Non parametric mixture priors based on an exponential random scheme[*]

Sonia Petrone and Piero Veronese [†]

April 9, 2001

Abstract

We propose a general procedure for constructing nonparametric priors for Bayesian inference. Under very general assumptions, the proposed prior selects absolutely continuous distribution functions, hence it can be useful with continuous data. We use the notion of *Feller-type approximation*, with a random scheme based on the natural exponential family, in order to construct a large class of distribution functions. We show how one can assign a probability to such a class and discuss the main properties of the proposed prior, named *Feller prior*. Feller priors are related to mixture models with unknown number of components or, more generally, to mixtures with unknown weight distribution. Two illustrations relative to the estimation of a density and of a mixing distribution are carried out with respect to well known data-set in order to evaluate the performance of our procedure. Computations are performed using a modified version of an MCMC algorithm which is briefly described.

Summary

Keywords: Bernstein Polynomials, density estimation, Feller operators, Hierarchical models, Mixture Models, Non-parametric Bayesian Inference.

1 Introduction

In many applications the researcher has not enough information for specifying a parametric model for the random mechanism generating the data, thus a semiparametric or a nonparametric approach seems more appropriate. For example,

[*] *Acknowledgments:* S.P. was partially supported by grants from MURST and Università dell'Insubria (FAR 2000) , while P.V. by grants from CNR (99.01688CT10) and L. Bocconi University.

[†] Sonia Petrone, Università dell'Insubria, Dipartimento di Economia, Via Ravasi 2, 21100, Varese, Italy. Email: spetrone@mail.uninsubria.it and Piero Veronese, Università L. Bocconi, Milano, Istituto di Metodi Quantitativi, viale Isonzo 25, 20135 Milano, Italy; E-mail piero.veronese@uni-bocconi.it

with heterogeneous data it might be reasonable to use a model where the number of modes is not fixed.

In a Bayesian approach, a nonparametric model requires to specify a prior on the family of all the distributions on the sample space. Many nonparametric priors proposed in the literature are based on a partition of the sample space, such as the Ferguson-Dirichlet process or Polya trees; see Cifarelli et al.(1999) for a recent review.

An alternative way of defining a nonparametric prior is using mixture models, i.e. to model the unknown distribution of the data as a mixture of parametric kernel distributions. Early proposals in this sense are due to Ferguson (1983) and Lo (1984). However, in this context, the literature is mainly focused on applications in density estimation and regression, without a careful analysis of the theoretical properties of a nonparametric "mixture prior". Indeed this would require a preliminary study of the family $\mathcal{F}$ of mixture models that is used. For the prior to be "nonparametric", it is necessary that the family $\mathcal{F}$ is "large". Roughly speaking, this means that any distribution function H on the sample space can be approximated, in some sense, by a sequence of distributions in $\mathcal{F}$. Different approximation properties can lead to different properties of the mixture prior.

For example, if any H can be weakly approximated by distributions in $\mathcal{F}$, then we can expect that the prior has full weak support. If the approximation holds in a stronger sense, we can expect stronger properties of the prior, such as consistency. A clear illustration of this connection can be seen in the Bernstein prior (Petrone, 1999), defined for exchangeable data in $[0, 1]$. Informally, a Bernstein prior selects a finite mixture of beta densities with a random number of components. The fact that any distribution function on $[0, 1]$ can be weakly approximated by mixtures of beta distributions is used for showing that the Bernstein prior has full weak support. For bounded continuous densities, the approximation by beta mixtures is uniform. Petrone and Wasserman (2001) use this stronger property for proving weak consistency of the Bernstein prior.

The aim of this paper is first to discuss a general scheme for defining a "large" family of probability distributions on a real set and to show the approximation properties of such a class (section 3). The presentation will be informal and we refer to Petrone and Veronese (2001) for more details.

The proposed scheme is based on Feller operators (section 2). These are a generalization of Bernstein polynomials. Even if extensions to the multivariate case are possible, here we do not deal with them, and we shall restrict our attention to the univariate case. An attractive of the proposed procedure is that it is fairly simple, since in most cases it leads to absolutely continuous distribution functions, with density given by a (infinite or finite) mixture of given kernel densities.

In section 4 we show how to construct a nonparametric prior based on Feller operators. Because of their approximation properties we conjecture that the proposed prior has good theoretical properties. In section 6 we present two applications in order to illustrate the satisfying performance of the *Feller prior* in Bayesian nonparametric inference. In particular, we consider a density estimation problem (section 6.1) and parameters and mixing distribution estimation in a problem related to combining different experiments (section 6.2). We use two well studied data sets for sake of comparison. An MCMC algorithms for simulating from the posterior distribution is provided in section 5.

2 Feller-type approximations.

Feller (1971, chapter VII) defines a constructive way of approximating a given bounded and continuous function on a (bounded or unbounded) interval $E \subseteq \Re$. The idea is as follows. Let $U : E \to \Re$ be a bounded continuous function, and let us consider a family of random variables $\{Z_{k,x}, x \in E, k = 1, 2, \ldots\}$ such that $E(Z_{k,x}) = x$ and $Var(Z_{k,x}) \to 0$ for $k \to \infty$. Then, for large k, $Z_{k,x}$ will be close to x and $U(Z_{k,x})$ will be close to $U(x)$.

More precisely, we call *Feller operator* of order k for U the function $B(x; k, U) = E(U(Z_{k,x}))$. The following theorem is a slight generalization of Feller's result.

Theorem 1 *Suppose that U is bounded on E and that $\{Z_{k,x}, x \in E, k = 1, 2, \ldots\}$ is such that, for $k \to \infty$, $E(Z_{k,x}) \to x$ and $Var(Z_{k,x}) \to 0$, for each $x \in E$. Then*

$$\lim_{k \to \infty} B(x; k, U) = U(x)$$

at any continuity point x of U. If U is continuous, the convergence is uniform in every closed interval in which $E(Z_{k,x}) \to x$ and $Var(Z_{k,x}) \to 0$ uniformly.

The sequence of random variables $\{Z_{k,x}, x \in E, k = 1, 2, \ldots\}$ is called a *random scheme* for the approximation (Altomare e Campiti, 1994, page 283).

One example of Feller operators, where $kZ_{k,x}$ has a binomial distribution with parameters (k, x), are Bernstein polynomials. A more general example is obtained by considering a sequence of independent and identically distributed (i.i.d.) random variables $\{Y_1, Y_2, \ldots\}$, with expected value $E(Y_i) = x$ and finite variance. In this case, the mean $Z_{k,x} = \frac{1}{k}\sum_{i=1}^{k} Y_i$ satisfies the required properties, namely $E(Z_{k,x}) = x$ and $V(Z_{k,x}) \to 0$, for $k \to \infty$.

