Research Statement

Alexander D'Amour

In broadest terms, I am interested in developing new foundational principles for applied statistics. I intend to pursue this research agenda by tackling applied statistical problems in social science, business, sports, and natural science, and to formalize common themes that trickle up from these diverse application areas. This agenda is an extension of a template that was effective in my graduate work. I find that the most exciting theoretical extensions arise from generalizing familiar statistical ideas to problems that stretch existing frameworks by breaking convenient equivalences. For this reason, I intend to focus much of my applied work on problems with non-standard data (e.g., network, natural language, spatial, or point process) and specialized decision processes (e.g., causal inference or attribution as opposed to pure prediction).

These goals are intentionally general and are designed to dovetail with the growing focus on Data Science in both academic and industry settings. Problems categorized as Data Science generally have broader scope than traditionally formulated statistical estimation or prediction problems – in particular, Data Science problems usually involve the full stack of challenges that come with building a data-centric system: the goals of the decision-maker or investigator, the origin of the data, the characteristics of the models, the computational properties of the inference and prediction algorithms, and the most effective methods for communicating results. Unlike traditional information systems where modular components can be developed independently and fit together with simple interfaces, the relationships between components in data science systems are often more complex. For example, small changes in the analytical model can fundamentally change the requirements for compression schemes at the database layer. For applied statisticians to thrive in this ecosystem, a formal set of principles that can be used to customize, communicating, and justify methodological decisions is essential.

The need for foundational principles

I am motivated by three core axioms that I have found to characterize an effective relationship between applied and theoretical Statistics in my academic and consulting experience:

1. Statistical methods are a formalized extension of basic scientific and rhetorical logic.
2. Interfaces between methods and applications should be shaped by the logical argument underlying the problem rather than the format of the data.
3. The theoretical demands made of statistical methods should depend on the logical argument in which they are being deployed, rather than the argument being shaped by the properties of a standard model.

In real-world statistical applications, questions like "Which data-based procedure should I use to support this decision?" are commonplace, but underserved by the academic literature. The bulk of the Statistics and Machine Learning literatures answer questions that appear farther down the analytic pipeline, once the practitioner has selected a particular problem formulation, modeling framework, loss function, and asymptotic frame. The process of translating a decision problem into these terms requires strong statistical intuition, but the current literature (not to mention the growing marketplace for data science products and services) largely frames this process as an art rather than a science, which often results in practitioners choosing techniques based on the shape of the input data rather than the specifics of the problem they wish to solve (e.g., assuming that LASSO selected covariates represent causal effects).

To fill this gap, we require a better-formalized schema of model and method building that maps the particular context and constraints of a given estimation or decision problem to requirements for the statistical or machine learning method to be deployed. This framework should be agnostic to the particular perspective from which the method was developed (e.g., Bayesian, Frequentist, algorithmic), but should instead be able to measure these methods on their ability to satisfy sets of criteria that depend on the context of the problem. In my work as a graduate student and a consultant, I encountered several situations where such a schema was necessary.

Examples

This research agenda involves a number of abstractions that are better demonstrated in terms of completed, ongoing, and future projects. Each of these examples begins with a specific applied problem but leads to a more general principle.

Network modeling. In the theoretical chapter of my thesis, I discovered that many generative networks models with desirable theoretical properties were giving non-sensical results for a social scientific study that a collaborator and I hoped to conduct. Upon investigating, it became clear that most network models in the literature were being deployed to characterize single networks (single-sample problems) rather than to compare multiple networks with different sets of actors (superpopulation problems), and that drawing valid scientific conclusions in these two contexts placed very different requirements on the estimator. I made novel use of a theoretical result about the finite-sample behavior of the MLE with respect to its pseudo-true value to propose necessary criterion for the usefulness of a superpopulation estimator. Based on this criterion, I proposed a new modeling framework that gave stable answers and has made the original applied project feasible. In ongoing work, I am generalizing this criterion to contexts beyond networks.

Complex models for dependent data are particularly fertile ground for uncovering modeling principles because many equivalences that exist in simple models (for example, the equivalence between a superpopulation of units and the large sample limit for independent data) break down under additional complexity. Once formalized, these insights can often be mapped back to simpler problems such as standard regression modeling.

Player tracking. In joint work (to appear in JASA) with Dan Cervone, Luke Bornn, and Kirk Goldsberry, I contributed to building a framework for computing a continuously updated expected value for a basketball possession from newly available player tracking data from the NBA. We called this system expected possession value, or EPV. Beyond prediction, the main goal of developing EPV was to provide a tool for evaluating player decisions, even when those decisions did not directly result in points being scored or a loss of possession. To achieve this goal, it become clear that our estimator would itself need to have a Martingale structure so that changes in the estimated EPV would only reflect changes on the court rather than aggregation artifacts in the estimation. More generally, we established the criterion that, for attribution tasks, an estimator must be stochastically consistent to have the appropriate interpretation. Based on this criterion, we implemented a process model for a basketball possession, rather than, for example, implementing a more flexible machine learning method that may have exceeded our method in pure predictive accuracy.

Causal inference in business. A large proportion of statistical applications within businesses are causal they concern a decision to intervene in a process rather than simply predict its evolution. The critical differences between causal and predictive inference are already well-established in the literature, and in many cases causal inference is already being deployed (e.g., the ubiquity of so-called A/B testing in the tech sector). However, in cases where perfect randomized experiments are difficult to implement because of operational or ethical reasons (e.g., a company needs to maintain he continuity of a user's experience, or a microbank cannot give loans that they know will likely destroy a borrowers credit), modeling principles become enormously important because some amount of extrapolation is necessary to estimate the counterfactual distribution.

These concerns become still more pressing when data are collected from systems where actors are behaving strategically, as in sports. I have a strong interesting in developing model selection and optimization criteria to fit these circumstances, and in applying the machinery of causal inference to game-theoretic contexts, where the main inferential challenge is also to estimate counterfactual outcomes.

Conclusion

I am an applied statistician with a broad range of research interests, but my research agenda is unified by the goal of developing general principles that will facilitate the wide application of rigorous statistical methodology. I have the capacity to tackle these general problems in a disciplined way. I have a broad-based understanding of Statistical and computational methods and an ability to translate and frame key ideas from separate paradigms into the same language. I thrive in collaborative settings, whether my role is to bring rigor to an applied project where the collaborators are practitioners, or to bring context to theoretical investigations where my collaborators are more theoretically oriented. Throughout my graduate career, I was able to amplify the strengths of my collaborators and lab-mates and facilitate communication between them. This not only makes me an effective researcher; it also gives me insight into what formalism can be used to effectively communicate requirements between individuals who collaborate on building data-driven systems in the wide world.