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 © 2013 Elsevier B.V. All rights reserved.
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 nonparametrics and the motivation for this is explained and classes of model for regression and time series are presented in Section 8. Finally, Section 9 concludes with a brief discussion.

2. What is being learnt about?

What θ represents is usually overlooked with the immediacy of worrying about what $\pi(\theta)$ is. But, going back a step, what is θ? It is something we want to estimate. To the frequentist it is fixed. To the Bayesian it is random. But what is fixed? And what is random? What are we trying to estimate? Construct a prior for? When the model is chosen correctly we can say that we are trying to estimate the parameter value which makes the sequence independent and identically distributed or, equivalently, the θ value which minimizes

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

assuming that the parametric model is identifiable.

On the other hand, when the model is not chosen correctly, we must rely on this latter idea rather than the former; that is, we must rely on distances to define the targeted value of θ. For in terms of sampling there is now no connection between x and any $\theta\in\Theta$. Even connecting x and θ through the $f(x;\theta)$ as a density function could now be seen as problematic. Though we do not pursue this line here, this is what motivated Brown and Walker (2012) to connect x and θ through the loss function $-\log f(x;\theta)$, rather than through the density function $f(x;\theta)$. There is a hint of this idea in what follows.

The idea must be that hidden somewhere in Θ is a particular value of θ we should want to learn about (or estimate). This parameter value must be a matter of choice, one could claim to identify a number of possible values, but the one which clearly stands out from the rest is the one which minimizes

$$l(\theta)=-\int_{\mathbb{X}}\log f(x;\theta)F_0(\mathrm{d}x)$$

where F_0, as we have previously indicated, is the data generating mechanism. The θ minimizing this expression, and let us call it θ^*, therefore minimizes the Kullback–Leibler divergence (Kullback and Leibler, 1951) between the family $\{f(x;\theta),\ \theta\in\Theta\}$ and $f_0(x)$. In other words, it is the best parameter value in a clearly defined sense.

There are a number of distances or divergences that one could use and in each case a possibly different θ^* would be identified; so there would appear to be an element of arbitrariness about the Kullback–Leibler divergence. However, even though we have not yet moved on to how we update the prior, or even how we can specify the prior yet, if the intention is to use a Bayes theorem for the update then it is well known that under mild regularity conditions the posterior does accumulate at θ^* (see Berk, 1966). Hence, the use of the Kullback–Leibler divergence and the target θ^*. This is no circular argument; we are merely targeting the parameter value about which the Bayesian can learn about. The Bayesian specifying a different parameter value can indeed target it, but through the Bayesian update will not be learning about it.

This set-up is not really anything exclusively to do with Bayesian inference; indeed the maximum likelihood estimator is also typically going to converge to θ^* (see White, 1982). This is no surprise since the sample estimate of $l(\theta)$ is precisely

$$\widehat{l}_n(\theta)=-n^{-1}\sum_{i=1}^{n}\log f(X_i;\theta)=-\int_{\mathbb{X}}\log f(x;\theta)F_n(\mathrm{d}x)$$

where $F_n(\mathrm{d}x)$ is the empirical distribution function and is in many ways the best estimate of $F_0(x)$. And minimizing $\widehat{l}_n(\theta)$ yields $\widehat{\theta}$, the maximum likelihood estimator.

The picture emerging then is as follows. We are targeting θ^*, which minimizes $l(\theta)$, and we know if we construct a prior $\pi(\theta)$, which is used to express subjective beliefs about the location of θ^*, or could even be an objective prior, then using Bayes theorem, whether one felt this was justified or not, would be able to claim the targeted value was being learnt about through the updates. But this is a weak story at the moment and we can strengthen it.

3. Assigning probabilities

So now θ^* is a fixed quantity which can be seen, or observed, through some massive sampling plan; i.e. sampling (X_i) from F_0 for a long time. It is an observable. A first issue that needs to be tackled is whether probability is being correctly employed by assigning a probability distribution to what could be regarded as a fixed element which is simply not known. This issue has been raised by Cox (2006). "Because θ^* is typically an unknown constant, it is not in this setting meaningful to consider a probability distribution for θ^*". Note, Cox used θ_0 where I have changed this to θ^*.

What the weather will do tomorrow is fixed. It is a matter of science. But the complexities of weather systems mean that while the outcome is fixed, a human has no possibility of working out what will happen precisely. Hence, the human introduces, for example, a 60% chance of rain for the next day. No one would claim this is an abuse of the use of probability. It is in fact a highly appropriate use. It is an approximation to the science. And most people would understand what it meant. More likely to rain than not, but not by much. And the current evidence is what the weather systems have been presenting up to the time at which the prediction of rain is made.

So the mechanism of generating an infinite number of the (X_i) from F_0 is not feasible or has not even yet started when a guess at θ^* using $\pi(\theta)$ is needed. But just as it is possible to assign a probability to a particular weather outcome, so it is possible to assign a probability distribution about the location of θ^*, with whatever evidence is currently available.

There are and should be a variety of interpretations of probability. For example, we have the probability of rain tomorrow, of a horse winning a race, of a party winning an election. It is also of interest to think of a coin flip here. Once flipped, the side the coin lands is determined; it is fixed, due to physical laws. According to Cox therefore it would be inappropriate to guess at a probability of a head or tail after the coin has been flipped but not yet landed. This is clearly not a healthy interpretation of probability. And to add to the mix, there is also a probability that a Brownian motion reaches a particular height in some given interval. So there is no absolutism about the use of probability. It does not need, certainly as far as statistics is concerned, to be revered and one sole interpretation required.

There has been a large amount of criticism about Bayesian methods recently, see Gelman (2008), with associated discussion, unjustifiably, since the criticisms are based on a misunderstanding of what Bayes is. Bayesian statistics is about the statistician, for whatever reason they may have, of guessing or estimating the distribution of the next outcome. It is a worthy desire since decisions may be needed at interim stages and in this case the expected utility rule may be employed.

The correct starting point for the Bayesian is the desire to construct $m_n(x)$, for $n=1,2,\ldots$, where $m_n(\cdot)$ serves as a guess as to where the next observed value X_n could be regarded as coming from in a stochastic sense. This sequence would serve as a vehicle with which to make decisions, for example. For suppose there is an action $a\in\mathbb{A}$ to be made with utility function $U(a,X)$. If $X=x$ is known then the appropriate action is to maximize $U(a,x)$, whereas if X is unknown and has a current guess as to its distribution as $m_n(x)$, then the coherent action now is to find the a which maximizes the expected utility

$$U(a)=\int_{\mathbb{X}}U(a,x)m_n(\mathrm{d}x).$$

See von Neumann and Morgenstern (1944) and Hirshleifer and Riley (1992). As information arrives this $m_n(\cdot)$ would need updating so that improved decisions can be made. Hence, there is a desire to think about how to construct $\{m_n(\cdot)\}_{n\ge1}$. This is the subject of the next section.