3 An exponential family random scheme.

In this section we consider a random scheme where $Z_{k,x}$ is the mean of i.i.d. random variables with common distribution belonging to the natural exponential family. In particular, we focus on the case where U is a real distribution function. This situation is studied in Petrone and Veronese (2001), to which we refer for a brief review on the exponential family, and for an extended discussion of the results of this section.

Let

$$p_\theta(y) = \exp(\theta y - M(\theta)) \tag{1}$$

be a density with respect to a σ-finite measure ν, where $M(\theta) = \ln \int \exp(\theta x)\nu(dx)$ and $\theta \in \Theta = \{\theta : M(\theta) < +\infty\}$. The family of probability measures which admit a density of the form (1) is called Natural Exponential Family (NEF). In the sequel we assume that ν is absolutely continuous with respect to Lebesgue or counting measure.

From well known results on the exponential family, we have that $\mu \equiv E_\theta(X) = M'(\theta)$, where M' denotes the first derivative of the function M. Furthermore, the family can be parametrized in terms of the mean parameter μ and $V(\mu) \equiv Var_\mu(X) = M''(\theta(\mu))$, where M'' is the second derivative of the M. $V(\mu)$ is called *variance function*.

Now let E be the convex hull of the support (briefly, the convex support) of a distribution function U, $E^o \equiv (a, b)$ be the interior of E and take x in E^o. In order to construct the random scheme for the Feller-type approximation of U, let us consider a sequence of i.i.d. variables $\{Y_1, Y_2, \ldots\}$ from a density of the form (1) such that $M'(\theta) = x$. Then, the probability law of the mean $Z_{k,x} = \frac{1}{k}\sum_{i=1}^k Y_i$ still belongs to the exponential family, and we have $E(Z_{k,x}) = x$ and $Var(Z_{k,x}) \to 0$ for $k \to \infty$.

Let $F_{k,x}$ be the distribution function of $Z_{k,x}$, parametrized by the mean parameter x and with variance function $V(x)$ and define

$$B(x; k, U) = \begin{cases} 0 & x < a \\ U(a) & x = a \\ \int_E U(z) dF_{k,x}(z) & a < x < b \\ 1 & x \geq b. \end{cases} \tag{2}$$

The following proposition summarizes some results proved in Petrone and Veronese (2001).

Proposition 1 *If U is a distribution function with convex support E, then*

(i) $B(\cdot; k, U)$ defined in (2) is still a distribution function, with support E;

(ii) $B(\cdot; k, U)$ is a mixture of a distribution function degenerate on $\{a\}$ (if a is finite), with weight $U(a)$, and an absolutely continuous distribution function, with density

$$b(x; k, U) = \int_{E^o} g_k(x; z) dU(z), \quad x \in (a, b)$$

where $g_k(x; z) = \int_{[z,\infty)} (t - x) dF_{k,x}(t)/V(x)$.

(iii) For each $z \in E^o$, $g_k(\cdot; z)$ is a density with respect to Lebesgue measure with support E.

(iv) For each $x \in E^o$, $g_k(x, \cdot)$ is a density with respect to Lebesgue measure, with support E, whose expected value converges to x and whose variance converges to zero as $k \to \infty$.

From (i) of Proposition 1, we call $B(\cdot; k, U)$ *Feller distribution function* with parameters (k, U). Point (ii) shows that, if $U(a) = 0$, $B(\cdot; k, U)$ is an absolutely continuous distribution function. Results (ii) and (iii) show that the *Feller density* $b(x; k, U)$ is a mixture of densities $g_k(\cdot; z)$, with mixing distribution U.

From point (iv), if U is absolutely continuous with bounded density u, then $b(x; k, U)$ can be written as $E(u(\tilde{Z}_{k,x}))$, where $\tilde{Z}_{k,x}$ has density $g_k(x, \cdot)$. Furthermore, $\{\tilde{Z}_{k,x}, k = 1, 2 \ldots, x \in E^o\}$ is a random scheme and $b(\cdot; k, U)$ is the Feller operator which approximates u. Therefore, for the previous considerations and for theorem 1, we have the following

Proposition 2 (Approximation property). *Any distribution function U with convex support E can be weakly approximated by Feller distribution functions $B(\cdot; k, U)$.*

Furthermore, if U is absolutely continuous with a bounded and continuous density u, then u can be pointwise approximated by the Feller densities $b(\cdot; k, U)$.

In some relevant cases, $g_k(\cdot; z)$ has the form of a conjugate density for $p_\theta(x)$. In particular, the examples below show that z has the role of a location parameter and k has the role of a scale, or smoothing, parameter. Thus our construction is related to results by Dalal and Hall (1983) e Diaconis and Ylvisaker (1985), who use mixtures of conjugate priors for approximating an arbitrary prior density.

Furthermore our general scheme includes connections with the nonparametric techniques based on kernel functions, or more generally with delta sequences (Prakasa Rao, 1983, p. 136).

For illustration, we report some examples which are discussed in details in Petrone and Veronese (2001).

Example 1. (Bernstein polynomials). Suppose we have to approximate a distribution function U describing data in $[0, 1]$. Then a *binomial random scheme* can be used, where $kZ_{k,x}$ has a binomial distribution with parameters (k, x). In this case, $B(\cdot; k, U)$ is the Bernstein polynomial of order k for U. The kernel density $g_k(\cdot; z)$ is a beta density $\beta(\cdot; [kz], k - [kz] + 1)$, where $[a]$ denotes the integer part of a and $z \in [0, 1]$. This is a conjugate prior for the binomial model. Note that $g_k(\cdot; z)$ is a piecewise constant function of z. Consequently, the density $b(x; k, U)$ is a finite mixture of beta densities

$$b(x; k, U) = \sum_{j=1}^{k} w_{j,k} \beta(x; j, k - j + 1),$$

where $w_{j,k} = U(j/k) - U((j-1)/k), j = 1, \ldots, k$. Clearly, k has the role of smoothing parameter and, for a given k, j has the role of a translation parameter. The Bernstein density $b(\cdot; k, U)$ has been used as a flexible statistical model for data in $[0, 1]$ (for example, Tenbusch, 1995; Petrone, 1999).

