IMS Collections
Contemporary Developments in Bayesian Analysis and Statistical Decision Theory: A Festschrift for William E. Strawderman
Vol. 8 (2012) 190–199

DOI: 10.1214/11-IMSCOLL813

A nonparametric Bayesian method for estimating a response function

Scott Brown[1] **and Glen Meeden**[2]

The Integra Group and University of Minnesota

Abstract: Consider the problem of estimating a response function which depends upon a non-stochastic independent variable under our control. The data are independent Bernoulli random variables where the probabilities of success are given by the response function at the chosen values of the independent variable. Here we present a nonparametric Bayesian method for estimating the response function. The only prior information assumed is that the response function can be well approximated by a mixture of step functions.

Contents

1. Introduction

We are interested in the problem where a response variable has a Bernoulli distribution and where the unknown probability of success depends on the level of some factor. Usually the link function, which gives the relationship between levels of the factor and the corresponding probabilities, is not known. A common approach to such problems is logistic regression where a convenient mathematical form is assumed to describe this relationship. A recent nonparametric approach using kernel estimation to estimate the link function is discussed in Signorini and Jones (2004). Various Bayesian nonparametric approaches to the problem can be found in Mallick and Gelfand (1994), Newton et al. (1996) and Wood and Kohn (1998). A more truly nonparametric Bayesian approach is presented in the work of Coram and Lalley (2006) and the consistency of their Bayes estimator is demonstrated. Here we will suggest another nonparametric Bayesian approach to this problem. It is based only on the assumption that the relationship is given by a continuous function which can be well approximated by a mixture of step functions. Although similar in spirit to the estimator of Coram and Lalley (2006) our estimator has

[1]The Integra Group, Brooklyn Park, MN 55443, e-mail: sbrown@integracts.com
[2]School of Statistics, University of Minnesota, Minneapolis, MN 55455, e-mail: glen@stat.umn.edu
AMS 2000 subject classifications: Primary 62C10, 62C15; secondary 62G05
Keywords and phrases: binary regression, nonparametric Bayes, stepwise Bayes

a stepwise Bayes justification which we will use to show its admissibility. Brown (1981) showed that the family of stepwise Bayes procedures forms a complete class for this problem.

In section 2 we introduce some notation and describe our estimator. In section 3 we briefly review the stepwise Bayes argument for proving the admissibility of the maximum likelihood estimator for the probability of success in the Binomial model. In section 4 we show how this stepwise Bayes argument can be extended to prove the admissibility of our estimator. In section 5 we compare our estimator to some others and section 6 contains some concluding remarks.

2. Notation for the problem and our estimator described

Let x be the dependent variable whose range is D, and which for convenience we assume is the unit interval. Given a value of x let Y_x be a Bernoulli random variable where $\theta(x)$ denotes the probability that Y_x is equal to one. The data of our experiment are (X, Y) where $X = (x_1, \ldots, x_n)$ and $Y = (Y_{x_1}, \ldots, Y_{x_n})$ for some positive integer n. We assume that the x_j's are fixed and under our control and are increasing in j. We assume that the Y_{x_j}'s are independent.

Let $\theta : [0,1] \rightarrow [0,1]$ denote the response function. The function θ represents the probability of success of a Bernoulli experiment at each predictor value $x \in [0,1]$. θ is the unknown parameter and we wish to estimate its values at some finite subset of points of [0.1]. We will assume that θ belongs to some subset of the set of all piecewise linear and continuous functions from the unit interval into itself. This reflects our basic assumption that θ is more or less smooth. We will give the exact definition of this set in just a bit.

Let $\vec{\gamma} = (i_1, \ldots, i_k)$ denote a set of positive integers where $1 < i_1 < i_2 < \cdots < i_k < n$. Then $\vec{\gamma}$ can be used to define a partition of X given by

$$\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_{k+1}) = (\{x_1, \ldots, x_{i_1}\}, \{x_{i_1+1}, \ldots, x_{i_2}\}, \ldots, \{x_{i_k+1}, \ldots, x_n\}).$$

If we set $i_0 = 0$, $i_{k+1} = n$ then $n_{\vec{\gamma}}(j) = i_j - i_{j-1}$, is the number of elements belonging to the jth element of the partition. We let $Y_{\vec{\gamma}}(j)$ be the subset of Y which belongs to the jth element of the partition. We can also use the members of the partition to define $k+1$ disjoint subintervals of the unit interval in the natural way.