4. Bayesian motivation

Having seen $X_1=x_1,\ldots,X_{n-1}=x_{n-1}$, the aim is to produce $m_n(\cdot)$. But there are rules and conventions to adhere to. One cannot blindly construct any sequence. And how about $m_1(\cdot)$; the guess for the density from which the first observation is coming? Undoubtedly a hard task, but not even a unique problem. It is not even a Bayesian problem. See Hirshleifer and Riley (1992) for the general theory. Decision making with uncertainty is all about constructing guesses for the density of an unknown outcome. It is essential for an application of the maximum expected utility rule. To be asked and to provide a guess for the density of an outcome is the stuff of decision making with uncertainty and, as we have indicated, is not a Bayesian problem per se.

The fundamental idea for the Bayesian is about constructing a sequence of density functions $(m_n(x))_{n\ge1}$ which acts as a guess or, more technically, a predictive density for the observation X_n once $(X_1=x_1,\ldots,X_{n-1}=x_{n-1})$ has been observed. There is no need for a philosophical debate here to agree that whatever set of densities emerge, or are chosen, to pretend that one can assert X_n comes from $m_n(x)$ for all n is an absurd position. So one must accept that this sequence of densities is not more than one of guesses.

The sequence must adapt as n changes. It is an obstinate Bayesian who persistently guesses $m_1(x)$ as the data generating mechanism for X_n despite having seen $(x_1,\ldots,x_{n-1})$. These observations would lead a Bayesian to modify their guess and for this to be done properly it must necessarily depend on $(x_1,\ldots,x_{n-1})$. Hence, we should write the guesses in the usual notation of a conditional density, as

$$m_n(x)=m_n(x|x_1,\ldots,x_{n-1}).$$

These guesses are density functions which can generate a "set of data" themselves, starting with a piece of "data" being generated from $m_1(x)$, and so on. These, of course, need not look anything like the observed data. Therefore, there is as usual a joint density given by

$$m(x_1,\ldots,x_n).$$

If it is to be anticipated that the order in which the data arrive is to be regarded as irrelevant then the mathematical results of de Finetti (1937) and Hewitt and Savage (1955) state that the sequence of guesses must be of the type, for some family of density functions $f(x;\theta)$, with $\theta\in\Theta$, and for some probability density $\pi(\theta)$ on Θ:

$$m(x_1,\ldots,x_n)=\int_{\Theta}\prod_{i=1}^{n}f(x_i;\theta)\pi(\mathrm{d}\theta).$$

The choice of $f(x;\theta)$ is not discussed here, save to say and to reiterate the $m(x_1,\ldots,x_n)$ it generates is not to be regarded as correct.

So one needs to specify a probability $\pi(\theta)$ to complete the picture. To think about this let us go back to $m_1(x)$. Now

$$m_1(x)=\int_\Theta f(x;\theta)\pi(\mathrm{d}\theta).$$

There may be a density function denoted by $f_0(x)$ from which x is genuinely coming, and of course $f_0(x)$ is not known. But once $f(x;\theta)$ has been chosen one needs to identify a specific $\theta^*\in\Theta$ which is being targeted. To identify what this θ^* should be, we note that the best guess possible for $m_1(x)$ will arise if we take $\pi(\theta)$ as the point mass at θ^*, where θ^* minimizes the Kullback–Leibler divergence from $f_0(x)$ to $f(x;\theta)$. Hence, the prior must be targeting θ^* and so $\Pi(A)$ has the meaning that it is θ^* which is believed to lie in the set A with probability $\Pi(A)$.

Under this scenario, it is unclear whether Bayes theorem is applicable. But the Bayes update emerges from the de Finetti representation theorem. Since

$$m_{n+1}(x)=\int f(x;\theta)\pi_n(\mathrm{d}\theta)$$

where

$$\pi_n(\theta)=\frac{\prod_{i=1}^n f(x_i;\theta)\pi(\theta)}{\int_\Theta\prod_{i=1}^n f(x_i;\theta)\pi(\mathrm{d}\theta)}.$$

So the only change in the system from 1 to n is the update of $\pi(\theta)$ to $\pi_n(\theta)$, which is the usual Bayes update.

Thus, whether $f(x;\theta)$ has been chosen to be $f_0(x)$, is an approximation to $f_0(x)$, or has been purposely chosen to be nowhere near $f_0(x)$, the parameter of interest can only be the one θ^* which minimizes the Kullback–Leibler divergence from $f_0(x)$ to $f(x;\theta)$. The prior targets this value. And Bayes theorem is the appropriate update when the order of guesses is regarded as irrelevant.

Moreover, it is well known (see for example Berk, 1966) that the posterior distributions accumulate at θ^*, and this is the real coherence: what we are expressing beliefs about is where the learning machine takes us.

5. Model selection

Bayesian model selection typically proceeds using Bayes factors; see for example Kass and Raftery (1995). As is the theme of the paper, we will assume that from the models under consideration, none of them represent the correct data generating mechanism. This set-up has been referred to as the M-open view; see Bernardo and Smith (1994), and see also Key et al. (1999). Yet articles such as the one by Kass and Raftery (1995) state things like “We begin with data D, assumed to have arisen under one of the two hypotheses H_1 and H_2 according to a probability density $\mathrm{pr}(D|H_1)$ or $\mathrm{pr}(D|H_2)$.” This scenario is highly unlikely to occur in practice.

Suppose we have two models and for each one of them we wish to learn about the parameter which takes us the closest, with respect to the Kullback–Leibler divergence, to the true data generating mechanism. One would then choose the best model to be the one which has the parameter minimizing this Kullback–Leibler divergence. The objective is clear and is a natural development of the procedure involving a single model.