Example 2. (Mixtures of gammas). For data in $[0, \infty)$, we can use a *Poisson random scheme*, where $kZ_{k,x}$ has a Poisson distribution with mean parameter kx. Then the kernel density $g_k(\cdot; z)$ is gamma, with shape parameter $[kz]$ and mean $[kz]/k$; in symbols, $Ga(\cdot; [kz], k)$. The gamma is the conjugate prior for the Poisson model. The Feller density is a countable mixture of gamma densities,

$$b(x; k, U) = \sum_{j=1}^{k} w_{j,k} Ga(x; j, k),$$

where $w_{j,k} = U(j/k) - U((j-1)/k), j = 1, 2, \ldots$.

Example 3. (Mixture of inverse gammas). For data in the open interval $(0, \infty)$, a *Gamma random scheme* can be used, where $Z_{k,x} \sim Ga(k, k/x)$. The kernel density

$g_k(\cdot;z)$ is an inverse-gamma with parameters (k,kz), denoted by In-Ga$(\cdot;k,kz), k > 0$. (We recall that, if $X \sim Ga(\cdot;a,b)$, then $1/X \sim$ In-Ga$(\cdot,a,b)$).

The Feller density is a continuous mixture of inverse gamma densities,

$$b(x;k,U) = \int \text{In-Ga}(x;k,kz)dU(x).$$

An application of this model is given in section 6.1.

Example 4. (Mixture of normals). For data in the real line, a *Gaussian random scheme* can be used, where $Z_{k,x}$ has a normal distribution $N(\cdot;x;\sigma^2/k)$ with σ^2 known. The kernel density $g_k(\cdot;z)$ is $N(\cdot;z,\sigma^2/k)$, so that the Feller density is a continuous mixture of normal densities

$$b(x;k,U) = \int N(x;z,\sigma^2/k)dU(z).$$

Mixture of normals are used in many contexts; a recent application in Bayesian nonparametrics is by Ghosal, Ghosh, and Ramamoorthi (1999).

4 Feller priors for Bayesian nonparametric inference.

Because of their approximation properties, Feller distributions with an exponential family random scheme can be used as flexible models for inference on an unknown distribution function. As shown in the previous section, a Feller distribution is a mixture of given kernel densities. Let us consider the smoothing parameter of the kernel density and the mixing distribution as random quantities and assign them a prior distribution. This induces a prior on the space of all the distributions on the sample space, which we will call *Feller prior.* As suggested by the examples of section 3, many popular mixture models can be interpreted as particular cases of a Feller prior. We describe now, in an informal way, the procedure for constructing a Feller prior for Bayesian nonparametric inference.

Let $\{X_1, X_2, \ldots\}$ be a sequence of exchangeable random variables with values in a set $E \subset \Re$. Then, for de Finetti's representation theorem, the n-dimensional distribution function can be written as

$$P(X_1 \leq x_1, \ldots, X_n \leq x_n) = \int_{\mathcal{P}} \prod_{i=1}^{n} F(x_i) d\pi(F), \tag{3}$$

where π is the prior on the class $\mathcal{P}$ of all the distribution functions on E.

We say that a prior *is supported* on a class $\mathcal{F}$ if it selects distribution functions in $\mathcal{F}$ with probability one. A *Feller prior* with random scheme $\{Z_{k,x}, x \in E^o, k = 1, 2, \ldots\}$ is a prior π supported on the class of Feller distribution functions

$$\mathcal{F} = \{B(\cdot;k,U) : k = 1,2,\ldots;\ U \in \mathcal{U}\},$$

where $\mathcal{U}$ is the class of all the distribution function with convex support E and B has the random scheme $\{Z_{k,x}, x \in E^o, k = 1, 2, \ldots\}$.

One way of constructing a prior π on $\mathcal{F}$ is to regard (k, U) as random quantities and to assign them a probability law $\mathbf{P}$. This induces a probability law π on the operator $B(\cdot; k, U)$. In particular, we will focus on the case when $\mathbf{P}$ is such that k and U are independent, k has probability function p and U is a Dirichlet process with parameters $\alpha, U_0(\cdot)$, in symbols $U \sim \mathcal{D}(\alpha, U_0)$, where α is the scale parameter, and U_0 is a distribution function expressing the prior guess on U. In this case, we say that π is a *Feller-Dirichlet* prior with parameters (p, α, U_0).

In the following, we will assume that $\mathbf{P}(U(a) = 0) = 1$, so that the random distribution function $B(\cdot; k, U)$ is almost surely absolutely continuous, with density $b(\cdot; k, U)$. In this case, choosing a Feller prior with random scheme $\{Z_{k,x}, x \in E^o, k = 1, 2, \ldots\}$ in (3) is equivalent to the following hierarchical model.

(i) For any n, $X_1, \ldots, X_n | k, U$ are conditionally i.i.d, with common density

$$b(x; k, U) = \int g_k(x; z) dU(z),$$

where $g_k(x; z)$ is the kernel density associated to the given random scheme.

(ii) (k, U) have joint probability law $\mathbf{P}$.

Conditionally on k, the Feller prior corresponds to a mixture model with a random mixing distribution, where the kernel density $g_k(\cdot; z)$ is suggested by the random scheme adopted in the Feller operator. The parameter k appears as a smoothing parameter in the kernel. This prior generalizes the Bernstein prior proposed by Petrone (1999) for data in $[0, 1]$

Computations in mixture models is in general analytically complicated. However, MCMC approximations of the posterior or of other quantities of interest are possible. In order to semplify computations, it is usually convenient to introduce auxiliary random variables $Y_1, Y_2, \ldots$ which have the role of "labels", and rewrite the model above as

(i) $X_1, \ldots, X_n \mid k, U, y_1, \ldots, y_n$ are independent, with joint density $\prod_{i=1}^n g_k(x_i; y_i)$;

(ii) $Y_1, \ldots, Y_n \mid k, U$ are i.i.d. according to U;

(iii) k and U have joint probability law $\mathbf{P}$.

Furthermore, the labels are useful for studying clusters in the population. In particular, for the Feller-Dirichlet prior, the common distribution U of the labels is almost surely discrete. This produces ties in the labels $Y_1, \ldots, Y_n$, which suggest the presence of clusters in the population. For example, if $Y_1 = \cdots = Y_m \neq Y_{m+1} = \cdots = Y_n$, we can think that there are two clusters, with the first m observations coming from one cluster, while the remaining from the other one. Therefore, the heterogeneity in the population can be studied by looking at the distribution of ties among $(Y_1, \ldots, Y_n)$. This aspect will be illustrated in example 6.1. The probability law on the cluster partition induced by the Dirichlet process is studied by Antoniak, 1974; see also Green and Richardson, 1999.

