The Annals of Statistics
1998, Vol. 26, No. 6, 2104–2127

BAYESIAN BOOTSTRAP CREDIBLE SETS FOR MULTIDIMENSIONAL MEAN FUNCTIONAL[1]

BY NIDHAN CHOUDHURI

Michigan State University

This paper shows that the Bayesian bootstrap (BB) distribution of a multidimensional mean functional based on i.i.d. observations has a strongly unimodal Lebesgue density provided the convex hull of the data has a nonempty interior. This result is then used to construct the finite sample BB credible sets. The influence of an outlier on these credible sets is studied in detail and a comparison is made with the empirical likelihood ratio confidence sets in this context.

1. Introduction. Let $X_1, \ldots, X_n$ be i.i.d. d-dimensional random variables having an arbitrary unknown distribution F_0 with finite expectation. Let $\mathscr{F}$ denote the class of all distribution functions on d-dimensional Euclidean space $\mathbb{R}^d$ and $\tilde{\mathscr{F}}$ denote the subclass of $\mathscr{F}$ with finite expectation, that is,

$$\tilde{\mathscr{F}} = \left\{ F \in \mathscr{F} \colon \int_{\mathbb{R}^d} \|x\| \, dF(x) < \infty \right\},$$

and μ denote the mean functional on $\tilde{\mathscr{F}}$ defined as

$$\mu(F) = \int_{\mathbb{R}^d} x \, dF(x), \qquad F \in \tilde{\mathscr{F}}. \tag{1.1}$$

The focus of this paper is to construct a set estimate for $\mu(F_0)$.

Bayes' approach to this problem is to construct a prior probability on $\mathscr{F}$. One assumes F to be a random element of $\mathscr{F}$ according to this prior probability, F_0 to be a particular realization of F and given F, $X_1, \ldots, X_n$ are i.i.d. F. Then one uses the posterior distribution of F and $\mu(F)$ to infer about F_0 and $\mu(F_0)$. A nonparametric prior often used in the literature is a Dirichlet process prior with a finite shape measure α [Ferguson (1973)]. A probability on $\mathscr{F}$ is said to be a Dirichlet process with shape measure α if for every measurable finite partition $B_1, \ldots, B_k$ of $\mathbb{R}^d$, the random variable $(F(B_1), \ldots, F(B_k))$ has a Dirichlet distribution on $\mathbb{R}^k$ with parameter $(\alpha(B_1), \ldots, \alpha(B_k))$. In this case, the posterior distribution of F is also a Dirichlet process with the shape measure $\alpha + \sum_1^n \delta_{X_i}$. A choice of α with

$$\int_{\mathbb{R}^d} \|x\| \, d\alpha(x) < \infty$$

ensures that both the prior and the posterior probabilities are concentrated on $\tilde{\mathscr{F}}$. [See Ferguson (1973) for more on Dirichlet processes.]

Received June 1996; revised July 1998.

[1]Supported in part by the NSF Grant DMS-94-02904.

AMS 1991 *subject classifications*. 62G09, 62G15.

Key words and phrases. Bayesian bootstrap distribution, posterior distribution, noninformative prior, Dirichlet process prior, empirical likelihood, outlier.

But most often one does not have enough initial information about F to construct any kind of prior. Besides, quantifying the prior knowledge in the form of a Dirichlet process is not an easy task. The need for a noninformative prior to represent vague initial information in nonparametric Bayesian statistics is thus well justified.

Rubin (1981) introduces the concept of Bayesian bootstrap to express the posterior knowledge about F and its functionals in the absence of any prior information. Replacing the mass $1/n$ of the empirical distribution F_n by random weights, he defines a random distribution function on $\mathbb{R}^d$ as

$$D_n = \sum_1^n W_i \delta_{X_i}, \tag{1.2}$$

where the joint distribution of $(W_1, \ldots, W_n)$ is uniform on the simplex

$$\Omega_n = \left\{ w \in \mathbb{R}^n \colon \sum_{i=1}^n w_i = 1,\ w_i \geq 0 \right\} \subset \mathbb{R}^n \tag{1.3}$$

and is independent of the sample $X_1, \ldots, X_n$. The Bayesian bootstrap (BB) distribution of any functional θ on $\mathscr{T}$ is the conditional distribution of $\theta(D_n)$, given $X_1, \ldots, X_n$.

Rubin (1981) argues that for a fixed finite sample, the BB distribution of F can be obtained as a weak limit of the posterior distribution under Dirichlet priors when the total mass of the shape measure α tends to zero, that is, $\alpha(\mathbb{R}^d) \to 0$. The results thus obtained are then comparable to standard frequentist results, as illustrated by the applications in Section 5 of Ferguson (1973). Gasparini [(1995), Theorem 3], proves that if $\alpha = \alpha(\mathbb{R}^d)Q$ where Q is a probability measure with

$$\int_{\mathbb{R}^d} \|x\|^2 \, dQ(x) < \infty,$$

then as $\alpha(\mathbb{R}^d) \to 0$, the posterior distribution of $\mu(F)$ converges weakly to $\mu(D_n)$. These facts establish the role of Bayesian bootstrap as a noninformative prior in nonparametric Bayesian statistics. Hence, in the absence of any prior knowledge, using the $(1-p)$ central part of the BB distribution of $\mu(F)$ as a posterior credible set and in turn using it as a $(1-p)$ level Bayesian set estimate for $\mu(F_0)$ is natural.

The concept of using the BB distribution to produce credible sets has been used before. Example 1.1 of Lo (1987) uses the BB distribution to obtain a 95% probability band for a univariate distribution function F. For the multidimensional mean functional, the difficulty lies in selecting the central $(1-p)$ part of the BB distribution. In the one-dimensional case, the interval between the $(p/2)$th and the $(1-p/2)$th quantiles represents the central $(1-p)$ part of the distribution. Hence this interval can be used as a credible set. But this quantile approach does not extend to higher dimensions due to the lack of a proper definition of quantile. To identify the central part of a multivariate distribution, it is important to know the nature of the distribution.

This paper establishes the existence of a strongly unimodal Lebesgue density for the exact BB distribution of the multidimensional mean functional under a mild condition. This result is then used in the construction of credible sets. The construction procedure is then extended for the cases when the condition fails. Then this paper finds the influence of an outlier on BB credible sets and compares these credible sets with the empirical likelihood confidence sets of Owen (1990) in this context.

The plan for the rest of the paper is as follows. A brief literature survey ends Section 1. Section 2.1 presents the main strong unimodality results and identifies the BB credible sets. A two-step procedure for constructing the credible sets is presented in Section 2.2. Section 3 compares these sets with the empirical likelihood confidence sets in connection with an outlier. The proofs of the four main theorems are deferred to the Appendix.

1.1. *Related work.* In the case F has finite support $\{d_1, \ldots, d_k\}$, a vector θ with $\theta_j = F\{d_j\}$, $F\{d_j\}$ being the probability of the singleton set $\{d_j\}$, uniquely identifies F. Hence the space of all probability measures on $\{d_1, \ldots, d_k\}$ can be parameterized by the k-variate unit simplex. Now a prior on θ with density proportional to $\prod \theta_j^{p_j}$ leads to the posterior density proportional to $\prod \theta_j^{p_j+n_j}$, where n_j's are the number of observations equal to d_j's. A noninformative prior (improper) with all $p_j = -1$ leads to the fact that $\theta_i = 0$ with posterior (improper) probability 1 for any unobserved d_i, and the posterior distribution becomes the BB distribution. An important fact is that one does not need to know the values of unobserved d_i's, as pointed out in Owen (1990). This gives a justification for using BB as an noninformative prior in the finite support case.

Owen (1990) introduces the concept of empirical likelihood as a nonparametric generalization of the well-studied parametric likelihood and uses it to construct confidence sets and test statistics for several nonparametric functionals. He observes that in the finite support case the empirical likelihood is proportional to the BB density. He thus argues in favor of connecting empirical likelihood with the posterior density under a noninformative prior as in the parametric case.

Asymptotic equivalence of the BB distribution and the posterior distribution under a Dirichlet prior with nonzero α has been noticed earlier. Lo (1987) shows that in one dimension, the posterior distribution of F under a Dirichlet prior and the BB distribution, conditional on the data, are first-order asymptotically equivalent in the sense that for almost all sample sequences and subject to proper centering and $n^{1/2}$ scaling, they achieve the same limiting conditional distribution. Weng (1989) points out that for the one-dimensional mean functional, the two distributions are equivalent up to a second-order asymptotic if

$$\int_{\mathbb{R}^d} \|x\|^3 \, d\alpha(x) < \infty,$$

and F_0 has finite third moment. This helps one to approximate the posterior distribution under a Dirichlet prior through a simulation of the BB distri-

bution. The approximation thus is useful since it is easier to simulate a BB distribution than a posterior Dirichlet process.

The operational and structural similarities between BB and the bootstrap of Efron (1979) are mentioned in Rubin (1981) and Efron (1982). Rubin has shown that the ordinary bootstrap is the same as BB except the very fact that the weights $(W_1, \ldots, W_n)$ are continuous in the BB, whereas they are replaced by some discrete weights in the ordinary bootstrap. Further, he gives an example in which the histogram of 1000 BB correlation coefficients is similar to, but smoother than, a histogram of 1000 ordinary bootstrap correlation coefficients. Lo (1987) proves the first order asymptotic equivalence of the two procedures for a variety of functionals, including the mean functional and the identity functional. Similar results for the finite population case are obtained in Lo (1988). This gives a frequentist perspective on the BB method.

