This is to certify that the

dissertation entitled

Bayesian Bootstrap Credible Sets for

Multidimensional Mean Functional

presented by

Nidhan Choudhuri

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Statistics

HKoul
Major professor
Hira L. Koul

Date June 15, 1998

Bayesian Bootstrap Credible Sets for Multidimensional Mean Functional

By

Nidhan Choudhuri

A Dissertation

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Statistics and Probability

1998

Abstract

Bayesian Bootstrap Credible Sets for Multidimensional Mean Functional

By

Nidhan Choudhuri

Let $X_1, \ldots, X_n$ be i.i.d. observations from an unknown d-dimensional distribution function F with finite expectation. The aim is to obtain a Bayesian set estimate for the mean of F in a nonparametric setup. Results using several kinds of nonparametric priors can be found in the literature for this purpose. But quantifying the prior information in the form of a nonparametric prior distribution is not an easy task. Besides this, the main difficulty is that often one does not have enough initial information to construct any prior. Hence there is a need for a non-informative nonparametric prior.

Rubin (1981) introduced the concept of Bayesian bootstrap (BB) to express the posterior knowledge about F and its functionals in the absence of any prior information in a nonparametric setup. The justification behind using Bayesian bootstrap as the posterior under a non-informative prior can be found in Rubin (1981) and Gasparini (1995). Hence in the absence of any prior knowledge using the $(1-p)$ central part of BB distribution of the mean functional as a $(1-p)$ level posterior credible set is natural.

This dissertation establishes the existence of a strongly unimodal Lebesgue density for the exact BB distribution of the multidimensional mean functional provided the convex hull of the data has nonempty interior. This result is then used to identify the posterior credible sets at different levels of coverage. Then a two step procedure is described for constructing a credible set. First a Monte-Carlo procedure is used to simulate observations from the BB distribution. Then a histogram smoothing approach is adopted to approximate the posterior credible set. A theorem there proves that for almost every simulation sequence, the simulation based credible set converges to the exact BB credible set at the rate $O(m^{-1/(d+1)} \log m)$ with respect to the metric defined by the Lebesgue measure of the symmetric difference. Here m is the simulation size. The results are then extended to the case when the interior of the convex hull of the data is empty.

The shapes of the credible sets are also investigated. It is found that the shape of a BB credible set is completely determined by data alone and reflects the presence of any skewness in the underlying distribution F. The influence of an outlier to these sets is discussed in great detail. A theorem quantifies the extent of non-robustness by considering how much an outlier can deform a credible set. The effect is proportional to the distance of the outlier from the data cloud and inversely proportional to the sample size n. In this outlier context, a comparison is made with the empirical likelihood ratio (ELR) confidence set (Owen: 1990) and a normal approximation set estimate. Another theorem shows that the effect of an outlier on a ELR confidence set is of the same type as a BB credible set. But the constants of proportionality for the BB credible sets are found to be smaller than those of the ELR confidence sets at every level of coverage and whatever be dimension of the data. The dissertation ends with an argument showing that Bayesian bootstrap can be viewed as the Bayesian counter part of empirical likelihood in a general nonparametric setup.

To My Parents

Acknowledgments

I would like to express my sincere gratitude to my dissertation advisor, Professor Hira L. Koul, for his constant help, advice, encouragement, guidance, mentorship and extreme patience. His caring personality and friendly nature made the whole doctoral experience enjoyable.

I would also like to thank Professor Raoul LePage, Professor Dennis Gilliland and Professor W.T. Sledd for serving on my guidance committee, Professor V. Mandrekar, Processor R.V. Ramamoorthi and Mr. Alex White for their encouragement, suggestions and many helpful conversations. I would also like to thank Professor A. Dasgupta for many informal discussions and suggestions. I would also like to thank Mr. Aditya Vailaya, Mr. Visal Thakkar, Mr. Prasun Sinha and Mr. Samik Sengupta for helping me in writing the C program for the simulations and processing the simulated images.

I cannot thank my parents and my brother enough, for the support and encouragement provided by them. This has been the main motivating force behind all my endeavors. I would also like to thank Professor P. Bhimasankaran and Professor D. Sengupta for the care and interest they showed in my progress as a student at Indian Statistical Institute, and for motivating me into the research.

Table of Contents

List of Tables

LIST OF FIGURES

Chapter 1

Introduction

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

$$\tilde{\mathcal{F}} = \{F \in \mathcal{F} : \int_{\mathbb{R}^d} \|x\| dF(x) < \infty\},$$

and μ be the mean functional on $\tilde{\mathcal{F}}$ defined as

$$\mu(F) = \int_{\mathbb{R}^d} x dF(x) \qquad F \in \tilde{\mathcal{F}}. \tag{0.1}$$

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 $\mathcal{F}$. One assumes F to be a random element of $\mathcal{F}$ according to this prior probability, F_0 to be a particular realization of F and given F, $X_1, \dots, X_n$ 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 non-parametric prior often used in the literature is a Dirichlet process prior with a finite shape measure α (Ferguson: 1973).

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 prior is not an easy task. The need for a non-informative prior to represent vague initial information in non-parametric Bayesian statistics is thus well justified.

Rubin (1981) introduced 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 defined a random distribution function on $\mathbb{R}^d$ as

$$D_n = \sum_{1}^{n} W_i \delta_{X_i}, \tag{0.2}$$

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

$$\Omega_n = \{w \in \mathbb{R}^n : \sum_{i=1}^{n} w_i = 1, w_i \geq 0\} \subset \mathbb{R}^n \tag{0.3}$$

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

Rubin (1981) argued that for a fixed finite sample, the BB distribution of F can be obtained as a weak limit of the posterior distributions under Dirichlet priors when the total mass of the shape measure α tends to zero, i.e. $\alpha(\mathbb{R}^d) \longrightarrow 0$. The results thus obtained are then comparable to standard frequentist results, as illustrated by the applications in Section 5 of Ferguson (1973). Gasparini (1995) proves that the posterior distribution of $\mu(F)$, for large class of Dirichlet priors, converges weakly to $\mu(D_n)$ when $\alpha(\mathbb{R}^d) \longrightarrow 0$. These facts establish the role of Bayesian bootstrap as an non-informative prior in nonparametric Bayesian statistics. Hence, in the absence of any prior knowledge, using the $(1-p)$ central part of 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 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 $(1-p/2)^{th}$ quantile 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 the higher dimension due to the lack of proper definition of quantiles. To identify the central part of a multivariate distribution, it is important to know the nature of the distribution.

This thesis establishes the existence of a strongly unimodal Lebesgue density for the exact BB distribution of the multidimensional mean functional under some mild conditions. 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 (Owen: 1990) in this context. An argument is presented to show that the BB can be thought as a Bayesian counterpart of the empirical likelihood.

The plan for the thesis is as follows. General Bayesian procedure is presented in section 1.1. Section 1.2 and 1.3 establishes the role of BB as a non-informative prior. A brief literature survey ends Section 1.4. Section 2.2 presents the main strong unimodality result and identifies the credible sets. A two step procedure of constructing the credible sets is presented in Section 2.3. The strong unimodality result is then extended to more general case in Section 2.4 and 2.5. Section 3.2 compares BB credible sets with the empirical likelihood confidence sets and the connection between the two procedure is presented in Section 3.5.

Chapter 2

Bayesian Bootstrap in the Nonparametric Bayesian Inference: A Historic Review

2.1 General Bayesian inference and credible set

In a statistical experiment, data is collected following a probability model with an unknown parameter θ, lying in a parameter space Θ. A Bayesian would use a prior probability on Θ, representing his/her prior belief about the unknown parameter θ. Then the prior belief is updated using the data to obtain a posterior belief. For a Bayesian, the posterior distribution encapsulates all that is known about θ following the observations and any inference about θ should be made by analyzing the posterior distribution.

Let $X_1, \ldots, X_n$ be i.i.d. d-dimensional random variable having unknown distribution F_0. Let $\mathcal{B}^d$ be the Borel σ-field in $\mathbb{R}^d$. Consider a family of distribution function $\{F_\theta : \theta \in \Theta\}$ on $(\mathbb{R}^d, \mathcal{B}^d)$ which describes the probability model of the statistical experiment generating each of the X_i's and let θ_0 be the true parameter value such that $F_{\theta_0} = F_0$. A Bayesian approach here is to first obtain an σ-field $\mathcal{A}$ on Θ such that the map $\theta \longrightarrow F_\theta\{B\}$ is $\mathcal{A}$-measurable for every $B \in \mathcal{B}^d$. Then one constructs a prior probability m on $(\Theta, \mathcal{A})$ and assumes that the unknown parameter takes a value according to this prior probability and conditional on θ, $X_1, \ldots, X_n$ are i.i.d. with common distribution F_θ. So one can think that $(\theta_0, X_1, \ldots, X_n)$ is a particular realization from the joint distribution of the parameter and the data, where the parameter value θ_0 is missing. The Bayesian idea is to use $\hat{m}_n(\theta | X_1, \ldots, X_n)$, the conditional distribution of θ given $X_1, \ldots, X_n$, to infer about θ_0. This conditional distribution is known as the posterior distribution.

Let $\gamma : \Theta \longrightarrow \mathbb{R}^q$ be a $\mathcal{A}$-measurable map. The aim is to obtain a set estimate for $\gamma(\theta_0)$. First one obtains the the distribution in $\mathbb{R}^q$ induced by $\hat{m}_n$ under the map γ, known as the posterior distribution of $\gamma(\theta)$. Then one analyses this posterior distribution to obtain a credible set for $\gamma(\theta)$ as is defined below.

