Journal of Statistical Planning and Inference 143 (2013) 1621–1633

Contents lists available at SciVerse ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Bayesian inference with misspecified models

Stephen G. Walker

Department of Mathematics & Division of Statistics and Scientific Computation, University of Texas, Austin, USA

ARTICLE INFO

Available online 29 May 2013

Keywords:
Bayesian asymptotics
Exchangeable
Independent and identically distributed
Learning model
Misspecified model
Mixture of Dirichlet process model
Regression model
Time series model

ABSTRACT

This article reviews Bayesian inference from the perspective that the designated model is misspecified. This misspecification has implications in interpretation of objects, such as the prior distribution, which has been the cause of recent questioning of the appropriateness of Bayesian inference in this scenario. The main focus of this article is to establish the suitability of applying the Bayes update to a misspecified model, and relies on representation theorems for sequences of symmetric distributions; the identification of parameter values of interest; and the construction of sequences of distributions which act as the guesses as to where the next observation is coming from. A conclusion is that a clear identification of the fundamental starting point for the Bayesian is described.

1. Introduction

One of the consequences of adopting a Bayesian approach to statistical inference is the necessary assignment of a probability distribution (the prior) on a parameter which indexes a family of distribution functions. A parametric family $\{f(x;\theta)\}$, with $\theta\in\Theta$, and Θ as the parameter space, has been chosen to model a sequence of observations. For the moment we will adopt the notion that the order in which the sequence is observed does not alter how one learns, and we will use the notation $(X_1,\ldots,X_n)$ to represent an as yet unseen sequence, and $(x_1,\ldots,x_n)$ as the observed values.

The Bayesian makes inference about θ through probability distributions on Θ. Starting with the prior $\pi(\theta)$, after n samples have been observed, the posterior distribution is given by

$$\pi(\theta|x_1,\ldots,x_n)=\frac{l_n(\theta)\pi(\theta)}{\int_\Theta l_n(\theta)\pi(\mathrm{d}\theta)},$$

where

$$l_n(\theta)=\prod_{i=1}^{n} f(x_i;\theta).$$

There seems little explanation needed for this update as it represents an application of Bayes theorem. Hence, it would appear that as long as there is information about θ with which to construct a subjective prior, or failing this one can adopt one of the many choices of objective prior, the data are all that are needed to provide the posterior distribution.

There are a number of ways in which Bayesian statistics can be introduced. Bernardo and Smith (1994) develop an axiomatic approach based on the notion of rational behavior or choice. Others, for example Berry (1996), describe the Bayesian model as a family of density functions and a prior; the posterior distribution arising as an application of Bayes theorem, as we have previously mentioned. Having said this, there is and should be a concern with a Bayesian constructing

E-mail address: s.g.walker@math.utexas.edu

0378-3758/$ - see front matter
http://dx.doi.org/10.1016/j.jspi.2013.05.013

the model $f(x;\theta)$ and then assigning a prior distribution $\pi(\theta)$ only because that is what Bayesians do. We will be challenging this perception throughout the paper, from the realistic perspective that the choice of $f(x;\theta)$ is misspecified. By this we mean that there is no $\theta_0 \in \Theta$ for which it can be assumed that the (X_i) are independent and identically distributed from $f(x;\theta_0)$. This is referred to as the M-open case in Bernardo and Smith (1994).

Yet the view that the Bayesian can always apply Bayes theorem is widely held and practiced and has lacked any real challenge to the extent that the big debate in Bayes is really only about how to construct $\pi(\theta)$, i.e. the prior. The two well known positions are the *subjective Bayes* and the *objective Bayes* approaches; see Goldstein (2006) and Berger (2006) for recent reviews. This article is not about the debate between these ideas, as interesting as they may be.

The Bayesian also has the work of de Finetti (1937) to mention; namely the representation theorem for (0,1) *exchangeable* sequences, which was later extended by Hewitt and Savage (1955) to more general spaces. This theorem, a strictly mathematical result with no associated philosophy, is concerned with arbitrary long sequences of symmetric density functions. It is not the intention to discuss it further here as it will be returned to later on. In fact, it is fundamental to Bayes. The current idea here is that as long as an experimenter is willing to assume that the sequence of observations is exchangeable, or the densities from which they arise are symmetric, then the Bayesian model follows through the representation theorem. The point is that the representation theorem guarantees the existence of the $\pi(\theta)$. The Bayesian appeal to de Finetti is not universal due to the simple fact that not all data structures are as elegant or as simple as exchangeable.