Variants of BB can be obtained by replacing the uniform weights $(W_1, \ldots, W_n)$ by some other continuous exchangeable weights. Let $Y_1, \ldots, Y_n$ be i.i.d. positive continuous random variables and independent of the data. Then $(Y_1/\sum Y_i, \ldots, Y_n/\sum Y_i)$ give continuous exchangeable weights. The class of random distribution functions obtained by using different Y_i is called Bayesian bootstrap clone (BBC) in Lo (1991). Many asymptotic results on BBC for several functionals is found in Lo (1991) including the conditions on the random variable Y_i. Weng (1989) shows that using a Gamma (4,1) random variable for Y_i, a two-term Edgeworth expansion for the BBC distribution of the mean functional is identical to that of the sampling distribution of the sample mean, like the bootstrap; whereas the BB distribution is as accurate as the normal approximation. But the optimal choice of Y_i's depends on the functional of interest and is not universal.

2. The BB credible sets for the mean functional.

2.1. *Identifying the credible sets.* Let $\mathbb{X}$ denote the sample sequence $\{X_1, X_2, \ldots\}$, F_0^∞ denote the infinite product measure on $(\mathbb{R}^d)^\infty$, $\bar{X}_n = \mu(F_n)$ denote the sample mean and $\Lambda_{n,X}$ denote the BB distribution of the mean functional. The aim is to find a central high probability concentration set of $\Lambda_{n,X}$. When

$$\int_{\mathbb{R}^d} \|x\|^2 \, dF_0(x) < \infty, \tag{2.1}$$

a normal approximation of the BB distribution may be useful for this purpose.

THEOREM 2.1. *If* (2.1) *holds, then for almost every sample sequence* $\mathbb{X}$,

$$\sqrt{n}\{\mu(D_n) - \bar{X}_n\}|\mathbb{X} \Longrightarrow N_d(0, \Sigma), \tag{2.2}$$

where Σ *is the dispersion matrix of* F_0.

The result for one dimension is proved in Lo (1987), which can be extended to d-dimension by the Cramér–Wold device.

If Σ is of full rank, then one can substitute Σ by the sample dispersion matrix

$$S_n = \frac{1}{n-1}\sum_1^n (X_i - \bar{X}_n)(X_i - \bar{X}_n)^T$$

in the limiting normal distribution and obtain

$$A_p = \{x\colon n(x - \bar{X}_n)^T S_n^{-1}(x - \bar{X}_n) \le \chi^2_{d,1-p}\}$$

as an approximate central high probability concentration region of $\Lambda_{n,X}$. But the convergence in (2.2) is in the first-order sense as indicated by Weng (1989). So A_p cannot reflect any higher order moment structure of $\Lambda_{n,X}$ such as skewness. Besides, A_p is always elliptical in shape. Note that A_p is the same as the frequentist confidence set obtained by Hotelling's T^2-distribution up to a scale factor; that is, the cut-off point $\chi^2_{d,1-p}$ is replaced by a multiple of some quantile of an F-distribution. Naturally A_p cannot represent a BB credible set. If the posterior distribution under a noninformative prior is the prime object, then it is important to find the central part of the exact BB distribution.

There are some difficulties in identifying the central part of an arbitrary multivariate probability distribution. A probability on $\mathbb{R}^d$ is said to be *strongly unimodal* if it has a Lebesgue density g such that every high density contour $\{x \in \mathbb{R}^d\colon g(x) \ge c\}$ is a convex set. The existence of the density implies that any high density contour is the smallest set (in terms of Lebesgue measure) among all sets with the same probability. Strong unimodality implies that such a high density contour is a convex set and is surrounded by a low probability concentration region. Hence a high density contour in some sense represents a central high probability concentration region and can be used as a credible set. We shall prove the strong unimodality of $\Lambda_{n,X}$.

DEFINITION 2.1 [Prékopa (1973), equation (1.1)]. A nonnegative function f on $\mathbb{R}^d$ is said to be *logconcave* if for every $x, y \in \mathbb{R}^d$, $t \in [0,1]$,

$$f(tx + (1-t)y) \ge [f(x)]^t\,[f(y)]^{1-t}, \tag{2.4}$$

with the understanding that $0^0 = 1$.

PROPOSITION 2.1. *A probability with logconcave Lebesgue density is strongly unimodal.*

THEOREM 2.2. *If the convex hull of $X_1, \ldots, X_n$ has a nonempty interior, then $\Lambda_{n,X}$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$ and there is a logconcave version of the Lebesgue density.*

REMARK 2.1. The conclusion of Theorem 2.2 still holds if the joint distribution of $(W_1, \ldots, W_n)$ belongs to a general class of probability measures on Ω_n besides being uniform. The general class is identified in Corollary A1.1 of the Appendix.

Note that the convex hull of $X_1, \ldots, X_n$ has nonempty interior if and only if all X_i's are not confined in a hyperplane; that is, F_n is nonsingular. (F on $\mathbb{R}^d$ is *nonsingular* if $F\{H\} < 1$ for every hyperplane H.) If F_0 is nonsingular, then for almost every sample sequence $\mathbb{X}$, there is an N, depending on $\mathbb{X}$, such that F_n is nonsingular for $n > N$. Hence $\Lambda_{n,X}$ is eventually strongly unimodal. Moreover if $F_0\{H\} = 0$ for any hyperplane H, then with F_0^∞ probability 1, F_n is nonsingular for every $n \geq (d+1)$. Hence the condition of Theorem 2.2 is satisfied in most of the cases.

One can proceed even if the interior of the convex hull of the data is empty. In this case, all the X_i's are confined in an affine of $\mathbb{R}^d$. [An affine in $\mathbb{R}^d$ is a subset $\mathscr{M}$ of $\mathbb{R}^d$ such that for every $x, y \in \mathscr{M}$ and $-\infty < t < \infty$ we have $tx + (1-t)y \in \mathscr{M}$; that is, the entire line passing through x and y are in $\mathscr{M}$. If $0 \in S$, then $\mathscr{M}$ is called a subspace. Affines are sometimes known as lower dimensional planes.] Define the affine hull of a set A in $\mathbb{R}^d$ as

$$\mathscr{H}(A) = \{tx + (1-t)y\colon x, y \in A,\ -\infty < t < \infty\}. \tag{2.5}$$

This is the smallest affine containing A. Let $\mathscr{H}_0$ be the affine hull of the data set $\{X_1, \ldots, X_n\}$ and s be the dimension of $\mathscr{H}_0$. Then we have the following theorem.

THEOREM 2.3. *If $0 < s < d$, then $\Lambda_{n,X}$ is absolutely continuous with respect to the s-dimensional Lebesgue measure restricted to $\mathscr{H}_0$ and there is a logconcave version of the corresponding density.*

PROOF. Let λ_s denote the Lebesgue measure on $\mathbb{R}^s$ and $\tilde{\lambda}_s$ denote the s-dimensional Lebesgue measure restricted to $\mathscr{H}_0$. Then there exists a bijective affine map $\mathbb{L}\colon \mathscr{H}_0 \to \mathbb{R}^s$ such that $\tilde{\lambda}_s\mathbb{L}^{-1} = \lambda_s$. ($\tilde{\lambda}_s\mathbb{L}^{-1}$ denotes the induced measure of $\tilde{\lambda}_s$ on $\mathbb{R}^s$.) Let $Y_i = \mathbb{L}(X_i)$. Linearity of $\mathbb{L}$ implies that the affine hull of $\{Y_1, \ldots, Y_n\}$ is the image of the affine hull of $X_1, \ldots, X_n$ under $\mathbb{L}$, that is, the entire $\mathbb{R}^s$. Hence the convex hull of $\{Y_1, \ldots, Y_n\}$ has nonempty interior in $\mathbb{R}^s$. Note that $\Lambda_{n,X}\mathbb{L}^{-1}$ is the same as the BB distribution on $\mathbb{R}^s$ obtained by the transformed data $\{Y_1, \ldots, Y_n\}$. Hence by Theorem 2.2, $\Lambda_{n,X}\mathbb{L}^{-1}$ has logconcave Lebesgue density g_s on $\mathbb{R}^s$. Since $\mathbb{L}$ is one-to-one, $\Lambda_{n,X}\mathbb{L}^{-1} \ll \lambda_s = \tilde{\lambda}_s\mathbb{L}^{-1}$ implies $\Lambda_{n,X} \ll \tilde{\lambda}_s$ and $\tilde{g}(x) = g(\mathbb{L}(x))$ defines a version of $d\Lambda_{n,X}/d\tilde{\lambda}_s$. Affine property of $\mathbb{L}$ and logconcave property of g_s implies $\tilde{g}$ is logconcave on $\mathscr{H}_0$. □

Note that the map $\mathbb{L}$ is not unique and g_s depends on $\mathbb{L}$. But $g_s \circ \mathbb{L}$ is a version of $d\Lambda_{n,X}/d\tilde{\lambda}_s$ and hence is independent of the choice of $\mathbb{L}$. Since $\mathbb{L}$ is one-to-one and affine, the inverse image of a high density contour of g_s will be a high density contour of $\tilde{g}$ in $\mathscr{H}_0$. The high density contours of g_s may depend on $\mathbb{L}$ but their inverse images in $\mathscr{H}_0$ under the map $\mathbb{L}$ will not depend on $\mathbb{L}$, as they are the high density contours of $\tilde{g}$. Hence these high density contours of $\tilde{g}$ can be used as BB credible sets in $\mathscr{H}_0$.

To apply this result, one needs to find an $\mathbb{L}$. As $\mathscr{H}_0$ has dimension s, the rank of the sample dispersion matrix S_n is s. So S_n has exactly s nonzero

eigenvalues. Take a spectral decomposition of S_n, and let $e_1, \ldots, e_s$ be the eigen vectors corresponding to the nonzero eigenvalues. Then a candidate for $\mathbb{L}$ is

$$\mathbb{L}(x) = [e_1, \ldots, e_s]^T(x - \bar{X}).$$

In this case the inverse function $\mathbb{L}^{-1}$: $\mathbb{R}^s \to \mathscr{H}_0$ is of the form $\mathbb{L}^{-1}(y) = [e_1, \ldots, e_s]y + \bar{X}$.

