A Festschrift for Herman Rubin
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 45 (2004) 291–308

On the strong consistency, weak limits and practical performance of the ML estimate and Bayesian estimates of a symmetric domain in R^k

Wen-Chi Tsai[1] **and Anirban DasGupta**[2]

National Cheng-Chi University and Purdue University

Abstract: This paper considers a problem of estimating an unknown symmetric region in R^k based on n points randomly drawn from it. The domain of interest is characterized by two parameters: size parameter r and shape parameter p. Three methods are investigated which are the maximum likelihood, Bayesian procedures, and a composition of these two. A modification of Wald's theorem as well as a Bayesian version of it are given in this paper to demonstrate the strong consistency of these estimates. We use the measures of symmetric differences and the Hausdorff distance to assess the performance of the estimates. The results reveal that the composite method does the best. Discussion on the convergence in distribution is also given.

1. Introduction

It is a pleasure to write this article for Professor Rubin's Festschrift. I cannot begin to enumerate the things I have learned from him, and the number of times I walked into his office or he walked into mine, drew up a chair, and started a conversation, and opened my eyes. This paper itself is a prime example of how much I benefitted from him in my student days at Purdue.

In biology, the size and shape of home range within a community of a species of animal are often a starting point for the analysis of a social system. In forestry, estimating the geographical edge of a rare species of plant based on sighting of individuals is an important issue as well. The need to estimate an unknown domain by using a set of points sampled randomly from it can also be seen in many other disciplines. See Macdonald et al. (1979), Seber (1986, 1992), and Worton (1987).

If one considers the shape of the unknown domain an infinite-dimensional parameter, the convex hull of the sample will be the maximum likelihood solution. Most of the literature hence focuses on the studies of the convex hull and the results are all for one dimensional and two dimensional regions. Refer to Ripley et al. (1977), Moore (1984), and Bräker et al. (1998).

However, if we use these results in some other applications, for example, recognizing the valid region of predictor variables, which usually involves more than two dimensions, we will then encounter some difficulties in implementation. As the dimensionality rises to higher than three dimensions, where a simple visual illustration is impossible, describing the convex hull of a sample becomes much more

[1]National Cheng-Chi University, 64, Sec. 2, Zhi-nan Rd., Wenshan, Taipei 116, Taiwan, Republic of China. e-mail: `wctsai@nccu.edu.tw`

[2]Purdue University, Department of Statistics, 150 N. University Street, West Lafayette, IN 47907-2068, USA. e-mail: `dasgupta@stat.purdue.edu`

Keywords and phrases: ball, Bayes, convex, domain, Euclidean space, extreme value, maximum likelihood, simulation, strong consistency, weak limit.

AMS 2000 subject classifications: 62F10, 62F12, 62F15.

difficult. Hence, a more practical approach for estimating a higher dimensional domain is necessary. Due to this consideration, we would like to characterize the shape of a domain by a finite-dimensional parameter rather than using a non-parametric model to which most literature is devoted. Besides, it is easier to establish properties of estimates of the set of interest under parametric modelling. This would make us more comfortable using these estimates.

Since the configuration of a roughly spherical object is easier characterized, we would like to start our investigation with a particular family of sets, the L_p balls, because of their richness in fitting roughly rounded objects and in deriving pilot theoretical inference.

Let $B_{p,r}$ denote the centered L_p ball with radius r with respect to the metric induced from p-norm in the k-dimensional Euclidean space, R^k; namely

$$B_{p,r} = \{\underset{\sim}{x} \in R^k : \|\underset{\sim}{x}\|_p \le r\}, \tag{1}$$

here

$$\|\underset{\sim}{x}\|_p = \begin{cases} \left(|x_1|^p + \cdots + |x_k|^p\right)^{1/p} & \text{when } p \text{ is finite,} \\ \max(|x_1|, \ldots, |x_k|) & \text{when } p \text{ is infinite.} \end{cases}$$

We call $\|\cdot\|_p$ the p-norm operator. The unknown set S we wish to estimate will be assumed to be an L_p ball; namely $S = B_{p_0,r_0}$ for some $0 < p_0 \le \infty$ and $0 < r_0 < \infty$.

Notice that in our approach, the center of symmetry of the domain S is assumed to be known. This will not be exactly true in practice. A short discussion is given in the last section.

Also notice that, when the dimension k equals one, the family of L_p balls becomes the family of closed intervals $[-r, r]$ in the real line. Our one dimensional version of estimating an L_p ball can be viewed as the well known "end-point" problem: estimating the end points a and b by using points randomly selected from $[a, b]$. Also, p does not play any role in characterizing the set S which we wish to estimate when $k = 1$. Therefore throughout this paper, we will take $k \ge 2$. However, the one dimensional case often lends much intuition to the case of higher dimensions.

Now let $\underset{\sim}{x}_1 = (x_{11}, \ldots, x_{1k})', \ldots, \underset{\sim}{x}_n = (x_{n1}, \ldots, x_{nk})'$ denote a realization of n points from the domain S. We would like to estimate S by using these observations $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$. We will assume that $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$ are independently uniformly drawn from S. It is possible in practice that $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$ are independently drawn from S not uniformly but following a measure μ on R^k other than the Lebesgue measure, truncated to S with finite $\mu(S)$; i.e. $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n \overset{i.i.d.}{\sim} \frac{\mu(\cdot)}{\mu(S)}$. There will be no problem in deriving similar results which we establish in this article if μ is known and for which $B_{p,r}$ is identifiable. However, if μ is unknown, estimating S becomes much more difficult. The reason is that we will be unable to distinguish between a rare event (e.g. the density with respect to μ at a point $\underset{\sim}{x}$ is small) and a null event (e.g. point $\underset{\sim}{x}$ is not in the support S); see Hall (1982).

To summarize, we have taken an interesting problem and analyzed an interesting parametric model. We have given two very general results on strong consistency, and additional results on weak convergence as well as practical evaluation by very detailed numerics. We have indicated how to possibly address more general cases and commented on application. These are the main contributions.

2. Estimation

As the domain S which we wish to estimate is characterized by parameters p and r, a plug-in method can be used to estimate S. We will consider three natural methods of estimation for p and r: maximum likelihood method, a Bayesian approach, and a combination of these two methods.

The maximum likelihood estimates have a drawback that they underestimate the volume of the true set with probability one and the magnitude of this bias is difficult to evaluate. The Bayesian approach does not have this underestimating problem. However, they are hard to calculate. That is not uncommon in a Bayesian analysis. An alternative approach which combines the maximum likelihood estimate and the Bayesian approach is therefore proposed. This approach treats the volume of the true set as a parameter and estimates it using a Bayesian method. Then it corrects the maximum likelihood estimates for their biases accordingly. We are excited about this approach.

Let us now look at the maximum likelihood method in detail first. Recall that $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$ are uniformly drawn from S. Thus the likelihood function of p and r is

$$L(p, r|\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n) = \frac{1}{\lambda(B_{p,r})^n} \mathbf{1}_{\{(p,r): \underset{\sim}{x}_i \in B_{p,r} \ \forall \ i=1,\ldots,n\}}; \tag{2}$$

here λ is the Lebesgue measure. The formula for the Lebesgue volume of $B_{p,r}$ is

$$\lambda(B_{p,r}) = 2^k r^k \frac{\Gamma(1+\frac{1}{p})^k}{\Gamma(1+\frac{k}{p})}; \tag{3}$$

(see Gradshteyn and Ryzhik (1994), p. 647). If we denote the maximum likelihood estimate of (p, r) by $(\hat{p}_{mle}, \hat{r}_{mle})$, then we have

$$(\hat{p}_{mle}, \hat{r}_{mle}) = \arg\max_{(p,r)} L(p, r|\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n) = \arg \min_{\{(p,r): \underset{\sim}{x}_i \in B_{p,r} \forall i=1,\ldots,n\}} \lambda(B_{p,r}). \tag{4}$$

Moreover, as $\lambda(B_{p,r})$ is an increasing function of r for any fixed p, $(\hat{p}_{mle}, \hat{r}_{mle})$ must satisfy

$$\hat{r}_{mle} = \max_{1 \le i \le n} \|\underset{\sim}{x}_i\|_{\hat{p}_{mle}},$$

and hence

$$\hat{p}_{mle} = \arg\max_p \left(2^k \left(\max_{1 \le i \le n} \|\underset{\sim}{x}_i\|_p \right)^k \frac{\Gamma(1+\frac{1}{p})^k}{\Gamma(1+\frac{k}{p})} \right)^{-1}.$$

The profile likelihood of p mostly appears to be unimodal and therefore it is usually not difficult to obtain $\hat{p}_{mle}$ and $\hat{r}_{mle}$ numerically.