DEFINITION 1.1. A *credible set* for $\gamma(\theta)$ of level $1-p$ is a set C contained in the support of the posterior distribution of $\gamma(\theta)$ such that

i) posterior probability of C is $1-p$ and

ii) C represents the central part of the posterior distribution.

Since a credible set represents the central probability concentration region of the posterior distribution, it is natural to use this set as a Bayesian set estimate for $\gamma(\theta_0)$. Note that the choice of C is not unique and depends on the definition on centrality in item (ii).

2.2 Non-parametric Bayes and Dirichlet Process Priors

In many statistical experiments, it is desirable to make fewer assumptions about the underlying population from where the data is obtained than are required for a parametric model.

Non-parametric models are constructed to provide support for more eventualities than are supported by a parametric model. Often one assumes that F_0, the common distribution of X_i's, can be any element of the set $\mathcal{F}$, the class of all distribution functions on $\mathbb{R}^d$, or an element of a large subset of $\mathcal{F}$. For a Bayesian, the first job is to construct a σ-field and a prior probability on $\mathcal{F}$.

The natural σ-field, $\mathcal{A}$, on $\mathcal{F}$ is the smallest σ-field that makes the map $F \longrightarrow F\{B\}$ measurable for every $B \in \mathcal{B}^d$. Since $\mathbb{R}^d$ is a complete separable metric space, one can think of the weak convergence topology on $\mathcal{F}$. A sequence $\{F_r\} \in \mathcal{F}$ converges weakly to $F \in \mathcal{F}$, if and only if, $\int gdF_r \longrightarrow \int gdF$ for all bounded continuous function g on $\mathbb{R}^d$. Under the weak convergence topology, $\mathcal{F}$ becomes a complete separable metric space and the corresponding Borel σ-field is the same as $\mathcal{A}$.

There are several classes of priors on $(\mathcal{F}, \mathcal{A})$ in the literature. The Dirichlet process priors plays a central role in the non-parametric Bayesian analysis.

DEFINITION 1.2. (*Ferguson 1973*) Let α be a non-zero finite measure on $(\mathbb{R}^d, \mathcal{B}^d)$. A probability D_α on $(\mathcal{F}, \mathcal{A})$ is said to be a Dirichlet process with shape measure α if, for every finite measurable partition $\{B_1, \cdots, B_k\}$ of $\mathbb{R}^d$, the random vector $(F\{B_1\}, \cdots, F\{B_k\})$ has a Dirichlet distribution on $[0,1]^k$ with parameters $(\alpha(B_1), \cdots, \alpha(B_k))$.

RESULT 1.1. (*Existence and uniqueness*) For every non-zero finite measure α on $(\mathbb{R}^d, \mathcal{B}^d)$, there exists an unique Dirichlet process measure on $(\mathcal{F}, \mathcal{A})$.

RESULT 1.2. (*Posterior Distribution*) Let F is a random element in $\mathcal{F}$ with a Dirichlet process prior with shape measure α and given F, $X_1, \ldots, X_n$ is i.i.d. F; then the posterior distribution of F is also a Dirichlet process with shape measure $\alpha + \sum_1^n \delta_{X_i}$.

Result 1.1 can be established in many ways. One proof using the Kolmogorov consistency result can be found in Ferguson (1973). The same paper contains a proof of Result 1.2.

2.3 Non-informative priors and Bayesian bootstrap

Non-informative priors are of some special interest in Bayesian literature. In a parametric case, non-informative priors are obtained as a limit of a sequence of priors when the prior information is decreasing through the sequence. In the nonparametric case, one can start with a sequence of Dirichlet priors.

Let F be a random element in $\mathcal{F}$ with a prior distribution D_α. Let $\alpha^\star = \alpha(\mathbb{R}^d)$, the total measure of α and $\bar{\alpha} = \alpha/\alpha^\star$, the normalized probability measure. Then for every $B \in \mathcal{B}^d$, $F\{B\}$ has a Beta distribution $B(\alpha(B), \alpha^\star - \alpha(B))$. Hence

$$\begin{aligned} \mathbb{E}F\{B\} &= \alpha(B)/\alpha^\star = \bar{\alpha}(B) \\ VarF\{B\} &= \bar{\alpha}(B)(1 - \bar{\alpha}(B))/(\alpha^\star + 1) \end{aligned}$$

So $\bar{\alpha}$ is the center of D_α in $\mathcal{F}$ and can be thought as a prior guess for F. $\alpha^\star$ controls the variance of F. A large value of $\alpha^\star$ implies D_α is concentrated near $\bar{\alpha}$ and a small value of $\alpha^\star$ implies D_α is widely distributed. With D_α priors, the posterior expectation of F is

$$\mathbb{E}(F|X_1, \ldots, X_n) = \frac{\alpha + \sum_1^n \delta_{X_i}}{\alpha^\star + n} = \frac{\alpha^\star}{\alpha^\star + n}\bar{\alpha} + \frac{n}{\alpha^\star + n}F_n,$$

which is a weighted average of the prior guess and empirical distribution. Hence $\alpha^\star$ can be thought as an index of confidence on the prior probability and letting $\alpha^\star$ tend to 0, one can expect to have a non-informative prior.

In parametric situation, by taking limits of a reasonable sequence of priors, one often ends up with an improper prior instead of a probability measure. But still the procedure become useful in the sense that the sequence of posteriors often result in a probability measure as their limit, which has no influence of prior guess and is completely determined by the data.

Since $\mathcal{F}$, equipped with weak convergence topology, is a complete separable metric space, one can introduce the notion of convergence in distribution of a sequence of prior and posterior probabilities. We need to see what happens to a sequence of posterior distributions under Dirichlet priors D_{α_r} when $\alpha_r^\star \longrightarrow 0$. The following convergence result will be useful in this regard.

RESULT 1.3. (*Sethuraman and Tiwari 1982*) Let $\{\alpha_r\}$ be a sequence of finite measures on $(\mathbb{R}^d, \mathcal{B}^d)$ such that

$$\sup_{B \in \mathcal{B}^d} |\alpha_r(B) - \alpha_0(B)| \longrightarrow 0$$

for some non-zero finite measure α_0 on $(\mathbb{R}^d, \mathcal{B}^d)$. Then D_{α_r} converges in distribution to D_{α_0} on $\mathcal{F}$.

RESULT 1.4. Let D_{α_r} be a sequence of Dirichlet priors on $(\mathcal{F}, \mathcal{A})$ such that $\alpha_r^\star \longrightarrow 0$. Then the sequence of posterior distributions of F converges in distribution to the BB distribution of F on $\mathcal{F}$ for any data set $X_1, \ldots, X_n$.

PROOF: Under D_{α_r} prior, the posterior is $D_{\alpha_r + nF_n}$. For every $B \in \mathcal{B}^d$,

$$|(\alpha_r + nF_n)(B) - (nF_n)(B)| = \alpha_r(B) \leq \alpha_r^\star.$$

This implies that

$$\sup_{B} |(\alpha_r + nF_n)(B) - (nF_n)(B)| \longrightarrow 0 \qquad \text{as} \qquad \alpha_r^\star \longrightarrow 0.$$

Hence by Result 1.3, the posterior distribution of F converges to D_{nF_n}. Note that D_{nF_n} is the BB distribution of F and this completes the proof. □

The above result proves the claim in Rubin (1981) that the BB distribution of F can be obtained as a weak limit of Dirichlet process priors when the faith on the prior is decreasing to zero and hence BB distribution can be thought of as a non-informative posterior in a nonparametric set up.

Since our objective is to infer about the mean functional, it would be interesting to know what happens to the posterior distributions of $\mu(F)$ in the above case. Since mean functional is not continuous on $\mathcal{F}$ with respect to the weak convergence topology, the limiting behavior of $\mu(F)$ cannot be obtained from Result 1.4. Gasparini (1995) has a result in this regard.

RESULT 1.5. Let $\alpha_r = \alpha_r^\star \bar{\alpha}$ be a sequence of non-zero finite measure on $(\mathbb{R}^d, \mathcal{B}^d)$ such that $\bar{\alpha}$ is a probability measure with

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

and $\alpha_r^\star \longrightarrow 0$. Then the posterior distribution of $\mu(F)$ under D_{α_r} priors converges in distribution to the BB distribution of $\mu(F)$ on $(\mathbb{R}^d, \mathcal{B}^d)$.

Result 1.4 and Result 1.5 establish the role of Bayesian bootstrap as a non-informative prior in a nonparametric setup.

2.4 Other perspective of Bayesian bootstrap

Asymptotic equivalence of BB distribution and the posterior distribution under a Dirichlet prior with non zero α has been noticed earlier. Lo (1987) showed that in the one dimension case, the posterior distribution of F for 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) pointed 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 finite has third moment. Thus one can approximate the posterior distribution under a Dirichlet prior by the BB distribution. This approximation becomes useful as it is very easy to simulate from BB distribution.

The operational and structural similarities between BB and 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 BB, whereas they are replaced by some discrete weights in ordinary bootstrap. Rubin (1981) 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) proved 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 finite population case were obtained in Lo (1988). This gives a frequentist perspective of BB method.

