Machine-Assisted System Covariance Design for Structured Knowledge Recovery & Inductive Bias Determination

Graduate Thesis Proposal

Thurston Sexton

3 February 2022

Introduction

We increasingly see the importance of applying inductive biases in the use of statistical and data-driven automation methods. This is especially relevant when analyzing and intervening within complex systems, where both units and interactions between them are of interest. In (Torres et al. 2021) and elsewhere, they point out that the prior *dependencies* we endow a system with (prior to our statistical analyses), whether explicitly or implicitly, play a huge role in the outcomes of these analyses. As such, considerable time is spent in various domains to provide mechanisms to describe, communicate, and effectively enforce these prior inductive biases about our knowledge of dependencies within the algorithms we use.

Still, assuming a given individual (or even an entire domain) *has a catalog of prepared inductive biases*, let alone in a format that our algorithms are prepared to utilize, is problematic at best. It is well known and studied the extent to which algorithm performance can directly depend on the quantity and quality of previous human labor. Consequently, the domains that are the most isolated, or "niche", will simultaneously need the most application of inductive bias to achieve good algorithmic performance, *and have the least ability to build and deploy those biases* in the first place. In fact, these domain experts are often expecting that the role Machine Learning, AI, and data-driven systems in general will play for them is that of *assistant* in discovering, iterating on, and validating key sets of dependencies within their complex systems, so that they can apply them as inductive biases in downstream analysis. Their "rude awakening" to the current state of data-driven systems is often realizing that the heavy labor load they wished to avoid is exactly the labor-intensive creation of computationally-accessible inductive biases that the algorithms will *need* to successfully operate: *garbage in, garbage out.*

What is lacking, and what we demonstrate in this research, is a more clearly defined problem statement: **Machine-assisted Inductive Bias Determination & Structured Knowledge Recovery**, and the beginnings of a cohesive approach toward solutions to these types of problems. We approach this through a principled modeling approach to extracting meaningful interaction dependencies, which in-turn enables what we term machine-assisted *covariance design*. Inspired by recent work in the social network analysis, quantitative finance, disease and contagion diffusion, electrical, and spectral graph communities, we improve upon recent work on the recovery of sparse covariance structures from observation data generated by Markov Random Fields. We also demonstrate that a Riemannian Geometric approach to covariance recovery can successfully bias our assistive techniques toward more human-desirable dependency structures, such as trees.

These methods are compared against other sparse covariance estimation techniques, as well as techniques from other unwittingly related fields like information-cascade network recovery, social network backboning, and spectral sparsification. Our online, observation-level covariance estimation scheme is shown to operate at scale with an annotator-in-the-loop, which is the first case the authors are aware of such a problem setting being demonstrated. Then, our riemannian approach to modifying this covariance estimation is shown to reduce distortion of the estimated dependency network when it is known a priori to be preferentially more "tree-like".

Torres, Leo, Ann S. Blevins, Danielle Bassett, and Tina Eliassi-Rad. 2021. "The Why, How, and When of Representations for Complex Systems." *SIAM Review* 63 (3): 435–85. https://doi.org/10.1137/20M1355896.