Despite this easy characterization of the maximum likelihood estimate, there is a disadvantage in using this estimate. Consider the end-point problem. Suppose $x_1, \ldots, x_n$ are iid Unif([a,b]). It is well known that the maximum likelihood set estimate of $[a, b]$, $[x_{(1)}, x_{(n)}]$, is always contained in the true interval. And therefore the length of the estimated support $[x_{(1)}, x_{(n)}]$ is always shorter than the true length $b - a$. Similarly, when the dimension $k > 1$, the volume of the maximum likelihood set estimate $B_{\hat{p}_{mle}, \hat{r}_{mle}}$ is always smaller than the true volume $\lambda(B_{p_0, r_0})$. The reason is that the maximum likelihood set estimate $B_{\hat{p}_{mle}, \hat{r}_{mle}}$ is the L_p ball which possesses the smallest volume among L_p balls containing all the observations. On the other hand, the true domain evidently contains all the observations. Therefore, we have $\lambda(B_{\hat{p}_{mle}, \hat{r}_{mle}}) \le \lambda(B_{\hat{p}_0, \hat{r}_0})$.

Here we would like to point out that unlike the end-point problem (or the nonparametric setting) where the maximum likelihood interval estimate is always contained in the true interval, the maximum likelihood set estimate $B_{\hat{p}_{mle}, \hat{r}_{mle}}$ does not need to be inside the true set all the time.

Now let us move to the Bayesian approach. We will choose the loss function being

$$l_\lambda(S, \widehat{S}\,) = \lambda(S \triangle \widehat{S}\,) \tag{5}$$

where $\triangle$ denotes the symmetric difference operator. If we denote the prior of (p, r) by π, the posterior of (p, r) after observing $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$ is

$$\pi(p, r|\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n) \propto \pi(p, r)\frac{1}{\lambda(B_{p,r})^n}\mathbf{1}_{\{(p,r)\,:\,\underset{\sim}{x}_i \in B_{p,r} \forall 1 \leq i \leq n\}}.$$

Thus the Bayesian estimate based on the loss function (5) is

$$(\hat{p}_{bayes}, \hat{r}_{bayes}) = \arg\min_{(\hat{p},\hat{r})} E_{\pi(p,r|\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_n)}\big(\lambda(B_{p,r} \triangle B_{\hat{p},\hat{r}})\big). \tag{6}$$

Though we are able to show theoretically that $(\hat{p}_{bayes}, \hat{r}_{bayes})$ is strongly consistent and does not have the underestimating problem like $(\hat{p}_{mle}, \hat{r}_{mle})$ does, however the computation of $(\hat{p}_{bayes}, \hat{r}_{bayes})$ is difficult. The reason is that we do not have a formula of $\lambda(B_{p,r} \triangle B_{\hat{p},\hat{r}})$ for any two general $B_{p,r}$ and $B_{\hat{p},\hat{r}}$ unless $B_{p,r} \subset B_{\hat{p},\hat{r}}$ or $B_{\hat{p},\hat{r}} \subset B_{p,r}$. So, in general, it seems we have to approximate numerically the Bayesian estimate. This is a formidable numerical problem and indeed we are not sure that a minimizer reported by the computer can be trusted.

Therefore an alternative approach is introduced to fix the drawback of the maximum likelihood set estimates which always underestimate the true volume and the disadvantage of the Bayesian estimates which have computational difficulty. The alternative approach tries to estimate the true volume using the Bayesian method, and then corrects the maximum likelihood estimate for bias, based on the estimated volume.

If we consider the loss function

$$l_{vol}(S, \widehat{S}\,) = \big|\lambda(S) - \lambda(\widehat{S}\,)\big|, \tag{7}$$

it can be analyzed easily. One notes that it only gives a penalty for inaccuracy of volume estimation. Therefore it provides us with only a decision on the volume of S. The following proposition characterizes the class of all Bayesian estimates in this situation.

Proposition 1. *Let $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$ be a random sample from $B_{p,r}$. Define the transformation $v(p, r) = \lambda(B_{p,r})$ and denote a median of posterior of $v(p, r)$ by v_m. Then all the L_p balls with volume v_m are Bayesian estimates under the loss* (7).

Proof. Let us denote by $\pi(v|\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n)$ the distribution of $v = v(p, r) = \lambda(B_{p,r})$ with (p, r) having distribution $\pi(p, r|\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n)$. The risk

$$\rho(\hat{p}, \hat{r}) = E_{\pi(p,r|\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_n)}\big(|\lambda(B_{p,r}) - \lambda(B_{\hat{p},\hat{r}})|\big) = E_{\pi(v|\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_n)}\big(|v - v(\hat{p}, \hat{r})|\big) \tag{8}$$

which depends only on $v(\hat{p}, \hat{r})$ and is minimized when $v(\hat{p}, \hat{r})$ equals v_m. Namely $B_{\hat{p},\hat{r}}$ is a Bayes estimate with respect to loss (7) for any $(\hat{p}, \hat{r})$ for which $\lambda(B_{\hat{p},\hat{r}}) = v_m$. $\square$

As there are infinitely many L_p balls with volume v_m, we need a criterion to help us to choose one among these as the estimate of S. A reasonable way to choose a specific L_p ball as an estimate of S could be the pair (p, r) that has the smallest Euclidean distance from $(\hat{p}_{mle}, \hat{r}_{mle})$ among the infinitely many pairs implied in Proposition 1. Thus, this composite approach is to find

$$(\hat{p}_{comb}, \hat{r}_{comb}) = \arg\min_{\{(p,r):\lambda(B_{p,r})=v_m\}} (p - \hat{p}_{mle})^2 + (r - \hat{r}_{mle})^2. \tag{9}$$

We characterize $(\hat{p}_{comb}, \hat{r}_{comb})$ below. It is nice that the characterization is as explicit as it turned out to be.

Proposition 2. *Let $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$ be a random sample from $B_{p,r}$. Then $(\hat{p}_{comb}, \hat{r}_{comb})$ in (9) exists. Furthermore, $\hat{p}_{comb}$ is the unique root of*

$$p^2(p - \hat{p}_{mle}) - r(p)\left(\psi\left(1 + \frac{k}{p}\right) - \psi\left(1 + \frac{1}{p}\right)\right)\big(r(p) - \hat{r}_{mle}\big) = 0 \qquad (10)$$

and $\hat{r}_{comb} = r(\hat{p}_{comb})$. Here ψ is the digamma function and

$$r(p) = \frac{\hat{v}_m^{1/k}}{2} \frac{\Gamma(1 + \frac{k}{p})^{1/k}}{\Gamma(1 + \frac{1}{p})}. \qquad (11)$$

Proof. It is clear from (3) that for any fixed $0 < p \leq \infty$, $r(p)$ is the unique solution in r of $\lambda(B_{p,r}) = v_m$. If we can show that $(p - \hat{p}_{mle})^2 + (r(p) - \hat{r}_{mle})^2$ has a unique minimum at some $p = \tilde{p}$, then $(\hat{p}_{comb}, \hat{r}_{comb}) = (\tilde{p}, r(\tilde{p}))$.