The above logconcavity result can be extended to the BB distribution of a linear functional. Let φ: $\mathbb{R}^d \to \mathbb{R}^q$ be a Borel measurable function and μ_φ be a linear functional on $\mathscr{F}$ defined as

$$\mu_\varphi(F) = \int_{\mathbb{R}^d} \varphi \, dF.$$

Let Y_i denote $\varphi(X_i)$. Then the BB distribution of μ_φ is the same as the BB distribution of the mean functional on $\mathbb{R}^q$ based on the transformed data $Y_1, \ldots, Y_n$. Hence all the logconcavity results follow for $\mu_\varphi(F)$. The case $q > d$ will be taken care of by Theorem 2.3.

2.2. *Constructing the confidence region.* When the convex hull of $X_1, \ldots, X_n$ has a nonempty interior, a Monte Carlo simulation can be used for constructing the high density contours of $\Lambda_{n,X}$. Throughout this subsection, the original sample size n and the data $X_1, \ldots, X_n$ are fixed. Let g be the logconcave Lebesgue density of $\Lambda_{n,X}$ and

$$\mathscr{C}_{\mathrm{BB}} = \{x \in \mathbb{R}^d \colon g(x) \geq \lambda\} \tag{2.6}$$

be the high density contour such that $\Lambda_{n,X}\{\mathscr{C}_{\mathrm{BB}}\} = 1 - p$. A two-step procedure for constructing $\mathscr{C}_{\mathrm{BB}}$ is described here.

First we need to generate uniform distribution on Ω_n. Two different procedures for that are described below.

Procedure 1. Let $U_{(1)}, \ldots, U_{(n-1)}$ be the order statistics of $n-1$ i.i.d. $U(0,1)$ and $U_{(0)} = 0$, $U_{(n)} = 1$. Define $W_i = U_{(i)} - U_{(i-1)}$, $i = 1, \ldots, n$. Then $(W_1, \ldots, W_n)$ is uniform on Ω_n.

Procedure 2. Define $W_i = Y_i / \sum_1^n Y_i$, $i = 1, \ldots, n$, where $Y_1, \ldots, Y_n$ are i.i.d. exponentials. Then $(W_1, \ldots, W_n)$ is uniform on Ω_n.

These two methods of generating uniform random variables on Ω_n have been known in the literature for a long time and proofs can be found in Devroye [(1986), pages 207–210]. Procedure 2 is much easier to perform on a computer, while Procedure 1 is useful in proving some theoretical results about BB distributions.

STEP 1. Simulate $w^1, \ldots, w^m$ i.i.d. with uniform distribution on Ω_n. Then obtain m points $\tilde{X}_1, \ldots, \tilde{X}_m$ in $\mathbb{R}^d$ with $\tilde{X}_j = \sum_{i=1}^n w_i^j X_i$, where w_i^j is the ith component of w^j.

Note that $\tilde{X}_1, \ldots, \tilde{X}_m$ obtained in Step 1 are i.i.d. $\Lambda_{n,X}$. In the next step, we shall use a histogram smoothing on $\tilde{X}_1, \ldots, \tilde{X}_m$ to obtain a density estimate g_m of g. Then we shall use the $(1-p)$ probability high density contour of g_m as an approximation of $\mathscr{C}_{\mathrm{BB}}$. A similar idea of using density estimation to bootstrap replicates for constructing likelihood-based confidence regions for a vector parameter is found in Hall (1987).

For $l = 1, \ldots, d$, let us define

$$a_l = \min\{X_i^{(l)}: 1 \le i \le n\},$$
$$b_l = \max\{X_i^{(l)}: 1 \le i \le n\},$$

where $X_i^{(l)}$ is the lth component of the ith observation X_i. Then the hyperrectangle

$$\mathscr{R} = \{x \in \mathbb{R}^d: a_l \le x^{(l)} \le b_l,\ 1 \le l \le d\}$$

contains all the data points $X_1, \ldots, X_n$ and hence contains the support of $\Lambda_{n,X}$. Define a hypercube of length $h > 0$ around a point $x \in \mathbb{R}^d$ as $R(x,h) = \{y \in \mathbb{R}^d: |y^{(l)} - x^{(l)}| \le h/2, \forall\ l = 1, \ldots, d\}$. Now we will partition the region $\mathscr{R}$ into small hypercubes. Fix an $h > 0$. For $l = 1, \ldots, d$, let $S_l = \{(i + 1/2)h: i = 0, \ldots, [(b_l - a_l)/h]\} \subset \mathbb{R}$, where $[c]$ denotes the largest integer less than or equal to c. Define the grid set on $\mathbb{R}^d$ as $\mathscr{R}_h = \prod_1^d S_l$, the Cartesian product of S_l's. Then the hypercubes $\{R(x,h): x \in R_h\}$ cover the set $\mathscr{R}$ and are disjoint except at the boundaries. For each $x \in \mathbb{R}^d$, define

$$\tau(x) = \sum_{j=1}^{m} \{\tilde{X}_j \in R(x,h)\} = \#\text{ of } \tilde{X}_j\text{'s belonging to } R(x,h). \tag{2.7}$$

STEP 2. For the data $X_1, \ldots, X_n$, obtain the set $\mathscr{R}$. Choose $h = m^{-1/(d+2)}$ and obtain the grid set $\mathscr{R}_h$. Calculate $\tau(x)$ for each $x \in \mathscr{R}_h$ and order the grid points according to the descending order of $\tau(x)$. Let $\{x_1, \ldots, x_k\}$ denote the ordered grid points, where k is the number of points in $\mathscr{R}_h$. Find the integer k_0 such that

$$\sum_1^{k_0-1} \tau(x_j) < (1-p)m \quad \text{and} \quad \sum_1^{k_0} \tau(x_j) \ge (1-p)m.$$

This can be done by adding $\tau(x_j)$'s one at a time until we reach $(1-p)m$. Now use the set

$$\mathscr{C}_m = \bigcup_1^{k_0} R(x_j, h)$$

as an approximation to $\mathscr{C}_{\mathrm{BB}}$.

To measure the performance of $\mathscr{C}_m$ in approximating $\mathscr{C}_{\mathrm{BB}}$, one needs to define a measure of proximity between sets. Define a metric d_1 on the subsets of $\mathbb{R}^d$ as

$$d_1(B_1, B_2) = \mathrm{Leb}(B_1 \triangle B_2), \qquad B_1, B_2 \subset \mathbb{R}^d,$$

where $\triangle$ defines the symmetric difference of sets and "Leb" denotes the Lebesgue measure on $\mathbb{R}^d$. Then we have the following convergence result on $\mathscr{C}_m$.

THEOREM 2.4. *If the convex hull of any $n-1$ data points has nonempty interior, then for a.e. simulation sequence,*

$$d_1(\mathscr{C}_m, \mathscr{C}_{\mathrm{BB}}) = O(m^{-1/(d+2)} \ln(m)) \quad \text{as } m \to \infty.$$

Some simulation results are presented in Figure 1 using the simulation size $m = 200,000$ for each case. Figure 1a shows the BB credible sets with confidence level 80%, 95% and 99% for the twelve observations from the bivariate normal distribution with mean 0 variance 1 for each component and the correlation coefficient -0.5. The credible sets are almost elliptical in shape as expected with data coming from an elliptically symmetric distribution. Figure 1b shows these credible sets based on twelve observations from the skewed bivariate distribution with density $f(u, v) = uve^{-(u+v)}$, $u, v > 0$, that is, a bivariate gamma distribution whose components are independent univariate gamma with the shape parameter 2 and the scale parameter 1. Note that these credible sets are able to reflect the skewness of the underlying distribution in their shape. The moderate sample BB credible sets with 40 observations from these two distributions are presented in Figure 1c and 1d. These credible sets for the gamma observations in Figure 1d are almost elliptical in shape with very little skewness. This is expected, because the standardized BB distribution is asymptotically normal.

Commenting on the computational aspect, Step 1 takes time proportional to the simulation size m. The number of grid points $k = \prod_1^d [(b_l - a_l)m^{-1/(d+2)} + 1] \simeq cm^{d/(d+2)}$. Hence the calculation of $\tau(x)$ for $x \in \mathscr{R}_h$ will take time proportional to km, which is of the order smaller than $o(m^2)$. Ordering of grid points will take time proportional to k^2, which is also of the order smaller than $o(m^2)$. Hence the magnitude of time taken for the entire procedure will be of the order smaller than $o(m^2)$. This allows one to perform the simulation with large m and thus to make the approximation more accurate. Another advantage here is that the computational time depends mainly on m and remains almost unchanged with a change in dimension d or in sample size n. However, the convergence rate of $\mathscr{C}_m$ decreases with an increase in the dimension, and one needs to use a larger m to achieve the same resolution.