Let m and k be positive integers. We assume that n is such that X can be partitioned into $k+1$ subsets each containing at least m members in more than one way. Typically, k will usually be quite small compared to m and m will be small compared to n. The values of m and k are fixed and are selected by the statistician. Given m and k, let

$$\Gamma(m,k) = \{\vec{\gamma} : i_1 \geq m,\ i_k \leq n - m \ \text{ and } i_j - i_{j-1} \geq m \text{ for } j = 2, \ldots, k\}$$

denote the class of all such partitions.

Given such a partition consider the situation where $\theta(x)$ is constant over each subinterval of the partition and linearly interpolates in the gaps between the x_{i_j} and $x_{i_{j+1}}$ for $j = 1, \ldots, k$. More formally we let

$$(\theta_{\vec{\gamma}}(1), \ldots, \theta_{\vec{\gamma}}(k+1))$$

denote these constant probabilities and use them to define a possible parameter value for θ. That is

$$\begin{aligned}\theta_{\vec{\gamma}}(x) &= \theta_{\vec{\gamma}}(1) \quad \text{for} \quad x \in [0, x_{i_1}] \\ &= \theta_{\vec{\gamma}}(j) \quad \text{for} \quad x \in [x_{i_{j-1}+1}, x_{i_j}] \text{ and } j = 2, \ldots, k \\ &= \theta_{\vec{\gamma}}(k+1) \quad \text{for} \quad x \in [x_{i_k}, 1] \\ &= \frac{x - x_{i_j}}{x_{i_j+1} - x_{i_j}} \theta_{\vec{\gamma}}(j+1) + \frac{x_{i_j+1} - x}{x_{i_j+1} - x_{i_j}} \theta_{\vec{\gamma}}(j) \\ &\qquad \text{for } x \in (x_{i_j}, x_{i_j+1}) \text{ and } j = 1, \ldots, k.\end{aligned}$$

If we assume that $\vec{\gamma}$ as an unknown parameter as well then the parameter space for our problem is

$$\Theta(m.k) = \{(\vec{\gamma}, \theta_{\vec{\gamma}}) \,:\, \vec{\gamma} \in \Gamma(m,k) \text{ and } 0 \le \theta_{\vec{\gamma}}(j) \le 1 \text{ for } j = 1, \ldots, k+1\}.$$

A natural way to define a prior distribution on $(\vec{\gamma}, \theta_{\vec{\gamma}})$ is the following

$$\begin{aligned}p(\vec{\gamma}, \theta_{\vec{\gamma}}) &= p(\vec{\gamma})\, p(\theta_{\vec{\gamma}} \mid \vec{\gamma}) \\ &= p(\vec{\gamma}) \prod_{j=1}^{k+1} p(\theta_{\vec{\gamma}}(j) \mid \vec{\gamma}).\end{aligned}$$

Here we are assuming that the priors for the $\theta_{\vec{\gamma}}(j)$'s are independent.

Then the joint probability distribution of the data and the parameters can be expressed as

$$\begin{aligned}p(\vec{\gamma}, \theta_{\vec{\gamma}}, y) &= p(\vec{\gamma}) \prod_{j=1}^{k+1} p(\theta_{\vec{\gamma}}(j) \mid \vec{\gamma})\, p(y_{\vec{\gamma}}(j) \mid \vec{\gamma}, \theta\vec{\gamma}) \\ &= p(\vec{\gamma}) \prod_{j=1}^{k+1} \Big(p(y_{\vec{\gamma}}(j) \mid \vec{\gamma}) p(\theta_{\vec{\gamma}}(j) \mid \vec{\gamma}, y_{\vec{\gamma}}(j)) \Big).\end{aligned}$$

A standard calculation yields

$$p(\vec{\gamma}, , \theta_{\vec{\gamma}} \mid y) = p(\vec{\gamma} \mid y) \prod_{j=1}^{k+1} p(\theta_{\vec{\gamma}}(j) \mid \vec{\gamma}, y_{\vec{\gamma}}(j)), \tag{2.1}$$

where

$$p(\vec{\gamma} \mid y) = p(\vec{\gamma}) \prod_{j=1}^{k+1} p(y_{\vec{\gamma}}(j) \mid \vec{\gamma}) / \sum_{\vec{\gamma}'} p(\vec{\gamma}') \prod_{j=1}^{k+1} p(y_{\vec{\gamma}}(j) \mid \vec{\gamma}'). \tag{2.2}$$

