Limits to Probability Modeling

Chris Sims

January 3, 2014

How this talk emerged

- In 2007 I gave a talk to the Econometric Society summer meetings here at Duke, whose subtitle was "**Why Econometrics Should Always and everywhere be Bayesian**".

How this talk emerged

- In 2007 I gave a talk to the Econometric Society summer meetings here at Duke, whose subtitle was "**Why Econometrics Should Always and everywhere be Bayesian**".

- In 2009 I prepared a critical comment, that appeared in the *Journal of Economic Perspectives*, on a paper by Angrist and Pischke titled "The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics". The ideas there are discussed in more detail in a paper on my website, "**Understanding Non-Bayesians**"

How this talk emerged

- In 2007 I gave a talk to the Econometric Society summer meetings here at Duke, whose subtitle was "**Why Econometrics Should Always and everywhere be Bayesian**".

- In 2009 I prepared a critical comment, that appeared in the *Journal of Economic Perspectives*, on a paper by Angrist and Pischke titled "The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics". The ideas there are discussed in more detail in a paper on my website, "**Understanding Non-Bayesians**"

- A subtitle for this talk might be, "**Why are there no real Bayesians**?"

Outline

1. Bayesian decision theory is not descriptive of real-world decision making.

Outline

1. Bayesian decision theory is not descriptive of real-world decision making.

2. It's not even normative.

Outline

1. Bayesian decision theory is not descriptive of real-world decision making.

2. It's not even normative.

3. Not that there's something better out there.

Outline

1. Bayesian decision theory is not descriptive of real-world decision making.
2. It's not even normative.
3. Not that there's something better out there.
4. So can we rescue it?

Not descriptive

- There are fewer than 2 truly Bayesian chess players (probably none).

Not descriptive

- There are fewer than 2 truly Bayesian chess players (probably none).
- We know the optimal form of the decision rule when two such players play each other: Either white resigns, black resigns, or they agree on a draw, all before the first move.

Not descriptive

- There are fewer than 2 truly Bayesian chess players (probably none).
- We know the optimal form of the decision rule when two such players play each other: Either white resigns, black resigns, or they agree on a draw, all before the first move.
- But picking which of these three is the right rule requires computations that are not yet complete.
- That is, Bayesian decision theory pays no attention to costs of computation or to the possibility that we can be uncertain about something just because we don't know how to perform a calculation in the available time.

But maybe approximately descriptive?

- Natural selection does not select for Bayesian behavior.
- When a herd of gnus has to decide where to cross the river, there can be life-or-death consequences.

The original lecture had here a photograph of a herd of gnus that had made a bad choice of river crossing point. Omitted in this publicly accessible version because of possible copyright issues.

Optimizing gnus?

- Natural selection does not want them each to weigh the probability of death at each crossing point and to choose the safest place to cross.

- That would have them all choosing the same place, and with some probability suffering massive loss of life.

Optimizing gnus?

- Natural selection does not want them each to weigh the probability of death at each crossing point and to choose the safest place to cross.

- That would have them all choosing the same place, and with some probability suffering massive loss of life.

- Nature will select for **probability-matching** behavior instead — individuals should randomize over crossing spots, choosing them with probability proportional to the probability of survival.

- In some experimental settings, pigeons, rats, and people all do probability-matching instead of Bayesian optimal behavior.

Reference

"The Origin of Behavior"

Thomas J. Brennan, Northwestern University School of Law

Andrew W. Lo, MIT

Maybe descriptive of people trained in probability?

- (This might even explain why Bayesians are still a minority — truly Bayesian behavior leads to limited reproductive success.)

Maybe descriptive of people trained in probability?

- (This might even explain why Bayesians are still a minority — truly Bayesian behavior leads to limited reproductive success.)

- We often use data that have been pre-processed by non-probabilistic methods to hide complexity.

Maybe descriptive of people trained in probability?

- (This might even explain why Bayesians are still a minority — truly Bayesian behavior leads to limited reproductive success.)

- We often use data that have been pre-processed by non-probabilistic methods to hide complexity.

- Macroeconomists use the national income accounts. Microeconomists use data on individual "consumption" and "saving". We use data on firm "investment" and "employment". Often seasonally adjusted.

- Financial empiricists use daily or monthly averages more often than actual "tick data" on transactions, to avoid complex modeling of behavior for which we have no widely accepted, manageable theory.

Maybe descriptive of people trained in probability?

- We simplify the state space.

Maybe descriptive of people trained in probability?

- We simplify the state space.

- A true Bayesian individual has only one decision problem to solve — actually she already solved it.

- There is no problem of what to have for breakfast on December 17, 2013.

Maybe descriptive of people trained in probability?

- We simplify the state space.
- A true Bayesian individual has only one decision problem to solve — actually she already solved it.
- There is no problem of what to have for breakfast on December 17, 2013.
- She had to solve that problem for every possible condition (how hungry is she, what's available, have the Greeks seceded from the EMU, etc.) yesterday as part of the decision problem of what to eat at the reception December 16.
- (Savage wrote about this issue.)

Simplified state space example

- At the end of the 1990's there had been a string of high productivity growth quarters, and there was low unemployment.
- One view was that the low unemployment implied accelerated inflation would soon emerge, and that monetary policy should therefore contract.
- Another view was that the high productivity growth, if sustained, would prevent inflation from emerging, so that no monetary contraction was appropriate.
- But was there reason to believe the high productivity growth was going to persist?