Sampling from the posterior of $(k, U, Y_1, \ldots, Y_n)$ from a Feller-Dirichlet prior can be obtained by MCMC. Since, as previously noted, conditionally on k our model is a mixture with a Dirichlet process mixing distribution, an MCMC algorithm can be constructed by using a Gibbs sampling from the conditional distribution given k and the data, and then adding a step for taking into account the randomness of k. In the literature, there are several algorithms that can be used for the first step. One class of algorithms works by integrating out U, then using a Gibbs sampling from the conditional law of $Y_1, \ldots, Y_n, k|x_1, \ldots, x_n$. Another class of algorithms is a based on a finite truncation of the Sethuraman's (1994) representation of the Dirichlet process as an infinite sum.

In the following section we present a *truncation MCMC algorithm* for the Feller-Dirichlet prior, which we use in section 6.1. A different computational strategy will be used for the hierarchical model of section 6.2.

5 Computational issues.

We give a brief description of an MCMC algorithm for simulating from the posterior corresponding to a Feller prior. The algorithm is based on the well known characterization of the Dirichlet process as an infinite sum (Sethuraman, 1994). If $U \sim \mathcal{D}(\alpha, U_0)$, then U is almost surely a discrete distribution function

$$U(x) = \sum_{j=1}^{\infty} p_j \delta_{(-\infty, Z_j]}(x) \;, \tag{4}$$

where $\delta_A(\cdot)$ is the indicator function of the set A, $(Z_1, Z_2, \ldots)$ are independent draws from U_0, $p_1 = V_1, p_j = (1 - V_1)(1 - V_2) \cdots (1 - V_{j-1})V_j, j \geq 2$, and $V_1, V_2, \ldots$ are i.i.d. according to a beta density with parameters $(1, \alpha)$. In the literature, there are several proposals of simulations algorithms which are based on a truncation

$$U_N(x) = \sum_{j=1}^{N} p_j \delta_{(-\infty, Z_j]}(x). \tag{5}$$

of the series (4) to a finite N, such that the residual probability is negligible; see for example Muliere and Tardella, 1998 and Ishwaran and Zarepour, 2000.

We extend, in particular, the algorithm proposed by Ishwaran and Zarepour for simulating from the posterior generated by a Feller prior.

A Gibbs sampling from the posterior of $(U_N, k, Y_1, \ldots, Y_n)$ can be obtained by iteratively drawing values from the conditional distributions of

(a) $(p_1, \ldots, p_N, Z_1, \ldots, Z_N)|k, j_1, \ldots, j_n, x_1, \ldots, x_n$,
where $(J_1, \ldots, J_N)$ are classifications variables that identify the Z_j associated to each Y_i, by $Y_i = Z_{J_i}$.

(b) $J_1, \ldots, J_n|k, p_1, \ldots, p_N, z_1, \ldots, z_N, x_1, \ldots, x_n$,

(c) $k|p_1, \ldots, p_N, z_1, \ldots, z_N, j_1, \ldots, j_n, x_1, \ldots, x_n$.

The full conditional corresponding to step (a) is proportional to

$$f(p_1,\ldots,p_N|k,j_1,\ldots,j_n)\prod_{i=1}^{n} u_0(z_i)g_k(x_i;z_{J_i}). \tag{6}$$

where f is the conditional probability function of $p_1,\ldots,p_N$ given $k,j_1,\ldots,j_n$ and u_0 is the density of U_0. For the Dirichlet process, $p_1,\ldots,p_N|k,j_1,\ldots,j_n$ have a generalized Dirichlet distribution with parameters $(a_1,b_1,\ldots,a_{N-1},b_{N-1})$, where $a_j=1+m_j$, $b_j=\alpha+m_{j+1}+\cdots+m_N$ and m_j is the number of $j_1,\ldots,j_n$ which are equal to j, $j=1,\ldots,N-1$.

Sampling $Z_1,\ldots,Z_N$ from (6) is straightforward if U_0 is the conjugate prior for the model $g_k(\cdot;z)$. This is the case of example 6.1 where we choose U_0 as the gamma prior for the inverse-gamma model.

Regarding the step (b), we have that $J_1,\ldots,J_n$ are conditionally independent, with

$$Y_i|k,p_1,\ldots,p_N,z_1,\ldots,z_N,x_1,\ldots,x_n\sim\sum_{j=1}^{N}p^*_{j,i}\delta_{(-\infty,z_j]}(\cdot),\ i=1,\ldots,n,$$

where $p^*_{j,i}\propto p_j g_k(x_i|z_j)$, $j=1,\ldots,N$.

Finally, the full conditional of k is proportional to

$$p(k)\prod_{i=1}^{n}g_k(x_i|y_i)$$

on $k=1,2,\ldots,K$. In fact, even if, in principle, k is unbounded, in practice it is truncated to a maximum value K. We shall discuss the choice of K in section 6.1.

Given an MCMC sample $\{(k^{(s)},U_N^{(s)},Y_1^{(s)},\ldots,Y_n^{(s)}),s=1,2,\ldots,S\}$ from the posterior of $k,U_N,Y_1,\ldots,Y_n$, we can approximate the density estimate $b_n(x)=E(b(x;k,U)|x_1,\ldots,x_n)$ by

$$\frac{1}{S}\sum_{s=1}^{S}b(x;k^{(s)},U_N^{(s)})=\frac{1}{S}\sum_{s=1}^{S}\sum_{j=1}^{N}p_j^{(s)}\,g_{k^{(s)}}(x;z_j^{(s)}) \tag{7}$$

The number of clusters in the population can be studied by the MCMC approximation of the posterior of $Y_1,\ldots,Y_n$. In particular, the empirical frequencies of the number of distinct values among $Y_1,\ldots,Y_n$ approximate the posterior distribution of the maximum number of clusters.

6 Applications.

6.1 Density estimation.

The Feller prior (with the restriction $U(a)=0$ almost surely) selects absolutely continuous distribution functions. Therefore, it can be appropriate for Bayesian inference on an unknown distribution in the case of continuous data. In particular, the Bayesian estimate of an unknown density, under a quadratic loss function, is given by the posterior expectation of the random density, which can be approximated by (7).

To illustrate the procedure, we reanalyse the galaxy data in Roeder (1990), representing the relative velocities of $n = 82$ galaxies from six well-separated conic sections of space. We choose this data set for sake of comparison. Among other authors, the data have been studied by Escobar and West (1995) using mixtures of normal densities, with a Dirichlet process prior for the mixing distribution of the mean and variance. Ishwaran and Zarepour (2000) use a similar model, but different computational strategies.