So, for the model

$$M_j=\{f_j(x;\theta_j),\pi_j(\theta_j),\theta_j\in\Theta_j\}$$

let us define the parameter minimizing the Kullback–Leibler divergence to be θ_j^*. That is, θ_j^* minimizes

$$l_j(\theta)=-\int_{\mathbb{X}}\log f_j(x;\theta)F_0(\mathrm{d}x).$$

All of these quantities can be found asymptotically, since the posterior $\pi_j(\theta|x_1,\ldots,x_n)$ accumulates at θ_j^* and hence these become known, as does $F_0(x)$. Letting θ^* be the minimizer of $\{l_1(\theta_1^*),l_2(\theta_2^*)\}$ therefore, it is clear that we can find θ^* and hence we can find the best model.

Thus, when a model probability is assigned, it is a meaningful one. So $\mathrm{P}(M_j)$ is the prior probability that model M_j contains θ^* or, more accurately, that Θ_j contains θ^*.

There is a special case when the models are nested so that $\Theta_1\subset\Theta_2$. Then, as before, we can interpret $\mathrm{P}(M_1)$ as the probability that $\theta^*\in\Theta_1$ but now $\mathrm{P}(M_2)$ has the interpretation of being the probability that $\theta^*\in\Theta_2-\Theta_1$. This principle can be extended to an arbitrary number of nested models and then $\mathrm{P}(M_j)$ is interpreted as the probability that $\theta^*\in\Theta_j-\cup_{l<j}\Theta_l$. For example, in the case of three nested models

$$\mathrm{P}(M_1)=\mathrm{P}(\theta^*\in\Theta_1),\quad \mathrm{P}(M_2)=\mathrm{P}(\theta^*\in\Theta_2-\Theta_1)$$

and

$$\mathrm{P}(M_3)=\mathrm{P}(\theta^*\in\Theta_3-(\Theta_2\cup\Theta_1)).$$

These probabilities are to do with which model is the best in the sense of the smallest model containing the best parameter. The idea being that this would be the selected model.

The update for these probabilities is also based on the notion that there is a sample mechanism which can determine which set θ^* exists in. One keeps sampling from F_0. The update would be determined by

$$\mathrm{P}(M_j|X_1,\ldots,X_n),$$

and this is an application of Bayes theorem as motivated by de Finetti (see Section 4), so

$$\mathrm{P}(M_j|X_1,\ldots,X_n)=\frac{\int_{\Theta_j}\prod_{i=1}^n f_j(X_i;\theta_j)\pi_j(\mathrm{d}\theta_j)\mathrm{P}(M_j)}{\sum_j\int_{\Theta_j}\prod_{i=1}^n f_j(X_i;\theta_j)\pi_j(\mathrm{d}\theta_j)\mathrm{P}(M_j)}.$$

In order to implement this model assessment, it is required to compute the marginal density functions

$$m_j(x_1,\ldots,x_n)=\int_{\Theta_j}\prod_{i=1}^n f_j(x_i;\theta_j)\pi_j(\mathrm{d}\theta_j).$$

Of course, there is no new methodology that is being presented here, merely an interpretation of what things mean. However, one might argue that interpretation is the most important, to properly understand what one is doing.

There are some necessary asymptotic studies that are needed here. If $\theta^*\in\Theta_j$ then we require

$$\mathrm{P}(M_j|X_1,\ldots,X_n)\to 1 \quad \text{a.s.}$$

For two choices of model, and assuming $\mathrm{P}(M_1)=\mathrm{P}(M_2)=1/2$, then

$$\mathrm{P}(M_1|X_1,\ldots,X_n)=\frac{I_{n1}}{I_{n1}+I_{n2}},$$

where

$$I_{nj}=\int_{\Theta_j}\prod_{i=1}^n f_j(x_i;\theta_j)\pi_j(\mathrm{d}\theta_j).$$

Under mild regularity conditions it is known that $-n^{-1}\log I_{nj}$ converges a.s. to δ_j^* which is the minimum Kullback–Leibler divergence between $f_0(x)$ and the family $\{f_j(x;\theta),\theta\in\Theta_j\}$. Hence, asymptotically

$$\frac{\mathrm{P}(M_1|X_1,\ldots,X_n)}{\mathrm{P}(M_2|X_1,\ldots,X_n)}\sim\exp\{n(\delta_2^*-\delta_1^*)\}.$$

See also Walker et al. (2004).

6. Asymptotics

It is only recently that Bayesian asymptotics, and most notably consistency, have started to be fully explored in great detail. This is particularly so of misspecified models, which is the scenario of interest here. The set-up is as follows: a model $f(x;\theta)$ has been chosen and the target is θ^*, the parameter value which minimizes

$$l(\theta)=-\int\log f(x;\theta)F_0(\mathrm{d}x).$$

And we can define δ^* as the value of

$$d_{KL}(f_0(\cdot),f(\cdot;\theta^*)),$$

assuming that

$$\delta^*=\inf_{\theta\in\Theta}d_{KL}(f_0(\cdot),f(\cdot;\theta)),$$

and the infimum is attained at $\theta^*\in\Theta$. The maximum likelihood estimator is typically a consistent sequence, i.e. $\widehat{\theta}\to\theta^*$ a.s., and the reason is quite clear, since $\widehat{\theta}$ minimizes

$$l_n(\theta)=-\int\log f(x;\theta)F_n(\mathrm{d}x).$$

See White (1982).

The Bayesian would expect the posterior distribution to accumulate at θ^* and would also expect the sequence of guesses to converge in some sense to $f(x;\theta^*)$. Formally, the appropriate consistency result can be written as

$$\Pi(A_\epsilon|X_1,\ldots,X_n)\to 0 \quad \text{a.s.}$$

where $A_\epsilon=\{\theta:d(\theta,\theta^*)<\epsilon\}$ and $d(\theta,\theta^*)$ denotes the Euclidean distance between parameter values.

In the case when the θ models the density itself, as in Bayesian nonparametrics, then $f(x;\theta)=\theta(x)$ and more commonly one then writes $f(x)$ as the density. Then we can talk about a $f^*(x)$ and now we need an alternative to the Euclidean distance which measures distances between density functions. The most appropriate distance, as it favors the mathematics, is the

Hellinger distance, given by

$$d_H(f,f^*)=\left\{\int(\sqrt{f}-\sqrt{f^*})^2\right\}^{1/2}.$$

Hence, now $A_\epsilon=\{f:d_H(f,f^*)<\epsilon\}$.