This follows on observing that $\lambda(B_{(\hat{p}_{mle}, \hat{r}_{mle})}) = \hat{v}_{mle} < v_m$ which implies that the point $(\hat{p}_{mle}, \hat{r}_{mle})$ is under the curve $(p, r(p))$ in the (p, r) plane. Furthermore, $r(p)$ is strictly convex and differentiable, therefore, we have the existence and uniqueness of $\tilde{p}$, and it must satisfy

$$(\tilde{p} - \hat{p}_{mle}) + \big(r(\tilde{p}) - \hat{r}_{mle}\big)r'(\tilde{p}) = 0. \qquad (12)$$

By some further calculations, we obtain $r'(p) = r(p)(\psi(1 + \frac{k}{p}) - \psi(1 + \frac{1}{p}))\frac{1}{p^2}$. From (12) it now follows that $\tilde{p}$ is the unique root of (10). □

3. Strong consistency of the estimates

Maximum likelihood and Bayesian estimates are the most widely used methods of estimation and there is an enormous amount of literature on it. However, a lot of the well known asymptotic theory applies only to those distributions satisfying certain "regularity" conditions. See Lehman and Casella (1998), Le Cam (1953), Huber (1967), and Perlman (1972). One of the conditions requires that the distributions have common support. Apparently, we cannot look for answers in these theories for our problem, as the support is the parameter itself. Consequently, a more direct approach would be necessary and the Wald theorem would be the core key.

3.1. Strong consistency of ML estimate

Let us consider the maximum likelihood estimate first. The most popular strong consistency theorem for the maximum likelihood estimate is due to Wald (1949). It can be applied to the non-regular case. In his paper, Wald gave several conditions to prove a main theorem first. Then he established, essentially through this main theorem, the strong consistency of the maximum likelihood estimate (in fact, of a more general family of estimates) provided that the distributions admit those conditions. Though our problem does not satisfy Wald's conditions, the main theorem, however, holds for our problem. Therefore, here we will try to combine his main theorem and his strong consistency theorem for our maximum likelihood estimate. For completeness, we provide the proof.

Theorem 1 (Wald). *Let $P_{\underset{\sim}{\theta}}$ be a distribution with density $f(\underset{\sim}{x}; \underset{\sim}{\theta})$, where $\underset{\sim}{\theta} \in \Theta$. Suppose the realizations $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$ come from $P_{\underset{\sim}{\theta}_0}$ independently for some $\underset{\sim}{\theta}_0 \in \Theta$. Let $\underset{\sim}{\hat{\theta}}_n$ be a function of $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$ satisfying*

$$\frac{f(\underset{\sim}{x}_1; \underset{\sim}{\hat{\theta}}_n) \cdots f(\underset{\sim}{x}_n; \underset{\sim}{\hat{\theta}}_n)}{f(\underset{\sim}{x}_1; \underset{\sim}{\theta}_0) \cdots f(\underset{\sim}{x}_n; \underset{\sim}{\theta}_0)} \geq c > 0 \quad \textit{for all } n \quad \textit{and} \quad \underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n \textit{ for some positive } c. \qquad (13)$$

If for any given neighborhood of θ_0, say U, it also holds that

$$P_{\underset{\sim}{\theta}_0}\left\{\lim_{n\to\infty}\frac{\sup_{\theta\in\Theta\setminus U} f(\underset{\sim}{x}_1;\underset{\sim}{\theta})\cdots f(\underset{\sim}{x}_n;\underset{\sim}{\theta})}{f(\underset{\sim}{x}_1;\underset{\sim}{\theta}_0)\cdots f(\underset{\sim}{x}_n;\underset{\sim}{\theta}_0)}=0\right\}=1, \tag{14}$$

then we have

$$P_{\underset{\sim}{\theta}_0}\left\{\lim_{n\to\infty}\hat{\underset{\sim}{\theta}}_{\mathrm{n}}=\underset{\sim}{\theta}_0\right\}=1. \tag{15}$$

This theorem basically states that if the likelihood ratio of $\underset{\sim}{\theta}$ to $\underset{\sim}{\theta}_0$ is uniformly small as $\underset{\sim}{\theta}$ falls outside any given neighborhood of the true parameter $\underset{\sim}{\theta}_0$, then the estimate $\hat{\underset{\sim}{\theta}}_n$ must be close to $\underset{\sim}{\theta}_0$ since by assumption its likelihood ratio to $\underset{\sim}{\theta}_0$ is always greater than or equal to c (which is greater than 0).

Proof. This theorem does not require that the coordinates of $\underset{\sim}{\theta}_0$ are finite (note that the shape parameter p in our problem can be infinity). But we will give the proof for θ_0 having finite coordinates only to avoid redundancy since the proofs are similar.

To prove (15), it suffices to show that for any neighborhood of $\underset{\sim}{\theta}_0$, say U, $\hat{\underset{\sim}{\theta}}_n$ will fall inside U eventually with probability one. But from (14), one sees that, with probability one, there exists N, which may depend on $\{\underset{\sim}{x}_i\}_{i=1}^{\infty}$, such that

$$\frac{\sup_{\theta\in\Theta\setminus U} f(\underset{\sim}{x}_1,\underset{\sim}{\theta})\cdots f(\underset{\sim}{x}_n,\underset{\sim}{\theta})}{f(\underset{\sim}{x}_1,\underset{\sim}{\theta}_0)\cdots f(\underset{\sim}{x}_n,\underset{\sim}{\theta}_0)}<\frac{c}{2}\ \forall\ n\geq N.$$

However, (13) claims that

$$\frac{f(\underset{\sim}{x}_1,\hat{\underset{\sim}{\theta}}_n)\cdots f(\underset{\sim}{x}_n,\hat{\underset{\sim}{\theta}}_n)}{f(\underset{\sim}{x}_1,\underset{\sim}{\theta}_0)\cdots f(\underset{\sim}{x}_n,\underset{\sim}{\theta}_0)}\geq c>\frac{c}{2}\ \forall\ n\quad\text{and}\quad \underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_n.$$

Thus, $\hat{\underset{\sim}{\theta}}_n\notin\Theta\setminus U$ when $n\geq N$. Therefore $\hat{\underset{\sim}{\theta}}_n$ belongs to U eventually with probability one, as claimed. □

Since a maximum likelihood estimate, if it exists, obviously satisfies (13) with $c = 1$, this theorem also proves the strong consistency of the maximum likelihood estimate provided (14) holds. Fortunately, our family of distributions $\{\mathrm{Unif}(B_{p,r})\}_{\{0<p\leq\infty,0<r<\infty\}}$ satisfies (14).

Lemma 1. *Let P_θ denote* $\mathrm{Unif}(\mathbf{B}_{p,r})$*, where $\underset{\sim}{\theta}=(p,r)$ and $\underset{\sim}{\theta}\in\Theta=\{(p,r):0<p\leq\infty,0<r<\infty\}$. Then $\{P_{\underset{\sim}{\theta}}\}_{\underset{\sim}{\theta}\in\Theta}$ satisfies* (14).

Proof. The proof is extremely lengthy and involved. To maintain the flow of this paper, we will only give a rough sketch here and refer the rigorous proof to Tsai (2000).

The basic idea of this proof is as follows. For any given $(p,r)\neq(p_0,r_0)$, one has either $B_{p_0,r_0}\subset_\lambda B_{p,r}$ or $\lambda(B_{p_0,r_0}\setminus B_{p,r})>0$, here $A\subset_\lambda B$ means A is contained in B properly in the Lebesgue measure; i.e. $A\subset B$ and $\lambda(B\setminus A)>0$. In the first situation, we will have the likelihood ratio equal to $(\frac{\lambda(B_{p_0,r_0})}{\lambda(B_{p,r})})^n$ which goes to 0 as n goes to ∞ since $\lambda(B_{p_0,r_0})<\lambda(B_{p,r})$. For the second case, we will, eventually, observe some $\underset{\sim}{x}_i$ not belonging to $B_{p,r}$, which results in the zero value of the likelihood ratio. As a result, (14) shall hold. □

Now by Theorem 1, and Lemma 1, we have the strong consistency of $(\hat{p}_{mle},\hat{r}_{mle})$.

