Stat260: Bayesian Modeling and Inference **Lecture Date: January 20, 2010**

Lecture 1: History and De Finetti's Theorem

Lecturer: Michael I. Jordan *Scribe: Tamara Broderick*

1 History

To answer the question "Why Bayesian statistics?", we start by looking at the history of statistics from a Bayesian point of view. Statistics has been around since the beginning of the 19th century, but even before that, probability with a Bayesian (subjective) flavor was being studied.

> The word "statistics" is of Italian origin. It is derived from *stato* (state), and a *statista* is a man who deals with affairs of the state. The original meaning of statistics is thus a collection of facts of interest to a statesman. (Hald, 2003)

Thomas Bayes, a reverend who was interested in probability, lived in England in the 18th century. At the time, there was no distinction between descriptive and inferential probability although probability used for inferential purposes would eventually come to be known as inverse probability. Another key player from this period was **Pierre-Simon Laplace**, a Frenchman who was particularly interested in data.

Both Bayes and Laplace were aware of a relation that is now known as Bayes' Theorem:

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \propto p(x|\theta)p(\theta) \tag{1}$$

Here, $x \in \mathcal{X}$ is an observable, where $\mathcal{X}$ is a sample space; that is, $\mathcal{X}$ has a probability structure. Also, $\theta \in \Theta$. While θ is just an index to a frequentist, a Bayesian will require that Θ have a probability structure as well. Indeed, we'll consider rich spaces, such as function spaces, for Θ later. In Eq. (1), the lowercase p in, e.g., $p(\theta|x)$ suggests a density; in this class, we'll work mostly with densities although the lowercase p can also represent a mass function in the discrete case. Eq. (1) decomposes into three principal terms:

$p(\theta\|x)$	posterior
$p(x\|\theta)$	likelihood
$p(\theta)$	prior

The proportionality ($\propto$) in Eq. (1) signifies that the $1/p(x)$ factor may be ignored for the purpose of inference on θ. We can see the origin of the phrase "inverse probability" by noting that Eq. (1) inverts the relationship between x and θ from $p(x|\theta)$ on the righthand side to $p(\theta|x)$ on the lefthand side.

Bayes and Laplace were "objective Bayesians" in that they viewed the prior, $p(\theta)$, with suspicion. They asked why one could trust any particular prior and, from there, why one could trust the resulting inference. We will later meet "subjective Bayesians," who embrace priors; the latter paradigm arose in the 1940's and 1950's. The question confronting Bayes and Laplace, though, was how to choose the prior so as not to bias their inference. Laplace's answer was to choose a flat (uniform) prior (Figure 1). Then $p(\theta)$ is constant, and Eq. (1) yields the relation *posterior* $\propto$ *likelihood.*

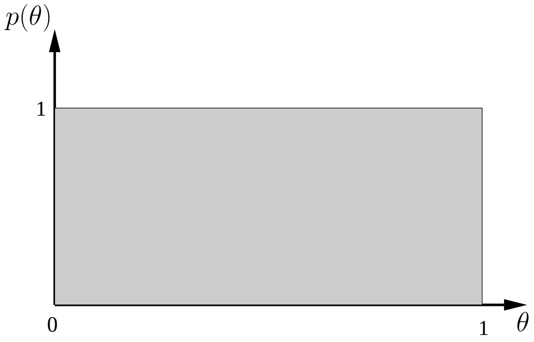

Figure 1: Uniform probability mass, i.e. constant density $p(\theta)$, on the interval $[0, 1]$.

For a century, only the Bayesian paradigm existed. Then, in the middle of the 19th century, there was a strong reaction against the prior. One objection was the observation that one-to-one transformations of the index may yield different prior densities; in particular, a uniform density for θ might yield a non-uniform density for some one-to-one function of θ. For example, consider the odds $\rho = \theta/(1-\theta)$ or the log-odds $r = \log[\theta/(1-\theta)]$. A uniform prior on θ yields densities on ρ and r that are each *not* uniform. Under Laplace's criterion above, we are not then encoding ignorance in the priors of ρ and r. Since these transformations are one-to-one, the choice to use θ, ρ, or r seems arbitrary. Therefore, it makes little sense to have ignorance in one case but not in the others.

In the 20th century, there was a search for a way to practice statistics without priors. Among the prominent figures of this period were **Sir Ronald A. Fisher**, **Jerzy Neyman**, and **Abraham Wald**, who each tried to create new principles upon which to found statistics. Fisher took the approach of studying the likelihood (and maximizing it as a function of the index). Neyman founded frequentism. Frequentism takes an entirely different approach from what we've encountered so far in that it stipulates a way of evaluating procedures, where a procedure can be any inference method (even a likelihood-based method or a method with a prior). A frequentist asks how the results would change if you ran a procedure over and over again, with the data changing each time. Type I and II errors are popular evaluations in this approach. This approach is particularly relevant for, e.g., considering software failure rates. In this class, we'll focus on the Bayesian paradigm while bringing in frequentist ideas where relevant. Like waves and particles in physics, Bayesian and frequentist methodologies have both been relevant to statistics for decades—though they don't always align. Wald was a mathematician who, inspired by game theory, developed decision theory. Decision theory formalizes what it means to do inference. In defining such quantities as loss and risk, it quantifies how "good" a method is.

