Stat 591: Prior-free Probabilistic Inference
Project Description

Objectives. The goal of the course project is to two-fold. First, it is a sort of introduction to the research process, i.e., reading existing literature, thinking about new ways to approach an interesting problem, carrying out the work, and presenting your findings in a written report. Second, I expect that the results of these course projects will be, at least, very good starts towards a journal publication and/or a PhD dissertation.

General Requirements. Individually (slightly preferred) or in pairs, students will select a research topic of interest. A list of potential topics is given below, but students are welcome to select a topic that is not in the list, *subject to instructor approval*. Roughly, after identifying a problem of interest, students will read some existing literature with the intention of finding a direction that would benefit from a new approach and, if possible, provide some first steps towards this new approach. Not all projects fit with this general structure, but the expectation is that some progress towards something new will be made. The instructor will be available to provide guidance at each step.

Specific Requirements. There are three specific requirements for the project. The first two are just natural steps along the way, designed to give the instructor an opportunity to provide some guidance.

1. A proposal document that simply lists your name(s), the general topic to be considered, and a few key books/papers that have been identified as relevant.
2. A progress report that describes how you have narrowed down your focus and, if possible, what specific problem you plan to address in the final report.
3. A final written report. Students are encouraged to prepare their report in LaTeX, but this is not required; a formal bibliography with proper references is required, however. The report should introduce the problem, explaining why it is important and/or interesting, review the existing literature on the problem, identifying a gap that is to be filled, and present the new ideas/results to fill the identified gap.

Timeline. (Subject to minor changes.)

October 2nd:	Proposal document due.
November 6th:	Progress report due.
December 9th:	Final written report due.

Potential Topics. Below is a rough list of project topics. They are arranged into three categories, but likely all projects would involve elements from each category.

- *Methodological problems.*
  - Survival analysis and other problems in reliability theory often involve censored data. These are practically important problems but have not been considered yet from an IM perspective.

  - Time-series and longitudinal problems involve dependent data, and the interesting problem would be to make inference on those parameters that help to describe the dependence structure. I can imagine that developments here could lead to some ideas for IMs in financial applications.
  - Discrete-data problems are particularly challenging for all approaches to statistical inference. Analysis of contingency tables, log-linear models, logistic regression, etc are all very important and yet to be addressed directly from an IM perspective. (Even the simple Poisson model is a challenge to do well; see Martin et al, `arXiv:1207.0105`.) Another interesting and related problem is the assessment of agreement.
  - Classification is an important problem that often appears in "machine learning" applications. This and other IM-based machine learning tools would be a very interesting and useful contribution.
  - Another class of problems often overlooked in introductory courses are those with non-trivial parameter constraints. Some work on IMs for constrained problems is available (Ermini Leaf and Liu, *IJAR* 2012), and it would be interesting to explore the use of these techniques in other problems. Interestingly, non-trivial constraints often pop up unexpectedly in certain problems (e.g., Stein's paradox) so a better understanding is needed in general.
  - Meta-analysis concerns the combination of separate analyses of several data sets concerning the same underlying problem. There are general rules for combining independent belief functions, e.g., "Dempster's rule of combination," but it is not clear if this is the most effective tool from an IM perspective.
  - Prediction is a fundamentally important problem in statistics, but often doesn't get much attention in statistics theory courses. A general IM approach for prediction is available (Martin and Lingham, *Technometrics* 2015+) but the emphasis is on independent models. Of course, prediction in dependent-data problems is potentially even more important, so it would be interesting to suitably extend the currently available IM approach.
  - There are a number of problems that involve some kind of "model selection," e.g., variable selection in regression, order selection in mixture models, order selection in autoregressive models, etc. Some basic tools for these problems are now available (Martin et al, `arXiv:1412.5139`), and applying these to other contexts would be interesting and useful; "model selection" is also a theoretical problem discussed below.
  - ...

- *Computational problems.*
  - Discrete-data problems pose some computational challenges. It would be interesting to see if tools from algebraic statistics can be useful for such cases.
  - A common summary of the IM output is the "plausibility function" and a corresponding "plausibility region" (see, e.g., Martin, *JASA* 2015+). This region corresponds to a level set of the plausibility function which, except for

in simple problems, can only be evaluated via Monte Carlo. Doing lots of separate Monte Carlo runs is not efficient—how to do it better?

- This is too vague, but it would be very helpful to have some "black-box" techniques for IM computation, analogous to the Metropolis–Hastings for Bayesian computations.
- IMs for big-data problems...
- ...

- *Theoretical problems.*
  - A Bayesian's criticism of the IM approach might be that it's not coherent in the betting sense while the Bayesian posterior distribution is. These coherence arguments, however, involve a sort of symmetry assumption which may be questionable. It would be interesting to see if relaxing the symmetry assumption can provide a coherence-like justification for the IM approach.
  - A key feature of the IM approach is conditioning, and a general technique for carrying out this conditioning is available (Martin and Liu, *JRSS–B* 2015). This technique is based on differential equations and, therefore, is only applicable in suitably smooth problems. An interesting question is how to extend this technique to some non-smooth problems, such as $\mathsf{Unif}\big(a(\theta), b(\theta)\big)$?
  - IM marginalization in so-called "regular" problems is easy (Martin and Liu, *JASA* 2015+), but there are important problems outside this class. A good example is in linear mixed-effect models, i.e., for some parameters the model is regular (Cheng, Gao, and Martin, *EJS* 2014) but for others it is not. Some approaches are available for marginalization in non-regular problems, but these likely are not optimal, so better understanding is needed.
  - High-dimensional problems all involve an implicit low-dimensional structure, otherwise, quality inference would be impossible. Typically, this low-dim structure is introduced via some form of regularization and, for example, estimation or model selection corresponds to a penalized optimization problem. However, these procedures are not "inference" in the sense we discussed in class, so this fundamentally important area is *completely open*. In the paper (Martin et al, `arXiv:1412.5139`) mentioned above, the regularization is incorporated via a "simultaneous validity" condition, which is like an IM version of the notion of multiplicity correction. This is actually for a low-dimensional problem, so there is lots of room for improvement and extension.
  - ...