An alternate procedure to the one described in Step 2 can be found using the excess mass approach. The idea here is to chose the set with the smallest Lebesgue measure among all convex sets containing $(1-p)m$ or more simulated points $\tilde{X}_j$. This is the multidimensional extension of Hartigan's

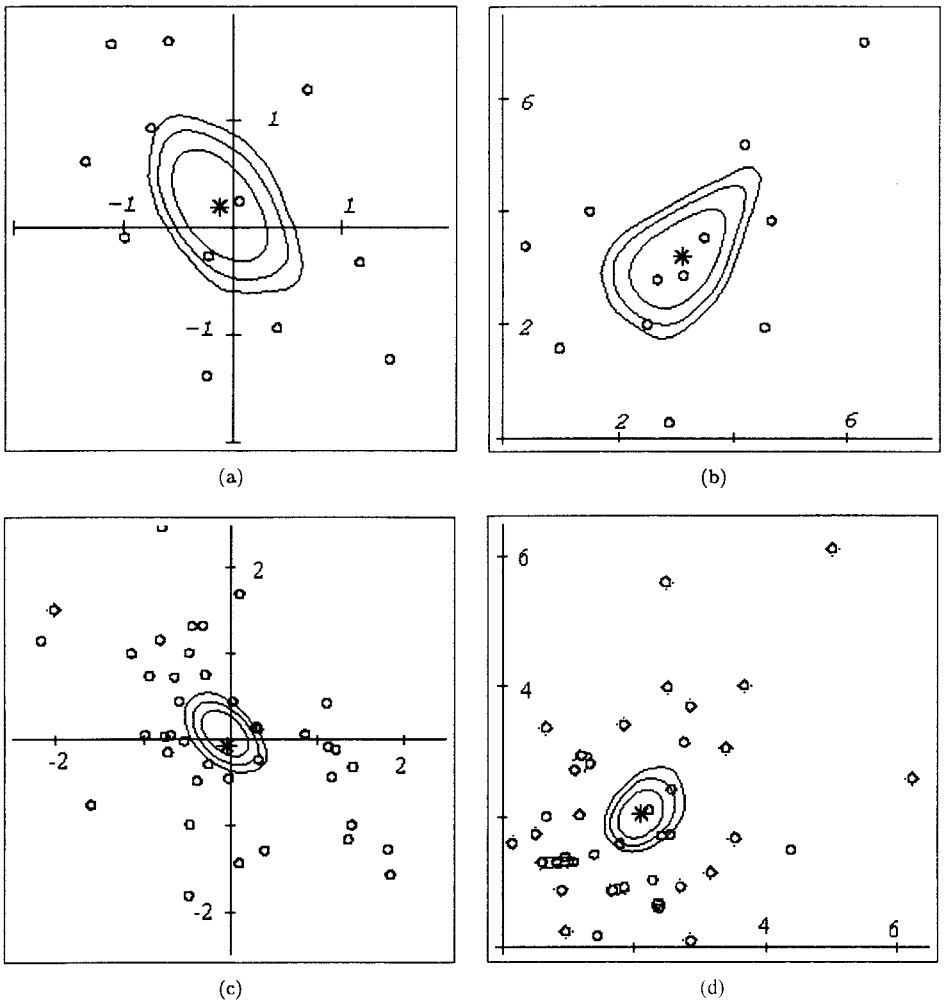

FIG. 1. *The* 80%, 95% *and* 99% *level BB credible sets for the mean functional, based on small sample as well as moderate sample observations from two different bivariate distributions using a histogram smoothing approach with simulation size* 200,000. (a) *Bivariate normal with* $n = 12$; (b) *bivariate gamma with* $n = 12$; (c) *bivariate normal with* $n = 40$; (d) *bivariate gamma with* $n = 40$.

(1987) idea in two dimensions. Lemma A2.1 (see Appendix) shows that g satisfies the second equation in Section 5 of Tsybakov (1997) with regularity parameter 1. Then by the results in Section 5 of Tsybakov [(1997), second paragraph on page 957], the rate of convergence for this procedure in d_1 metric is $o(m^{-2/(d+5)})$. Though the convergence rate is better than the histogram smoothing approach, the problem here lies in the computational part. In dimension two, the computational time of the excess mass approach is $O(m^3)$, which is much larger compared to that of the density estimation approach,

whereas the achievement in the convergence rate over the histogram smoothing approach is not that significant. Besides this fact, the calculation of the area of a region is so time-consuming that the excess mass approach takes a long time even for moderate m. On the other hand, in a given amount of time, the histogram smoothing approach can handle a much larger simulation size, resulting in more accuracy.

3. Comparison of BB credible sets with empirical likelihood ratio confidence sets. For i.i.d. data $X_1, \ldots, X_n$, Owen (1990) defines the empirical likelihood of a distribution function $F \in \mathscr{F}$ as

$$L(F) = \prod_{i=1}^{n} F\{X_i\}, \tag{3.1}$$

where $F\{x\}$ denote the probability of the singleton set $\{x\}$ under F. This likelihood function is maximized at the empirical distribution function F_n, the nonparametric MLE of F_0. In some cases, the empirical likelihood ratio function,

$$R(F) = \frac{L(F)}{L(F_n)} = n^n \prod F\{X_i\} \tag{3.2}$$

can be used to construct confidence sets and test statistic for a functional θ on $\mathscr{F}$. Consider sets of the form $\{\theta(F)\colon F \ll F_n,\ R(F) \geq r\}$ for $0 < r < 1$. Owen (1990) gives conditions on θ and F_0 under which these sets can be used as confidence sets for $\theta(F_0)$. For $\theta = \mu$, the mean functional, define

$$\mathscr{C}_{\mathrm{EL}} = \{\mu(F)\colon F \ll F_n,\ R(F) \geq r\}. \tag{3.3}$$

Then we have the following result by Owen (1990), Theorem 1.

RESULT 3.1. *If F_0 has finite second moment, that is,* (2.1) *holds, and the dispersion matrix Σ is of rank $s > 0$, then for every $0 < r < 1$, $\mathscr{C}_{\mathrm{EL}}$ is a convex set and*

$$\lim_{n\to\infty} P_{F_0}(\mathscr{C}_{\mathrm{EL}} \ni \mu(F_0)) = P(\chi_s^2 \leq -2\log r).$$

If one chooses $r = \exp\{-\frac{1}{2}\chi^2_{s,\,p}\}$, then $\mathscr{C}_{\mathrm{EL}}$ serves as a confidence set for $\mu(F_0)$ with the (frequentist) asymptotic coverage probability $1 - p$. Theorem 1 of Owen (1990) also contains some results related to $O(n^{-1/2})$ rate of convergence of the above limit. DiCiccio, Hall and Romano (1991) have shown that the rate is $O(n^{-1})$ if the assumptions justifying Edgeworth expansions are met and the Bartlett factor improves the rate to $O(n^{-2})$. Results related to some other functionals can be found in Owen (1990).

One advantage of both the BB and ELR methods for constructing set estimates is that the shapes of these sets are determined by the data. These sets are also able to incorporate the skewness of the data in their shapes and hence in turn capture the skewness of the underlying distribution. Figure 2a shows

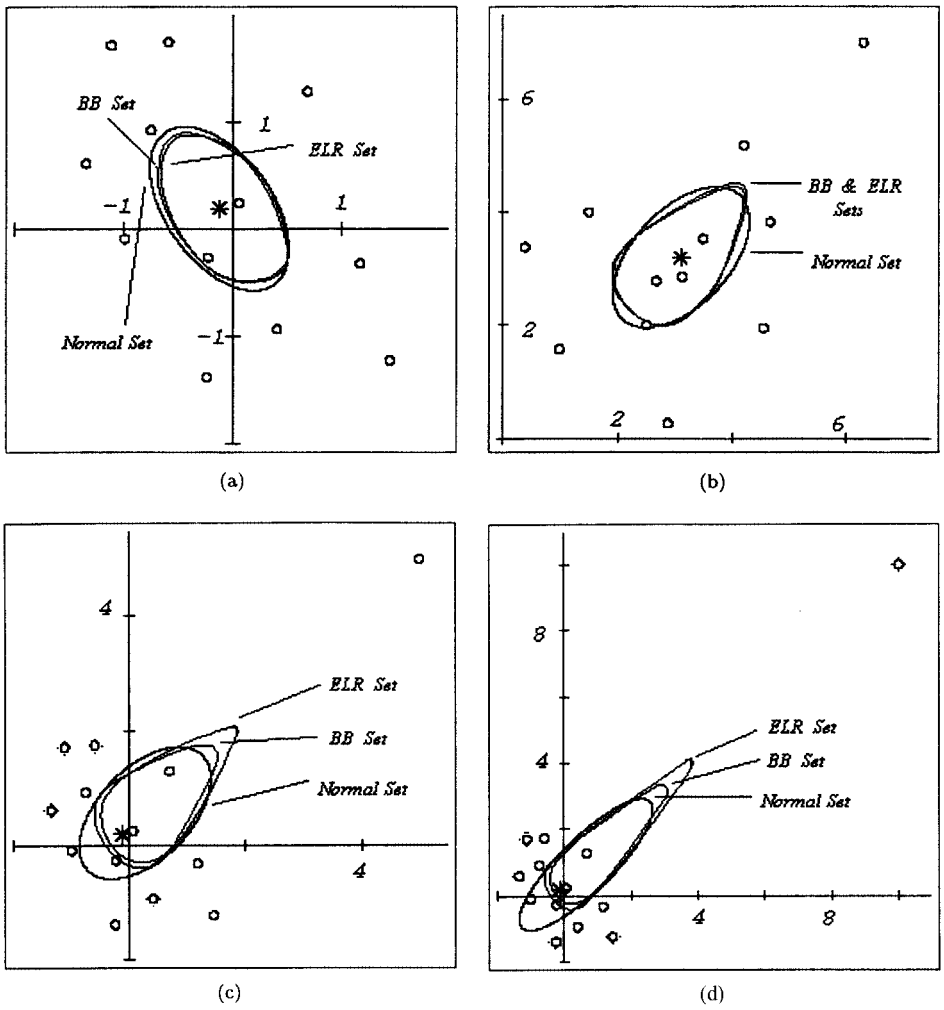

FIG. 2. *The BB credible sets, the empirical likelihood confidence sets and the normal approximation credible sets with coverage level* 95% *based on four different types of data sets.* (a) *Bivariate normal*; (b) *bivariate gamma*; (c) *data with small outlier*; (d) *data with large outlier.*

95% confidence sets using the BB, ELR and normal approximation methods based on twelve normal observations used in Section 2.2. Figure 2b shows these three sets based on the twelve observations from the skewed distribution used in Section 2.2. For the normal data, all three sets behave similarly, whereas in the skewed distribution case, the BB credible set and the ELR confidence set are able to reflect the skewness in the underlying distribution while the normal approximation method fails to do so.