Corollary 1. *Let $\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_n$ be a random sample from $B_{p,r}$. Then the maximum likelihood estimate $(\hat{p}_{mle},\hat{r}_{mle})$ is strongly consistent.*

3.2. Strong consistency of Bayesian estimate

Let us now move to the consistency of the Bayesian estimate. The following is a general result on the strong consistency of the Bayesian estimate under a general assumption on the distribution family and the loss function. Basically, this theorem and its proof are very similar to the Wald Theorem given in the previous section except that we have to include the prior and the loss which are the other elementary components for Bayesian analysis. The generality of this theorem makes it an attractive result of independent interest.

Theorem 2. *Suppose $P_{\utilde{\theta}}$ denotes a distribution with density $f(\utilde{x};\utilde{\theta})$, where $\utilde{\theta} \in \Theta$. Assume the observations $\utilde{x}_1, \ldots, \utilde{x}_n$ are iid with probability $P_{\utilde{\theta}_0}$ for some $\utilde{\theta}_0 \in \Theta$. Let $\pi(\utilde{\theta})$ be a prior of $\utilde{\theta}$ and $l(\utilde{\theta}, \hat{\utilde{\theta}})$ be a loss function such that*

$$\int_\Theta \pi(\utilde{\theta})\, \mathrm{d}\utilde{\theta} < \infty \quad \textit{and} \quad \int_\Theta l(\utilde{\theta}, \utilde{\theta}_0)\pi(\utilde{\theta})\, \mathrm{d}\utilde{\theta} < \infty. \tag{16}$$

Then the Bayesian estimate will converge to $\utilde{\theta}_0$ with probability one (under $P_{\utilde{\theta}_0}$) provided that for any neighborhood of $\utilde{\theta}_0$, say U, there exist sets $W \subset V \subset U$ satisfying

(i) $$P_{\utilde{\theta}_0}\left\{\lim_{n\to\infty} \frac{\sup_{\utilde{\theta}\in\Theta\setminus V} f(\utilde{x}_1;\utilde{\theta})\cdots f(\utilde{x}_n;\utilde{\theta})}{\inf_{\utilde{\theta}\in W} f(\utilde{x}_1;\utilde{\theta})\cdots f(\utilde{x}_n;\utilde{\theta})} = 0\right\} = 1,$$

(ii) $$\int_W \pi(\utilde{\theta})\mathrm{d}\utilde{\theta} > 0, \quad \textit{and}$$

(iii) $$\inf_{\hat{\utilde{\theta}}\in U^c, \utilde{\theta}\in V} l(\utilde{\theta}, \hat{\utilde{\theta}}) - l(\utilde{\theta}, \utilde{\theta}_0) \geq \epsilon \quad \textit{for some } \epsilon > 0.$$

Remark 1. In this theorem, there is a condition on all components of the problem (likelihood, prior, and loss). Condition (i) states that the likelihood ratio for $\utilde{\theta}$ far away from $\utilde{\theta}_0$ versus $\utilde{\theta}$ near $\utilde{\theta}_0$ is uniformly small. Condition (ii) requires that the prior puts a positive mass around the true $\utilde{\theta}_0$. Condition (iii) says that the loss function does punish for bad decisions. These conditions are all quite mild.

Proof. We divide the proof into several steps for clarity and ease of understanding.
Step 1: Let us denote the posterior of $\utilde{\theta}$ given $\utilde{x}_1, \ldots, \utilde{x}_n$ by

$$\pi(\utilde{\theta}|\utilde{x}_1, \ldots, \utilde{x}_n) \propto \Pi_{i=1}^n f(\utilde{x}_i, \utilde{\theta})\, \pi(\utilde{\theta}).$$

Then the posterior expected loss for decision $\hat{\utilde{\theta}}$ is $\rho(\hat{\utilde{\theta}}) = E_{\pi(\utilde{\theta}|\utilde{x}_1,\ldots,\utilde{x}_n)}\left(l(\utilde{\theta}, \hat{\utilde{\theta}})\right)$, and the Bayesian estimate is $\hat{\utilde{\theta}}_{bayes} = \arg\min_{\hat{\utilde{\theta}}\in\Theta} \rho(\hat{\utilde{\theta}})$.

To prove the strong consistency of $\hat{\utilde{\theta}}_{bayes}$, it suffices to show that for any neighborhood of $\utilde{\theta}_0$, say U, $\hat{\utilde{\theta}}_{bayes}$ will fall inside U eventually with probability one (under $P_{\utilde{\theta}_0}$). Now, let V, W, and ϵ be as defined in condition (i), (ii), and (iii). We will show that

$$P_{\utilde{\theta}_0}\left\{\inf_{\hat{\utilde{\theta}}\in U^c} \rho(\hat{\utilde{\theta}}) \geq \rho(\utilde{\theta}_0) + \frac{1}{4}\epsilon \quad \text{eventually}\right\} = 1. \tag{17}$$

This will imply

$$P_{\utilde{\theta}_0}\left\{\arg\min_{\hat{\utilde{\theta}}} \rho(\hat{\utilde{\theta}}) \in U \ \ \text{eventually}\right\} = 1,$$

proving this theorem.

Step 2: In this step, we will break $\rho(\hat{\underset{\sim}{\theta}})-\rho(\underset{\sim}{\theta}_0)$ into several terms whose magnitudes are easier to investigate. Note that

$$\begin{aligned}
&\rho(\hat{\underset{\sim}{\theta}}) - \rho(\underset{\sim}{\theta}_0)\\
&\quad= E_{\pi(\underset{\sim}{\theta}|\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_n)}(l(\underset{\sim}{\theta},\hat{\underset{\sim}{\theta}}) - l(\underset{\sim}{\theta},\underset{\sim}{\theta}_0))\\
&\quad= \frac{\int_\Theta (l(\underset{\sim}{\theta},\hat{\underset{\sim}{\theta}}) - l(\underset{\sim}{\theta},\underset{\sim}{\theta}_0))\Pi_{i=1}^n f(\underset{\sim}{x}_i|\underset{\sim}{\theta})\pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta}}{\int_\Theta \Pi_{i=1}^n f(\underset{\sim}{x}_i|\underset{\sim}{\theta})\pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta}}\\
&\quad= \frac{\int_V (l(\underset{\sim}{\theta},\hat{\underset{\sim}{\theta}}) - l(\underset{\sim}{\theta},\underset{\sim}{\theta}_0))\Pi_{i=1}^n f(\underset{\sim}{x}_i|\underset{\sim}{\theta})\pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta}}{\int_V \Pi_{i=1}^n f(\underset{\sim}{x}_i|\underset{\sim}{\theta})\pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta}} \cdot \frac{\int_V \Pi_{i=1}^n f(\underset{\sim}{x}_i|\underset{\sim}{\theta})\pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta}}{\int_\Theta \Pi_{i=1}^n f(\underset{\sim}{x}_i|\underset{\sim}{\theta})\pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta}}\\
&\qquad+ \frac{\int_{V^c} l(\underset{\sim}{\theta},\hat{\underset{\sim}{\theta}})\Pi_{i=1}^n f(\underset{\sim}{x}_i|\underset{\sim}{\theta})\pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta}}{\int_\Theta \Pi_{i=1}^n f(\underset{\sim}{x}_i|\underset{\sim}{\theta})\pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta}} - \frac{\int_{V^c} l(\underset{\sim}{\theta},\underset{\sim}{\theta}_0)\Pi_{i=1}^n f(\underset{\sim}{x}_i|\underset{\sim}{\theta})\pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta}}{\int_\Theta \Pi_{i=1}^n f(\underset{\sim}{x}_i|\underset{\sim}{\theta})\pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta}}\\
&\quad= (I)\cdot(II) + ((III) - (IV)). \qquad (18)
\end{aligned}$$

Step 3: In this step, we will show that (I) is always greater than or equal to ϵ. From condition (iii), it is easy to see that

$$(I) \geq \frac{\int_V \ \epsilon \cdot \Pi_{i=1}^n f(\underset{\sim}{x}_i|\underset{\sim}{\theta})\pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta}}{\int_V \Pi_{i=1}^n f(\underset{\sim}{x}_i|\underset{\sim}{\theta})\pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta}} = \epsilon \ \text{ for all } \ \hat{\underset{\sim}{\theta}} \in U^c. \qquad (19)$$

