtained by exploiting a combination of elements from convex analysis and the classical theory of moments.

(iii) ***Transformative Impact:*** The recent exponential growth in data collection and actuation capabilities has the potential to profoundly impact society. Aware sensors endowed with tracking and scene analysis capabilities can prevent crime and reduce time response to emergency scenes. Enhanced imaging methods can substantially reduce the amount of radiation required in medical procedures. Moreover, the investment required to accomplish these goals is relatively modest, since a large number of sensors are already deployed and networked. Arguably, a major road-block to realizing this vision stems from the curse of dimensionality. Simply put, existing techniques are ill-equipped for robustly processing the resulting vast amount of data, often noisy and fragmented, within the constraints imposed by the need for real time operation in dynamic, partially stochastic scenarios. The approach described in this talk covers the first steps towards a new paradigm that seeks to address these issues by exploiting very recent developments at the confluence of dynamical systems theory, semi-algebraic geometry, dynamic vision and machine learning.

The introduction of robust dynamic systems tools in this context is an altogether new direction in computational thinking. It addresses acute needs and complements and expands tools from existing algebraic, geometric and statistical approaches, to create a new, far-reaching framework. Indeed, while methods from each of these constituent fields will be subsumed as key components, none alone is sufficient. It is precisely the unfolding of the proposed new, integrative methods that is expected to lead to the new discipline of *dynamics based compressive information extraction* paradigm

Expanded Vision

Our vision builds on a key commonality among diverse problems where reliable decision–making requires extracting sparsely encoded information from massive data flows: Information that reflects spatiotemporal dependencies can be compactly encapsulated in difference equations. *Dynamic Sparsity* is reflected by low model rank, a measure of the dimension of useful *information* that is often dramatically lower than the dimension of the raw *data*. Dynamic structures can be tractably discovered from the data in a way which leverages this inherent sparsity. One key feature of this approach is the ability of these dynamic representations to produce quantifiable measures of uncertainty as *provable error bounds* on the validity of the data interpretation suggested by the model. Another is their relative computational simplicity, which can be critical for feasibility: it is vastly easier to quantify the difference between two dynamic models (which in many contexts requires only a rank comparison) than to search for elusive overlapping time sequences and then compare two such very high dimensional data streams.

As we will illustrate in the talk, the application of these ideas leads to tractable solutions to some very challenging problems. One such example is video, image and genomic data segmentation, where the goal is to detect changes, for instance in scenes, activities, texture, or gene promoter expressions. A second example is the detection of dynamically correlated actions by a few elements of a large population such as co-promoted genes, or coordinated activities by a few persons within a crowd. A third one is event detection from incomplete data sequences, where the missing elements are precisely the ones that mediate the changes. We plan to conclude the talk by exploring the connection between information extraction, hybrid systems identification and machine learning, and discuss new open problems and research directions in optimization, machine learning and systems theory motivated by these problems.