Fix t at one of the x_{i_j} values and consider estimating $\theta(t)$. For a given partition $\vec{\gamma}$, let $\theta_{\vec{\gamma}}(j_t)$ denote the constant probability on the subinterval which contains t and $y_{\vec{\gamma}}(j_t)$ denote the corresponding y values. If we put the uniform prior over $\Gamma(m,k)$, the space of possible partitions, then we have from equations (2.1) and (2.2) that

$$\sum_{\vec{\gamma}} \frac{\prod_{j=1}^{k+1} p(y_{\vec{\gamma}}(j) \mid \vec{\gamma})}{\sum_{\vec{\gamma}'} \prod_{j=1}^{k+1} p(y_{\vec{\gamma}}(j) \mid \vec{\gamma}')} E(\theta_{\vec{\gamma}}(j_t) \mid \vec{\gamma}, y_{\vec{\gamma}}(j_t)) \tag{2.3}$$

is the Bayes estimate of $\theta(t)$ under squared error loss. We see that for a given partition, $\vec{\gamma}$, the Bayes estimator of $\theta_{\vec{\gamma}}(j_t)$ is just its posterior expectation given $y_{\vec{\gamma}}(j_t)$. Then one uses the marginal probabilities of data under each partition to average over all possible partitions to get the final estimate. We can now indicate how our noninformative Bayesian procedure will work.

Given the data y and a partition $\vec{\gamma}$, let $s_{\vec{\gamma}}(j)$ be the total number of y's equal to 1 on the jth interval. Given $\vec{\gamma}$, a natural estimate of $\theta_{\vec{\gamma}}(j)$ is just $s_{\vec{\gamma}}(j)/n_{\vec{\gamma}}(j)$. That is on each subinterval we use the maximum likelihood estimator (mle) to estimate $\theta_{\vec{\gamma}}(j)$. It remains to decide how the $p(y_{\vec{\gamma}}(j) \mid \vec{\gamma})$'s should be handled. Although for the binomial problem the the mle is not a Bayes estimator it does have a stepwise Bayesian justification. We will see that this fact will allow us to compute the $p(y_{\vec{\gamma}}(j) \mid \vec{\gamma})$'s in a sensible manner. Although the resulting procedure is not Bayes, it is stepwise Bayes and its admissibility will follow from this fact. In the next section we will briefly review the stepwise Bayes argument for proving the admissibility of the mle in the Binomial setup since it is crucial for what follows.

3. Admissibility for the binomial problem

We let W denote a binomial(n, θ) random variable with $\theta \in [0, 1]$. We wish to prove the admissibility of the mle, W/n, as an estimate of θ under squared error loss.

For an $\alpha > 0$ let π denote the Beta(α, α) prior distribution for θ. The resulting Bayes estimator is $\delta_\pi(w) = (w + \alpha)/(n + 2\alpha)$ and is admissible and approaches the mle as α approaches 0. This suggest that the mle might be Bayes against the improper prior $1/(\theta(1 - \theta))$. This is wrong because under this "prior" $E(\theta|w)$ is undefined when $w = 0$ and $w = n$. The stepwise Bayes argument gets around this problem by showing that the mle becomes Bayes in two steps or stages.

In the first step we consider the prior π_1 which puts mass $1/2$ at $\theta = 0$ and $\theta = 1$. Under π_1 the marginal probability function is given by

$$\begin{aligned} p(w; 1, n) &= 1/2 \qquad \text{for } w = 0, n \\ &= 0 \qquad \text{for } w = 1, 2, \ldots, n-1 \end{aligned} \tag{3.1}$$

Under this prior the only possible w values are 0 and n and the corresponding posterior expectations of θ are 0 and 1. Any estimator which behaves in this way is Bayes against π_1. Many estimators have this property including the mle but not all of them will be admissible because they may behave badly at other values of w. To identify admissible ones we need to proceed to the second step or stage.