For these data, density estimation is of interest, as a tool for investigating the number of clusters in the galaxies.

Since the data are in $(0, \infty)$, we find appropriate to use a Feller prior with a Gamma random scheme. In example 3 of section 3, we showed that the corresponding Feller density is a mixture of inverse-gamma densities. In particular, we use a Feller-Dirichlet prior, with $\alpha = 2$ and $U_0 = Ga(20, 1)$. The gamma has been chosen because it is the conjugate prior for the inverse-gamma, while the specific values of the parameters guarantee that the probability mass is concentrated on a range of values which is reasonable for this kind of data.

The prior on k is uniform on $\{1, 2, \ldots, K\}$. In order to choose K, we suggest an empirical procedure. If K is too small, the posterior will tend to concentrate on K. This case is illustrated in Figure 1(a) where we plotted the sample values of k in the MCMC iterations, with $K = 500$. Notice that the samples often hit the boundary K. We can therefore increase K until this behavior disappears (see Figure 1(b), in which $K = 1000$.

We used the *truncation MCMC algorithm* described in the previous section, having fixed $N = 100$. Figure 2 represents the Feller-Dirichlet density estimate for the galaxy data. We see that there appear to be 5 or 6 distinct modes, according to the results in the literature. Figure 3 shows the empirical distribution of the number of distinct Y_i values, which gives information on the maximum number of clusters.

Sensitivity of the results to prior assumptions has been studied. In particular we observe that the choice of U_0 has influence on the posterior of k. A flat U_0 favours small values of k and consequently a smoother density estimate will be obtained.

6.2 Estimating a mixing distribution.

In this section we illustrate the use of a Feller prior at the second stage of a hierarchical model. Specifically we consider the problem of combining results from different binomial experiments in order to estimate the different mean parameters, $(\theta_1, \ldots, \theta_n)$, say. In frequentist analysis, Stein-kind shrinkage estimators are used in order to shrink the maximum likelihood estimates towards a lower dimensional space. In many applications, however, it is not clear whether to shrink towards the overall mean or towards different points, representing the means of possible clusters in the data. In this case it is preferable to combine different shrinkage estimators (George, 1986).

In a Bayesian approach, one could specify a hierarchical model taking into account different forms of partial exchangeability structures for $(\theta_1,\dots,\theta_n)$. Partition models, where a prior is explicitly given on the unknown partition, have been proposed by Consonni and Veronese (1995). However all the above procedures require to specify coefficients or hyperparameters with respect to which are quite sensitive. As a consequence, it seems reasonable to explore a nonparametric approach.

Let us assume that $(\theta_1,\dots,\theta_n)$ is a random sample from a distribution function H, and use a Feller-Dirichlet prior for H, with binomial random scheme, i.e. a Bernstein-Dirichlet prior. The motivation for these assumptions can be easily explained. The simplest choice for H would be the conjugate prior for the binomial model, which is a beta distribution. However, since we are uncertain about the presence of clusters in the experiments, it is more reasonable to model H as a possibly multimodal distribution. The Bernstein-Dirichlet prior reaches this aim since it is equivalent to model H as a mixture of beta densities of the kind $\sum_{j=1}^{k} w_{j,k}\beta(j,k-j+1)$, where the number of components k and the mixture weights $w_{j,k}$ are unknown (see example 1). Notice that one could alternatively use a mixture of the type $\sum_{j=1}^{k} w_{j,k}\beta(a_j,b_j)$, with a_j and b_j random. However our choice has the advantage of not requiring the estimates of the parameters a_j, b_j which can be troublesome for identifiability problems.

Adopting a Bernstein-Dirichlet prior with parameters (p,α,U_0), the model can be presented as follows.

i) $X_1,\dots,X_n|\theta_1,\dots,\theta_n,k,w_k \sim \prod_{i=1}^{n}\mathrm{bi}(x_i;m_i,\theta_i)$, where $\mathrm{bi}(\cdot;m_i,\theta_i)$ denotes a binomial probability function with m_i trials and success probability θ_i), and $w_k=(w_{1,k},\dots,w_{k,k})$;

ii) $\theta_1,\dots,\theta_n|k,w_k$ are i.i.d. according to a Bernstein density $b(\cdot;k,w_k)=\sum_{j=1}^{k} w_{j,k}\beta(\cdot;j,k-j+1)$;

iii) k and w_k are independent; $k\sim p(k)$ and $w_k \sim D(\alpha_{1,k},\dots,\alpha_{k,k})$, a Dirichlet distribution with parameters $\alpha_{j,k}=\alpha[U_0(j/k)-U_0((j-1)/k)], j=1,\dots,k$.

Notice that, conditionally on (k,U), $X_1,\dots,X_n$ are i.i.d. from a mixture of beta-binomial distributions, precisely

$$p(x|k,U)=\sum_{j=1}^{k} w_{j,k}f_{j,k}(x) \qquad (8)$$

where

$$f_{j,k}(x)=\binom{m_j}{x}\frac{\mathbf{B}(x+j,m_j-x+k-j+1)}{\mathbf{B}(j,k-j+1)}, \qquad (9)$$

and $\mathbf{B}$ is the beta special function.

In this context there are two canonical problems. The first is to estimate the mixing distribution H, the density $b(\cdot;k,w_k)$ in our case. The other is to estimate the binomial parameters $(\theta_1,\dots,\theta_n)$.

A natural Bayesian estimate of the mixing density is the predictive density. This is given by

$$\tilde{b}(\theta;k,w_k) \;=\; E(b(\theta;k,w_k)|k,x_1,\dots,x_n)$$

$$= \sum_{k=1}^{\infty} b(\theta; k, E(w_k|k, x_1, \ldots, x_n))p(k|x_1, \ldots, x_n) \tag{10}$$

where $E(w_k|k, x_1, \ldots, x_n)$ is the vector of the conditional expectations of $w_k|k, x_1, \ldots, x_n$ and $p(k|x_1, \ldots, x_n)$ is the posterior probability function of k.

The binomial parameter θ_i can be estimated by its posterior expectation, given by

$$\tilde{\theta}_i = E(\theta_i|x_1, \ldots, x_n) = \sum_{k=1}^{\infty} \left[\frac{m_i}{m_i + k + 1} \hat{\theta}_i + \frac{k+1}{m_i + k + 1} \sum_{j=1}^{k} \frac{j}{k+1} q_{j,k} \right] p(k|x_1, \ldots, x_n) \tag{11}$$

where $\hat{\theta}_j = x_j/m_j$ is the MLE,

$$q_{j,k} = E\left(\frac{w_{j,k} f_{j,k}(x_i)}{\sum_{t=1} w_{j,k} f_{t,k}(x_i)} |k, x_1, \ldots, x_n \right)$$

and $f_{j,k}(x)$ is the beta-binomial probability function (9).