After this intensive effort to circumvent the prior, the pendulum swang back even further in the Bayesian direction to subjective Bayesianity with the work of **Leonard J. Savage** and **Bruno de Finetti**. These two were uncomfortable with p-values and Type I/II errors. They found paradoxes and incoherencies in the frequentist framework, and, in reaction, embraced priors. Emphasizing the subjectivity of the prior, a subjective frequentist will, in practice, sit down with a domain expert to find an appropriate prior for any problem. There were massive conflicts between Fisher, Neyman, Wald, Savage, and de Finetti.

In a move that may be seen as coming full circle, the emphasis on subjective Bayesianity was followed by the rise of objective Bayes, a paradigm that will be promoted in this course. This new movement, featuring physicist-turned-statistician **Harold Jeffreys**, was about going back to Laplace's work but trying to improve upon it. Notably, objective Bayesians are willing to use frequentist analytic tools to guide their choice of priors. The appeal of these tools is that they are automatic (especially useful when there are many parameters) and do not require a domain expert. Examples of such tools, which we will talk about in more detail later, include consistency Does an estimator (random variable) converge, in a probabilistic sense, to the right answer?

rates of convergence Sometimes rates slower than $n^{-1/2}$ are not good.

unbiasedness All Bayesian procedures are biased, but most frequentists now agree that unbiasedness is not a primary concern.

admissibility Is there another procedure that dominates the one in question everywhere in the parameter space?

More information

For more information on objective Bayes, search online for the 0B09 workshop (2009 International Workshop on Objective Bayes Methodology). For further reading on the history of statistics, a nice reference is Steven Stigler's book (1986).

2 De Finetti's Theorem

In this section, we'll motivate the use of priors on parameters and indeed motivate the very use of parameters. We begin with a definition.

Definition 1 (Infinite exchangeability)**.** We say that $(x_1, x_2, \ldots)$ is an infinitely exchangeable sequence of random variables if, for any n, the joint probability $p(x_1, x_2, \ldots, x_n)$ is invariant to permutation of the indices. That is, for any permutation π,

$$p(x_1, x_2, \ldots, x_n) = p(x_{\pi_1}, x_{\pi_2}, \ldots, x_{\pi_n})$$

There exists a finite exchangeability concept as well, but it's less elegant, and we won't focus on it in this course.

A key assumption of many statistical analyses is that the random variables being studied are independent and identically distributed (iid). Note that iid random variables are always infinitely exchangeable. However, infinite exchangeability is a much broader concept than being iid; an infinitely exchangeable sequence is not necessarily iid. For example, let $(x_1, x_2, \ldots)$ be iid, and let x_0 be a non-trivial random variable independent of the rest. Then $(x_0 + x_1, x_0 + x_2, \ldots)$ is infinitely exchangeable but not iid. The strength of infinite exchangeability lies in the following theorem.

Theorem 2 (De Finetti, 1930s)**.** *A sequence of random variables $(x_1, x_2, \ldots)$ is infinitely exchangeable iff, for all n,*

$$p(x_1, x_2, \ldots, x_n) = \int \prod_{i=1}^{n} p(x_i|\theta) P(d\theta),$$

for some measure P on θ.

If the distribution on θ has a density, we can replace $P(d\theta)$ with $p(\theta)d\theta$, but the theorem applies to a much broader class of cases than just those with a density for θ.

Clearly, since the product $\prod_{i=1}^{n} p(x_i|\theta)$ is invariant to reordering, we have that any sequence distribution that can be written as $\int \prod_{i=1}^{n} p(x_i|\theta)P(d\theta)$ for all n must be (infinitely) exchangeable. The other direction, though, is much deeper. It says that if we have exchangeable data,

- There must exist a parameter θ.

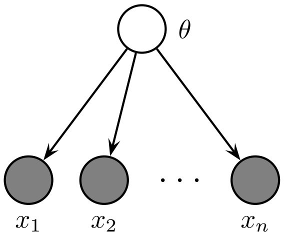

Figure 2: Graphical model illustrating the relationship between the parameter θ and the observables $(x_1, x_2, \ldots, x_n)$.

- There must exist a likelihood $p(x|\theta)$.
- There must exist a distribution P on θ.
- The above quantities must exist so as to render the data $(x_1, x_2, \ldots, x_n)$ conditionally independent.

Thus, the theorem provides an answer to the questions of why we should use parameters and why we should put priors on parameters. The graphical model of this scenario is shown in Figure 2.

Example 3 (Document processing and information retrieval)**.** To highlight the difference between iid and infinitely exchangeable sequences, consider that search engines have historically used bag-of-words models to model documents. That is, for the moment, pretend that the order of words in a document does not matter. Even so, the words are definitely not iid. If we see one word and it is a French word, we then expect that the rest of the document is likely to be in French. If we see the French words *voyage* (travel), *passeport* (passport), and *douane* (customs), we expect the rest of the document to be both in French and on the subject of travel. Since we are assuming infinite exchangeability, there is some θ governing these intuitions. Thus, we see that θ can be very rich, and it seems implausible that θ might always be finite-dimensional in Theorem 2. In fact, it is the case that θ can be infinite-dimensional in Theorem 2. For example, in nonparametric Bayesian work, we will see that θ can be a stochastic process.

References

Hald, A. (2003). *A History of Probability and Statistics and Their Applications Before 1750.* John Wiley & Sons, Hoboken, NJ.

Stigler, S. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900.* Harvard University Press, Cambridge, MA.