Here we consider the restricted problem where the sample space just consists of the values $\{1, 2, \ldots, n-1\}$. For this restricted problem the probability function is just the renormalized binomial likelihood function. This likelihood function, the second stage prior we need and the corresponding marginal probability function are

$$\begin{aligned} p(w \mid \theta) &= \frac{\binom{n}{w}\theta^w(1-\theta)^{n-w}}{1 - \theta^n - (1-\theta)^n} \quad \text{for} \quad w = 1, \ldots, n-1, \\ p_2(\theta) &= c_n \frac{1 - \theta^n - (1-\theta)^n}{\theta(1-\theta)} \quad \text{for} \quad \theta \in (0, 1), \\ p(w; 2, n) &= c_n \frac{\Gamma(w)\Gamma(n-w)}{\Gamma(n)} \quad \text{for} \quad w = 1, 2, \ldots, n-1, \end{aligned} \tag{3.2}$$

where c_n is the constant that makes π_2 integrate to one. It is easy to check that under the resulting posterior, the posterior expectation of θ is just the mle for all

the values in the sample space of the restrict problem. This shows the the mle is the unique stepwise Bayes estimator against these two priors and hence is admissible. For more details see Ghosh and Meeden (1997). We note in passing that the stepwise Bayes method for proving admissibility was introduced in Johnson (1971).

Since for our problem, as described in the previous section, for a given partition $\vec{\gamma}$, we just have a collection of binomial problems we will be able to adapt the above admissibility argument to our more complicated situation. This will be done in the next section.

4. Admissibility for our problem

Consider the procedure which for a given y and given $\vec{\gamma}$ estimates $\theta_{\vec{\gamma}}(j)$ by $s_{\vec{\gamma}}(j)/n_{\vec{\gamma}}(j)$ and uses the probability model given in equation 3.1 to compute the factors in $\prod_{j=1}^{k+1} p(y_{\vec{\gamma}}(j) \mid \vec{\gamma}')$ whose $y_{\vec{\gamma}}(j)$ are either all 0's or all 1's and uses the probability model in (3.2) to compute the remaining factors. Finally, one must use (2.3) to average over all partitions to get the estimate. In other words we are letting the data select which probability model we want to use on each subinterval for each partition. Clearly this cannot be a Bayes procedure but we will now demonstrate that it is unique stepwise Bayes. Before we formally define our estimator we need some additional notation.

For $r = k+1, k, \ldots, 1, 0$ let

$$\Theta(m,k)(r) = \{(\vec{\gamma}, \theta_{\vec{\gamma}}) \,:\, \text{where exactly } r \text{ of the } \theta_{\vec{\gamma}}(j)\text{'s equal 0 or 1}\}.$$

For $r > 0$ members of this set must produce data y which contains at least m consecutive members which are either all 0's or all 1's. Formally, we denote such a subsequence as a **run**. When defining the estimator we need not consider all possible runs. Let i_b and i_e be the beginning and ending indices of a run. Remember $i_e - i_b > m - 1$. We say that a run is **permissible** if $i_b = 1$ or if $i_b > m$ and $i_e \leq n - m$ or if $i_b = n$. For a given y, let $r(y)$ be the maximum number of permissible runs that can be formed simultaneously in y. Note the possible values for $r(y)$ are $0, 1, \ldots, k+1$.

To help clarify this notion consider the special case where $n = 20$, $k = 2$ and $m = 4$. If y consists of 3 1's followed by 4 0's followed by 13 1's then $r(y) = 1$ since the only permissible runs are those that start at $i_2 = 9, 10, \ldots, 17$.

Next we define

$$\Gamma(y; m, k) = \{\vec{\gamma} \,:\, \vec{\gamma} \in \Gamma(m,k) \text{ and } r(y) \text{ is the number of permissible runs in the partition defined by } \vec{\gamma}\}.$$

This set can vary from a single member to all of $\Gamma(m,k)$. Given a $\vec{\gamma} \in \Gamma(y; m, k)$, we let

$$\Lambda(y, \vec{\gamma}) = \{\gamma_j \,:\, \gamma_j \text{ is not a run}\}.$$

Note this is the empty set when $r(y) = k+1$.