Therefore, conditionally on k, the estimate $\tilde{\theta}_i$ is a weighted average between the MLE and a mixture of the prior guesses $j/(k+1)$ corresponding to the $\beta(\cdot; j, k-j+1)$ priors, for $j = 1, \ldots, k$. The weights in the mixture depend on the conditional distribution of $w_k|k, x_1, \ldots, x_n$. Roughly speaking, conditionally on k and $w_{j,k}$, the j^{th} beta component of the prior receives more weight if $w_{j,k}$ or $f_{j,k}(x_i)$ is large.

It is interesting to note that the structure of the estimate (11) recalls the one obtained by Consonni and Veronese (1995) using a partition model. In particular, the smoothing parameter k has a role somehow similar to that played by the random partition.

For simulating from the posterior distribution of (k, w_k) one can use the truncation MCMC algorithm presented in section 5.

Alternatively, since equation (8) shows that, conditionally on (k, U), $X_1, \ldots, X_n$ are i.i.d from a finite mixture of beta-binomial distributions, a reversible jump MCMC can be used. However, because the weights $w_{j,k}$ corresponding to different values of k are related, being the increments of the same distribution function U (see example 1), we prefer an algorithm which takes explicitly this fact into account. The MCMC algorithm developed, for Bernstein-Dirichlet prior, in Petrone (1999) can be used to this aim. Indeed the only modification required is the substitution of a beta distribution with a beta-binomial distribution.

Given an MCMC sample from the posterior of (k, w_k), the predictive density (10) and the estimates (11) of the binomial parameters can be approximated by their sample means.

Here we illustrate the procedure for a well studied data set, from George (1986). The problem consists in estimating the batting averages $(\theta_1, \ldots, \theta_{26})$ of all 26 major league baseball teams in USA, in the 1986 season, starting from the first 300 at bats. Since the remaining at bats in the seasons, for each team, is greater than 5000, these are adopted to evaluate the *true values* of θ. Consequently, we can compare different estimators by the corresponding mean-squared errors.

We choose a Bernstein prior with $\alpha = 2$, U_0 uniform on $[0, 1]$ and $p(k)$ uniform on $\{1, 2, \ldots, 1500\}$. These assumptions can be reasonably considered as non informative.

The estimates (11) of the binomial mean parameters $(\theta_1, \ldots, \theta_{26})$ are compared with alternative estimates through the mean square errors reported in Table 1. Our estimates behave fairly well, in comparison to estimates obtained by much more structurated models, such as George's multiple shrinkage estimates and Consonni and Veronese's hierarchical partition model estimates.

Table 1: Mean square errors for different estimators

Estimator	MSE
Maximum Likelihood	26.57
Multiple Shrinkage	from 4.4 to 5.37
Hierarchical Partition Model	from 5.18 to 9.05
Feller prior	5.41

Multiple Shrinkage estimators are taken from George (1986),
hierarchical partition models estimates are taken from Consonni-Veronese (1995)

Figure 4(a) shows an estimate of the mixing density, given by the the predictive density (9). This is unimodal, showing that the data are basically exchangeable. Therefore, the shrinkage of the MLEs is towards the overall sample mean. This is illustrated in Figure 4(b), where we plot our estimates against the maximum likelihood estimates.

References

Altomare, F. and Campiti, M. (1994), *Korovkin-type approximation theory and its application*, W. de Gruyter, Berlin.

Antoniak, C. E. (1974), Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems, *Ann. Statist.*, **2**, 1152-1174.

Cifarelli, D.M., Muliere, P. and Secchi, P. (1999), Prior processes for Bayesian nonparametrics,*Technical Report* n. **377/P**, Dip. Mat. F. Brioschi, Politecnico di Milano

Consonni, G. and Veronese, P. (1995), A Bayesian Method for Combining Results from Several Binomial Experiments, *J. Amer. Statist. Assoc.*, **90** ,935-944.

Dalal, S.R. and Hall, W. J. (1983), Approximating priors by mixtures of natural conjugate priors, *J. Roy. Statist. Soc. Ser. B*, **45**, 278-286.

Diaconis, P. and Ylvisaker, D. (1985), Quantifying prior opinion, *Bayesian statistics 2*, (J. M. Bernardo, M. H. deGroot, D. V. Lindley and A. F. M. Smith Eds), Elsevier Science Publishers B.V., North Holland, 133-156.

Escobar, M. D. and West, M. (1995), Bayesian density estimation and inference using mixtures, *J. Amer. Statist. Assoc.*, **90**, 577-587.

Feller, W. (1971), *An introduction to probability theory and its applications*, Vol. II, John Wiley, New York.

Ferguson, T. S. (1983), Bayesian density estimation by mixtures of normal distributions, *Recent advances in statistics*,(H. Rizvi and J. Rustagi eds), Academic Press, New York, 287-302.

George, E.I. (1986), Combining Minimax Shrinkage Estimators, *J. Amer. Statist. Assoc.*, **81**, 437-445.

Ghosal, S. Ghosh, J.K. and Ramamoorthi, R.V. (1999), Posterior consistency of Dirichlet mixtures in density estimation, *Ann. Statist.*, **27**, 143-158.

Green and Richardson, (1999), Modelling heterogeneity with and without the Dirichlet process,*Technical report*, University of Bristol.
http://www.stats.bris.ac.uk/ peter/.

Ishwaran, H. and Zarepour, M. (2000), Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models, *Biometrika*, **87**, 371-390.

Lo, A.Y. (1984), On a class of Bayesian nonparametric estimates, Density estimates, *Ann. Statist.*, **12**, 351-357.

Muliere, P. and Tardella, L. (1998), Approximating distributions of functionals of Ferguson-Dirichlet priors, *Can. J. Statist.*, **26**, 283-297.

Petrone S. (1999), Random Bernstein polynomials, *Scand. J. Statist.*, **26**, 373-393.

Petrone, S. and Veronese, P. (2001), Feller operators based on natural exponential families. *Studi Statistici*, n. **60**, Istituto Metodi Quantitativi, Università L. Bocconi, Milano.

Petrone, S. and Wasserman, L. (2001), Consistency of Bernstein posterior (under revision for *J. Roy. Stat. Soc.,* Ser.B).

Prakasa Rao, B.L.S. (1983), *Nonparametric Functional Estimation*, Academic Press, Orlando.