However, a problem with the BB and ELR methods is that both regions are sensitive to outliers. This can be seen from Figure 2c and 2d. An outlier (not

random) is added to the twelve normal observations used earlier. Figure 2c and 2d show 95% confidence sets using the three methods based on all the thirteen observations for two different values of the outlier. One can see that the outlier has deformed all the three regions and inflated them towards itself. But the extent of inflation in the BB credible set is less than that of the ELR confidence set, while the normal approximation method is least affected. A quantitative study of the extent of the sensitivity of BB and ELR confidence sets is done here.

We shall define two measures of nonrobustness by considering how much an outlier can deform a set estimate. Let $X_1, \ldots, X_{n-1}$ be the first $n-1$ observations and $\bar{X}_{n-1}$ denote their average. Let the nth observation X_n be such that $\|X_n - \bar{X}_{n-1}\|$ is large compared to

$$\eta = \sup\{\|X_i - \bar{X}_{n-1}\|: 1 \le i \le n-1\}. \tag{3.4}$$

Then call $X_1, \ldots, X_{n-1}$ the data cloud and X_n an outlier. A diagram in two dimensions ($d = 2$) is presented in Figure 3. Let $\mathscr{C}$ be an arbitrary set estimate for μ based on all observations including the outlier. To measure the inflation of $\mathscr{C}$, introduce the quantity

$$U = \sup\{\|x - \bar{X}_{n-1}\|: x \in \mathscr{C}\}, \tag{3.5}$$

which is the distance of the farthest point in $\mathscr{C}$ from $\bar{X}_{n-1}$. (See Figure 3a.) Large U signifies $\mathscr{C}$ has a long nose toward the outlier X_n, and one can conclude that the outlier has inflated the region $\mathscr{C}$ towards itself, whereas small U implies less effect of X_n on $\mathscr{C}$.

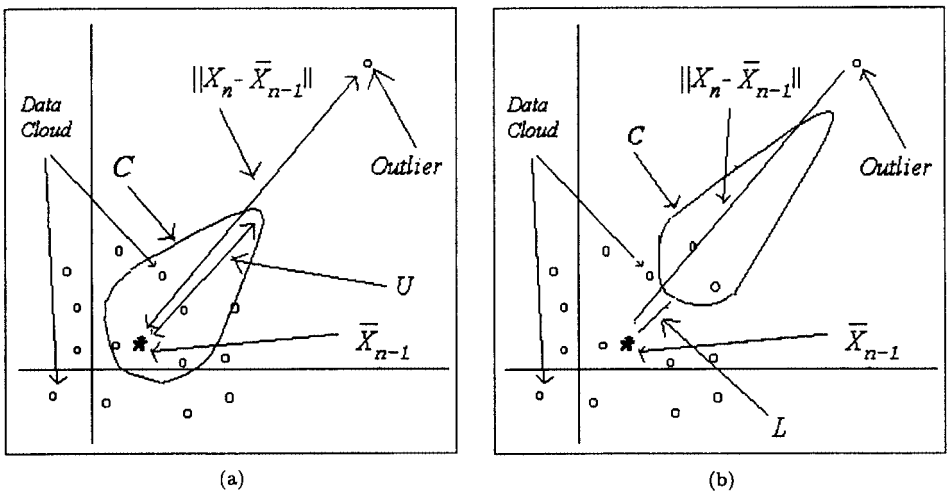

FIG. 3. *Diagram to identify the inflation effect and the shift effect of an outlier.* (a) *Measuring inflation effect*; (b) *measuring shift effect.*

Sometimes the influence of an outlier is so much that the whole region shifts away from the data cloud towards the outlier. (See Figure 3b.) We say that $\mathscr{C}$ has shifted from the data cloud if $\bar{X}_{n-1} \notin \mathscr{C}$ and we measure the shift by the quantity

$$L = \inf\{\|x - \bar{X}_{n-1}\|: x \in \mathscr{C}\}. \tag{3.6}$$

Note that $L = 0$ implies no shift and the reverse. Large L implies the region $\mathscr{C}$ has largely shifted from the data cloud, indicating the large influence of the outlier.

Let U_{BB} and U_{EL} denote the extent of inflation of the BB credible set and the ELR confidence set with coverage level $(1 - p)$. Let L_{BB} and L_{EL} denote the shifts for these two sets. Note that in Figure 2c and 2d, both L_{BB} and L_{EL} are zero, indicating that there is no shift effect of the outlier. However, there is a large inflation effect. The following two theorems give theoretical bounds on U_{BB}, U_{EL} and L_{EL}.

THEOREM 3.1. *For any data set* $X_1, \ldots, X_n$,

$$U_{\mathrm{EL}} \geq u_n \frac{\|X_n - \bar{X}_{n-1}\|}{n} \tag{3.7}$$

and

$$L_{\mathrm{EL}} \geq l_n \frac{\|X_n - \bar{X}_{n-1}\|}{n} - \eta\, 2^{3/2}(-\log r)^{1/2}(n-1)^{-1/2}, \tag{3.8}$$

where l_n *and* u_n *are the smallest and the largest roots of the equation*

$$f_n(h) := h\left(1 + \frac{1-h}{n-1}\right)^{n-1} = r, \qquad 0 \leq h \leq n, \tag{3.9}$$

with $r = \exp\{-\frac{1}{2}\chi^2_{d,\,p}\}$. *Moreover, as* $n \to \infty$,

$$\begin{aligned} l_n &= l_0 + o(n^{-1}), \\ u_n &= u_0 + o(n^{-1}), \end{aligned} \tag{3.10}$$

where l_0 *and* u_0 *are the smallest and the largest roots of the equation*

$$f(h) := he^{1-h} = r, \qquad 0 \leq h < \infty. \tag{3.11}$$

OBSERVATION 3.1. The function f_n is continuous, strictly increasing on $[0, 1)$, strictly decreasing on $(1, n]$ and $f_n(0) = 0$, $f_n(1) = 1$ and $f_n(n) = 0$. So for every $0 < r < 1$, $f_n(h) = r$ has exactly two solutions, l_n in $(0, 1)$ and u_n in $(1, n)$ and $\{h: f_n(h) \geq r\} = [l_n, u_n]$.

OBSERVATION 3.2. For a fixed coverage level $(1 - p)$, the quantity r in (3.9) and (3.11) decreases with an increase in the dimension d, as the percentiles of

a χ^2 distribution increase with an increase in the degrees of freedom. Thus u_n increases with an increase in the dimension d as both f_n and f are decreasing for $x > 1$.

THEOREM 3.2. *Let the outlier X_n satisfies $\|X_n\| = O(n)$. Then*

$$U_{\mathrm{BB}} \approx (-\log p)\frac{\|X_n - \bar{X}_{n-1}\|}{n}. \tag{3.12}$$

Theorems 3.1 and 3.2 help one in comparing the nonrobustness of the BB credible sets and the ELR confidence sets. The extent of inflation on both types of sets is proportional to the distance of the outlier from the data cloud and is inversely proportional to the sample size n. Here u_n describes the constant of proportionality for an ELR set which increases with an increase in the dimension of the data as well as with an increase in the coverage level. On the other hand, $-\log p$ in (3.12), the BB constant, does not depend on the dimension of the data, and most importantly, the BB constants are always smaller than the ELR constants at every level of coverage and whatever the dimension. Table 1 presents the values of u_n for $n = 13$ and $d = 2, 3$ along with the values of $-\log p$ at four different levels of noncoverage probability p. Since $u_n \to u_0$, as $n \to \infty$, the values of u_0 are also attached. These observations indicate some robustness advantage for the BB method over the ELR method, but neither method is robust.

No theoretical bound is found for L_{BB}. The BB credible sets usually contain $\bar{X}_{n-1}$ and $L_{\mathrm{BB}} = 0$ unless the outlier is too big. For small magnitude of the outlier, L_{EL} is also equal to zero. The first term in the lower bound of (3.8) is $O(n^{-1})$, whereas the second term is $O(n^{-1/2})$. Hence the right-hand side is often negative, making the inequality trivial. This implies that the shift effect of an outlier on both of these set estimates is negligible, but the inflation effect is very prominent.

The data diameter η is stochastically increasing in n and the magnitude is $O(n^{1/r})$ if the rth moment is finite. Hence an observation $O(n)$ is rare. But the effect of an outlier is inversely proportional to n and so only an outlier with magnitude $O(n)$ or more will be of concern.

TABLE 1
Values of u_n in (3.7) along with u_0 for dimensions 2 and 3, and values of $(-\log p)$ in (3.12) for four values of p

	$d = 2$		$d = 3$		
p	u_{13}	u_0	u_{13}	u_0	$(-\log p)$
0.01	5.954	7.638	6.609	8.853	4.605
0.05	4.796	5.744	5.480	6.828	2.9957
0.10	4.214	4.890	4.899	5.901	2.3026
0.20	3.560	3.994	4.230	4.912	1.6094

APPENDIX

A1. Proof of Theorem 2.2. For a nonempty set $A \in \mathbb{R}^k$, let $\mathscr{H}(A)$ denote the affine hull of A defined in (2.5). For a $t \in [0, 1]$ and two nonempty sets, A and B, define a convex combination of these two sets as

$$tA + (1-t)B = \{tx + (1-t)y\colon x \in A,\ \ y \in B\}.$$

DEFINITION A1.1 [Prékopa (1973), equation (1.2)]. A probability P on Borel sets of $\mathbb{R}^k$ is said to be *logconcave* if for all Borel measurable sets A, B and every $t \in [0, 1]$,

$$P\{tA + (1-t)B\} \geq [P(A)]^t\,[P(B)]^{1-t}.$$

A probability P in $\mathbb{R}^d$ is nonsingular if $P(H) < 1$ for every hyperplane H. To prove Theorem 2.2, first we shall show that the distribution $\Lambda_{n,X}$ is logconcave and nonsingular and then we shall use the standard logconcavity results to prove the existence of a logconcave density. Lemmas A1.1–A1.4 develop the required machinery for this purpose.