And for the sequence of guesses accumulating at the right place, this is automatic from the above result for accumulation at θ^* or f^*. Writing, for brevity

$$m_n(x)=m_n(x|x_1,\ldots,x_{n-1}) \quad \text{and} \quad \pi_n(\mathrm{d}f)=\pi(\mathrm{d}f|x_1,\ldots,x_n),$$

we have

$$d_H(m_n,f^*)\leq\int d_H(f,f^*)\pi_{n-1}(\mathrm{d}f)=\int_{A_\epsilon}d_H(f,f^*)\pi_{n-1}(\mathrm{d}f)+\int_{A_\epsilon^c}d_H(f,f^*)\pi_{n-1}(\mathrm{d}f),\leq\epsilon+\Pi_{n-1}(A_\epsilon^c)$$

and hence $d_H(m_n,f^*)$ must go to 0 due to the arbitrariness of ϵ, and on the assumption that $\Pi_{n-1}(A_\epsilon^c)\to 0$ a.s. for all $\epsilon>0$.

Hence, the aim is to demonstrate the asymptotics for convergence and accumulation at θ^*. This is often referred in the literature to the misspecified model scenario. It should be called the realistic model scenario.

First work in this direction has been done by Berk (1966) and more recently by Bunke and Milhaud (1998), Kleijn and van der Vaart (2006) and De Blasi and Walker (2012). These papers employ techniques to establish the appropriate asymptotic result of posterior accumulation at θ^*, or f^* in the nonparametric case. However, this type of asymptotics is rarely discussed or studied, yet with the emphasis of the wrong model and for the reasons discussed in the paper, makes it the most appropriate.

Another issue to discuss is how to determine the asymptotics and for this it is necessary to determine the correct distribution of the sample $X_1,X_2,\ldots$. The Bayesian model clearly induces a dependence structure in the sequence, but this does not imply there is a stochastic dependence in reality. The dependence is so that learning can take place and so that guesses can be refined and improved. For there to be a stochastic dependence, one would really be setting up a function connecting the outcomes, more like a time series model. An explicit example here is something of the type $X_n=\rho X_{n-1}+\varepsilon_n$. The type of observation we are considering here is that they arise from some fixed distribution and are independent in the sense that there is no functional form connecting any two observations.

This is not in conflict with the model setting up a stochastic dependence. More explicitly, X_1 provides information about θ which obviously then provides further information about X_2, which was not available prior to the observation of X_1. Hence, it is the obvious reason why guesses change and adapt according to what has been observed. This does not amount to a functional form of dependence but corresponds to a learning form of stochastic dependence. The learning form of dependence is represented by the de Finetti representation theorem.

So it is reasonable to conduct the asymptotic studies assuming that the observations are independent and identically distributed from some fixed but unknown distribution function $F_0(x)$. For some reason, which is far from clear, studying a Bayesian model under this scenario is described as a frequentist study of the Bayesian model. This seems to be a simple misunderstanding. The difference between Bayes and frequentist methodology is not about assumptions to do with the observations. Assuming observations is independent and identically distributed should be nothing to do with any designated statistical inference technique. As we have previously mentioned, the separation is that the Bayesian is seeking to guess $m_n(x)$ for each $n\geq 1$.

Moreover, the alternative to the so-called frequentist asymptotic study of a Bayesian model is the Doob (1949) style which assumes that everything about the Bayesian model is correctly assigned, even down to the prior. This asymptotics would suggest that the Bayesian model is better than it actually is.

7. Non-symmetric models

So now let us think about alternative notions where the order of observations is regarded as relevant. Suppose one has a model for which one has

$$m(x_1,\ldots,x_n)=\int_\Theta\prod_{i=1}^n f(x_i|x_{i-1},\ldots,x_i;\theta)\pi(\mathrm{d}\theta).$$

This model is coherent with the choice of guesses as

$$m_n(x|x_{n-1},\ldots,x_1)=\int_\Theta f(x|x_{n-1},\ldots,x_1;\theta)\pi(\mathrm{d}\theta|x_1,\ldots,x_{n-1})$$

where

$$\pi(\mathrm{d}\theta|x_{n-1},\ldots,x_1)=\frac{\prod_{i=1}^{n-1}f(x_i|x_{i-1},\ldots,x_1;\theta)\pi(\theta)}{\int_\Theta\prod_{i=1}^{n-1}f(x_i|x_{i-1},\ldots,x_1;\theta)\pi(\mathrm{d}\theta)}.$$

Making various assumptions about $m(x_1,\ldots,x_n)$, such as stationarity, see Maitra (1977); one has a representation theorem from which the Bayes update and learning process can be extracted.

However, there are plenty of data structures for which there is no such representation theorem. And now Bayes needs to be motivated in a different way. The argument in Section 4 is no longer available, as the sequence of guesses has no representation theorem. Now, to proceed as a Bayesian one must start with the model

$$f(x_n|x_1,\ldots,x_{n-1};\theta) \quad \text{and} \quad \pi(\theta).$$

In this case, the marginal densities

$$m_n(x|x_1,\ldots,x_{n-1}) = \int_\Theta f(x_n|x_1,\ldots,x_{n-1};\theta)\pi(\mathrm{d}\theta)$$

do not follow a structure from which f and π can be deduced. Given the state of the wrong model and interest focusing on θ^*, the direct application of Bayes theorem is unmotivated.

We must now look for an alternative plan and one can be found by treating the update of $\pi(\theta)$ to $\pi(\theta|x_1,\ldots,x_n)$ as a decision problem. Hence, if the action is denoted by $\nu(\theta)$, the aim is to construct a loss function $L(\nu;x^{(n)},\pi)$, where we now write $x^{(n)}=(x_1,\ldots,x_n)$. The most appropriate loss function is using notions of cumulative loss treating the information provided by the $n+1$ pieces of information $(\pi,x_1,\ldots,[x_n|x^{(n-1)}])$ as mutually independent. Hence, the form of loss function will be of the type

$$L(\nu;x^{(n)},\pi) = \sum_{i=1}^{n} l_1([x_i|x^{(i-1)}],\nu) + l_2(\pi,\nu),$$

where l_1 and l_2 are as yet unspecified loss functions.