Step 4: Now, we claim

$$P_{\underset{\sim}{\theta}_0}\{(II) \longrightarrow 1\} = 1. \qquad (20)$$

Note that

$$\begin{aligned}
|1-(II)| &= \frac{\int_{V^c} \Pi_{i=1}^n f(\underset{\sim}{x}_i|\underset{\sim}{\theta})\pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta}}{\int_\Theta \Pi_{i=1}^n f(\underset{\sim}{x}_i|\underset{\sim}{\theta})\pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta}} \leq \frac{\int_{V^c} \Pi_{i=1}^n f(\underset{\sim}{x}_i|\underset{\sim}{\theta})\pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta}}{\int_W \Pi_{i=1}^n f(\underset{\sim}{x}_i|\underset{\sim}{\theta})\pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta}}\\
&\leq \frac{\sup_{\underset{\sim}{\theta}\in V^c} f(\underset{\sim}{x}_1;\underset{\sim}{\theta})\cdots f(\underset{\sim}{x}_n;\underset{\sim}{\theta})}{\inf_{\underset{\sim}{\theta}\in W} f(\underset{\sim}{x}_1;\underset{\sim}{\theta})\cdots f(\underset{\sim}{x}_n;\underset{\sim}{\theta})} \frac{\int_{V^c}\pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta}}{\int_W \pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta}}. \qquad (21)
\end{aligned}$$

From condition (i), together with condition (ii) and (16), we get that the upper bound (21) converges to 0 with probability one. Consequently, claim (20) is proved.
Step 5: Now let us look at the term $(III) - (IV)$. We would like to show that

$$P_{\underset{\sim}{\theta}_0}\left\{\inf_{\hat{\underset{\sim}{\theta}}\in U^c}\{(III)-(IV)\} \geq -\frac{1}{4}\epsilon \ \text{ eventually}\right\} = 1. \qquad (22)$$

Since (III) is nonnegative, we have $(III) - (IV) \geq -(IV)$ which does not depend on $\hat{\underset{\sim}{\theta}}$. Moreover

$$0 \leq (IV) \leq \frac{\sup_{\underset{\sim}{\theta}\in V^c} f(\underset{\sim}{x}_1;\underset{\sim}{\theta})\cdots f(\underset{\sim}{x}_n;\underset{\sim}{\theta})}{\inf_{\underset{\sim}{\theta}\in W} f(\underset{\sim}{x}_1;\underset{\sim}{\theta})\cdots f(\underset{\sim}{x}_n;\underset{\sim}{\theta})} \frac{\int_{V^c} l(\underset{\sim}{\theta},\underset{\sim}{\theta}_0)\pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta}}{\int_W \pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta}}$$

Again from conditions (i) and (ii), and (16), we get that (IV) converges to 0 with probability one. Therefore (22) is true.
Step 6: Finally, as an immediate consequence of (18), (19), (20), and (22) together, we obtain (17). This theorem therefore follows. $\square$

Now we would like to apply Theorem 2 to our problem. The following lemma says that the distribution family Unif$(B_{p,r})$, and the loss function $l((p,r),(\hat{p},\hat{r})) = \lambda(B_{p,r}\triangle B_{\hat{p},\hat{r}})$ satisfy condition (i) and (iii) of Theorem 2.

Lemma 2. *Let P_θ denote the distribution* Unif$(B_{p,r})$*, where $\underset{\sim}{\theta} = (p,r)$ and $\underset{\sim}{\theta} \in \Theta = \{(p,r) : 0 < p \leq \infty, 0 < r < \infty\}$. Let $l((p,r),(\hat{p},\hat{r})) = \lambda(B_{p,r} \triangle B_{\hat{p},\hat{r}})$ be the loss function and let π be the prior on $\underset{\sim}{\theta}$. Suppose $\underset{\sim}{\theta}_0 = (p_0, r_0)$ is a fixed point in Θ and π is positive in some neighborhood of (p_0, r_0). Then for any neighborhood of (p_0, r_0), say U, there exist sets $W \subset V \subset U$ such that the conditions* (i), (ii), *and* (iii) *in Theorem* 2 *hold.*

Proof. The idea of the proof is not difficult. However, the proof is very lengthy. Refer to Tsai (2000). □

Now, as an application of Theorem 2, we have the strong consistency of $(\hat{p}_{bayes}, \hat{r}_{bayes})$ as follows:

Corollary 2. *Let $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$ be iid with distribution* Unif$(B_{p,r})$*. Suppose the true value of (p,r) is denoted by (p_0, r_0). Let π be a proper prior on (p,r) such that π is positive in some neighborhood of (p_0, r_0). Assume also that $E_{\pi(p,r)}(\lambda(B_{p,r}))$ is finite. Then the Bayesian estimate under the loss $l((p,r),(\hat{p},\hat{r})) = \lambda(B_{p,r} \triangle B_{\hat{p},\hat{r}})$ converges to (p_0, r_0) with probability one.*

Proof. From the assumption on π, one has

$$E_{\pi(p,r)}\left[l((p,r),(\hat{p},\hat{r}))\right] \leq E_{\pi(p,r)}\left[\lambda(B_{p,r}) + \lambda(B_{p_0,r_0})\right] < \infty.$$

Thus, the corollary follows from Theorem 2 and Lemma 2 immediately. □

3.3. Strong consistency of combined estimate

Now we discuss the strong consistency of a combined estimate $(\hat{p}_{comb}, \hat{r}_{comb})$. Recall that it is the pair (p,r) closest to the initial guess $(\hat{p}_{mle}, \hat{r}_{mle})$ with $\lambda(B_{p,r})$ equal to v_m, the posterior median of $v = \lambda(B_{p,r})$. From Corollary 1 and Corollary 3 below, $(\hat{p}_{mle}, \hat{r}_{mle})$ and v_m are both strongly consistent in the respective parameters. One may expect, therefore, that the combined estimate will be strongly consistent as well. We give a general theorem in this direction below. Again the generality makes it an appealing theorem of independent interest.

Theorem 3. *Let $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$ be a sample from a distribution P_θ, $\underset{\sim}{\theta} \in \Theta$. Let Θ be a metric space with a metric d. Let $\hat{\underset{\sim}{\theta}}_n$ and $\hat{\underset{\sim}{\beta}}_n$ be functions of the observations $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$ such that $\hat{\underset{\sim}{\theta}}_n$ and $\hat{\underset{\sim}{\beta}}_n$ converge almost surely to $\underset{\sim}{\theta}$ and $\underset{\sim}{\beta}(\underset{\sim}{\theta})$, respectively, under $P_{\underset{\sim}{\theta}}$, where $\underset{\sim}{\beta}(\underset{\sim}{\theta})$ is a function of $\underset{\sim}{\theta}$. Suppose $\tilde{\underset{\sim}{\theta}}_n \overset{def}{=} \arg\min_{\{\underset{\sim}{\theta}:\underset{\sim}{\beta}(\underset{\sim}{\theta})=\hat{\underset{\sim}{\beta}}_n\}} d(\hat{\underset{\sim}{\theta}}_n, \underset{\sim}{\theta})$ exists and is unique. Then $\tilde{\underset{\sim}{\theta}}_n$ converges to $\underset{\sim}{\theta}$ with probability one if for any $\epsilon > 0$, there exists a neighborhood of $\underset{\sim}{\beta}(\underset{\sim}{\theta})$ contained in $\underset{\sim}{\beta}(B_d(\underset{\sim}{\theta}, \epsilon))$, where $B_d(\underset{\sim}{\theta}, \epsilon)$ is the ϵ-ball centered at $\underset{\sim}{\theta}$ with respect to the metric d.*