In the case of a finite support set $\{d_1, \cdots, d_k\}$ for F, a vector θ with $\theta_j = F\{d_j\}$, $F\{d_j\}$ being the probability of singleton set $\{d_j\}$, uniquely identifies F. Hence the space of all probability measure on $\{d_1, \cdots, d_k\}$ 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 non-informative 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 value of unobserved d_i's, as pointed out in Owen (1990). This gives a justification behind using BB as an non-informative prior in finite support case.

Owen (1990) introduced the concept of empirical likelihood as a nonparametric generalization of the well studied parametric likelihood and used this concept to construct confidence sets and test statistics for several nonparametric functionals. He observed that in the finite support case the empirical likelihood is proportional to the BB density. He thus argued in favor of connecting empirical likelihood with the posterior under a non-informative prior as in the parametric case. This argument is extended for the general nonparametric set up in the Section 3.5 of this Thesis.

Chapter 3

The BB distribution and credible sets for the mean functional

The role of Bayesian bootstrap in the nonparametric inference may inspire one to use the BB distribution of $\mu(F)$ for constructing a set estimate for $\mu(F_0)$ by means of a BB credible set. As we have seen in the introduction, there are some difficulties in choosing the central probability concentration region of a multivariate distribution. Besides, one is also concerned about the shape of the region. Since one is using this credible set as a set estimate of the unknown mean, a connected region is preferred than a union of some disjoint sets. Since mean is a convex sum of the points in the support of a distribution function, convexity of a set estimate is desired. Besides desire on the shape, the size of the credible set should be as small as possible (in terms of Lebesgue measure) to make the estimate precise. Along with these, one needs to remember that a credible set should have the required $(1-p)$ posterior coverage probability. These are the issues to be remembered while constructing the Bayesian bootstrap credible set.

3.1 A normal approximation credible set

Let $\mathbb{X}$ denote the sample sequence $\{X_1, X_2, \dots\}$, 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

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

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* $\mathbb{X}$,

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

where Σ *is the dispersion matrix of* F_0.

PROOF: The proof is just a multidimensional extension of its one dimensional version in Lo (1987), Theorem 4.1, which says that if $\int x^2 dF_0 < \infty$, then for almost every sample sequence $\mathbb{X}$,

$$(2.3) \qquad \sqrt{n}\{\mu(D_n) - \bar{X}_n\}|\mathbb{X} \Longrightarrow N(0, \sigma^2),$$

where σ^2 is the variance of F_0.

Now for every $l \in \mathbb{R}^d$, by (2.1)

$$\int_{\mathbb{R}^d} (l^T x)^2 dF_0(x) \leq \|l\|^2 \int_{\mathbb{R}^d} \|x\|^2 dF_0(x) < \infty.$$

So by (2.3) for almost every sample sequence $\mathbb{X}$,

$$\begin{aligned} l^T(\sqrt{n}\{\mu(D_n) - \bar{X}_n\})|\mathbb{X} &= \sqrt{n}\{\sum W_i(l^T X_i) - l^T \bar{X}_n\}|\mathbb{X} \\ &\Longrightarrow N(0, \sigma_l^2), \end{aligned}$$

where $\sigma_l^2 = \text{Var}(l^T X_1) = l^T \Sigma l$.

Hence by the Cramér-Wold device, the proof is complete. □

If Σ is of full rank, then one can substitute for Σ with 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

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

as an approximate central high probability concentration region of $\Lambda_{n,X}$. This set is the smallest set (in terms of Lebesgue measure) with limiting coverage probability $(1-p)$ and also has the desired convex shape. Hence A_p can be used as a credible set of level $(1-p)$ for $\mu(F)$.

But there are some limitations of this asymptotic procedure. First, the convergence in (2.2) is in the first order sense as indicated by Weng (1989). So A_p does not reflect any higher order moment structure of $\Lambda_{n,X}$ such as skewness. Besides A_p is always elliptical in shape. If the data shows a clear skewness, it may be difficult to have faith on an elliptical set estimate for the mean. A data determined shape is thus more desirable. Note that A_p is the same as the frequentist confidence set obtained by Hotelling's T^2-distribution up to a scale factor, i.e. the cut off point $\chi^2_{d,1-p}$ is replaced by a multiple of some quantile of an F-distribution. If the posterior distribution under a non-informative prior is the prime object, then it is important to find the central part of the exact BB distribution.

3.2 An exact BB credible set

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 $\{x \in \mathbb{R}^d : g(x) \geq c\}$ 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 high density contours are convex and surrounded by low probability region. Hence a high density contour in some sense represent the central high probability concentration region and can be used as a credible set. If we can prove the strong unimodality of $\Lambda_{n,X}$, than a high density contour can be used as a BB credible set for the mean.

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

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

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

PROOF: Fix an $c > 0$. To prove strong unimodality, we have to show that $\{x : f(x) \geq c\}$ is a convex set. Let $x, y \in \{x : f(x) \geq c\}$. Then $f(x) \geq c$ and $f(y) \geq c$. So for any $t \in [0,1]$, by (2.4),

$$f(tx + (1-t)y) \geq c^t c^{1-t} = c.$$

Hence $tx + (1-t)y \in \{x : f(x) \geq c\}$ which implies that $\{x : f(x) \geq c\}$ is convex. This completes the proof. □

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

A probability distribution F on $\mathbb{R}^d$ is said to be *non-singular* if $F\{H\} < 1$ for every hyperplane H. 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 i.e. F_n is non-singular. If F_0 is non-singular then for almost every sample sequence $\mathbb{X}$, there is an N, depending on the sample sequence $\mathbb{X}$, such that F_n is non-singular 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 one, F_n is non-singular for every $n \geq (d+1)$. Hence the condition of Theorem 2.2 is satisfied in most of the cases and a BB credible set could be obtained through high density contours. The rest of section establishes the proof of Theorem 2.2.

An affine in $\mathbb{R}^d$ is a subset $\mathcal{M}$ of $\mathbb{R}^d$ such that for every $x, y \in \mathcal{M}$ and $-\infty < t < \infty$ we have $tx + (1-t)y \in \mathcal{M}$ i.e. the entire line passing through x and y are in $\mathcal{M}$. If $0 \in S$, then $\mathcal{M}$ 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

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

$\mathcal{H}(A)$ is the smallest affine containing A. For a $t \in [0,1]$ and two nonempty sets A and B, define a convex combination of these two sets as

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

DEFINITION 2.2. [*Prékopa (1973), eqno 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 $t \in [0,1]$

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

To prove Theorem 2.2, first we shall show that the distribution $\Lambda_{n,X}$ is logconcave and non-singular and then we shall use the standard logconcavity results to get the logconcave density. Lemma 2.1-2.4 develops the required machinery for this purpose.

PROPOSITION 2.2. P is nonsingular on $\mathbb{R}^k$ if and only if the affine hull of its support is the whole $\mathbb{R}^k$.

LEMMA 2.1. *Let P be a logconcave probability on $\mathbb{R}^k$ and $L : \mathbb{R}^k \to \mathbb{R}^s$ be an affine transform. Then PL^{-1} is logconcave on $\mathbb{R}^s$.*

PROOF: See Dharmadhikari and Joag-dev (1988), Lemma 2.1, page 47. □

LEMMA 2.2. *Let P be a non-singular probability on $\mathbb{R}^k$. Then P is logconcave if and only if P has a logconcave Lebesgue density on $\mathbb{R}^k$.*

PROOF: See Dharmadhikari and Joag-dev (1988), Theorem 2.8, page 51. □

LEMMA 2.3. *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. Hence the joint distribution of $(W_1, \cdots, W_{n-1})$ has a Lebesgue density

$$f(u) = I(u_1 + \cdots + u_{n-1} \leq 1), \qquad u = (u_1, \cdots, u_{n-1}) \in \mathbb{R}^{n-1}.$$

The function f is logconcave as the indicator function of any convex set is logconcave. The support of f is $\{u_1 + \cdots + u_{n-1} \leq 1\}$ whose affine hull is the whole $\mathbb{R}^{n-1}$. Hence by proposition 2.2, the joint distribution of $(W_1, \cdots, W_{n-1})$ is non-singular. Then by Lemma 2.2, the joint distribution of $(W_1, \cdots, W_{n-1})$ is logconcave on $\mathbb{R}^{n-1}$. Since the map g :

$\mathbb{R}^{n-1} \longrightarrow \mathbb{R}^n$, defined as

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

is an affine transform and $g(W_1, \cdots, W_{n-1}) = (W_1, \ldots, W_n)$, the proof is complete by Lemma 2.1. □

Let us define a maps on Ω_n of (0.3) as

$$\tilde{\mu}(w) = \sum w_i X_i = \mu(F_w),$$

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

LEMMA 2.4. *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. Assume that the convex hull of $\{X_1, .., X_n\}$ has a non-empty interior. Then $P_n \tilde{\mu}^{-1}$ is non-singular 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 $H_n = \{w \in \mathbb{R}^n : \sum w_i = 1\}$

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

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

where $\tilde{H}^n = \{w \in \mathbb{R}^n : \sum (a^T X_i) w_i = c\}$ is a hyperplane in $\mathbb{R}^n$. Since the affine hull of the support of P_n is the hyper plane 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, $\tilde{H}^n = \{w \in \mathbb{R}^n : \sum(a^T X_i)w_i = 0\}$ 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 $\tilde{H}^n = \{w \in \mathbb{R}^n : \sum(a^T X_i/c)w_i = 1\}$. Hence $\tilde{H}^n = H_n$ iff $(a^T X_i)/c = 1$ for all $i = 1, \dots, n$. Thus $\tilde{H}^n = H_n$ implies that $X_i \in \{x \in \mathbb{R}^d : a^T x = c\}$, which again implies that the convex hull of $\{X_1, .., X_n\}$ lies inside a hyperplane. This contradicts the assumption that convex hull of $\{X_1, .., X_n\}$ has 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 2.3) and $\tilde{\mu}$ is a linear map from $\mathbb{R}^n$ to $\mathbb{R}^d$, by Lemma 2.1, $\Lambda_{n,X}$ is logconcave on $\mathbb{R}^d$.