The choices of losses here are now automatic. As shown in Bissiri and Walker (2012), for coherence, it is necessary for $l_2(\pi,\nu)$ to be the Kullback–Leibler divergence; i.e.

$$l_2(\pi,\nu) = \int_\Theta \nu(\mathrm{d}\theta)\log\{\nu(\mathrm{d}\theta)/\pi(\mathrm{d}\theta)\}.$$

Here coherence means that the solution after n observations can serve as the prior for future observations. The loss function $l(\theta,[x_i|x^{(i-1)}])$ can only be the self-information (the honest) loss function; i.e.

$$l(\theta,[x_i|x^{(i-1)}]) = -\log f(x_i|x^{(i-1)};\theta).$$

And since ν is to represent beliefs about θ, we can and should take $l_1([x_i|x^{(i-1)}],\nu)$ as the expected value of the self-information loss function; i.e.

$$l_1([x_i|x^{(i-1)}],\nu) = \int_\Theta -\log f(x_i|x^{(i-1)};\theta)\nu(\mathrm{d}\theta).$$

Therefore

$$L(\nu;x^{(n)},\pi) = -\sum_{i=1}^{n}\int_\Theta \log f(x_i|x^{(i-1)};\theta)\nu(\mathrm{d}\theta) + \int_\Theta \nu(\mathrm{d}\theta)\log\{\nu(\mathrm{d}\theta)/\pi(\mathrm{d}\theta)\}.$$

Surprisingly now, the solution, that is the minimizer of $L(\nu;x^{(n)},\pi)$ is given by

$$\nu(\mathrm{d}\theta) = \pi(\mathrm{d}\theta|x^{(n)}),$$

the Bayesian posterior.

8. Bayesian nonparametrics

An attempt to overcome the problem of working with the wrong model is to make the model very large. So large in fact that it is reasonable to assume that there is in the support of the prior a density which can be accepted as generating the observations. For example, suppose one has independent and identically distributed observations on the real line, then a common nonparametric prior is based on an infinite mixture of normal distributions

$$f(x) = \sum_{j=1}^{\infty} w_j \in (x|\mu_j,\sigma_j^2).$$

This can be written as

$$f(x) = \int \mathrm{N}(x|\mu,\sigma^2)\mathrm{d}P(\mu,\sigma)$$

which describes the mixture of Dirichlet process (MDP) model if the prior for P is assigned as a Dirichlet process (Ferguson, 1973). Then, according to Sethuraman (1994), and writing $\theta=(\mu,\sigma)$, we can construct

$$P = \sum_{j=1}^{\infty} w_j\,\delta_{\theta_j}$$

where the weights form a stick-breaking sequence and the (θ_j) are i.i.d from density function $g(\theta)$. And so, specifically for i.i.d. (v_j) from a beta $(1,c)$ density, for some $c>0$, $w_1=v_1$ and, for $j>1$, $w_j=v_j\prod_{l<j}(1-v_l)$. Other stick-breaking constructions are allowed based on alternative beta distributions; see Ishwaran and James (2001) for conditions.

Thus, the density model for the data arises as

$$f(x)=\sum_{j=1}^{\infty} w_j \mathrm{N}(x|\theta_j)$$

which is an infinite mixture model. See Lo (1984) and Escobar and West (1995) for original ideas for these models and Hjort (2010) for a recent review of Markov chain Monte Carlo methods to implement inference for these types of models.

It is acceptable to use Bayes theorem in this scenario as it is reasonable to assume that there is in the support of the prior a density $f_0(x)$ which generates the observations as independent and identically distributed. Yet Bayes theorem that can be used is not the end of the story since as always it is necessary to check that we are indeed learning about $f_0(x)$. This notion has generated a substantial amount of literature recently, and the papers which ignited the productive phase include Schwartz (1965), Barron (1988), Barron et al. (1999) and Ghosal et al. (1999). We do not need to discuss posterior consistency, or even posterior convergence rates, in any detail. The basic idea is that if $f_0(x)$ is in the Kullback–Leibler support of the prior, that is

$$\Pi(f : d_{KL}(f_0,f)<\epsilon)>0,$$

for all $\epsilon>0$, then strong consistency is guaranteed with some additional condition on the prior distribution.

The infinite mixture model is the bedrock of modern Bayesian nonparametrics. Since the work of Lo (1984), who introduced the mixture of Dirichlet process mixture model, and the advent of Bayesian posterior inference via simulation techniques (see Escobar, 1988, and Smith and Roberts, 1993), Bayesian nonparametric methods have developed at a rapid pace and the Dirichlet mixture model is one of the most popular among these methods. The models have now moved away from the standard set-up, namely i.i.d. observations, to cover more complex data structures involving regression and time series data. There are numerous works and papers over the last decade and it is therefore convenient to cite the book of Hjort et al. (2010), which contains references and discussions of many nonparametric models.

The idea is to now model the dependent variable y on regression variable x, using the idea of the infinite dimensional mixture model. Hence, the obvious first attempt would be of the type

$$f(y|x)=\int k(y|x,\theta)\,\mathrm{d}P_x(\theta).$$

Here $P_x(\theta)$ is similar in construction to $P(\theta)$, but the weights and locations can both depend on x. That is

$$P_x(\theta)=\sum_{j=1}^{\infty} w_j(x)\delta_{\theta_j(x)}.$$

Exactly how to define the $(w_j(x),\theta_j(x))$ suitably is an interesting problem and a number of attempts have been tried; see Hjort et al. (2010) for a review. But the vast number of published models here is testament to the difficulty in knowing how to choose the x-dependent weights and locations. It is a very difficult task to specify these components; i.e. the $(w_j(x))$, $(\theta_j(x))$ and $K(y|x,\theta(x))$. There are limitless possibilities and over-fitting and un-identifiability are serious issues. It is argued that some sort of guidance is needed in order to justify certain specifications.

The complexity of how to construct these key functions over the x-space can be avoided by modeling the joint density as a mixture model: temporarily assuming that the (y,x) are generated from some independent stochastic process, then we would model the joint density as

$$f(y,x)=\int k(y,x|\theta)\,\mathrm{d}P(\theta).$$

Here a single P is required. In this case the regression model provides simply through the definition of a conditional density

$$f(y|x)=\frac{f(y,x)}{f(x)}=\frac{\int k(y,x|\theta)\,\mathrm{d}P(\theta)}{\int k(x|\theta)\,\mathrm{d}P(\theta)}$$

so

$$f(y|x)=\frac{\sum_{j=1}^{\infty} w_j\, k(y,x|\theta_j)}{\sum_{j=1}^{\infty} w_j\, k(x|\theta_j)},$$

Such a model has not yet been entertained due to the difficulty of dealing with the denominator.

This attractively simple and intuitive approach to Bayesian nonparametric modeling has previously been proposed by Müller et al. (1996). Yet, presumably due to the difficulty of the denominator, they employed the likelihood function

$$\prod_{i=1}^{n} f(y_i,x_i).$$

However, the aim is regression rather than modeling the (y,x) and hence the appropriate likelihood function is given by the one involving the conditional rather than the joint density. The joint density is merely being used as a vehicle to construct a motivated form for the nonparametric regression model.