We now write

$$p(y \mid \vec{\gamma}) = (1/2)^{m(y)} \prod_{j \in \Lambda(y,\vec{\gamma})} c_{n_{\vec{\gamma}}(j)} \frac{\Gamma(s_{\vec{\gamma}}(j))\Gamma(n_{\vec{\gamma}}(j) - s_{\vec{\gamma}}(j))}{\Gamma(n_{\vec{\gamma}}(j))},$$

where when $\Lambda(y, \vec{\gamma})$ is empty the product is set to 1. If for fixed $\vec{\gamma} \in \Gamma(y; m, k)$ we use the probability model in equation 3.2 for the $\gamma_j \in \Lambda(y, \vec{\gamma})$ and the probability model

in equation 3.1 for the $\gamma_j \notin \Lambda(y, \vec{\gamma})$ then the above is the conditional probability of y given $\vec{\gamma}$.

If we put the uniform prior distribution over the set $\Gamma(y; m, k)$ then we can formally express our stepwise Bayes estimator as

$$\delta_{sb}(y) = \sum_{\vec{\gamma} \in \Gamma(y;m,k)} \frac{p(y \mid \vec{\gamma})}{\sum_{\vec{\gamma}' \in \Gamma(y;m,k)} p(y \mid \vec{\gamma}')} \hat{\theta}_{\vec{\gamma}}(j_t), \tag{4.1}$$

where $\hat{\theta}_{\vec{\gamma}}(j_t) = s_{\vec{\gamma}}(j_t)/n_{\vec{\gamma}}(j_t)$ is the proportion of 1's in the member of the partition which contains t.

We will now prove the admissibility of δ_{sb} in the special case when $k = 2$. Once this is understood the more general argument should be clear.

In the first stage of the proof we only put equal mass on the points belonging to the set $\Theta(m, 2)(3)$. In this case the only y's we can see are those are those consisting of 3 runs, each at least as long as m.

For example, suppose we observe a run of 1's followed by a run of 0's followed by another run of 1's. There can only be one member of $\Theta(m, 2)(3)$ which produces such a y so our model assigns a posterior probability of 1 to this point.

Another possibility is a run of at least $2m$ 1's followed by a run of at least m 0's. If the run of 1's is longer than $2m$ than there can be more than one member of $\Theta(m, 2)(3)$ which is consistent with the data. But each of them will give the same estimate of $\theta(t)$.

Another possibility is that y is either all 0's or all 1's. Again $\Theta(m, 2)(3)$ contains more than one member which is consistent with the data, but all lead to the same estimate.

For the second stage of the argument we consider $\Theta(m, 2)(2)$. Now the argument must proceed in several steps. The first step considers the subset $\Theta(m, 2)(2; 1, 2)$ where $\theta_{\vec{\gamma}}(1)$ and $\theta_{\vec{\gamma}}(2)$ are the two that are either 0 or 1. The next step assumes that the first and third are either 0 or 1 while the last step assumes that the last two are either 0 or 1. (Actually, these three steps can be taken in any order. On $\Theta(m, 2)(2; 1, 2)$ our prior distribution for $\vec{\gamma}$ is the uniform distribution and given $\vec{\gamma}$ the priors for the $\theta_{\vec{\gamma}}(j)$'s are independent. For the first two we use the prior from the probability model in (3.1) and for the last the prior in equation (3.2). We only need consider data points which can be generated from these parameter points which were not taken care of at the first stage.

One possibility is data that begins with a run of $v_1 \geq m$ 1's followed by a run of $v_2 \geq m$ 0's where $v_1 + v_2 < n$. If $v_2 = m$ or if $v_1 + v_2 = n - m$ then the only members of $\Theta(m, 2)(2)$ which could produce this data are $(\vec{\gamma}, \theta_{\vec{\gamma}}) = ((v_1, v_2), (1, 0, \theta_{\vec{\gamma}}(3)))$ where $0 < \theta_{\vec{\gamma}}(3) < 1$. For all other values of $v_1 \geq m$ and $v_2 \geq m$ there will be more than one choice for $\vec{\gamma}$ which could produce the data. To get the estimate one needs to average over all possible partitions.

Data which ends begins and ends in a run or ends in two runs is handled in the same manner. It is important to note that there exist data which has two runs but cannot arise in this stage.

For example consider the data which begins with $m - 1$ 0's followed by m 1's and ends with $n - 2m - 1$ 0's. There is no member of $\Theta(m, 2)(2)$ which gives positive probability to these data because each member of $\vec{\gamma}$ must contain at least m successive units.