As the affine hull of the support of Γ_n is H_n, by Lemma 2.4, $\Lambda_{n,X}$ is non-singular on $\mathbb{R}^d$. Hence by Lemma 2.2, the proof is complete. □

COROLLARY 2.1. *Let the weights $(W_1, \dots, W_n)$ in (0.2) be replaced by some other weights $(W_1^*, \dots, W_n^*)$, 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 $X_1, \dots, X_n$ has non-empty interior, the distribution of $\mu(D_n)$ has logconcave Lebesgue density.*

3.3 Constructing the confidence region

When the convex hull of $X_1, \dots, X_n$ has 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 $X_1, \dots, X_n$ are fixed. Let g be the logconcave Lebesgue density of $\Lambda_{n,X}$ and

$$C_{BB} = \{x \in \mathbb{R}^d : g(x) \geq \lambda\} \tag{2.6}$$

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

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

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

Procedure 2. Define $W_i = Y_i / \sum_1^n Y_i, \quad i = 1, \ldots, n$; 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 distribution on Ω_n are known in the literature for long time and proofs can be found in Devroye (1986) page 207-210. Procedure 2 is much easier to perform on a computer while Procedure 1 is useful in proving some theoretical results on BB distribution.

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 i^{th} 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 will use the $(1-p)$ high density contour of g_m as an approximation to $\mathcal{C}_{BB}$.

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

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

where $X_i^{(l)}$ be the l^{th} component of the i^{th} observation X_i. Then the hyper rectangle

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

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

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

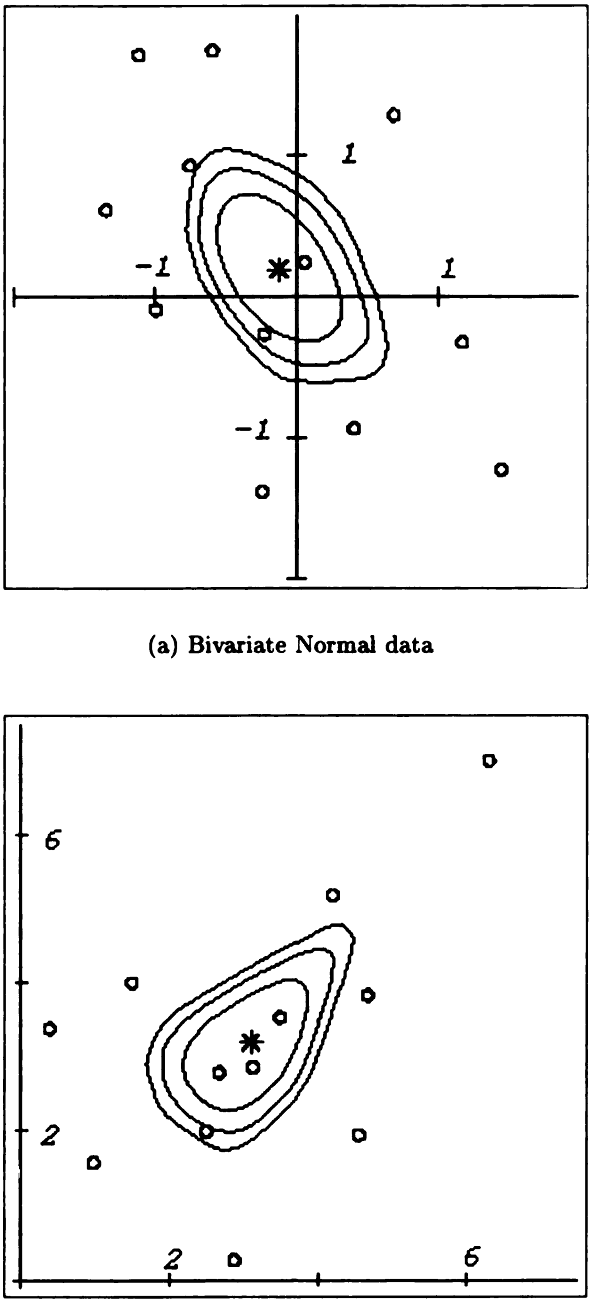
STEP 2. *For the data $X_1, \ldots, X_n$, obtain the set $\mathcal{R}$. Choose $h = m^{-1/(d+2)}$ and obtain the cells $\mathcal{R}_h$. Calculate $\tau(x)$ for each $x \in \mathcal{R}_h$ and order the cells according the descending order of $\tau(x)$. Let $\{x_1, \cdots, x_k\}$ denote the ordered cells,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

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

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

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

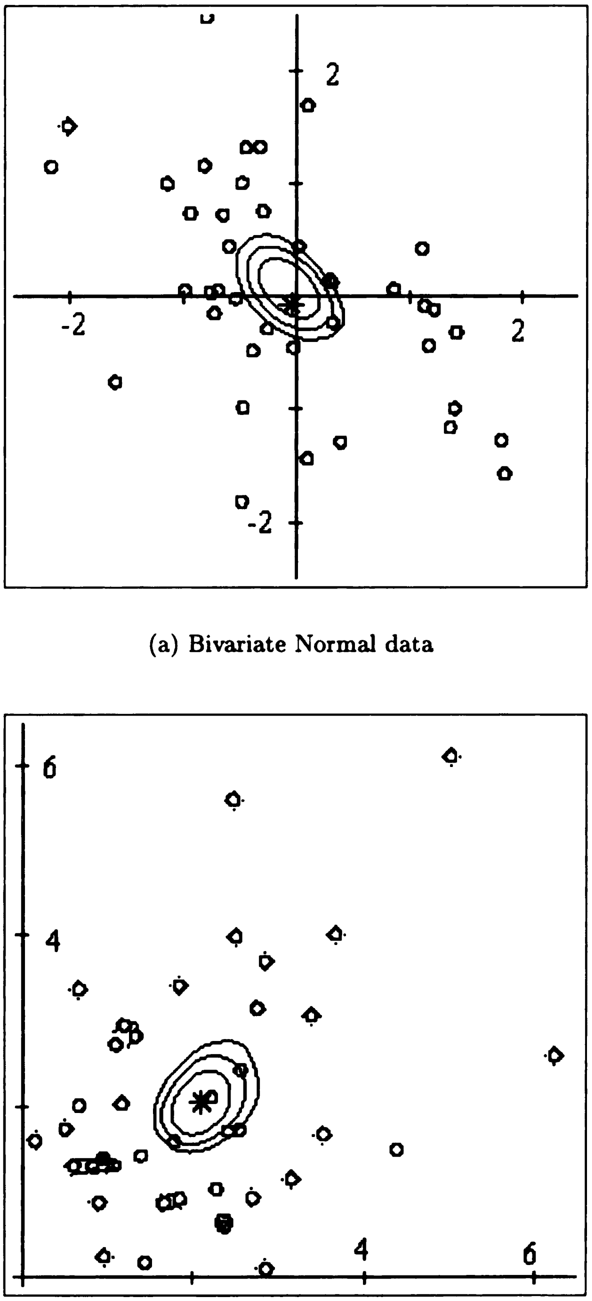
as an approximation to C_{BB}.



(a) Bivariate Normal data

(b) Bivariate Gamma data

Figure 3.1: *Small sample BB credible sets of levels* 80%, 95%, *and* 99% *for the mean.*



(a) Bivariate Normal data

(b) Bivariate Gamma data

Figure 3.2: *Large sample BB credible sets of levels* 80%, 95%, *and* 99% *for the mean.*

Some simulation results are presented in Figure 1 using the simulation size $m = 200,000$ for each case. Figure 2.1(a) shows the BB credible sets with confidence level 80%, 95% and 99% for the twelve observation from a 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 2.1(b) shows these credible sets based on twelve observation form a skewed bivariate distribution with density $f(u,v) = uve^{-(u+v)}$, $u, v > 0$, i.e bivariate Gamma. Note that these credible sets are able to reflect the skewness of the underlying distribution in their shape. The moderate sample BB credible sets taking $n = 40$ from these two distribution are presented in Figure 2.2(a) and 2.2(b). These credible sets for Gamma observations in Figure 2.2(b) are almost elliptical in shape with very little skewness. This is expected as 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 cells $k = \prod_1^d [(b_l - a_l)m^{-1/(d+2)} + 1] \simeq cm^{d/(d+2)}$. Hence the calculation of $\tau(x)$ for $x \in \mathcal{R}_h$ will take time proportional to km which is of the order smaller than $o(m^2)$. Ordering of cells 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 making 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. But the convergence rate of C_m decreases with an increase in the dimension and one needs to use a larger m to achieve the same resolution.

To measure the performance of C_m in approximating C_{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' denote the Lebesgue measure on $\mathbb{R}^d$. Then we have the following convergence result on $\mathcal{C}_m$.