Proof. To prove the strong consistency of $\tilde{\underset{\sim}{\theta}}_n$, it is enough to show that for any $\epsilon > 0$,

$$P_{\underset{\sim}{\theta}}\left\{ d(\underset{\sim}{\theta}, \tilde{\underset{\sim}{\theta}}_n) < 3\epsilon \text{ eventually}\right\} = 1. \tag{23}$$

By assumption, there exists a neighborhood of $\underset{\sim}{\beta}(\underset{\sim}{\theta})$, say B, contained in $\underset{\sim}{\beta}(B_d(\underset{\sim}{\theta}, \epsilon))$; so, if $\hat{\underset{\sim}{\beta}}_n \in B$, there exists $\tilde{\underset{\sim}{\theta}}$ within ϵ distance of $\underset{\sim}{\theta}$ such that $\underset{\sim}{\beta}(\tilde{\underset{\sim}{\theta}}) = \hat{\underset{\sim}{\beta}}_n$. Then, one has

$$d(\hat{\underset{\sim}{\theta}}_n, \tilde{\underset{\sim}{\theta}}_n) = \min_{\{\underset{\sim}{\theta}:\underset{\sim}{\beta}(\underset{\sim}{\theta})=\hat{\underset{\sim}{\beta}}_n\}} d(\hat{\underset{\sim}{\theta}}_n, \underset{\sim}{\theta}) \leq d(\hat{\underset{\sim}{\theta}}_n, \tilde{\underset{\sim}{\theta}}) \leq d(\hat{\underset{\sim}{\theta}}_n, \underset{\sim}{\theta}) + d(\underset{\sim}{\theta}, \tilde{\underset{\sim}{\theta}}) \leq d(\underset{\sim}{\theta}, \hat{\underset{\sim}{\theta}}_n) + \epsilon,$$

which implies

$$d(\underset{\sim}{\theta}, \tilde{\underset{\sim}{\theta}}_n) \le d(\underset{\sim}{\theta}, \hat{\underset{\sim}{\theta}}_n) + d(\hat{\underset{\sim}{\theta}}_n, \tilde{\underset{\sim}{\theta}}_n) \le 2d(\underset{\sim}{\theta}, \hat{\underset{\sim}{\theta}}_n) + \epsilon.$$

Furthermore, if $d(\underset{\sim}{\theta}, \hat{\underset{\sim}{\theta}}_n) < \epsilon$, then we have $d(\underset{\sim}{\theta}, \tilde{\underset{\sim}{\theta}}_n) < 3\epsilon$. On the other hand, $\hat{\underset{\sim}{\theta}}_n$ and $\hat{\underset{\sim}{\beta}}_n$ are strongly consistent for $\underset{\sim}{\theta}$ and $\underset{\sim}{\beta}(\underset{\sim}{\theta})$ respectively. This implies

$$P_{\underset{\sim}{\theta}}\left\{ d(\underset{\sim}{\theta}, \hat{\underset{\sim}{\theta}}_n) < \epsilon \text{ and } \hat{\underset{\sim}{\beta}}_n \in B \text{ eventually}\right\} = 1.$$

This proves (23) and hence the theorem. □

To apply the above general theorem to our problem, we need the strong consistency of v_m. This will be implied by the following theorem which generalizes Theorem 2.

Theorem 4. *Let $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$ be a sample from P_θ with density $f(\underset{\sim}{x}; \underset{\sim}{\theta})$, where $\underset{\sim}{\theta} \in \Theta$. Suppose we are interested in estimating a function $\underset{\sim}{\beta}(\underset{\sim}{\theta})$ (rather than $\underset{\sim}{\theta}$) itself and the loss is a function of $\underset{\sim}{\theta}$ through $\underset{\sim}{\beta}(\underset{\sim}{\theta})$, say $l(\underset{\sim}{\beta}(\underset{\sim}{\theta}), \hat{\underset{\sim}{\beta}})$. Denote the true value of $\underset{\sim}{\theta}$ by $\underset{\sim}{\theta}_0$ and the prior of $\underset{\sim}{\theta}$ by π. Assume $\int \pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta} < \infty$ and $\int l(\underset{\sim}{\beta}(\underset{\sim}{\theta}), \underset{\sim}{\beta}(\underset{\sim}{\theta}_0))\pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta} < \infty$. Then the Bayesian estimate of $\underset{\sim}{\beta}(\underset{\sim}{\theta})$, $\arg\min_{\hat{\underset{\sim}{\beta}}} E_{\pi(\underset{\sim}{\theta}|\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_n)}(l(\underset{\sim}{\beta}(\underset{\sim}{\theta}), \hat{\underset{\sim}{\beta}}))$, converges to $\underset{\sim}{\beta}_0 \equiv \underset{\sim}{\beta}(\underset{\sim}{\theta}_0)$ with probability one under $P_{\underset{\sim}{\theta}_0}$ provided that for any neighborhood of $\underset{\sim}{\beta}_0$, say B, there exists sets $W \subset V \subset \underset{\sim}{\beta}^{-1}(B)$ satisfying*

(i) $$P_{\underset{\sim}{\theta}_0}\left\{\lim_{n\to\infty} \frac{\sup_{\underset{\sim}{\theta}\in\Theta\setminus V} f(\underset{\sim}{x}_1; \underset{\sim}{\theta}) \cdots f(\underset{\sim}{x}_n; \underset{\sim}{\theta})}{\inf_{\underset{\sim}{\theta}\in W} f(\underset{\sim}{x}_1; \underset{\sim}{\theta}) \cdots f(\underset{\sim}{x}_n; \underset{\sim}{\theta})} = 0\right\} = 1,$$

(ii) $$\int_W \pi(\underset{\sim}{\theta})\mathrm{d}\underset{\sim}{\theta} > 0, \quad \textit{and}$$

(iii) $$\inf_{\hat{\underset{\sim}{\beta}}\in B^c, \underset{\sim}{\theta}\in V} l(\underset{\sim}{\beta}(\underset{\sim}{\theta}), \hat{\underset{\sim}{\beta}}) - l(\underset{\sim}{\beta}(\underset{\sim}{\theta}), \underset{\sim}{\beta}(\underset{\sim}{\theta}_0)) \ge \epsilon \quad \textit{for some } \epsilon > 0.$$

Remark 2. Theorem 2 is a special case of Theorem 4 when we take $\underset{\sim}{\beta}(\underset{\sim}{\theta}) = \underset{\sim}{\theta}$. Moreover, in this theorem, $\underset{\sim}{\beta}$ does not have to be one-to-one and $\underset{\sim}{\beta}^{-1}(B)$ is defined as $\{\, \underset{\sim}{\theta} : \underset{\sim}{\beta}(\underset{\sim}{\theta}) \in B\}$.

Proof. The proof is exactly the same as that of Theorem 2. □

We now apply Theorem 4 to prove the strong consistency of v_m.

Corollary 3. *Let $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$ be a random sample from $B_{p,r}$. Define $v = v(p, r) = \lambda(B_{p,r})$. Let π be a prior on (p, r) and v_m the posterior median of v. Let also (p_0, r_0) denote the true value of (p, r). If π is positive in a neighborhood of (p_0, r_0), then v_m converges to $v(p_0, r_0)$ with probability one.*

Proof. Denote $v(p_0, r_0)$ by v_0. Let B be a neighborhood of v_0. Without loss of generality, we can assume $B = (v_0 - \delta, v_0 + \delta)$ for some $\delta > 0$. Since $v(p, r)$ is a continuous function of (p, r), there exists a neighborhood of (p_0, r_0), say U, such that $U \subset v^{-1}(B')$, where $B' = (v_0 - \frac{\delta}{3}, v_0 + \frac{\delta}{3})$. Then by Lemma 2, there exist sets $W \subset V \subset U$ such that conditions (i) and (ii) in Theorem 4 hold.

Furthermore, if $(p, r) \in V$, one has $v(p, r) \in B'$, which implies

$$|v(p, r) - v(p_0, r_0)| < \frac{\delta}{3} \quad \text{and} \quad |v(p, r) - \hat{v}| > \frac{2\delta}{3} \quad \text{for all } \hat{v} \in B.$$

This gives us condition (iii) of Theorem 4.