PROPOSITION A1.1. *P is nonsingular on $\mathbb{R}^k$ if and only if the affine hull of its support is the whole $\mathbb{R}^k$.*

LEMMA A1.1. *Let P be a logconcave probability on $\mathbb{R}^k$ and $L\colon \mathbb{R}^k \to \mathbb{R}^s$ be an affine transform. Then PL^{-1} is logconcave on $\mathbb{R}^s$.*

For the proof, see Dharmadhikari and Joag-dev [(1988), Lemma 2.1, page 47].

LEMMA A1.2. *Let P be a nonsingular probability on $\mathbb{R}^k$. Then P is logconcave if and only if P has a logconcave Lebesgue density on $\mathbb{R}^k$.*

For the proof, see Dharmadhikari and Joag-dev [(1988), Theorem 2.8, page 51].

LEMMA A1.3. *The joint distribution of $(W_1, \ldots, W_n)$ is logconcave on $\mathbb{R}^n$.*

PROOF. The joint distribution of $(W_1, \ldots, W_n)$ is uniform on Ω_n of (1.3). Hence the joint distribution of $(W_1, \ldots, W_{n-1})$ has a Lebesgue density

$$f(u) = I\{u_1 + \cdots + u_{n-1} \leq 1\}$$

on $\mathbb{R}^{n-1}$. The function f is logconcave as the indicator function of any convex set is logconcave. Hence by Lemma A1.2, the joint distribution of $(W_1, \ldots, W_{n-1})$ is logconcave on $\mathbb{R}^{n-1}$. Since the map $g\colon \mathbb{R}^{n-1} \to \mathbb{R}^n$, defined as

$$g(u) = (u_1, \ldots, u_{n-1}, 1 - u_1 - \cdots - u_{n-1}), \qquad u \in \mathbb{R}^{n-1},$$

is an affine map and $g(W_1, \ldots, W_{n-1}) = (W_1, \ldots, W_n)$, the proof is complete by Lemma A1.1. □

Define a map on Ω_n as

$$\tilde{\mu}(w) = \sum w_i X_i = \mu(F_w), \tag{A1.1}$$

where $F_w = \sum w_i \delta_{X_i}$. Let H_n denote the affine hull of Ω_n.

LEMMA A1.4. *Let P_n be a probability on $\mathbb{R}^n$ with support in Ω_n and let the affine hull of its support be H_n. Let the convex hull of $\{X_1, \ldots, X_n\}$ have a nonempty interior. Then $P_n\tilde{\mu}^{-1}$ is nonsingular in $\mathbb{R}^d$.*

PROOF. Let $Q_n = P_n\tilde{\mu}^{-1}$. We need to show that $Q_n(H) < 1$ for any hyperplane H of $\mathbb{R}^d$. Note that $H_n = \{w \in \mathbb{R}^n\colon \sum w_i = 1\}$.

For any hyperplane H in $\mathbb{R}^d$, there exists a vector $a \in \mathbb{R}^d$ and a real constant c such that $H = \{x \in \mathbb{R}^d\colon a^T x = c\}$. Hence,

$$\begin{aligned} Q_n(H) &= P_n\left\{w \in \mathbb{R}^n\colon a^T\left(\sum w_i X_i\right) = c\right\} \\ &= P_n\left\{w \in \mathbb{R}^n\colon \sum (a^T X_i) w_i = c\right\} \\ &= P_n\{\tilde{H}^n\}, \end{aligned}$$

where $\tilde{H}^n = \{w \in \mathbb{R}^n\colon \sum (a^T X_i) w_i = c\}$ is a hyperplane in $\mathbb{R}^n$. Since the affine hull of the support of P_n is the hyperplane H_n, so $P_n(\tilde{H}^n) = 1$ iff $\tilde{H}^n = H_n$. We will prove that $\tilde{H}^n \neq H_n$ for the two cases, $c = 0$ and $c \neq 0$, separately.

CASE 1. $c = 0$. Then $\tilde{H}^n = \{w \in \mathbb{R}^n\colon \sum (a^T X_i) w_i = 0\}$ is passing through the origin. Thus, it can never be equal to H_n, as H_n does not pass through the origin.

CASE 2. $c \neq 0$. Then $\tilde{H}^n = \{w \in \mathbb{R}^n\colon \sum (a^T X_i/c) w_i = 1\}$. Hence $\tilde{H}^n = H_n$ iff $(a^T X_i)/c = 1$ for all $i = 1, \ldots, n$. Thus $\tilde{H}^n = H_n$ implies that $X_i \in \{x \in \mathbb{R}^d\colon a^T x = c\}$, which in turn implies that the convex hull of $\{X_1, \ldots, X_n\}$ lies inside a hyperplane. This contradicts the assumption that the convex hull of $\{X_1, \ldots, X_n\}$ has a nonempty interior. □

PROOF OF THEOREM 2.2. Let Γ_n denote the uniform measure on Ω_n. Then $\Lambda_{n,X} = \Gamma_n\tilde{\mu}^{-1}$. As Γ_n is logconcave on $\mathbb{R}^n$ (Lemma A1.3) and $\tilde{\mu}$ is a linear map from $\mathbb{R}^n$ to $\mathbb{R}^d$, by Lemma A1.1, $\Lambda_{n,X}$ is logconcave on $\mathbb{R}^d$.

As the affine hull of the support of Γ_n is H_n, by Lemma A1.4, $\Lambda_{n,X}$ is nonsingular on $\mathbb{R}^d$. Hence by Lemma A1.2, the proof is complete. □

COROLLARY A1.1. *Let the weights $(W_1, \ldots, W_n)$ in (1.2) be replaced by some other weights $(W_1^*, \ldots, W_n^*)$, such that their joint distribution is logconcave on Ω_n and the affine hull of its support is H_n. Then also, under the condition that the convex hull of $X_1, \ldots, X_n$ has nonempty interior, the distribution of $\mu(D_n)$ has logconcave Lebesgue density.*

A2. Proof of Theorem 2.4. Throughout this proof, the data $X_1, \ldots, X_n$ are fixed and the randomness comes from the simulation. Recall that the hypercubes $\{R(x, h)\colon x \in R_h\}$ cover $\mathscr{R}$ and are disjoint except at the boundaries. Then the function

$$g_m(x) = (mh^d)^{-1} \sum 1_{R(x_i, h)}(x)\tau(x_i), \qquad x \in \mathbb{R}^d,$$

is a histogram smoothing density estimate of g and the set $\mathscr{C}_m$ is the same as $\{x\colon g_m(x) \geq \lambda_m\}$ with $\lambda_m = (mh^d)^{-1}\tau(x_{k_0})$. Hence $\mathscr{C}_m$ is a high density contour of g_m. First we shall show that $g_m \to g$ and $\lambda_m \to \lambda$ a.e, where λ is defined in (2.6).

Note that one can identify the BB density g as a multivariate B-spline function with knots $X_1, \ldots, X_n$ from the probabilistic definition of B-spline in Section 2 of Karlin, Micchelli and Rinott (1986). Hence, whenever the convex hull of every $n-1$ data points has nonempty interior, by Corollary 3 and its extension in the adjacent paragraph of Micchelli (1980), g has a continuous bounded derivative and the derivative is nonzero in the interior of the support of g except at the mode of g. □

LEMMA A2.1. *Let λ be as in* (2.6). *Then there exist constants $b_i > 0$, $i = 1, 2, 3$ and $\delta_0 > 0$, possibly depending on λ, such that for all $\delta \leq \delta_0$, we have*

$$\text{(i)} \qquad \text{Leb}\{x\colon |g(x) - \lambda| \leq \delta\} \leq b_1\delta,$$

$$\text{(A2.1) (ii)} \qquad \text{Leb}\{x\colon 0 < \lambda - g(x) \leq \delta\} \geq b_2\delta,$$

$$\text{(iii)} \qquad \text{Leb}\{x\colon 0 < g(x) - \lambda \leq \delta\} \geq b_3\delta.$$

PROOF. The condition $0 < 1 - p < 1$ along with (2.6) implies that $\{y\colon g(y) = \lambda\}$ is in the interior of the support of g and does not contain the mode of g. Hence $\|\text{grad}\, g(x)\|$ is bounded away from zero in a neighborhood of $\{x\colon g(x) = c\}$. Thus, by the example next to Theorem 3.6 in Polonik (1995), for all small δ,

$$\int_{\{x\colon |g(x)-c|\leq\delta\}} g(x)\,dx \leq a_1\delta$$

for a constant a_1. Since $g(x) \geq c - \delta$ on $\{x\colon |g(x) - c| \leq \delta\}$, (A2.1i) follows.