Roeder, K. (1990), Density estimation with confidence sets exemplified by superclusters and voids in the galaxies, *J. Amer. Statist. Assoc.* **85**, 617-624.

Sethuraman, J. (1994), A constructive definition of Dirichlet priors, *Statist. Sinica*, **4**, 639-650.

Tenbusch, A. (1995), *Nonparametric curve estimation with Bernstein estimates*, Universitätsverlag Rasch, Osnabrück, Germany.

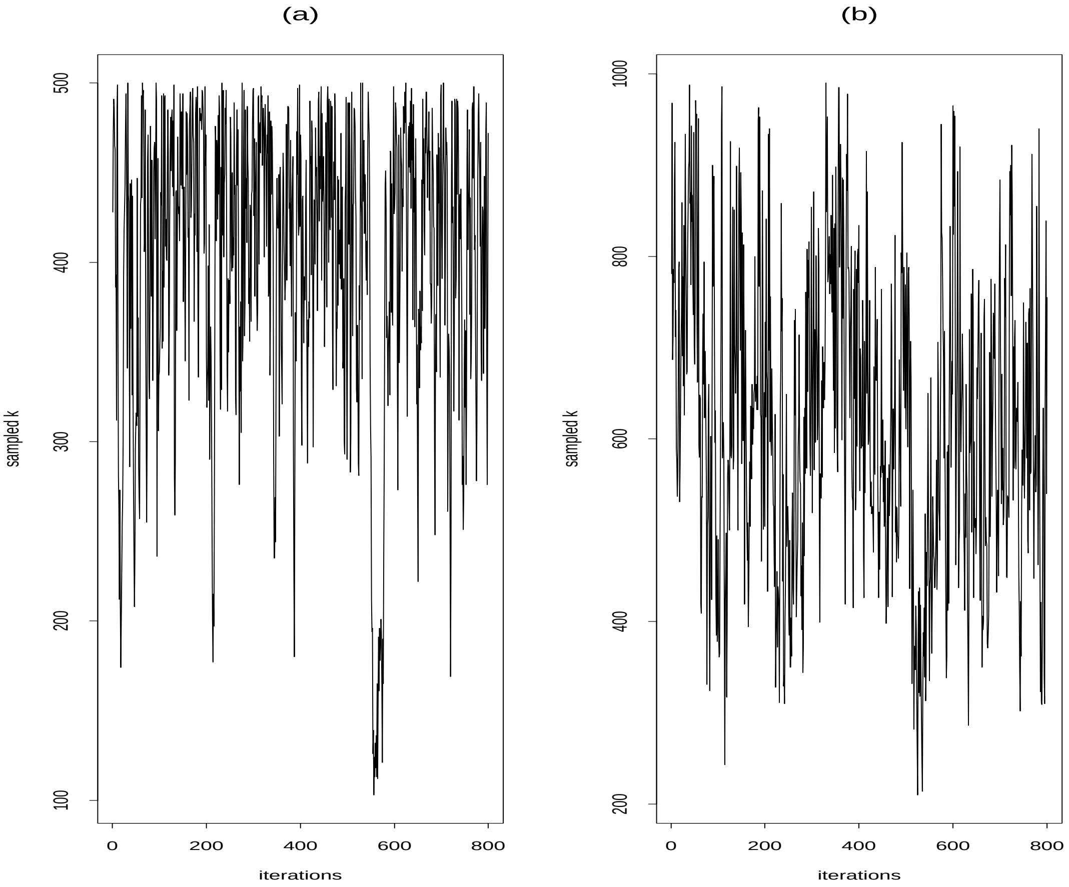

Figure 1: MCMC samples from the posterior of k, with $K = 500$ and with $K = 1000$.

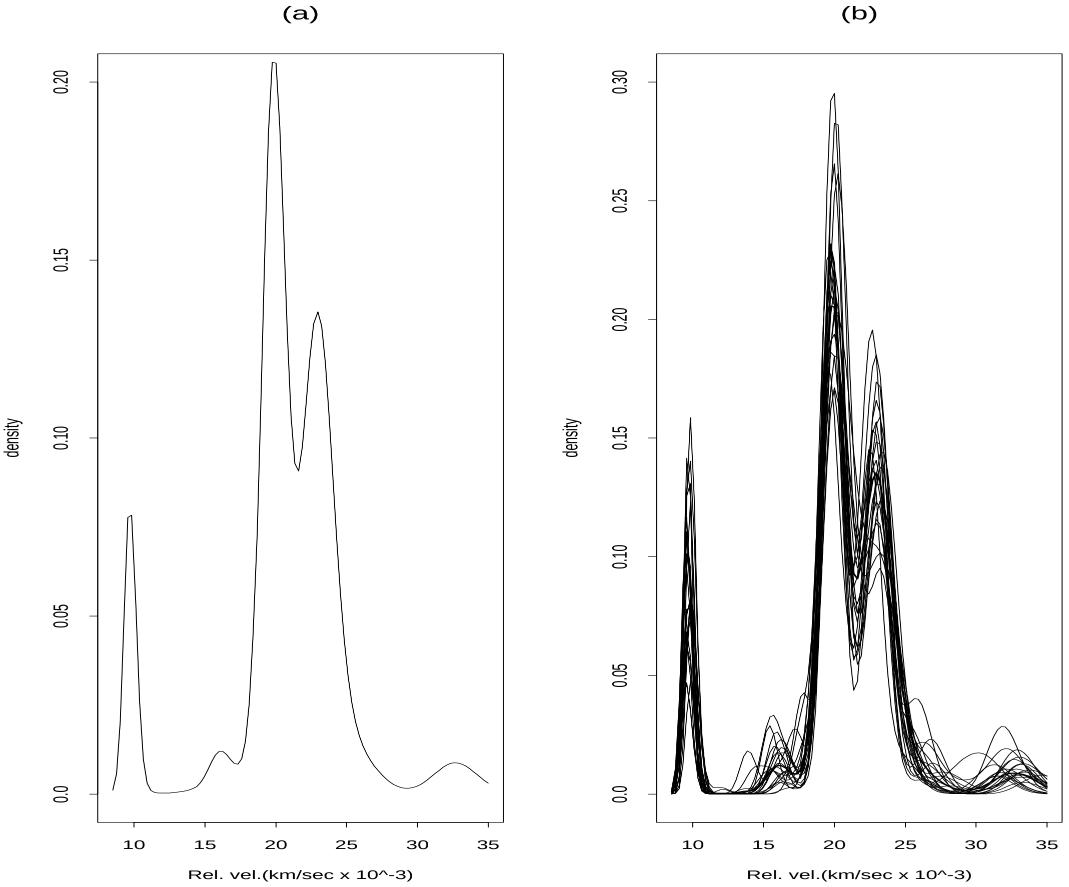

Figure 2: Density estimate for relative velocities in thousands of kilometers/second for 82 galaxies (Roeder,1990). (a) MCMC approximation of the predictive density. (b) Twenty densities, randomly selected from the posterior of the random density.