It is then possible to note that the weights $w_j(x)$ take a particular form which has not appeared in the literature

$$w_j(x)=\frac{w_j\,k(x|\theta_j)}{\sum_j w_j k(x|\theta_j)}.$$

The reason why such a simple, motivated and useful regression model has not appeared in the literature is due to the fact that the posterior distribution has an intractable normalizing constant. Inference is complicated by the need to evaluate the uncomputable integrals.

The aim here is to show that it is possible to use the correct likelihood for regression by showing how to deal with the problem of the normalizing constant. This uses ideas of latent variables (Besag and Green, 1993; Damien et al., 1999) and specifically ideas recently introduced in Walker (2011). However, we will do this for a first order stationary time series model as the inference will be very similar. The plan then is to construct a transition density $f(y|x)$ which would suitably capture irregular transition dynamics and then study the posterior distribution based on the likelihood function

$$\prod_{i=1}^{n} f(y_i|y_{i-1}).$$

A prior is assigned to the transition density and will be written as $\Pi(\mathrm{d}f)$. This will be based on the Dirichlet process mixture model.

The modeling of first order time series data is also a vast area in the literature. The aim is to use a Bayes nonparametric mixture model to construct $f(\cdot|\cdot)$; so, specifically, as with the regression setting

$$f(y|x)=\int k(y|x,\theta)\,\mathrm{d}P_x(\theta),$$

where $k(y|x,\theta)$ is a density for every $\theta\in\Theta$ and P_x is a probability measure that depends on x. This type of model will be able to capture a wide class of transition functions, and the infinite dimensional aspect to the model means that any surprise or change that arises in the future will be taken into account.

Let us start with a parametric first order stationary time series model $k(y,x|\theta)$ which has identical marginals

$$k(x|\theta)=\int k(y,x|\theta)\,\mathrm{d}y \quad \text{and} \quad k(y|\theta)=\int k(y,x|\theta)\,\mathrm{d}x.$$

Also, $k(y|\theta)$ is the stationary density:

$$k(y|\theta)=\int k(y|x,\theta)k(x|\theta)\,\mathrm{d}x.$$

This is now ready to be extended to the nonparametric setting by taking

$$f_P(y,x)=\int k(y,x|\theta)\,\mathrm{d}P(\theta)=\sum_{j=1}^{\infty} w_j\,k(y,x|\theta_j).$$

We can obtain the nonparametric model; the stationary density is

$$f_P(x)=\int k(x|\theta)\,\mathrm{d}P(\theta),$$

and the transition density is

$$f_P(y\Big|x)=\frac{\int k(y|x,\theta)k(x|\theta)\,\mathrm{d}P(\theta)}{\int k(x|\theta)\,\mathrm{d}P(\theta)},$$

so

$$\mathrm{d}P_x(\theta)=\frac{k(x|\theta)\,\mathrm{d}P(\theta)}{\int k(x|\theta)\,\mathrm{d}P(\theta)}.$$

Equivalently

$$w_j(x)=\frac{w_j k(x|\theta_j)}{\sum_{j=1}^{\infty} w_j k(x|\theta_j)}.$$

For the new model we will need to estimate the parameters of k and P. To make this concrete we will present a particular model. Assume $k(x|\theta,\sigma^2)$ to be normal with mean θ and variance σ^2 and let

$$P(\theta)=\sum_{j=1}^{\infty} w_j\delta_{\theta_j}(\theta).$$

Here the weights (w_j) sum to one and the (θ_j) are real numbers and the set-up assumes that the means can change from component to component but the variance will remain the same for all components. We will then be interested in estimating $((w_j,\theta_j),\sigma)$ and the prior for $\lambda=\sigma^{-2}$ will be denoted by $\pi(\lambda)$.

Now it can be seen that the likelihood function based on a sample $(y_1,\ldots,y_n)$ is given by

$$\prod_{i=1}^{n}\frac{\sum_{j=1}^{\infty}w_j\;k(y_i|y_{i-1},\theta_j,\lambda)k(y_{i-1}|\theta_j,\lambda)}{\sum_{j=1}^{\infty}w_j\;k(y_{i-1}|\theta_j,\lambda)}.$$

This looks like an insurmountable likelihood to deal with. Our aim then is to show how to undertake Bayesian inference for this model using well designed latent variables which result in a viable latent model.

The numerator has a common and standard technique for simplification and this is to introduce the allocation variables (d_i) which lead to the latent model

$$\prod_{i=1}^{n}\frac{w_{d_i}k(y_i|y_{i-1},\theta_{d_i},\lambda)k(y_{i-1}|\theta_{d_i},\lambda)}{\sum_{j=1}^{\infty}w_j\;k(y_{i-1}|\theta_j,\lambda)}.$$

Summing over the independent (d_i) returns the original likelihood. The issue now is to deal with the denominator.

First we can remove the λ from each k; so define

$$m(y|\theta,\lambda)=\exp\{-\tfrac{1}{2}\lambda(y-\theta)^2\}.$$

Now we write the latent likelihood model as

$$\lambda^{n/2}\prod_{i=1}^{n}\frac{w_{d_i}m(y_i|y_{i-1},\theta_{d_i},\lambda)m(y_{i-1}|\theta_{d_i},\lambda)}{\sum_{j=1}^{\infty}w_j\;m(y_{i-1}|\theta_j,\lambda)}.$$

We now focus on the denominator and the term

$$\frac{1}{\sum_{j=1}^{\infty}w_j\;m(y|\theta_j,\lambda)}.$$

This has been written with a generic y, and it is simpler to consider this first, and then put the product back together later.

Since the denominator is now between (0,1) we can write it as

$$\sum_{k=0}^{\infty}\left[\sum_{j=1}^{\infty}w_j(1-m(y|\theta_j,\lambda))\right]^k.$$

This suggests that we should introduce the latent variable k yielding the latent model

$$\left[\sum_{j=1}^{\infty}w_j(1-m(y|\theta_j,\lambda))\right]^k.$$

Finally, we can introduce the latent variables $(z_l:l=1,\ldots,k)$ and the latent model

$$\prod_{l=1}^{k}w_{z_l}(1-m(y|\theta_{z_l},\lambda)).$$

Putting this with the latent model for the numerator, and recalling we have a k_i for each i, the final latent model is given by

$$\lambda^{n/2}\prod_{i=1}^{n}w_{d_i}k(y_i|y_{i-1},\theta_{d_i},\lambda)k(y_{i-1}|\theta_{d_i},\lambda)\prod_{l=1}^{k_i}w_{z_{il}}(1-m(y_{i-1}|\theta_{z_{il}},\lambda)).$$

It is easy to see that summing over all the latent variables $((d_i),(k_i),(z_{il}))$ over their respective spaces returns the original likelihood.

