Bayesian Analysis (2008) **3**, Number 3, pp. 463–466

Comment on Article by Gelman

Larry Wasserman*

1 Introduction

Brad Carlin invited me to comment on Andrew Gelman's article because Brad considers me an "ex-Bayesian." It's true that my research moved away from Bayesian inference long ago. But I am reminded of a lesson I learned from Art Dempster over 20 years ago which I shall paraphrase:

> A person cannot be Bayesian or frequentist. Rather, a particular *analysis* can be Bayesian or frequentist.

My research is very frequentist but I would not hesitate to use Bayesian methods for a problem if I thought it was appropriate. So perhaps it is unwise to classify people as Bayesians, anti-Bayesians, frequentists or whatever. With the caveat, I will proceed with a frequentist tirade.

2 Coverage

I began to write this just a few minutes after meeting with some particle physicists. They had questions about constructing confidence intervals for a particular physical parameter. The measurements are very subtle and the statistical model is quite complex. They were concerned with constructing intervals with guaranteed frequentist coverage.

Their desire for frequentist coverage seems well justified. They are making precision measurements on well defined physical quantities. The stakes are high. Our understanding of fundamental physics depends on knowing such quantities with great accuracy. The particle physicists have left a trail of such confidence intervals in their wake. Many of these parameters will eventually be known (that is, measured to great precision). Someday we can count how many of their intervals trapped the true parameter values and assess the coverage. The 95 percent frequentist intervals will live up to their advertised coverage claims. A trail of Bayesian intervals will, in general, not have this property. Being internally coherent will be of little consolation to the physics community if most of their intervals miss the mark.

Frequentist methods have coverage guarantees; Bayesian methods don't. In science, coverage matters.

*Department of Statistics and Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA, http://www.stat.cmu.edu/~larry

 DOI:10.1214/08-BA318D

3 The Bayesian Quandry

Suppose we observe $X_1, \ldots, X_n$ from a distribution F. The usual frequentist estimator of F is the empirical distribution

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(X_i)$$

has two virtues. First, it is simple. Second, it comes with a frequency guarantee, namely,

$$\mathbb{P}(\sup_x |\widehat{F}_n(x) - F(x)| > \epsilon) \leq 2e^{-2n\epsilon^2}. \tag{1}$$

What is the Bayesian approach to estimating F? We put a prior π on the set of all distribution functions $\mathcal{F}$ and then we find the posterior. A common example is the Dirichlet process prior. The Bayes' estimator (under squared error loss) is $\overline{F}_n = \mathbb{E}(F|X_1, \ldots, X_n)$.

Does $\overline{F}_n$ have a guarantee like (1)? If yes, that's nice, but then we might as well just use $\widehat{F}_n$. If not, then why use $\overline{F}_n$? Isn't it better to use something with a strong guarantee like (1)? This is the basic quandry in Bayesian inference. If the Bayes estimator has good frequency behavior then we might as well use the frequentist method. If it has bad frequency behavior then we shouldn't use it.

4 The Bayesian Quandry Part II

We observe training data $(X_1, Y_1), \ldots, (X_n, Y_n)$ and we want to predict a new Y from a new X. The catch is that X has dimension p much larger than n. For example, we might have $n = 100$ and $p = 10,000$. A popular predictor is $\ell(X) = \widehat{\beta}^T X$ where $\widehat{\beta}$ is the lasso estimator (Tibshirani 1996) which minimizes

$$\widehat{\beta} = \sum_{i=1}^{n}(Y_i - X_i^T\beta) + \lambda\sum_{j=1}^{p}|\beta_j|$$

and λ is chosen by cross-validation. Not only can $\widehat{\beta}$ be computed very quickly but it has excellent theoretical properties. Specifically, the prediction risk of $\ell(X) = \widehat{\beta}^T X$ is close to optimal among all sparse linear predictions $\mathcal{L}$; see Greenshtein and Ritov (2004) for details.

It is important to understand that nowhere does this result assume that the true regression function $m(x) = \mathbb{E}(Y|X = x)$ is linear. How does Bayesian inference proceed here? Constructing a model for the 10,000 dimensional regression function would be absurd. Again, we can seek a formal Bayesian method with good frequency behavior but then, why not just use the lasso? A pure Bayesian approach lacking any frequency guarantee would be dangerous.

5 Adding Randomness

Some of the greatest contributions of statistics to science involve adding additional randomness and leveraging that randomness. Examples are randomized experiments, permutation tests, cross-validation and data-splitting. These are unabashedly frequentist ideas and, while one can strain to fit them into a Bayesian framework, they don't really have a place in Bayesian inference. The fact that Bayesian methods do not naturally accommodate such a powerful set of statistical ideas seems like a serious deficiency.

6 A Uniter Not a Divider

I've taken a strident and extreme tone to make my point. I reiterate that I would not hesitate to use Bayesian methods for a problem if I thought it was appropriate. There is room for both frequentist and Bayesian methods. Excepting a few special cases, frequency guarantees are essential even for Bayesian methods.

References

Greenshtein, E. and Ritov, Y. (2004). "Persistence in high-dimensional linear predictor selection and the virtue of overparametrization." *Bernoulli*, 10: 971–988.

Tibshirani, R. (1996). "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B*, 58: 267–288.