The corollary, therefore, follows from Theorem 4 immediately. □

Corollary 3 endows us with the strong consistency of v_m needed to apply the general result of Theorem 3. We are now ready to prove the strong consistency of $(\hat{p}_{comb}, \hat{r}_{comb})$.

Corollary 4. *The estimator* $(\hat{p}_{comb}, \hat{r}_{comb})$ *defined in Section* 2.3 *is strongly consistent.*

Proof. To prove this proposition, we will apply Theorem 3 for the case when the true value, p_0, of p is finite. When p_0 is infinity, we will prove this proposition directly. Recall that

$$(\hat{p}_{comb}, \hat{r}_{comb}) = \arg \min_{\{(p,r):\lambda(B_{p,r})=v_m\}} \left[(p-\hat{p}_{mle})^2 + (r-\hat{r}_{mle})^2\right].$$

Also Corollary 1 and Corollary 3 give us the strong consistency of $(\hat{p}_{mle}, \hat{r}_{mle})$ and v_m respectively.

Case 1 $p_0 = \infty$: By (10) and the fact that $v_m \geq \lambda(B_{\hat{p}_{mle},\hat{r}_{mle}})$, $\hat{p}_{comb}$ must be greater than $\hat{p}_{mle}$. As $\hat{p}_{mle}$ converges to $p_0 = \infty$ with probability one, so does $\hat{p}_{comb}$. Furthermore, Proposition 2 also gives $\hat{r}_{comb} = \frac{{v_m}^{1/k}}{2} \frac{\Gamma^{1/k}(1+\frac{k}{\hat{p}_{comb}})}{\Gamma(1+\frac{1}{\hat{p}_{comb}})}$. Thus the strong consistency of $\hat{r}_{comb}$ follows from the strong consistency of $\hat{p}_{comb}$ and v_m immediately.

Case 2 $p_0 < \infty$: We will prove this case as an application of Theorem 3. For any given $\epsilon > 0$, let us take $B = (\lambda(B_{p_0,(r_0-\epsilon)^+}), \lambda(B_{p_0,r_0+\epsilon}))$. It is easy to see that for any $b \in B$, there exists $(r_0 - \epsilon)^+ < r < r_0 + \epsilon$ such that $v(p_0, r) = b$ and certainly the distance between (p_0, r) and (p_0, r_0) is smaller than ϵ. Therefore, the assumptions in Theorem 3 are all satisfied. This proposition for the case when $p_0 < \infty$ follows. $\square$

4. Discussion

This section will first compare the performance of the maximum likelihood estimate with the combined estimate, especially when the sample size is small. Recall that the calculation of Bayes estimate is difficult. Then, some simulation and conjectures on the asymptotic distribution of the estimates will be given as, unfortunately, they are very hard. We end with a brief discussion for the case when the center of symmetry of the true set is unknown.

4.1. Comparison of $(\hat{p}_{mle}, \hat{r}_{mle})$ *and* $(\hat{p}_{comb}, \hat{r}_{comb})$

We remarked that the combined estimate can be principally considered as a dilation of the maximum likelihood estimate. Our simulation will try to examine: (i) in what fashion the combined estimate dilates the maximum likelihood estimate, (ii) if it indeed helps with regard to underestimation of the volume of the true set, and (iii) if the choice of the prior on p and r affects the performance of the combined estimate.

The tables and figures referenced below are based on a simulation of size 750 with true $(p, r) = (2, 1)$, dimension $k = 2$, and sample size $n = 10$. We consider three respective priors on (p, r). They are $\pi_1(p, r) = pe^{-p}re^{-r}$, $\pi_2(p, r) = \frac{1}{2}p^2e^{-p}re^{-r}$, and $\pi_3(p, r) = \frac{2}{\pi(1+p^2)}re^{-r}$ respectively. We denote each of the corresponding combined estimates by $(\hat{p}_{comb1}, \hat{r}_{comb1})$, $(\hat{p}_{comb2}, \hat{r}_{comb2})$, and $(\hat{p}_{comb3}, \hat{r}_{comb3})$, respectively.

Table 1 gives the mean and the standard error of the volume of $B_{\hat{p}_{mle},\hat{r}_{mle}}$, $B_{\hat{p}_{comb1},\hat{r}_{comb1}}$, $B_{\hat{p}_{comb2},\hat{r}_{comb2}}$, and $B_{\hat{p}_{comb3},\hat{r}_{comb3}}$, and their symmetric difference as well as their Hausdorff distances to the true set. This table shows that the volumes of the

Table 1: The mean and standard error (in parentheses) from a size 750 simulation of the volume of the maximum likelihood estimate and the combined estimates with respect to three different priors on (p, r) and the symmetric difference distances and the Hausdorff distances to the true set.

true $(p,r) = (2,1)$	$k = 2,\ n = 10$		
$(\hat{p}, \hat{r}) =$	$\lambda(B_{\hat{p},\hat{r}})$	$d_\lambda(B_{p,r}, B_{\hat{p},\hat{r}})$	$d_H(B_{p,r}, B_{\hat{p},\hat{r}})$
$(\hat{p}_{mle}, \hat{r}_{mle})$	2.70519(0.294328)	0.466593(0.290792)	0.118659(0.077068)
$(\hat{p}_{comb1}, \hat{r}_{comb1})$	3.10145(0.321972)	0.336655(0.200423)	0.095912(0.065939)
$(\hat{p}_{comb2}, \hat{r}_{comb2})$	3.10321(0.332331)	0.342807(0.201751)	0.097780(0.066906)
$(\hat{p}_{comb3}, \hat{r}_{comb3})$	3.13368(0.344959)	0.351907(0.202675)	0.098770(0.065463)

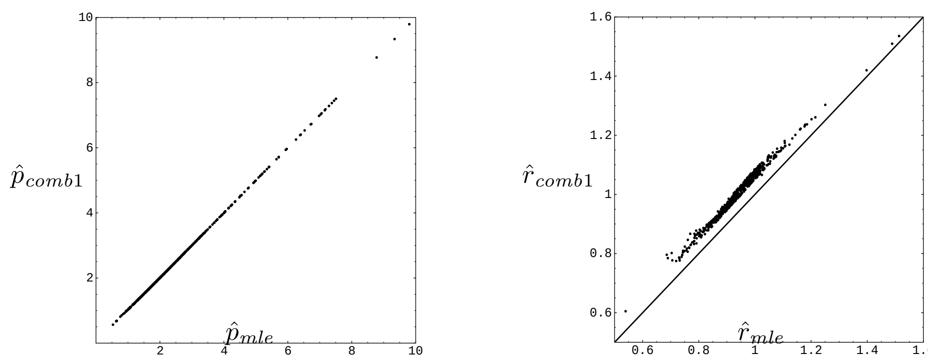

Figure 1: Scatter plots of $(\hat{p}_{mle}, \hat{p}_{comb})$ and $(\hat{r}_{mle}, \hat{r}_{comb})$.

combined estimates are much closer to the true volume (which is $\pi = 3.14159$), but with a higher variance, than that of the maximum likelihood estimate. Moreover, the distances, either one, of the combined estimates to the true set are about 20% to 30% less compared to the maximum likelihood estimate. It also appears that the selection of the prior does not affect the performance of the combined estimate very much.

Figure 1 plots $\hat{p}_{comb1}$ against $\hat{p}_{mle}$ and $\hat{r}_{comb1}$ against $\hat{r}_{mle}$. We see that the scatter plot of $(\hat{p}_{mle}, \hat{p}_{comb1})$ is virtually the 45 degree line; $(\hat{r}_{mle}, \hat{r}_{comb1})$s', on the other hand, all fall above the 45 degree line. We have similar results for the other two combined estimators. So, $B_{\hat{p}_{comb},\hat{r}_{comb}}$ may indeed be considered as if it was dilated from $B_{\hat{p}_{mle},\hat{r}_{mle}}$ by enlarging only the radius r while keeping p essentially fixed at $\hat{p}_{mle}$. This is interesting.