We are now in a position where the latent model is similar to standard latent models for mixture of Dirichlet process models; see Kalli et al. (2010). The form above suggests that there is a solution to the problem. Hence, we start to describe the sampling MCMC algorithm. As it stands, if we attempted to sample the d_i or the z_{il} we would face the problem that they are to be taken from the positive integers, and it would not be possible to evaluate all the relevant probabilities. We can therefore truncate this choice using ideas from Kalli et al. (2010) whereby we introduce latent variables δ_i and ζ_i which are combined with the latent model via

$$\mathbf{1}(\delta_i<e^{-\xi d_i})e^{\xi d_i}\quad\text{and}\quad\mathbf{1}(\zeta_{il}<e^{-\xi z_{il}})e^{\xi z_{il}}.$$

Here $\xi>0$ and its value is not a modeling issue. A discussion on its choice and its role is given in Kalli et al. (2010). Therefore

$$\mathrm{P}(d_i=j|\cdots)\propto e^{\xi j}w_jk(y_i|y_{i-1},\theta_j,\lambda)k(y_{i-1}|\theta_j,\lambda)\mathbf{1}(1\le j\le N_i),$$

where $N_i=\lfloor-\xi^{-1}\log\delta_i\rfloor$. Also

$$\mathrm{P}(z_{il}=j|\cdots)\propto e^{\xi j}w_j(1-m(y_{i-1}|\theta_j,\lambda))\mathbf{1}(1\le j\le N_{il}),$$

where $N_{il} = \lfloor -\xi^{-1} \log \zeta_{il} \rfloor$. The maximum value $N = \max\{N_i, N_{il}\}$ will then tell us exactly how many of the (θ_j, w_j) need to be sampled at each iteration of the MCMC algorithm. The weights are easy to sample and the conditional for each v_j is a straightforward extension of the usual mixture of Dirichlet process model, and is given by

$$v_j = \text{beta}\left(1 + \sum_{i=1}^{n} \mathbf{1}(d_i = j) + \sum_{i=1,l=1}^{n,k_i} \mathbf{1}(z_{il} = j),\ c + \sum_{i=1}^{n} \mathbf{1}(d_i > j) + \sum_{i=1,l=1}^{n,k_i} \mathbf{1}(z_{il} > j)\right).$$

These (v_j) can then be transformed to get the (w_j).

The (θ_j) are best sampled by introducing a latent variable u_{il} for each i and l. This enters the model via

$$\mathbf{1}(u_{il} < 1 - m(y_{i-1} | \theta_{z_{il}}, \lambda)).$$

These are standard slice random variables; see Damien et al. (1999). Hence

$$p(\theta_j | \cdots) \propto \pi(\theta_j) \prod_{d_i = j} m(y_i | y_{i-1}, \theta_j, \lambda) m(y_{i-1} | \theta_j, \lambda) \prod_{z_{il} = j} (m(y_{i-1} | \theta_j, \lambda) < 1 - u_{il}).$$

The conditional for λ is given by

$$p(\lambda | \cdots) \propto \lambda^{n/2} \prod_{i=1}^{n} m(y_i | y_{i-1}, \theta_{d_i}, \lambda) m(y_{i-1} | \theta_{d_i}, \lambda) \prod_{l=1}^{k_i} (m(y_{i-1} | \theta_{z_{il}}, \lambda) < 1 - u_{il}).$$

Finally, we need to update each k_i. We do this independently and so we consider a generic k with relevant model part given by

$$\prod_{l=1}^{k} w_{z_l} \psi_{z_l},$$

where we have written $\psi_{z_l} = 1 - m(y | \theta_{z_l}, \lambda)$.

We deal with this apparent changing dimension part of the model using ideas in Godsill (2001), which is based on the reversible jump MCMC methodology of Green (1995). If we write

$$p(k, z_1, \ldots, z_k) \propto \prod_{l=1}^{k} w_{z_l}\, \psi_{z_l},$$

then we extend the model to

$$p(k, z_1, \ldots, z_k, z_{k+1}, \ldots) \propto \left\{ \prod_{l=1}^{k} w_{z_l}\, \psi_{z_l} \right\} \prod_{l=k+1}^{\infty} w_{z_l}.$$

From k we can propose a move to $k+1$ with probability $\frac{1}{2}$, or to $k-1$ with probability $\frac{1}{2}$. The probability of accepting a move to $k+1$ is given by

$$\min\{1, \psi_{z_{k+1}}\},$$

where z_{k+1} has been sampled from the weights (w_j). On the other hand, the probability of accepting a move to $k-1$ is given by

$$\min\{1, \psi_{z_k}^{-1}\}.$$

This concludes a description for the MCMC algorithm for estimating the model.

9. Discussion

The key to the ideas presented in this paper is about the Bayesian properly defining the parameter value of interest. Coherence is then about the Bayesian updates providing learning about this value and this is necessarily checked by what is yielded asymptotically. This does not mean that one must have an asymptotic sample, it means that the target is what the Bayesian update is providing evidence for. This does suggest a gear change in Bayesian methodology; that when a model is proposed the following becomes mandatory:

1. The parameter value of interest is identified, and in the independent and identically distributed case this is most sensibly taken to be the parameter value minimizing the Kullback–Leibler divergence between the model and the true density generating the data.
2. Prior specifications are provided for this parameter of interest.
3. The mathematics is performed to determine that this parameter value of interest can be discovered through the Bayesian updating machinery. (If it cannot, the target is misspecified and in fact the machine is learning or moving to something different to what is of interest. The model and experimenter are incoherent. The experimenter and model can be coherent if the target and asymptotics match.)

When there is a representation theorem for the sequence of guesses $m_n(\cdot)$ for the distributions of the observations; then Bayes theorem applies, even when it is acknowledged the model is wrong; i.e. the Bayes update is to be found in the representation theorem. When there is no representation theorem and, due to the wrongness of the model, there is no direct application of Bayes theorem, then a decision theoretic approach to the construction of the posterior is enough to justify the use of the Bayes update.