For the third stage of the argument we consider $\Theta(m, 2)(1)$. As in the second stage we need to consider three possible steps depending on which $\theta_{\vec{\gamma}}(j)$ is either 0 or 1. Here we only have to consider data which contains one possible run. Just

as in the previous stages for some data there will only be one possible choice for $\vec{\gamma}$ while in other case there can be many. For each possible $\vec{\gamma}$ we will be using the probability models in equations (3.1) and (3.2) as needed.

In the fourth and final stage we consider $\Theta(m,2)(0)$ where none of the $\theta_{\vec{\gamma}}(j)$'s equal 0 or 1. Again we will use the uniform distribution over $\vec{\gamma}$ but we take independent versions of the prior from the probability model in equation 3.1 for the $\theta_{\vec{\gamma}}(j)$'s.

This completes the proof in the special case when $k = 2$. The proof for the general case just has $k+2$ stages. At each stage we keep reducing the the number of $\theta_{\vec{\gamma}}(j)$'s which equal 0 or 1 by 1 until we reach the final stage where all the $\theta_{\vec{\gamma}}(j)$'s are strictly between 0 and 1.

Theorem 4.1. *Given (X,Y) let m, k be positive integers such that X can be partitioned into $k+1$ subsets each containing at least m members in more than one way. Let $0 < t_1 < t_2 < \cdots < t_r < 1$ be fixed. Then under the sum of squared error loss the unique stepwise Bayes estimators of $(\theta(t_1), \theta(t_2), \ldots, \theta(t_r))$ are admissible.*

Proof. In the proof given above when $r = 1$ we assumed that t_1 was equal to one of the x_i's. This was done only for notational convenience. If t_1 is not equal to one of the x_i's then for some partitions $\vec{\gamma}$ we will have $t_1 \in \gamma_j$ for some j while for other partitions it will fall into one of the gaps. In either case for any partition $\vec{\gamma}$ it is uniquely defined and hence its average over all possible partitions is also uniquely defined and the result follows. Note the general case when $r > 1$ follows directly. □

It is not surprising that the stepwise Bayes technique can be used to prove admissibility in a nonparametric setting. Using this method the admissibility of several common nonparametric estimators was demonstrated in Meeden et al. (1985).

As we noted earlier values for m and k must be chosen by the statistician. In many applications where we do not expect to see long runs, the value for m does not matter much as long as it is big enough. It is true however that choosing a small value for m will generally result in a rougher estimate. Selecting a good value for k can be a more sensitive issue. We will say more about this in the next section.

For many problems we would not expect to see many long runs. This means in practice most samples would belong to those considered in the last couple of stages of the proof. The situations described in the earlier stages would rarely come into play.

5. Some examples

Thus far, we have discussed the construction and theoretical properties of the estimator δ_{sb}. For a concrete example, we consider a dataset drawn from a trial conducted by Acorn Cardiovascular (St. Paul, Minn.) investigating the performance of a therapy for congestive heart failure. Subjects in this trial were assigned randomly to optimal medical therapy or to implant of an investigational device, with 148 patients randomized to the device arm of the trial.

Among demographic data collected during the trial were a measure of heart size called left ventricular end diastolic dimension (LVEDD), measured to the nearest millimeter; a subset of these patients were then designated for further investigation based upon the ratio of their baseline LVEDD to body surface area (BSA), the latter measured in square meters. Membership in the subset of interest was declared for subjects with LVEDD/BSA between 30 and 40 mm/m^2; this constitutes the