THEOREM 2.3 *If the convex hull of any $n-1$ data pints has nonempty interior, then for a.e. simulation sequence,*

$$d_1(\mathcal{C}_m, \mathcal{C}_{BB}) = O(m^{-1/(d+2)}\ln(m)).$$

The rest of this section is dedicated to the proof of Theorem 2.3. Throughout this proof, the data $X_1, \ldots, X_n$ is fixed and the randomness is coming from the randomness in the simulation. Recall that the hyper cubes $\{R(x,h) : x \in R_h\}$ cover $\mathcal{R}$ and are disjoint except at the boundaries. Then the function g_m on $\mathbb{R}^d$ defined as

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

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

We shall identify the BB density g with a multivariate B-spline to obtain some smoothness result. Here is a probabilistic definition of B-spline, taken from Section 2 of Karlin, Micchelli and Rinott (1986).

DEFINITION 2.3. Let $X_1, \ldots, X_n$ be n points in $\mathbb{R}^d$ such that their convex hull has

a non-empty interior. Then there is a function $M_n(\cdot|X_1, \ldots, X_n) : \mathbb{R}^d \longrightarrow [0, \infty)$ with $X_1, \ldots, X_n$ as parameters, such that for every bounded continuous function $h : \mathbb{R}^d \longrightarrow \mathbb{R}$,

$$(2.8) \qquad (n-1)! \int_{\Omega_n} h(w_1 X_1 + \cdots + w_n X_n) dw = \int_{\mathbb{R}^d} h(x) M_n(x|X_1, \ldots, X_n) dx.$$

Here $\Omega_n = \{w \in \mathbb{R}^n : \sum w_i = 1, w_i \geq 0\}$ as defined in (0.3) and the LHS integration is with respect the $n-1$ dimensional Lebesgue measure on Ω_n. The function $M_n(\cdot|X_1, \ldots, X_n)$ is called a B-spline function with knots $X_1, \ldots, X_n$.

One can see that the LHS of (2.8) is $\mathbb{E}[h(Z)|X_1, \ldots, X_n]$, where $Z = W_1 X_1 + \cdots + W_n X_n$ and $(W_1, \ldots, W_n)$ are uniform on Ω_n. Thus the RHS of (2.8) implies that, $M_n(\cdot|X_1, \ldots, X_n)$ is a version of the Lebesgue density of the distribution of Z on $\mathbb{R}^d$. Note that Z is nothing but a BB mean and hence $M_n(\cdot|X_1, \ldots, X_n)$ is also a version of the BB density g. Thus one can use standard B-splines results on g. By Corollary 3 of Micchelli (1980) and its extension in the adjacent paragraph, we deduce the following result.

RESULT 2.1. *If the convex hull of any $n-1$ data points $X_1, \ldots, X_n$ has nonempty interior, then g has a continuous derivative on $\mathbb{R}^d$. The derivative vector $\bigtriangledown g$ is bounded above and is also non-zero in the interior of the support of g except at the mode of g.*

LEMMA 2.5. *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*

$$(2.9) \qquad \begin{array}{rlcl} (i) & \mathrm{Leb}\{x : |g(x) - \lambda| \leq \delta\} & \leq & b_1\delta, \\ (ii) & \mathrm{Leb}\{x : 0 < \lambda - g(x) \leq \delta\} & \geq & b_2\delta, \\ (iii) & \mathrm{Leb}\{x : 0 < g(x) - \lambda \leq \delta\} & \geq & b_3\delta. \end{array}$$

PROOF: First we shall prove (2.9.*ii*) and (2.9.*iii*) using the fact that $\| \bigtriangledown g \|$ is bounded above (Result 2.1). Let $A = \{y : \ g(y) = \lambda\}$ and $k = \sup_x \| \bigtriangledown g(x)\|$. Then for any x, y,

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

Thus,

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

and

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

Theorem 2.2 says that $\{x : \ g(x) - \lambda \geq 0\}$ is a convex set and A is its boundary. Hence $\cup_{y \in A}\{x : \|x - y\| \leq \delta/k\} \cap \{x : \ g(x) - \lambda < 0\}$ is the thin region of width (δ/k) outside the set $\{x : \ g(x) - \lambda \geq 0\}$. Convexity of $\{x : \ g(x) - \lambda \geq 0\}$ implies that the Lebesgue measure of $\cup_{y \in A}\{x : \|x - y\| \leq \delta/k\} \cap \{x : \ g(x) - \lambda < 0\}$ divided by δ has a positive limit as $\delta \longrightarrow 0$. Hence for small δ, there exists a b_2 such that

$$\text{Leb}[\cup_{y \in A}\{x : \|x - y\| \leq \delta/k\} \cap \{x : \ g(x) - \lambda < 0\}] \geq b_2\delta.$$

This completes the proof of (2.9.*ii*). The proof of (2.9.*iii*) is similar.

To prove (2.9.*i*), we shall use the fact that $\bigtriangledown g$ is nonzero. Continuity of g implies that A is a closed set and $\lambda > 0$ implies that A is in the interior of the support of g, which is a bounded set. Hence A is a compact set. Since A does not contain the mode of g, by Result 2.1, $\| \bigtriangledown g(y)\|$ is continuous and positive on A. So the compactness of A implies that

$\| \bigtriangledown g(y)\|$ is bounded away from zero on A.

Let $\inf_{y\in A} \| \bigtriangledown g(y)\| = k_1$. First we shall show, by contradiction, that

$$\{x : \ |g(x) - \lambda| \leq \delta\} \subseteq \bigcup_{y\in A} \{x : \|x - y\| \leq 3\delta/k_1\}.$$

Suppose this is not true. Then there exists an x_0 such that $|g(x_0)-\lambda| \leq \delta$ but $\|x_0-y\| > 3\delta/k_1$ for all $y \in A$. Then either $g(x_0) < \lambda$ or $g(x_0) > \lambda$. Assume $g(x_0) < \lambda$. Let y_0 be the closest point on A such that the line joining y_0 and x_0 is perpendicular to the tangential plane to A passing through y_0. Then the vector $\bigtriangledown g(y_0)$ has the same direction as the vector (y_0-x_0). Hence $(\bigtriangledown g(y_0))^T(y_0-x_0) = \|\bigtriangledown g(y_0)\| \cdot \|y_0-x_0\| \geq k_1\|y_0-x_0\|$. The continuity of $\bigtriangledown g$ implies that, when x is in a neighborhood of y_0, $(\bigtriangledown g(x))^T(y_0-x_0) > (k_1/2)\|y_0-x_0\| > (3/2)\delta$.

By the fundamental theorem of calculus,

$$\begin{aligned} g(y_0) - g(x_0) &= \int_0^1 \frac{\partial}{\partial t} g(ty_0 + (1-t)x_0)dt \\ &= \int_0^1 (\bigtriangledown g(ty_0 + (1-t)x_0))^T(y_0 - x_0)dt \\ &> \int_0^1 (3/2)\delta dt \\ &> \delta. \end{aligned}$$

Hence a contradiction. The case $g(x_0) > \lambda$ is handled similarly.

Note that $\bigcup_{y\in A}\{x : \|x-y\| \leq 3\delta/k_1\}$ is a thin region of width $3\delta/k_1$ around A. Convexity of $\{x : \ g(x) - \lambda \geq 0\}$ implies that the Lebesgue measure of $\cup_{y\in A}\{x : \|x - y\| \leq 3\delta/k_1\}$ divided by δ has a positive limit as $\delta \longrightarrow 0$. Hence for small δ, there exists a b_1 such that

$$\text{Leb}[\cup_{y\in A}\{x : \|x - y\| \leq 3\delta/k_1\}] \leq b_1\delta.$$

The proof of (2.9.i) follows from the fact that

$$\{x : |g(x) - \lambda| \leq \delta\} \subseteq \bigcup_{y \in A}\{x : \|x - y\| \leq 3\delta/k_1\}. \qquad \square$$

LEMMA 2.6. *Let* $\gamma_m = \sup_x |g_m(x) - g(x)|$. *Then for a.e. simulation sequence*

$$\gamma_m = o(m^{-1/(d+2)}\ln(m)). \tag{2.9}$$

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

LEMMA 2.7. $\lambda_m - \lambda = O(\gamma_m)$ *a.e.*

PROOF: Because $\int gI(g \geq \lambda) = 1 - p = \int g_m I(g_m \geq \lambda_m)$, we obtain

$$\begin{aligned} &\int g\{I(g \geq \lambda, g_m < \lambda_m) - I(g < \lambda, g_m \geq \lambda_m)\} \\ &\quad = \int g\{I(g \geq \lambda) - I(g_m \geq \lambda_m)\} \\ &\quad = \int gI(g \geq \lambda) - \int g_m I(g_m \geq \lambda_m) + \int (g_m - g)I(g_m \geq \lambda_m) \\ &\quad = \int (g_m - g)I(g_m \geq \lambda_m). \end{aligned} \tag{2.11}$$

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

$$\{g < \lambda, g_m \geq \lambda_m\} \subset \{\lambda_m - \gamma_m < g \leq \lambda\}. \tag{2.12}$$