To prove (A2.1ii) and (A2.1iii) we shall use the fact that $\|\text{grad}\, g\|$ is bounded above. Let $A = \{y\colon g(y) = \lambda\}$ and $k = \sup_x \|\text{grad}\, g(x)\|$. Then for any x, y,

$$|g(y) - g(x)| \leq \|y - x\|k.$$

Thus,

$$\{x\colon |\lambda - g(x)| \leq \delta\} \supseteq \bigcup_{y\in A} \{x\colon \|x - y\| \leq \delta/k\}$$

and

$$\begin{aligned}\{x\colon 0 < \lambda - g(x) \leq \delta\} &= \{x\colon |\lambda - g(x)| \leq \delta\} \cap \{x\colon g(x) - \lambda < 0\} \\ &\supseteq \bigcup_{y\in A} \{x\colon \|x - y\| \leq \delta/k\} \cap \{x\colon g(x) - \lambda < 0\}.\end{aligned}$$

Theorem 2.2 says that $\{x\colon g(x)-\lambda \geq 0\}$ is a convex set and A is its boundary. Hence $\bigcup_{y\in A}\{x\colon \|x-y\| \leq \delta/k\} \cap \{x\colon g(x)-\lambda < 0\}$ is the thin region of width (δ/k) outside the set $\{x\colon g(x)-\lambda \geq 0\}$. Convexity of $\{x\colon g(x)-\lambda \geq 0\}$ implies that the Lebesgue measure of $\bigcup_{y\in A}\{x\colon \|x-y\| \leq \delta/k\} \cap \{x\colon g(x)-\lambda < 0\}$ divided by δ has a positive limit as $\delta \to 0$. Hence (A2.1ii). The proof of (A2.1iii) is similar. □

LEMMA A2.2. *Let* $\gamma_m = \sup_x |g_m(x) - g(x)|$. *Then for a.e. simulation sequence*,

$$\gamma_m = o(m^{-1/(d+2)}\ln(m)).$$

Since g has bounded support, continuous bounded derivative, the result follows as a multidimensional extension of Theorem 3 in Révész (1972).

LEMMA A2.3. $\lambda_m - \lambda = O(\gamma_m)$ *a.e.*

PROOF.

$$\begin{aligned}
&\int g[I\{g \geq \lambda,\ g_m < \lambda_m\} - I\{g < \lambda,\ g_m \geq \lambda_m\}] \\
&\quad = \int g[I\{g \geq \lambda\} - I\{g_m \geq \lambda_m\}] \\
&\quad = \int gI\{g \geq \lambda\} - \int g_m I\{g_m \geq \lambda_m\} + \int (g_m - g)I\{g_m \geq \lambda_m\} \\
&\quad = \int (g_m - g)I\{g_m \geq \lambda_m\} \quad \text{as the first two terms are equal to } 1-p.
\end{aligned} \tag{A2.2}$$

Suppose $(\lambda_m - \lambda)/\gamma_m$ is not bounded above. Then there is a subsequence such that $(\lambda_m - \lambda)/\gamma_m > 1$ through that subsequence. By definition of γ_n,

$$\{g < \lambda,\ g_m \geq \lambda_m\} \subset \{\lambda_m - \gamma_m < g \leq \lambda\}. \tag{A2.3}$$

Hence through that subsequence $\lambda_m - \gamma_m > \lambda$, making the right-hand side of (A2.3) a null set and by (A2.2),

$$\int gI\{g \geq \lambda,\ g_m < \lambda_m\} = \int (g_m - g)I\{g_m \geq \lambda_m\} \leq \text{Leb}(\mathscr{R})\,\gamma_m, \tag{A2.4}$$

since $\{g_m \geq \lambda_m\} \subset \mathscr{R}$. Again,

$$\begin{aligned}
\int gI\{g \geq \lambda,\ g_m < \lambda_m\} &\geq \int gI\{\lambda \geq g < \lambda_m - \gamma_m\} \\
&\geq \lambda \text{Leb}(\{\lambda \leq g < \lambda_m - \gamma_m\}) \\
&\geq b(\lambda_m - \gamma_m - \lambda) \quad \text{by (A2.1ii)}.
\end{aligned}$$

Dividing the above inequality by γ_m, the left-hand side is bounded above by (A2.4), whereas the right-hand side diverges to $+\infty$ through that subsequence. Hence a contradiction to the assumption that $(\lambda_m - \lambda)/\gamma_m$ is not bounded above. Similarly, one can prove that $(\lambda_m - \lambda)/\gamma_m$ is bounded below. Hence the proof is complete. □

PROOF OF THEOREM 2.4. Note that

$$\mathscr{C}_m \triangle \mathscr{C}_{\mathrm{BB}} = \{g_m \geq \lambda_m,\ g < \lambda\} \cup \{g_m < \lambda_m,\ g \geq \lambda\}.$$

By the definition of γ_m,

$$\{g_m \geq \lambda_m,\ g < \lambda\} \subset \{\lambda < g \leq \lambda_m + \gamma_m\},$$
$$\{g_m < \lambda_m,\ g \geq \lambda\} \subset \{\lambda_m - \gamma_m < g \leq \lambda\}.$$

Thus by Lemma A2.2, there is a $K < \infty$ such that $|\lambda_m - \lambda| \leq K\gamma_m$. Hence by Lemma A2.1,

$$\mathrm{Leb}(\mathscr{C}_m \triangle \mathscr{C}_{\mathrm{BB}}) \leq b(K+1)\gamma_m.$$

Hence the proof is complete by (A2.2i). □

A3. Proof of Theorem 3.1. Define a likelihood function and a likelihood ratio on Ω_n, respectively, as

$$\begin{aligned} \tilde{L}_n(w) &= \prod w_i, \\ \tilde{R}_n(w) &= n^n \prod w_i. \end{aligned} \tag{A3.1}$$

LEMMA A3.1. *For every* $0 < r < 1$, *the set*

$$\tilde{\mathscr{C}}_{\mathrm{EL}} = \{\tilde{\mu}(w)\colon w \in \Omega_n,\ \tilde{R}_n(w) \geq r\} \tag{A3.2}$$

is the same as the set $\mathscr{C}_{\mathrm{EL}}$ *defined in* (3.3).

PROOF. By Lemma 1 of Owen (1988), we have

$$\tilde{R}_n(w) \geq r \Longrightarrow R(F_w) \geq r \tag{A3.3}$$

and

$$F \ll F_n,\ R(F) \geq r \Longrightarrow \begin{cases} \text{There is a } w \in \Omega_n \text{ s.t.} \\ \tilde{R}_n(w) \geq r \text{ and } F_w = F, \end{cases} \tag{A3.4}$$

where $F_w = \sum w_i \delta_{X_i}$. Hence $\tilde{\mathscr{C}}_{\mathrm{EL}} \subset \mathscr{C}_{\mathrm{EL}}$ is straightforward from (A3.3). To prove the converse, let $z \in \mathscr{C}_{\mathrm{EL}}$. So there is an $F \ll F_n$ such that $z = \mu(F)$ and $R(F) \geq r$. By (A3.4) there is a $w \in \Omega_n$ such that $\tilde{R}_n(w) \geq r$ and $F_w = F$. Hence $\tilde{\mu}(w) \in \tilde{\mathscr{C}}_{\mathrm{EL}}$. But $\tilde{\mu}(w) = \mu(F_w) = z$. So $\mathscr{C}_{\mathrm{EL}} \subset \tilde{\mathscr{C}}_{\mathrm{EL}}$ and the proof is complete. □

By Lemma A3.1, it is enough to consider $\tilde{\mathscr{C}}_{\mathrm{EL}}$ instead of $\mathscr{C}_{\mathrm{EL}}$. Define

$$f_n(h) = \sup\{\tilde{R}_n(w)\colon w \in \Omega_n,\ w_n = h/n\}.$$

It is easy to see that the supremum in the right-hand side is attained at w^h, where the n-vector w^h is defined as $w_i^h = (1 - h/n)/(n-1)$, $i = 1, \ldots, (n-1)$ and $w_n^h = h/n$. So the f_n defined here is the same as that in (3.9) and

$$\{f_n(h) \geq r\} \Longrightarrow \{\tilde{\mu}(w^h) \in \tilde{\mathscr{C}}_{\mathrm{EL}}\}.$$

Now

$$\tilde{\mu}(w^h) = (h/n)X_n + \{(1-h/n)/(n-1)\}\sum_1^{n-1} X_i$$
$$= (h/n)(X_n - \bar{X}_{n-1}) + \bar{X}_{n-1},$$

so that

$$\|\tilde{\mu}(w^h) - \bar{X}_{n-1}\| = \frac{h}{n}\|X_n - \bar{X}_{n-1}\|.$$

Hence, using the definition of U in (3.5) for U_{EL} and by Observation 3.1,

$$U_{\mathrm{EL}} \geq \sup\{\|\tilde{\mu}(w^h) - \bar{X}_{n-1}\|\colon f_n(h) \geq r\}$$
$$= u_n \frac{\|X_n - \bar{X}_{n-1}\|}{n}.$$

Hence (3.7) is proved.

To prove (3.8), let $h_w = nw_n$ and $\tilde{w} = (w_1/(1-w_n), \ldots, w_{n-1}/(1-w_n)) \in \Omega_{n-1}$. Then for any $w \in \Omega_n$,

$$\|\tilde{\mu}(w) - \bar{X}_{n-1}\| \geq w_n\|X_n - \bar{X}_{n-1}\| - \left\|\sum_1^{n-1} \tilde{w}_i(X_i - \bar{X}_{n-1})\right\| \tag{A3.5}$$

and

$$\tilde{R}_n(w) = f_n(h_w)\tilde{R}_{n-1}(\tilde{w}).$$

As $f_n \leq 1$ and $\tilde{R}_{n-1} \leq 1$,

$$\tilde{R}_n(w) \geq r \Longrightarrow \begin{cases} \text{(i) } \tilde{R}_{n-1}(\tilde{w}) \geq r, \\ \text{(ii) } f_n(h_w) \geq r. \end{cases} \tag{A3.6}$$

The definition of η in (3.4) yields

$$\left\|\sum_1^{n-1} \tilde{w}_i(X_i - \bar{X}_{n-1})\right\| \leq \eta K_{n-1}(r), \tag{A3.7}$$

where

$$K_{n-1}(r) = \sup\left\{\sum_1^{n-1}\left|\tilde{w}_i - \frac{1}{n-1}\right|\colon \tilde{w} \in \Omega_{n-1},\ \tilde{R}_{n-1}(\tilde{w}) \geq r\right\}.$$

By equation (5.1) of Owen (1988),

$$K_{n-1}(r) \leq 2(-2\log r)^{1/2}(n-1)^{-1/2}.$$

Again by (A3.6ii) and Observation 3.1, $\tilde{R}_n(w) \geq r$ implies $h_w \geq l_n$ and

$$w_n\|X_n - \bar{X}_{n-1}\| \geq l_n \frac{\|X_n - \bar{X}_{n-1}\|}{n}.$$

Hence for any $w \in \Omega_n$ with $\tilde{R}_n(w) \geq r$,

$$\|\tilde{\mu}(w) - \bar{X}_{n-1}\| \geq l_n \frac{\|X_n - \bar{X}_{n-1}\|}{n} - \eta 2^{3/2}(-\log r)^{1/2}(n-1)^{-1/2},$$

and (3.8) is proved. □

The proof of (3.10) is routine calculus and is omitted.