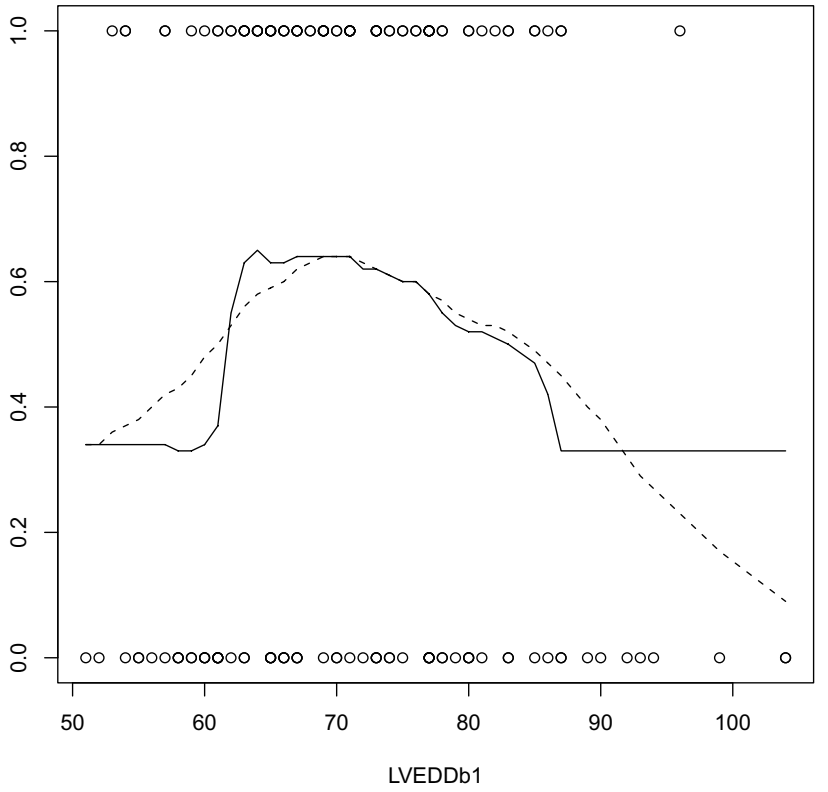

FIG 1. *For the LVEDD data the solid line is the plot of the stepwise Bayes estimator and the dotted line is the plot of the kernel estimator.*

response variable in our dataset, with subjects in the subset coded as a response of 1 and others as 0.

We make the following observations about the dataset at hand. A continuous parameter function is to be anticipated, with the probability of membership initially increasing, leveling off and then decreasing again as LVEDD increases. Since a typical BSA is around 2 mm/m^2, a sharp rise in the proportion of members might be expected around an LVEDD of 60 mm and a decrease around 80 mm. We can then consider the problem of estimating the probability of being a member of the subset conditional upon baseline LVEDD.

Results using both the stepwise Bayes estimator δ_{sb} and kernel estimation (for which we take the so-called *triangular* kernel function) are displayed in Figure 1, using $k = 2$ and $m = 10$ for δ_{sb} and a bandwidth b of 0.2 for the kernel method. Note that δ_{sb} captures a sudden rise in the probability of subset membership around an LVEDD of 60 mm, as anticipated, and a gentler decrease around an LVEDD of 80 mm. Comparatively, the kernel method fails to spot the sharp rise, and the decrease is observed but more gently. Note also that the kernel estimator is highly sensitive to a small number of data points on the high end of the LVEDD domain, while δ_{sb} is flat over that same region due to the selection of m as 10. Although not plotted here δ_{sb} for $m = 10$ and $k = 3$ yields a very similar picture. If for the kernel estimator one decreases the bandwidth b to 0.1 or 0.05 the resulting estimators become very jagged and unrealistic and do not capture the up tick found by δ_{sb}.

TABLE 1
MRMSE for θ, n and k for the stepwise Bayes method

		$n = 50$	$n = 100$	$n = 200$
θ_1	$k = 1$	0.108	0.089	0.076
	$k = 2$	0.125	0.087	0.068
	$k = 3$	0.148	0.094	0.070
θ_2	$k = 1$	0.119	0.088	0.064
	$k = 2$	0.135	0.095	0.071
	$k = 3$	0.160	0.108	0.079

TABLE 2
MRMSE for θ, n and b for the kernel method

		$n = 50$	$n = 100$	$n = 200$
θ_1	$b = 0.2$	0.126	0.089	0.062
	$b = 0.3$	0.103	0.075	0.055
	$b = 0.4$	0.092	0.068	0.053
	$b = 0.5$	0.087	0.070	0.054
	$b = 0.6$	0.088	0.074	0.060
θ_2	$b = 0.2$	0.134	0.108	0.090
	$b = 0.3$	0.128	0.110	0.099
	$b = 0.4$	0.132	0.116	0.109
	$b = 0.5$	0.136	0.127	0.121
	$b = 0.6$	0.147	0.141	0.136

For a different sort of example, we compare the error performance of the stepwise Bayes method to kernel estimation for selected hypothesized true parameter functions $\theta : [0, 1] \to [0, 1]$ and various sample sizes, focusing on the mean error rates of each method alongside the impact of the kernel bandwidth b and the parameter k in the stepwise Bayes estimator.