The aim of this article then is to expand upon a certain issue about the prior $\pi(\theta)$ once a family of density functions $\{f(x;\theta)\}$ has been selected to model the sequence of observations. We will adopt the position which is that the family of density functions is a model only and, consequently, there is no claim that for all sample sizes n, there is a $\theta_0 \in \Theta$ for which the $(X_i)_{i=1}^n$ are independent and identically distributed from $f(x;\theta_0)$. The quote of Box (1980) is well known; "All models are wrong, some are useful". This assumption is acceptable; but what is not acceptable to many commentators is that the Bayesian does not deviate as a consequence; a prior $\pi(\theta)$ is constructed targeting "θ_0", as though such a parameter value still existed, and Bayes theorem is applied. Such actions are at odds with this reasonable assumption that the model is wrong. This issue remains largely unresolved. An objective prior which attempts to target no θ and merely represents a function on Θ space is a partial solution. However, Bayes theorem may now lack formal justification and it is difficult to assess what the posterior means when sample sizes are small. Is the posterior also merely a function on Θ space?

Let us now consider the family of density functions $\{f(x;\theta)\}$, having adopted the assumption that the model is wrong in that there does not exist a θ_0. We do not now advocate the Bayesian carries on as though such a parameter value does exist, but rather acknowledges there needs to be an alternative parameter value which needs to be targeted. To illustrate what is meant here; we could if the model were correct define θ_0 as the θ which minimizes

$$d_1(f(\cdot;\theta), f_0(\cdot))$$

where, for example, d_1 denotes the L_1 distance between density functions, and we use $f_0(\cdot)$ to indicate the density function from which the observations are independent and identically distributed. Outside of the case of the existence of θ_0 we need an alternative target.

It is essential for the subjective Bayesian to know what they are talking about and targeting. For there is, for all appropriate sets or events A, a

$$\mathrm{P}(\theta \in A) = \Pi(A) = \int_A \pi(\mathrm{d}\theta)$$

to be specified. Presumably, in the subjective approach, this must mean something. So, according to the experimenter, something called θ lies in a set A with probability $\Pi(A)$. Unfortunately, merely specifying $f(x;\theta)$ is not sufficient for $\mathrm{P}(\theta \in A) = \Pi(A)$ to mean anything.

Hence, we need to target a θ which has a well defined meaning. It is also needed that this specific parameter value is being learnt about through an application of Bayes theorem and, taking this point to its logical conclusion, we would also need that asymptotically the sequence of posterior distributions accumulate at this selected parameter value. In the next section we specify such a parameter value and explain its interest.

Before this, we need to discuss something pertinent. Whatever we come up with, there cannot be a distinction between the experimenter who comes up with a model, i.e. the parametric family of densities $f(x;\theta)$, which is hopeless, not thought out and not even trying to approximate $f_0(\cdot)$; and the experimenter who has carefully crafted a suitable approximate model, but wrong all the same. Unfortunately, the maths of the update cannot be made aware of these two characteristics. Hence, we need a motivation for Bayes which works in both of these scenarios.

The layout of the article is as follows: In Section 2 we discuss what the target of the Bayesian is, or should be, in the absence of the notion of a true θ_0. In Section 3 we discuss aspects of probability pertinent to the Bayesian style of thinking and Section 4 provides motivation for the Bayesian approach using the idea of a coherent sequence of guesses as to the density from which the next observation is thought to come from. Section 5 discusses the important procedure of model selection from the point of view of the work developed in Section 4. Section 6 considers necessary asymptotic studies of the sequence of posterior distributions which actually form an integral part of the Bayesian idea. The work to this point relies on a representation theorem for the sequence of guesses alluded to in Section 4; without this an alternative derivation of the Bayesian posterior is needed, and this is provided in Section 7. A form of Bayesian inference is provided by Bayesian