Since $\lambda_m - \gamma_m > \lambda$ through that subsequence, the event in the RHS of (2.12) is a null set. Thus by (2.11), through that subsequence,

$$\int gI(g \geq \lambda, g_m < \lambda_m) = \int (g_m - g)I(g_m \geq \lambda_m) \leq \text{Leb}(\mathcal{R})\,\gamma_m, \tag{2.13}$$

as $\{g_m \geq \lambda_m\} \subset \mathcal{R}$. Again,

$$\begin{aligned}\int gI(g \geq \lambda, g_m < \lambda_m) &\geq \int gI(\lambda \leq g < \lambda_m - \gamma_m) \\ &\geq \lambda \text{Leb}(\{\lambda \leq g < \lambda_m - \gamma_m\}) \\ &\geq [b_2(\lambda_m - \lambda - \gamma_m)] \wedge \delta_0,\end{aligned}$$

where the last inequality follows from (2.9.ii). Dividing the above inequality by γ_m through that subsequence, by (2.13) we obtain that the LHS is bounded above, whereas by our assumption, the RHS diverges to $+\infty$. This contradicts our assumption that $(\lambda_m - \lambda)/\gamma_m$ is not bounded above. Similarly one can prove that $(\lambda_m - \lambda)/\gamma_m$ is also bounded below. Thus the proof is complete. □

Proof of Theorem 2.4. Note that

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

By definition of γ_n,

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

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

By Lemma 2.6, there is a $K > 0$ such that $|\lambda_m - \lambda| \leq K\gamma_m$. Hence by (2.9.i),

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

Hence the proof is complete by Lemma 2.6. □

3.4 The case of singular data

One can also proceed even if the data does not have a non-empty interior. In this case, all the X_i's are confined in an affine of $\mathbb{R}^d$. Let $\mathcal{H}_0$ be the affine hull of the data set $\{X_1, \ldots, X_n\}$ and s be the dimension of $\mathcal{H}_0$. Then we have the following theorem.

THEOREM 2.4 *If $0 < s < d$, then $\Lambda_{n,X}$ is absolutely continuous with respect to the s-dimensional Lebesgue measure restricted to $\mathcal{H}_0$ 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 on $\mathcal{H}_0$. Then there exists a bijective affine map $\mathbb{L} : \mathcal{H}_0 \longrightarrow \mathbb{R}^s$ 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 $\{Y_1, \cdots, Y_n\}$ is the image of the affine hull of $X_1, \ldots, X_n$ under $\mathbb{L}$, i.e. the entire $\mathbb{R}^s$. Hence the convex hull of $\{Y_1, \cdots, Y_n\}$ has non-empty 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 data $\{Y_1, \cdots, Y_n\}$. 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 $\frac{d\Lambda_{n,X}}{d\tilde{\lambda}_s}$. Affine property of $\mathbb{L}$ and logconcave property of g_s implies $\tilde{g}$ is logconcave on $\mathcal{H}_0$. □

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 $\frac{d\Lambda_{n,X}}{d\tilde{\lambda}_s}$ 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 $\mathcal{H}_0$. The high density contours of g_s may depend on $\mathbb{L}$ but their inverse images in $\mathcal{H}_0$ 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 $\mathcal{H}_0$.

To apply this result, one needs to find an $\mathbb{L}$. As the affine hull of data has dimension s, the rank of the sample dispersion matrix S_n is s and S_n has exactly s non-zero eigen-values. Take a spectral decomposition of S_n, and let $e_1, \cdots, e_s$ be the eigen vectors corresponding to the non-zero eigen-values. Then a candidate for $\mathbb{L}$ is

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

In this case the inverse function $\mathbb{L}^{-1} : \mathbb{R}^s \longrightarrow \mathcal{H}_0$ is of the form $\mathbb{L}^{-1}(y) = \sum_1^s y^i e_i + \bar{X}$.

3.5 Extension to linear functionals

The above logconcavity result can be extended to the BB distribution of a linear functional. Let $\varphi : \mathbb{R}^d \to \mathbb{R}^q$ be a Borel measurable function and μ_φ be a linear functional on $\mathcal{F}$ 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 data $Y_1, \cdots, Y_n$. Hence all the results follow for linear functionals. The case $q > d$ will be taken care of by Theorem 2.4.

Chapter 4

Connection with Empirical Likelihood

4.1 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 $F \in \mathcal{F}$ 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 well known 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 $\mathcal{F}$. Consider sets of the form $\{\theta(F) : F \ll F_n,\ R(F) \geq r\}$ 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

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

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

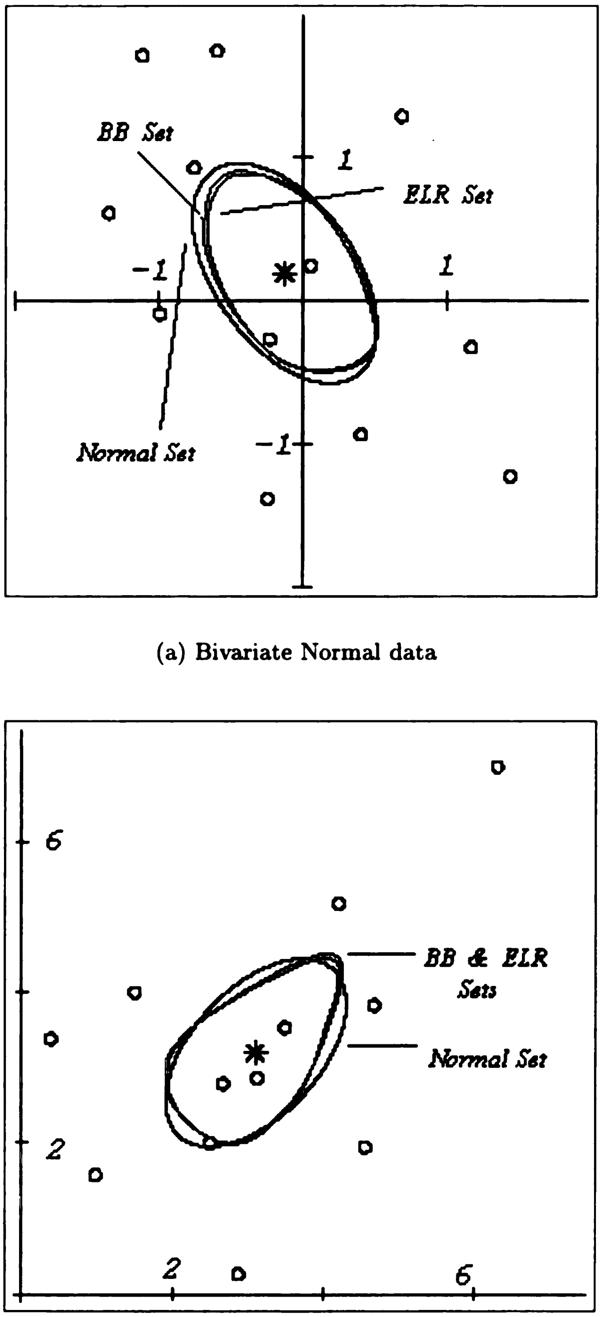
RESULT 3.1. *If F_0 has finite second moment, i.e. (2.1) holds, and the dispersion matrix Σ is of rank $s > 0$, then for every $0 < r < 1$, $\mathcal{C}_{EL}$ is a convex set and*

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

If one chooses $r = \exp\{-\frac{1}{2}\chi^2_{s,p}\}$, then $\mathcal{C}_{EL}$ serves as a confidence set for $\mu(F_0)$ with the (frequentist) asymptotic coverage $1 - \alpha$. 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 functional can be found in Owen (1990).

4.2 Comparison of BB credible sets with ELR Confidence sets in the presence of an outlier

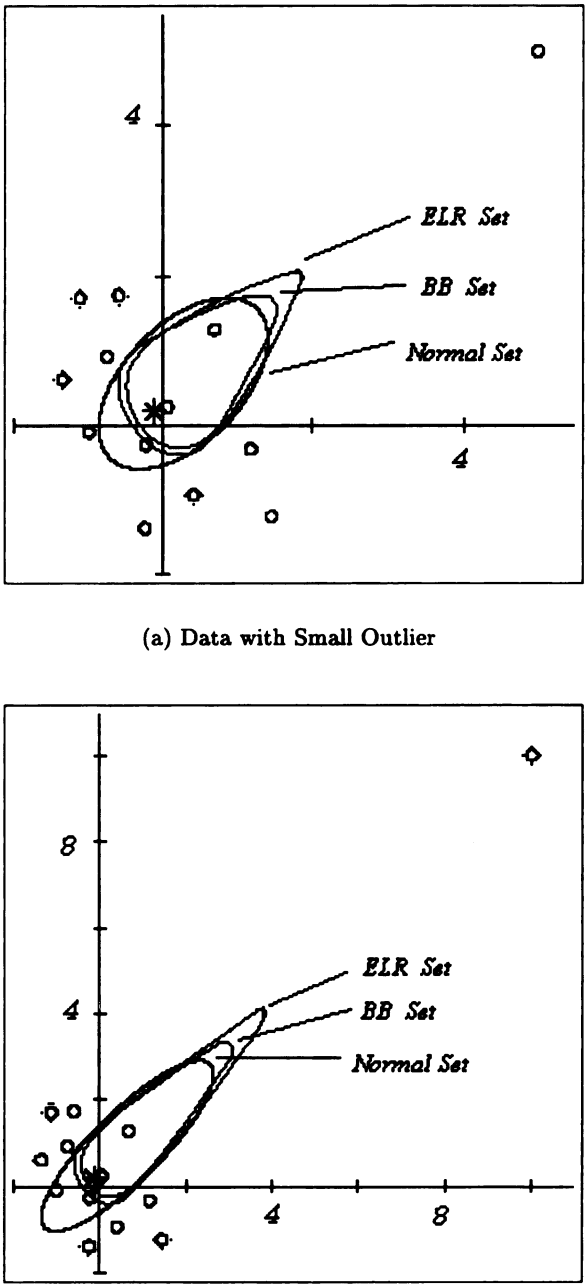
One advantage of both the BB and EL methods for constructing set estimate is that the shapes of these sets are completely determined by the data. These sets are also able to