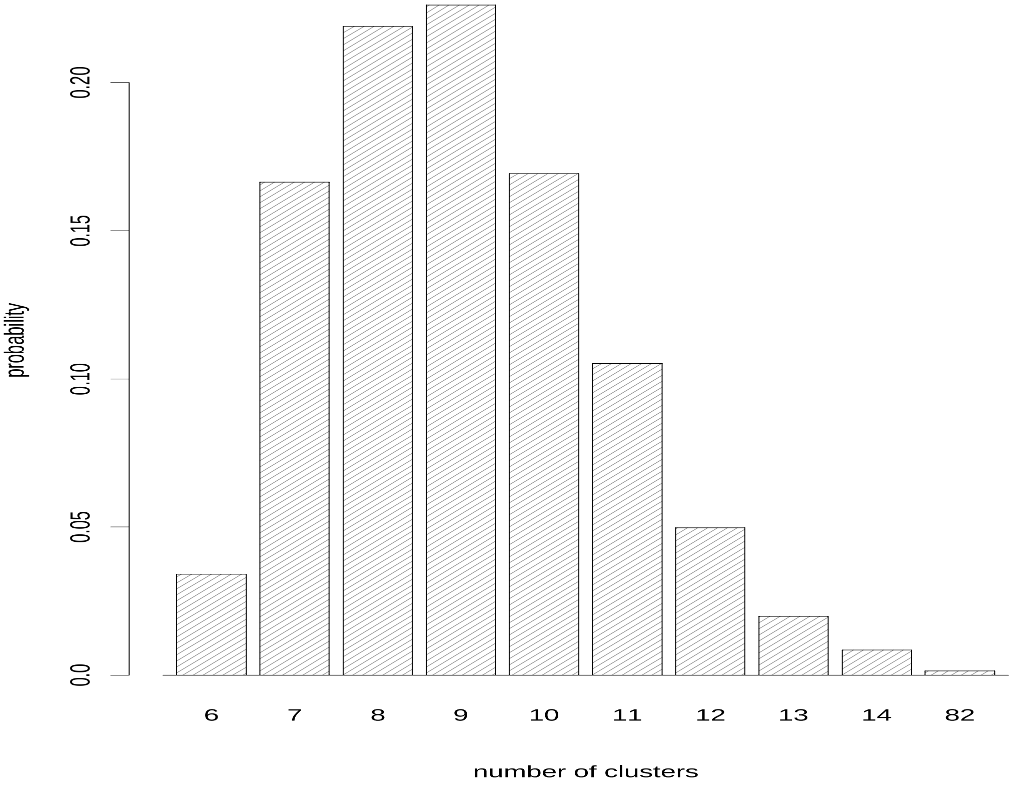

Figure 3: MCMC approximation of the posterior distribution of the number of distinct Y_i values, for the galaxy data. This gives an upper bound to the number of clusters.

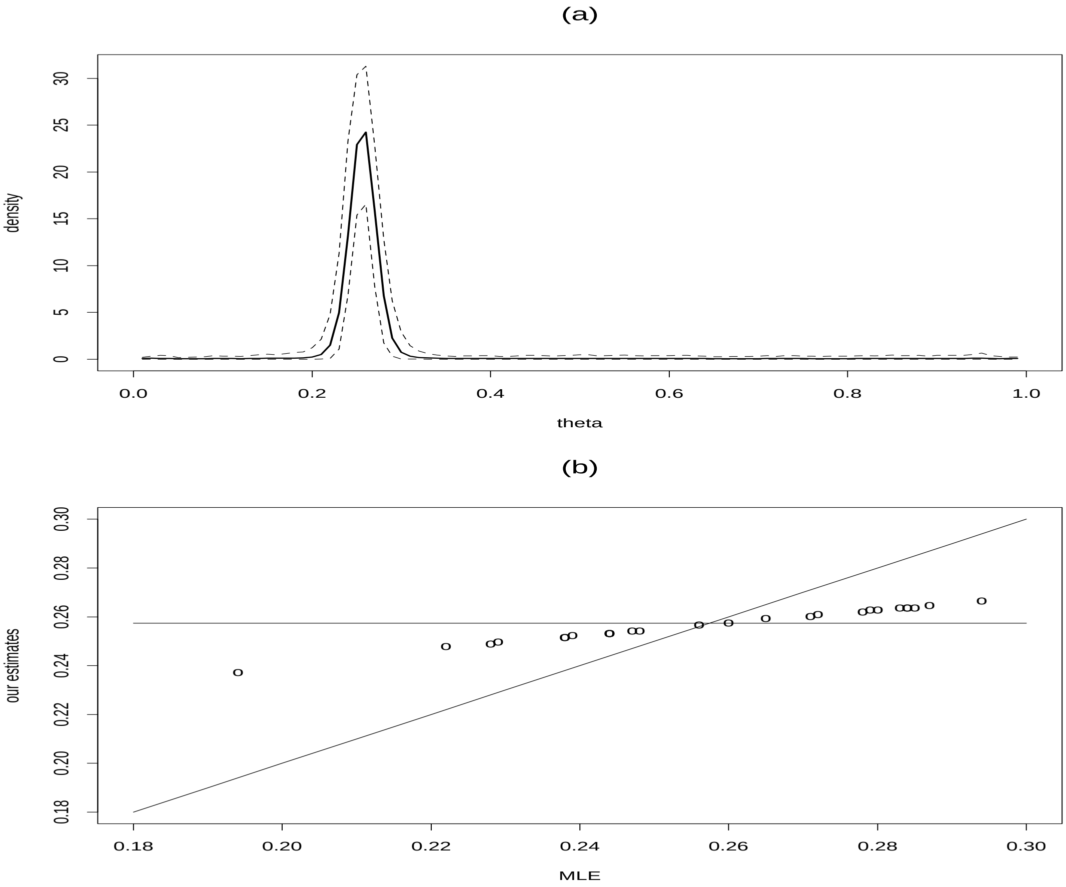

Figure 4: (a)Estimate of the mixing density for the baseball data. (b) Our estimates of the binomial parameters are plotted against the MLEs. The horizontal line is the overall mean.