Simplified state space example

- Alan Greenspan does know Bayesian decision theory, and to a remarkable degree used its language in public discussions of policy.
- But inside the Fed, the discussion was framed as that of deciding "whether there has been a permanent change in the rate of growth of productivity". A binary variable.
- (I heard some economists at the Fed at the time explain that "econometrics" could not contribute to the discussion, because it would have required them to "test the null hypothesis" of no change, and they realized that would not be useful.)

Actual Bayesian analysis does not use real priors or realistically complex models

- We use Gaussian, gamma, beta, binomial, multinomial, Dirichlet, Wishart, ...

Actual Bayesian analysis does not use real priors or realistically complex models

- We use Gaussian, gamma, beta, binomial, multinomial, Dirichlet, Wishart, ...

- Economists are using Bayesian methods on models with 30-dimensional (DSGE) or 150-dimensional (VAR) parameter spaces.

- Everyone knows that the "priors" they use are determined as much by tractability as by actual careful assessment of anyone's prior beliefs.

Recurrent heresies

- In macroeconomics, we have the “real business cycle school” of macroeconomics. It is “quantitative”, uses probability models, but, in most of its branches, is overtly hostile to any form of probabilistic inference.

- Their argument is that models should be adapted to their purpose, and that such models will be “rejected” by formal inference methods, and that the rejections are not interesting.

Recurrent heresies

- Angrist and Pitschke's book *Mostly harmless econometrics* can be seen as reflecting a similar impulse. It is framed as an argument for using "robust" inferential methods.

- But much of its appeal, in my view, is that it argues for using without apology simple models — OLS estimates of best linear predictors, linear instrumental variables of LATE effects.

- The idea is that if we promise to stick with a very simple model no matter how large the sample, we don't have to worry much about failure of asymptotic approximations.

Is there a principled response to this situation?

- We should recognize that Bayesian decision theory is only approximately descriptive, and only for a limited range of behavior. This is more a conclusion for economics and finance than for statisticians.

- We should recognize that simplicity of a model is perhaps more important than we usually acknowledge.

Insights from data compression

- Rissanen pointed out that one can arrive at something close to likelihood-based inference by considering the problem of compressing a data set.
- The idea is that you record a model and its residuals or shocks, and thereby economize on storage.
- Like Bayesian inference, this approach penalizes model complexity, since a more complex model requires more storage space.

Lossy compression

- Rissanen was thinking about lossless compression. His approach leads to comparing models based on likelihood with a model complexity penalty.
- Much, maybe most, of the data compression that is today ubiquitous is lossy. Consider image and sound storage formats.
- TIFF is a lossless compression method. More commonly used for photos on the web is JPEG, which is lossy.

Evaluating compression methods

- For a given dataset, evaluating lossless compression methods amounts to likelihood based comparison of models, one of which is correct.

- When one model gives probability p_1 to an event and another gives it probably p_2, with p_1/p_2 very large, this difference has a big impact on compression, even if p_1 and p_2 are both very small.

- When this rare event occurs, the p_2 model will use much more storage for it than the p_1 model, and both will use a lot of storage for it.

- But maybe we really don't care about accurately recovering these rare sorts of data?

Evaluating lossy compression

- Lossy compression will lose information. Evaluating it thus depends not only on how well it compresses, but also on how important is the information that is lost.
- A black-and-white version of an image may be all that is needed for some purposes, and certainly can be smaller than a color original.
- JPEG suppresses detail in large areas of similar color and shade. Whether we care about this depends on what images are being compressed and on what the compressed file is going to be used for.

Lessons for statistics

- An ideal approach, that we may sometimes approximate, attempts to find a “true” model, exploring increasingly complex models in increasingly large and informative samples.

- Even if we expect that our collection of models does not include the truth, so long as we remain open to considering new models and base model choice on likelihood, we are doing something like people devising methods for lossless compression.

Lessons for statistics

- An ideal approach, that we may sometimes approximate, attempts to find a "true" model, exploring increasingly complex models in increasingly large and informative samples.

- Even if we expect that our collection of models does not include the truth, so long as we remain open to considering new models and base model choice on likelihood, we are doing something like people devising methods for lossless compression.

- But we have to recognize that in some cases "lossy" models will be useful to people. These models are deliberately inaccurate, but simpler than a serious potentially true model.

How to evaluate lossy models?

- We have to recognize that simplicity of the model, the data to which it will be applied, and the use to which the model would be put will all be relevant.
- An ideal case (considered by Frank Schorfheide in his thesis and subsequent paper) is one where we have a candidate true model available (i.e. one that we don't see how to beat on likelihood-based measures of fit). Schorfheide calls this a "base model".
- Then one can evaluate simpler models by asking how they would behave in their intended use, with the base model generating the data.
- But we have to recognize that this will not always be possible.

Principled lossy model evaluation with no base model?

- There's no detailed formalism for this.
- We have to recognize that there is no way to evaluate such models without considering what it will be used for and what the true model might in fact be.
- Such "models" also arise in the form of decision rules or data summaries that are not probabilistic.
- We can try to formulate a probability model that would make the procedure non-lossy. This can help us understand what range of true models might make the distortions of the simple model unimportant for the model's application.

Conclusion

- The positive conclusions here may seem anti-climactic, unoriginal and common-sensical.
- Principled statisticians and econometricians should recognize that a model dominated in fit by another can nonetheless be useful.
- Users of practical models for real decisions should recognize that there can be value in talking with econometricians and statisticians that might threaten to show their models to be false. Probabilistic inference can show us what assumptions about the range of possible true models and about the range of applications of a practical model could justify it. Such analysis can guide practical model improvement without insisting that only true models are interesting.