(a) Bivariate Normal data

(b) Bivariate Gamma data

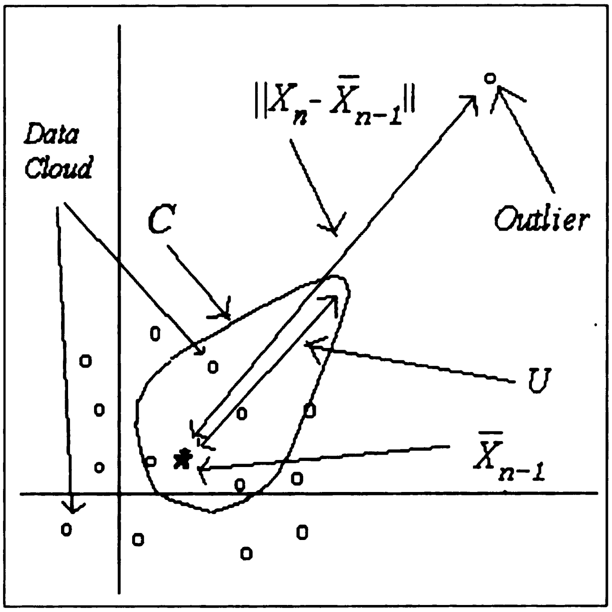
Figure 4.1: *The BB, ELR and Normal approximation credible sets of level* 95%.



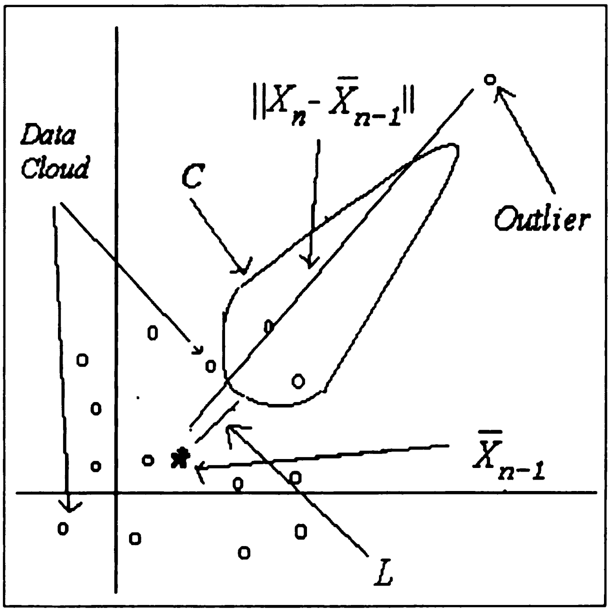
(a) Data with Small Outlier

(b) Data with Large Outlier

Figure 4.2: *The three credible sets of level* 95% *for a normal data set with an outlier.*



(a) Measuring inflation effect



(b) Measuring shift effect

Figure 4.3: *Diagram to identify the inflation effect and the sift effect of an outlier*

incorporate the skewness of the data in their shapes and hence in turn capture the skewness of the underlying distribution. Figure 3.1(a) shows 95% confidence sets using the BB, ELR and normal approximation methods based on twelve normal observations used in Section 2.3. Figure 3.1(b) shows these three sets based on the twelve observations from the skewed distribution used in Section 2.3. 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.

But a problem with the BB and ELR methods is that both the regions are sensitive to outliers. This can be seen from Figure 3.2(a) and 3.2(b). An outlier (not random) is added to the twelve normal observations used earlier. Figure 3.2(a) and 3.2(b) shows 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 BB credible set is less than that of the ELR confidence set, while the normal approximation method is least effected. A quantitative study of the extent of the sensitivity of BB and ELR confidence sets is done here.

We will define two measures of non-robustness 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 n^{th} observation X_n be such that $\|X_n - \bar{X}_{n-1}\|$ is large compared to

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

Then call $X_1, \ldots, X_{n-1}$ as the data cloud and X_n as an outlier. A diagram in two dimensions

($d = 2$) is presented in Figure 3.3. Let C be an arbitrary set estimate for μ based on all observations including the outlier. To measure the inflation of C, introduce the quantity

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

which is the distance of the farthest point in C from $\bar{X}_{n-1}$. (See Figure 3.3b). Note that, large U signifies C has a long nose towards the outlier X_n and one can conclude that the outlier has inflated the region C towards itself. Whereas small U implies less effect of X_n on C.

Sometimes the influence of an outlier is so much that the whole region shifts away from the data cloud towards the outlier. (See Figure 3.3b). We say that C has shifted from the data cloud if $\bar{X}_{n-1} \notin C$ and we measure the shift by the quantity

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

$L = 0$ implies no shift and the vice-verse. Large L implies the region C has largely shifted from the data cloud, indicating 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 3.2(a) and 3.2(b), both L_{BB} and L_{EL} are zero indicating that there is no shift effect of the outlier. But there is a large extent of inflation effect. Here we give some theoretical bounds on U_{BB}, U_{EL} and L_{EL}.

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

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

and

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

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

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

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

$$l_n = l_0 + o(n^{-1}) \qquad \text{and} \qquad u_n = u_0 + o(n^{-1}), \tag{3.10}$$

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 solution, 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 is an increasing function of the degrees of freedom. Thus u_n decreases 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*

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

Table 4.1: The values of u_n in (3.7) along with u_0 for dimension 2 and 3, and the 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	$(-\log p)$
.01	5.95392	7.63835	6.60939	8.85321	4.605170
.05	4.79608	5.74386	5.48027	6.82846	2.995732
.10	4.21403	4.88972	4.89876	5.90078	2.302585
.20	3.55979	3.99431	4.23016	4.91262	1.609438

Theorem 3.1 and 3.2 help one in comparing the non-robustness of the BB credible sets and the ELR confidence sets. The extent of inflation on both type of sets is proportional to the distance of outlier from the data cloud and is inversely proportional to the sample size n. u_n in (3.7) describes the constant of proportionality for an ELR set and increases with the increase in the dimension of the data as well as with the 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 be 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 level of coverage error p. Since $u_n \longrightarrow u_0$, as $n \longrightarrow \infty$, the values of u_0 are also attached. These observations indicates some robustness advantage for 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 $L_{BB} = 0$ unless the outlier is too big. L_{EL} is also equal to zero for small magnitude of the outlier. The first term in the lower bound (3.8) is of the order $O(n^{-1})$ whereas the second term is of the order $O(n^{-1/2})$. Hence the RHS is negative often making the inequality trivial. This implies that the shift effect of an outlier on both these set estimates are negligible, but the inflation effect is very much prominent.

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

4.3 Proof of Theorem 3.1.

Define a likelihood function and a likelihood ratio on Ω_n respectively as

$$\begin{aligned} \tilde{L}_n(w) &= \prod w_i \\ \tilde{R}_n(w) &= n^n \prod w_i. \end{aligned} \tag{3.13}$$

LEMMA 3.1. *For every* $0 < r < 1$, *the set*

$$\tilde{C}_{EL} = \{\tilde{\mu}(w) : w \in \Omega_n,\ \tilde{R}_n(w) \geq r\} \tag{3.14}$$

is the same as the set C_{EL} *defined in (3.3).*

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

$$\tilde{R}_n(w) \geq r \Longrightarrow R(F_w) \geq r \tag{3.15}$$

and

$$F \ll F_n,\ R(F) \geq r \Longrightarrow \begin{cases} \text{There is an } w \in \Omega_n\ s.t. \\ \tilde{R}_n(w) \geq r \text{ and } F_w = F, \end{cases} \tag{3.16}$$

where $F_w = \sum w_i \delta_{X_i}$. Hence $\tilde{C}_{EL} \subset C_{EL}$ follows from (3.15). To prove the converse, let $z \in C_{EL}$. So there is an $F \ll F_n$ such that $z = \mu(F)$ and $R(F) \geq r$. By (3.16) there is a $w \in \Omega_n$ such that $\tilde{R}_n(w) \geq r$ and $F_w = F$. Hence $\tilde{\mu}(w) \in \tilde{C}_{EL}$. But $\tilde{\mu}(w) = \mu(F_w) = z$. So $C_{EL} \subset \tilde{C}_{EL}$ and the proof is complete. □

By Lemma 3.1, it is enough to consider $\tilde{C}_{EL}$ instead of C_{EL}. Define

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

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

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

Now

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

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,

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

This proves (3.7). To prove (3.8), let $h_w = nw_n$ and $\tilde{w} = (w_1/(1-w_n), \ldots, w_{n-1}/(1-w_n)) \in$

Ω_{n-1}. Then for any $w \in \Omega_n$,

$$(3.17) \qquad \|\tilde{\mu}(w) - \bar{X}_{n-1}\| \geq w_n \|X_n - \bar{X}_{n-1}\| - \|\sum_1^{n-1} \tilde{w}_i (X_i - \bar{X}_{n-1})\|$$

and

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

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

$$(3.18) \qquad \tilde{R}_n(w) \geq r \Longrightarrow \begin{cases} i) & \tilde{R}_{n-1}(\tilde{w}) \geq r, \\ ii) & f_n(h_w) \geq r. \end{cases}$$

By using the definition of η in (3.4), the second term of (3.17) can be bounded above as

$$(3.19) \qquad \|\sum_1^{n-1} \tilde{w}_i (X_i - \bar{X}_{n-1})\| \leq \eta K_{n-1}(r)$$

where

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

By equation (5.1) of Owen (1988),

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

Again by (3.18.ii) and Observation 3.1, $\tilde{R}_n(w) \geq r$ implies $h_w \geq l_n$ and the first term of (3.17) can be bounded bellow as

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

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.