A4. Proof of Theorem 3.2. Since all the distances are measured from $\bar{X}_{n-1}$, without loss of any generality one can assume $\bar{X}_{n-1} = 0$. As the BB distribution is the conditional distribution of $\mu(D_n)$, given $X_1, \ldots, X_n$; throughout this proof, the sample sequence $\mathbb{X}$ is fixed and the randomness comes from $(W_1, \ldots, W_n)$. Let V_n denote $\sum_1^n W_i X_i$. Then

$$V_n = W_n X_n + (1 - W_n) \sum_1^{n-1} \tilde{W}_i X_i,$$

where $\tilde{W}_i = W_i/(1 - W_n)$. Identify W_i's in terms of $U_{(i)}$'s, the order statistics of i.i.d. $U(0, 1)$, as in Procedure 1 in Section 2.2. As the joint distribution of $(U_{(1)}/U_{(n-1)}, \ldots, U_{(n-2)}/U_{(n-1)})$ is independent of $U_{(n-1)}$, so the joint distribution of $(\tilde{W}_1, \ldots, \tilde{W}_{n-1})$ is independent of W_n. Let $\tilde{V}_{n-1}$ denote $\sum_1^{n-1} \tilde{W}_i X_i$. Then

$$V_n = W_n X_n + (1 - W_n)\tilde{V}_{n-1},$$

and $\tilde{V}_{n-1}$ is independent of W_n.

Let Z_n denote $(V_n^T X_n)/\|X_n\|$ and $\tilde{Z}_{n-1}$ denote $(\tilde{V}_{n-1}^T X_n)/\|X_n\|$. We shall find a $t_n > 0$ such that $P(Z_n > t_n) \approx \alpha$. To this effect, observe that $|\tilde{Z}_{n-1}| \leq \eta$ and $\|X_n\| > \eta$. Therefore, for a $t > \eta$, using the independence of W_n and $\tilde{Z}_{n-1}$,

$$\begin{aligned} P(Z_n > t) &= E\{P(W_n\|X_n\| + (1 - W_n)\tilde{Z}_{n-1} > t|\tilde{Z}_{n-1})\} \\ &= E\left\{1 - \frac{t - \tilde{Z}_{n-1}}{\|X_n\| - \tilde{Z}_{n-1}}\right\}^{n-1} \\ &= \left(1 - \frac{t}{\|X_n\|}\right)^{n-1} E\left\{1 - \frac{\tilde{Z}_{n-1}}{\|X_n\|}\right\}^{1-n}. \end{aligned}$$

Let $c_n = n(t/\|X_n\|)$. Then

$$\left(1 - \frac{t}{\|X_n\|}\right)^{n-1} \approx e^{-c_n}$$

and

$$\text{(A4.1)} \qquad \left\{1 - \frac{\tilde{Z}_{n-1}}{\|X_n\|}\right\}^{1-n} = 1 + (n-1)\frac{\tilde{Z}_{n-1}}{\|X_n\|} + \frac{(n-1)n}{2}\left\{\frac{\tilde{Z}_{n-1}}{\|X_n\|}\right\}^2 O(1).$$

Note that for every $i = 1, \ldots, (n-1)$,

$$\operatorname{Var}(\tilde{W}_i) = \frac{n-2}{n(n-1)^2},$$

and for $i \neq j$,

$$\operatorname{Cov}(\tilde{W}_i, \tilde{W}_j) = \frac{-1}{n-2} \operatorname{Var}(\tilde{W}_1).$$

Hence for an unit vector $l \in \mathbb{R}^{n-1}$,

$$\begin{aligned} l^T \operatorname{Var}(\tilde{V}_{n-1}) l &= \operatorname{Var}\left(\sum_1^{n-1} \tilde{W}_i l^T X_i \right) \\ &= \frac{n-2}{n(n-1)^2} \left[\sum_1^{n-1} (l^T X_i)^2 - \frac{1}{n-2} \sum_{i \neq j} (l^T X_i)(l^T X_j) \right] \\ &= \frac{n-2}{n(n-1)^2} \left[\frac{n-1}{n-2} \sum_1^{n-1} (l^T X_i)^2 \right] \\ &\leq \eta^2 n^{-1} \quad \text{as } |l^T X_i| \leq \eta. \end{aligned}$$

Again by the construction of $\tilde{Z}_{n-1}$, we obtain

$$E\tilde{Z}_{n-1} = \bar{X}_{n-1} = 0 \tag{A4.2}$$

and

$$E\tilde{Z}_{n-1}^2 = \frac{X_n^T \operatorname{Var}(\tilde{V}_{n-1}) X_n}{\|X_n\|^2} \leq \eta^2 n^{-1}. \tag{A4.3}$$

Using (A4.1) and (A4.2) in (A4.3),

$$E\left\{ 1 - \frac{\tilde{Z}_{n-1}}{\|X_n\|} \right\}^{1-n} = 1 + O(n^{-1})$$

and

$$P(Z_n > t) \approx e^{-c_n}.$$

Hence

$$t_n \approx (-\log \alpha) \frac{\|X_n - \bar{X}_{n-1}\|}{n}.$$

Note that $\tilde{V}_{n-1}$ is confined in a small region around 0 with the diameter η, X_n is far away from 0 and the random variable W_n is concentrated around zero with the density $(n-1)(1-u)^{n-2} I_{[0,1]}(u)$. Hence the density of V_n is high near zero and decreases as we approach X_n. So t_n will serve as an approximate upper bound for U_{BB}. Thus (3.12) is proved. □

Acknowledgments. The author thanks his academic advisor, Professor H. L. Koul, for continuously refining ideas and enriching subject matter. Thanks to Professor A. Dasgupta, Mr. A. White and Professor A. Cuevas for some interesting suggestions and helpful discussion. Thanks to two of the referees for indicating the computational problems in the excess mass approach.

REFERENCES

DEVROYE, L. (1986). *Nonuniform Random Variate Generation.* Springer, New York.

DHARMADHIKARI, S. and JOAG-DEV, K. (1988). *Unimodality, Convexity, and Applications.* Academic Press, New York.

DICICCIO, T., HALL, P. and ROMANO, J. (1991). Empirical likelihood is Bartlett-correctable. *Ann. Statist.* **19** 1053–1061.

EFRON, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.* **7** 1–26.

EFRON, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans.* SIAM, Philadelphia.

FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230.

GASPARINI, M. (1995). Exact multivariate Bayesian bootstrap distributions of moments. *Ann. Statist.* **23** 762–768.

HALL, P. (1987). On the bootstrap and likelihood-based confidence regions. *Biometrika* **74** 481–493.

HARTIGAN, J. A. (1987). Estimation of a convex density contour in two dimensions. *J. Amer. Statist. Assoc.* **82** 267–270.

KARLIN, S., MICCHELLI, C. A. and RINOTT, Y. (1986). Multivariate splines: a probabilistic perspective. *J. Multivariate Anal.* **20** 69–90.

LO, A. Y. (1987). A large sample study for the Bayesian bootstrap. *Ann. Statist.* **15** 360–375.

LO, A. Y. (1988). A Bayesian bootstrap for finite population. *Ann. Statist.* **16** 1684–1695.

LO, A. Y. (1991). Bayesian bootstrap clones and a biometry function. *Sankhyā Ser. A* **53** 320–333.

MICCHELLI, C. A. (1980). A constructive approach to Kergin interpolation in $\mathbb{R}^k$: multivariate B-spline and Lagrange interpolation. *Rocky Mountain J. Math.* **10** 485–497.

OWEN, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75** 237–249.

OWEN, A. B. (1990). Empirical likelihood ratio confidence regions. *Ann. Statist* **18** 90–120.

POLONIK, W. (1995). Measuring mass concentrations and estimating density contour clusters: an excess mass approach. *Ann. Statist.* **23** 855–881.

PRÉKOPA, A. (1973). On logarithmic concave measures and functions. *Acta Sci. Math.* **34** 335–343.

RÉVÉSZ, P. (1972). On empirical density function. *Period. Math. Hungar.* **2** 85–110.

RUBIN, D. B. (1981). A Bayesian bootstrap. *Ann. Statist.* **9** 130–134.

TSYBAKOV, A. B. (1997). On nonparametric estimation of density level sets. *Ann. Statist.* **25** 948–969.

WENG, C. S. (1989). A second-order property of the Bayesian bootstrap mean. *Ann. Statist.* **17** 705–710.

DEPARTMENT OF STATISTICS
UNIVERSITY OF MICHIGAN
1440 MASON HALL
ANN ARBOR, MICHIGAN 48109
E-MAIL: nidhan@umich.edu