Therefore, for a given dataset y we define the root-mean-squared error (RMSE) of δ_{sb} in the usual way. We then define the mean RMSE (MRMSE) of δ_{sb} as the mean of the various RMSE values for all possible realized vectors y, weighted by the probability function of y given a true parameter function θ. Analogous definitions are made for the kernel estimator, allowing us to assess the typical error of each method.

That is, given a particular true θ, we can compare kernel estimation to δ_{sb} for various values of b and k. Table 1 gives values of MRMSE for δ_{sb} for two possible θ. θ_1 is linear over $[0, 1]$ and ranges from 0.25 at $x = 0$ to 0.75 at $x = 1$. θ_2 takes the constant value of 0.25 from $x = 0$ to $x = 0.49$, the constant value of 0.75 from $x = 0.51$ to $x = 1$ and linear interpolates between $x = 0.49$ and $x = 0.51$

We apply three different sample sizes n (50, 100 and 200) and k equal to 1, 2 and 3 for δ_{sb} in Table 1, while Table 2 provides corresponding values for the kernel estimator with the same three sizes of n and various bandwidths b. For δ_{sb} we took $m = 12$ when $n = 50$, $m = 18$ when $n = 100$, and $m = 24$ when $n = 200$. (Note that although MRMSE can theoretically be computed directly since for a given problem the set of possible vectors y is finite, we use simulation to accomplish this for practical reasons.)

Naturally, MRMSE decreases as n increases for both methods; under kernel estimation, a larger n tends to correspond to a smaller optimal b, while for the stepwise Bayes technique the optimal k increases with n. For the linear parameter function θ_1, kernel estimation tends to outperform the stepwise Bayes estimator, but the reverse is true for the spiky θ_2.

This is consistent with the heuristic concepts behind both methods: kernel estimation uses a fixed portion of the available data for averaging, while δ_{sb} essentially seeks to identify points of change in $\theta : D \to [0, 1]$ and to average over such points. Hence, one would expect the stepwise Bayes method to be more successful when dealing with parameter functions such as θ_2; this is precisely what is observed in this example.

6. Final remarks

Here we have considered an objective, nonparametric Bayesian method for estimation a binary regression function. The only prior information assumed is that the response function can be well approximated by a mixture of step functions. The method depends on specifying two parameters m and k. The first is less important and gives a lower bound for the minimum length of the intervals defining an approximating step function. The second is the number of different values an approximating step function can take on. A good choice for it depends on the sample size and the smoothness of the regression function being estimated. But in the examples we considered our method seems to be more robust against this choice than kernel estimators are for the choice of band width. The method should work well for problems where the response function changes rapidly over some small interval of values for the independent variable.

References

Brown, L. (1981). A complete class theorem for statistical problems with finite sample spaces. *Annals of Statistics*, 9:1289–1300.

Coram, M. and Lalley, S. (2006). Consistency of Bayes estimators of a binary regression function. *Annals of Statistics*, 34:1233–1269.

Ghosh, M. and Meeden, G. (1997). *Bayesian Methods for Finite Population Sampling*. Chapman and Hall, London.

Johnson, M. (1971). On admissible estimators for certain fixed sample binomal problems. *Annals of Mathematical Statistics*, 42:1579–1587.

Mallick, B. and Gelfand, A. (1994). Generalized linear models with unknown link functions. *Biometrika*, 81:237–245.

Meeden, G., Ghosh, M., and Vardeman, S. (1985). Some admissible nonparametric and related finite population sampling estimators. *Annals of Statistics*, 17:811–817.

Newton, M., Czado, C., and Chappell, R. (1996). Bayesian inference for semiparametric binary regression. *Journal of the American Statistical Association*, 91:142–153.

Signorini, D. and Jones, M. (2004). Kernel estimators for univariate binary regression. *Journal of the American Statistical Association*, 99:119–126.

Wood, S. and Kohn, R. (1998). A bayesian approach to robust binary nonparametric regression. *Journal of the American Statistical Association*, 93:203–213.