4.4 Proof of Theorem 3.2.

Since all the distances are measured from $\bar{X}_{n-1}$, without loss of generality we 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 2.3. 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 will find a $t_n > 0$ such that $P(Z_n > t_n) \approx p$. To this effect observe that, $|\tilde{Z}_{n-1}| \leq \eta$ and for large n, $\|X_n\| > \eta$.

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

$$\left\{1 - \frac{\tilde{Z}_{n-1}}{\|X_n\|}\right\}^{1-n} = 1 + (n-1)\frac{\tilde{Z}_{n-1}}{\|X_n\|} + \frac{(n-1)n}{2}\left\{\frac{\tilde{Z}_{n-1}}{\|X_n\|}\right\}^2 \cdot O(1). \tag{3.20}$$

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

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

and for $i \neq j$,

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

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

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

$$
\begin{aligned}
&= \frac{n-2}{n(n-1)^2}\left[\frac{n-1}{n-2}\sum_1^{n-1}(l^T X_i)^2\right] \\
&\le \eta^2 n^{-1} \quad \text{as} \quad |l^T X_i| \le \eta.
\end{aligned}
$$

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

(3.21) $$E\tilde{Z}_{n-1} = \bar{X}_{n-1} = 0,$$

(3.22) $$E\tilde{Z}^2_{n-1} = \frac{X_n^T \mathrm{Var}(\tilde{V}_{n-1}) X_n}{\|X_n\|^2} \le \eta^2 n^{-1}.$$

Using (3.20) and (3.21) in (3.22), one obtains

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

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

Hence

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

Note that $\tilde{V}_{n-1}$ is confined in a small region around 0 with diameter η, X_n is far away from 0 and the random variable W_n is concentrated around zero with the density

$$(n-1)(1-u)^{n-2} I_{[0,1]}(u).$$

So 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}. Hence (3.12) is proved.

4.5 Connection Between EL and BB methodology.

Let $\mathcal{P}$ denote the class of all finite measurable partition of $\mathbb{R}^d$. For any $\pi_1, \pi_2 \in \mathcal{P}$, say $\pi_1 \preceq \pi_2$ (read π_2 is finer than π_1) if π_2 is a refinement of π_1, i.e. the elements of π_2 are obtained by partitioning some or all of the elements of π_1. Then $\preceq$ defines a partial order on $\mathcal{P}$ and $\mathcal{P}$ is a directed set under $\preceq$. (A directed set is a set along with a partial order $\preceq$ such that for every two elements π_1 and π_2, there is an element π with $\pi_1 \preceq \pi$ and $\pi_2 \preceq \pi$.) Now for every element $\pi = \{A_1, \cdots, A_k\} \in \mathcal{P}$, define a map $g_\pi : \mathcal{F} \longrightarrow \Omega_k$ as

$$g_\pi(F) = (F\{A_1\}, \cdots, F\{A_k\}),$$

where Ω_k is the unit simplex in $\mathbb{R}^k$ defined in (1.3). Then for every $F \in \mathcal{F}$, the collection of vectors $\{g_\pi(F) : \pi \in \mathcal{P}\}$ uniquely identifies F. Let $\mathcal{A}$ denote the Borel σ-algebra generated by the weak convergence topology on $\mathcal{F}$. Then g_π is $\mathcal{A}$-measurable and $\mathcal{A}$ is the smallest σ-algebra under which every g_π is measurable.

Now for i.i.d. data set $X_1, \ldots, X_n$ in $\mathbb{R}^d$, define a net of non-negative functions on $\mathcal{F}$ as

$$(3.23) \qquad L_\pi(F) = \prod_{i=1}^{n} \left\{ \sum_{j=1}^{k} F\{A_j\} 1_{\{X_i \in A_j\}} \right\}, \qquad F \in \mathcal{F},$$

where $\pi = \{A_1, \cdots, A_k\}$ is a k-partition of $\mathbb{R}^d$. (See Kelley: 1955, page 65, for the definition of net.) Then the net of functions L_π converges pointwise to the empirical likelihood function $L(F)$ in (3.1). To see this, take $\pi_0 = \{\{X_1\}, \cdots, \{X_n\}, \mathbb{R}^d - \{X_1, \ldots, X_n\}\}$. Then $L_{\pi_0}(F)$ is exactly equal to $L(F)$ and hence the limit is achieved at any finite stage π with $\pi_0 \preceq \pi$.

Now fix an arbitrary $\pi = \{A_1, \cdots, A_k\}$. If one is interested in knowing only the probabilities of $A_1, \cdots, A_k$ under F, then the problem reduces to a multinomial problem with cell probabilities equal to $g_\pi(F)$. Let n_j denote the number of X_i's in A_j. Then the multinomial likelihood of the parameters $g_\pi(F)$ is proportional to

$$\prod_{j=1}^{k}(F\{A_j\})^{n_j},$$

which is the same as L_π in (3.23). Since the collection of vectors $\{g_\pi(F) : \pi \in \mathcal{P}\}$ uniquely identifies F, one can think $g_\pi(F)$ as the finite stage parameter value of F and L_π as the finite stage likelihood function at stage π. So at a finite stage π, for the multinomial problem, one can put a prior density κ on $g_\pi(F)$ and obtain a posterior density proportional to $L_\pi(F)\kappa(g_\pi(F))$. A common choice of non-informative prior in multinomial case is an improper prior with Lebesgue density

$$\kappa(g_\pi(F)) = \prod_{j=1}^{k}(F\{A_j\})^{-1}.$$

Under this prior, the posterior density is proportional to

$$\prod_{j=1}^{k}(F\{A_j\})^{n_j-1} =: \tilde{\kappa}(g_\pi(F)), \tag{3.24}$$

and for any set A_j with $n_j = 0$, we have $F\{A_j\} = 0$ with posterior probability one. Thus we get a collection of posterior densities $\tilde{\kappa}(g_\pi(F))$. The question is, does this collection of densities lead to any probability on $(\mathcal{F}, \mathcal{A})$. The answer is affirmative.

One can use the Kolmogorov consistency result to prove the existence of a probability on $(\mathcal{F}, \mathcal{A})$. But the converse approach is easier here. Recall the $\mathcal{F}$ valued random variable D_n defined in (1.2), i.e. the BB distribution of F. Note that for any finite measurable partition $\pi = \{A_1, \cdots, A_k\}$, the distribution of $(D_n\{A_1\}, \cdots, D_n\{A_k\})$ is Dirichlet distribution with parameter $(n_1, \cdots, n_k)$ and the density is proportional to (3.24). So these collection of posteriors lead us to the BB distribution on $(\mathcal{F}, \mathcal{A})$. Since the empirical likelihood is obtained as the limit of the finite stage multinomial likelihood and the BB distribution is obtained from the collection of posterior probabilities under a non-informative prior of these multinomial problems, one can think of BB distribution as the posterior under a non-informative prior incorporated with empirical likelihood in the entire parameter space $\mathcal{F}$. This is the extension of Owens argument in finite support case to general case. Note that this arguments is valid if one has the observations on an arbitrary locally compact topological space $\mathcal{X}$.

Bibliography

[1] DHARMADHIKARI, S. and JOAG-DEV, K. (1988). *Unimodality,Convexity, and Applications.* Academic press.

[2] DEVROYE, L. (1986) *Nonuniform random variate generation.* Springer-Verlag, New York-Berlin.

[3] DICICCIO, T. and HALL, P. and ROMANO, J. (1991). Eempirical likelihood is Bartlett-correctable. *Ann. Statist.* **19** 1053-1061.

[4] EFRON, B. (1979). Bootstrap methods: Another look at the jackknife *Ann. Statist.* **9** 1-26.

[5] EFRON, B. (1982). *The jackknife, the Bootstrap and Other Resampling Plans.* SIAM, Philadelphia.

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

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

[8] HARTIGAN, J.A. (1987). Estimation of a convex density contour in two dimension. *J. Amer. Stat. Asso.* **82** 267-270.

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

[10] KELLEY, J.L. (1955). *General Topology.* Van Nostrand Reinhold Co., New York.

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

[12] LO, A.Y. (1988). A Bayesian bootstrap for finite population. *Ann. Statist.* **16** 1684-1695.

[13] LO, A.Y. (1991). Bayesian bootstrap clones and a biometry function. *Sankhya Ser. A* **53** 320-333.

[14] 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.

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

[16] OWEN, A.B. (1990). Empirical likelihood ratio confidence regions. *Ann. Statist* **18** 90-120.

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

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

[19] RÉVÉSZ, P. (1972). On empirical density function. *Period. math. Hunger.* **2** 85-110.

[20] RUBIN, D.B. (1981). A Bayesian bootstrap. *Ann. Statist.* **9** 130-134.

[21] SETHURAMAN,J. and TIWARI,R.C. (1882). Convergence of Dirichlet measures and the interpretation of their parameter. In *Statistical Decision Theory and Related Topics 3* (S.S.Gupta and J.O.Berger, eds.) **2** 305-315. Academic Press, New York.

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

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