Bayesian nonparametrics are about constructing models which are large enough so that the model cannot be claimed to be misspecified and hence a Bayes theorem is directly applicable.

References

Barron, A., 1988. The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions. Unpublished manuscript.

Barron, A., Schervish, M.J., Wasserman, L., 1999. The consistency of posterior distributions in nonparametric problems. Annals of Statistics 27, 536–561.

Berk, R.H., 1966. Limiting behavior of posterior distributions when the model is incorrect. Annals of Mathematical Statistics 37, 51–58. [Corrigendum **37**, 745–746].

Berger, J.O., 2006. The case for objective Bayesian analysis. Bayesian Analysis 1, 385–402.

Bernardo, J.M., Smith, A.F.M., 1994. Bayesian Theory. Wiley.

Berry, D., 1996. Statistics: A Bayesian Perspective. Duxbury, Belmont.

Besag, J., Green, P.J., 1993. Spatial statistics and Bayesian computation. Journal of the Royal Statistical Society, Series B 55, 25–37.

Bissiri, P.G., Walker, S.G., 2012. Converting information into probability measures with the Kullback–Leibler divergence. Annals of the Institute of Statistical Mathematics 64, 1139–1160.

Box, G.E.P., 1980. Sampling and Bayes' inference in scientific modeling and robustness. Journal of the Royal Statistical Society, Series A 143, 383–430.

Brown, P.J., Walker, S.G., 2012. Bayesian priors from loss matching (with discussion). International Statistical Review 80, 60–92.

Bunke, O., Milhaud, X., 1998. Asymptotic behavior of Bayes estimates under possibly incorrect models. Annals of Statistics 26, 617–644.

Cox, D.R., 2006. Frequentist and Bayesian statistics: a critique. At ⟨http://www.physics.ox.ac.uk/phystat05/proceedings/files/papb-ayesrev.pdf⟩.

Damien, P., Wakefield, J.C., Walker, S.G., 1999. Gibbs sampling for Bayesian non-conjugate and hierarchical models using auxiliary variables. Journal of the Royal Statistical Society, Series B 61, 331–344.

De Blasi, P., Walker, S.G., 2012. Bayesian asymptotics with misspecified models. Statistica Sinica 23, 169–187.

de Finetti, B., 1937. La prévision: ses lois logiques, ses sources subjectives. Annales de IÍnstitut Henri Poincaré 7, 1–68.

Doob, J.L., 1949. Application of the theory of martingales. In: Le Calcul des Probabilités et ses Applications, Colloques Internationaux du Centre National de la Recherche Scientifique, CNRS, Paris, vol. 13, pp. 23–37.

Escobar, M.D., 1988. Estimating the means of several normal populations by nonparametric estimation of the distribution of the means. Unpublished PhD Dissertation, Department of Statistics, Yale University.

Escobar, M.D., West, M., 1995. Bayesian density estimation and inference using mixtures. Journal of the American Statistical Association 90, 577–588.

Ferguson, T.S., 1973. A Bayesian analysis of some nonparametric problems. Annals of Statistics 1, 209–230.

Gelman, A., 2008. Objections to Bayesian Statistics (with discussion). Bayesian Analysis 3, 445–450.

Ghosal, S., Ghosh, J.K., Ramamoorthi, R.V., 1999. Posterior consistency of Dirichlet mixtures in density estimation. Annals of Statistics 27, 143–158.

Godsill, S.J., 2001. On the relationship between Markov chain Monte Carlo methods for model uncertainty. Journal of Computational and Graphical Statistics 10, 230–248.

Goldstein, M., 2006. Subjective Bayesian analysis: principles and practice. Bayesian Analysis 1, 403–420.

Green, P.J., 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82, 711–732.

Hewitt, E., Savage, L.J., 1955. Symmetric measures on Cartesian products. Transactions of the American Mathematical Society 80, 470–501.

Hirshleifer, J., Riley, J.G., 1992. The Analytics of Uncertainty and Information. Cambridge University Press.

Hjort, N.L., Holmes, C.C., Müller, P., Walker, S.G., 2010. Bayesian Nonparametrics. Cambridge University Press.

Ishwaran, H., James, L.F., 2001. Gibbs sampling methods for stick-breaking priors. Journal of the American Statistical Association 96, 161–173.

Kalli, M., Griffin, J.E., Walker, S.G., 2010. Slice sampling mixture models. Statistics and Computing 21, 93–105.

Kass, R., Raftery, A.E., 1995. Bayes factors. Journal of the American Statistical Association 90, 773–795.

Key, J.T., Pericchi, L.R., Smith, A.F.M., 1999. Bayesian model choice: What and why? (with discussion). In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F. M. (Eds.), Bayesian Statistics, vol. 6. . Oxford University Press, pp. 343–370.

Kleijn, B.J.K., van der Vaart, A.W., 2006. Misspecification in infinite dimensional Bayesian statistics. Annals of Statistics 34, 837–877.

Kullback, S., Leibler, R.A., 1951. On information and sufficiency. Annals of Mathematical Statistics 22, 79–86.

Lo, A.Y., 1984. On a class of Bayesian nonparametric estimates I. Density estimates. Annals of Statistics 12, 351–357.

Maitra, A., 1977. Integral representations of invariant measures. Transactions of the American Mathematical Society 229, 209–225.

Müller, P., Erkanli, A., West, M., 1996. Bayesian curve fitting using multivariate normal mixtures. Biometrika 83, 67–79.

Sethuraman, J., 1994. A constructive definition of Dirichlet priors. Statistica Sinica 4, 639–650.

Schwartz, L., 1965. On Bayes procedures. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 4, 10–26.

Smith, A.F.M., Roberts, G.O., 1993. Bayesian computations via the Gibbs sampler and related Markov chain Monte Carlo methods. Journal of the Royal Statistical Society, Series B 55, 3–23.

von Neumann, J., Morgenstern, O., 1944. Theory of Games and Economic Behavior. Princeton University Press.

Walker, S.G., Damien, P., Lenk, P.J., 2004. On priors with a Kullback–Leibler property. Journal of the American Statistical Association 99, 404–408.

Walker, S.G., 2011. Posterior sampling when the normalizing constant is unknown. Communications in Statistics: Simulation and Computation 40, 784–792 .

White, H., 1982. Maximum likelihood estimation of misspecified models. Econometrica 50, 1–25.