4.2. Convergence in distribution

In this section, some simulation and conjectures on the asymptotic distribution of the maximum likelihood estimate will be given. Figure 2 shows several scatter plots of $(n(\hat{p}_{mle} - p), n(\hat{r}_{mle} - r))$ with $p = 2$, $r = 1$, and various sample sizes. We believe that when the true value of p is finite, $(n(\hat{p}_{mle} - p), n(\hat{r}_{mle} - r))$ converges to some nondegenerate distribution which puts all its mass in the half plane: $\{(x, y)\ :\ y \leq$

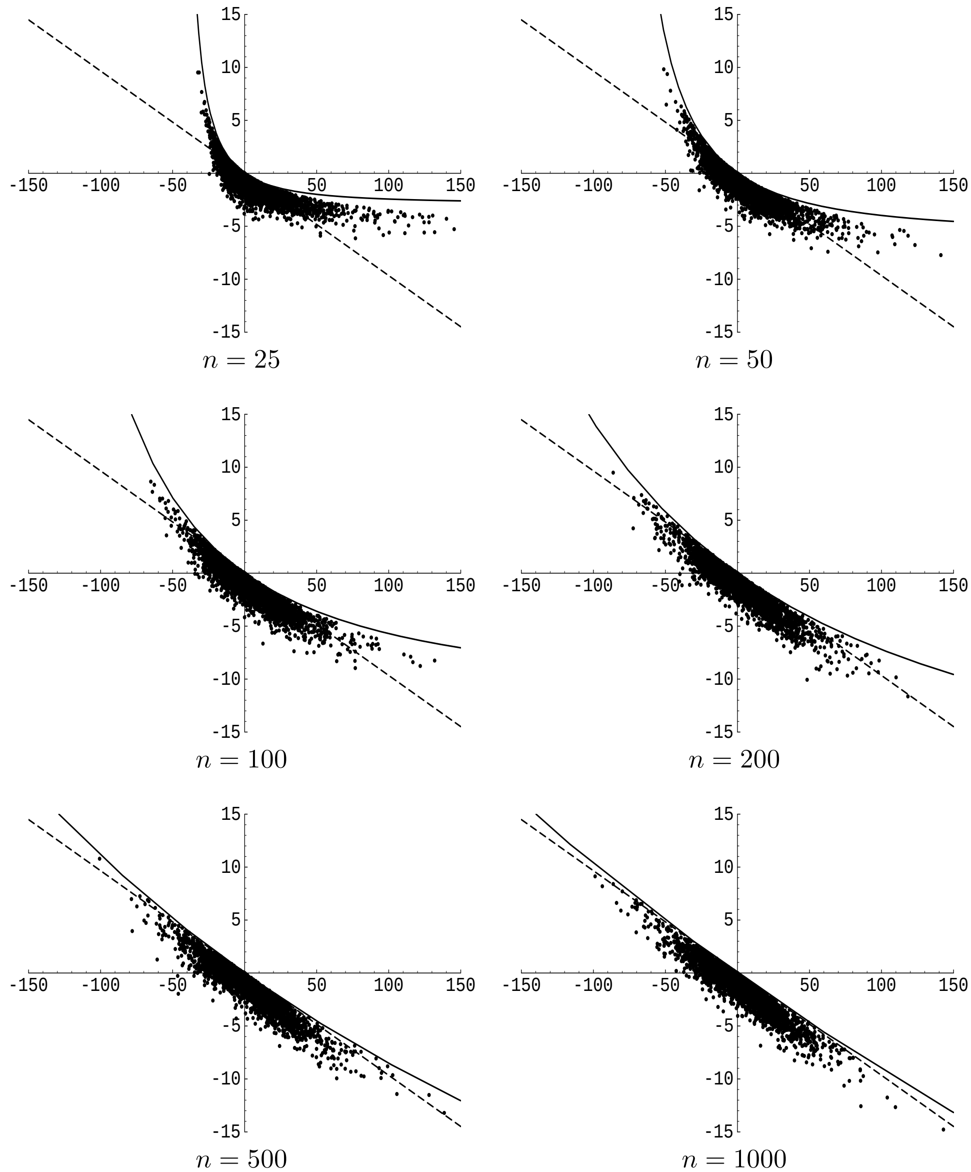

Figure 2: Scatter plots of $(n(\hat{p}_{mle} - p), n(\hat{r}_{mle} - r))$ with $(p, r) = (2, 1)$ and $k = 2$. Solid line is $(n(t-p), n(\frac{\Gamma^{1/k}(1+\frac{k}{t})}{\Gamma(1+\frac{1}{t})}\frac{\Gamma(1+\frac{1}{p})}{\Gamma^{1/k}(1+\frac{k}{p})} - 1)r)$, where t ranges from 0 to ∞. Broken line is the straight line through the origin with slope $\frac{-1}{p^2}(\psi(1+\frac{k}{p}) - \psi(1+\frac{1}{p}))$.

$\frac{-r}{p^2}\{\psi(1+\frac{k}{p})-\psi(1+\frac{1}{p})\}x\}$. It is also obvious that the correlation of $\hat{p}_{mle}$ and $\hat{r}_{mle}$ is negative. When $\hat{p}_{mle}$ overestimates the true p, the corresponding $\hat{r}_{mle}$ will then likely underestimate the true r, and vice versa.

Figure 3 gives scatter plots of $(\sqrt{n}(\frac{1}{\hat{p}_{mle}}), n(\hat{r}_{mle}-r))$ for the case where $(p,r) = (\infty, 1)$. It seems that $(\sqrt{n}(\frac{1}{\hat{p}_{mle}}), n(\hat{r}_{mle}-r))$ converges to some nondegenerate distribution having support in the fourth quadrant. Interestingly, the convergence rates seem dependent on the true value of p.

In fact, these conjectures were inspired by the case when one of the parameters (p or r) is known. A summary for the behavior of $\hat{p}_{mle}$ when r is assumed to be known is given below. Similar results can also be derived for the case when p is assumed to be known. See Tsai (2000) for details.

If we assume r is known, say, $r = r_0$, the characterization of the maximum likelihood estimate of p becomes very simple. We are in fact able to give the exact distribution of $\hat{p}_{mle}$ and therefore the weak convergence result for $\hat{p}_{mle}$. The idea of getting this result is very simple. Indeed this problem can be converted to an end-point problem if we consider the new random variables $z_i = \lambda(B_{p_{r_0}(\underset{\sim}{x}_i), r_0})$, where $B_{p_{r_0}(\underset{\sim}{x}_i), r_0}$ is the smallest L_p ball containing $\underset{\sim}{x}_i$ with radius r_0. It can be easily shown that $\underset{\sim}{Z}_i$'s are independently and identically distributed with value between 0 and the volume of the true domain and $\lambda(B_{\hat{p}_{mle}, r_0}) = \max_{1\le i\le n} z_i$, whose asymptotic distribution is well known. Thus we have the following weak convergence result for $\hat{p}_{mle}$ when the true r is known.

Proposition 3. *Suppose $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$ are iid from* $\mathrm{Unif}(B_{p,r_0})$, *where $0 < r_0 < \infty$ is known. Let G denote an exponential random variable with mean* 1. *Then*
(I) *when $p < \infty$,*

$$n(\hat{p}_{mle} - p) \xrightarrow{\mathcal{D}} -\frac{p^2}{k}\left(\frac{1}{\psi(1+\frac{k}{p})-\psi(1+\frac{1}{p})}\right)G, \tag{24}$$

where ψ is the digamma function, and
(II) *when $p = \infty$,*

$$\sqrt{n}\frac{1}{\hat{p}_{mle}} \xrightarrow{\mathcal{D}} \sqrt{\frac{12}{\pi^2 k(k-1)}}\sqrt{G}. \tag{25}$$

Remark 3. Note that when $p < \infty$, interestingly, the asymptotic variance, $(\frac{p^2}{k}(\frac{1}{\psi(1+\frac{k}{p})-\psi(1+\frac{1}{p})}))^2$, is a decreasing function of the dimension k. It appears that the curse of dimensionality does not show up in this problem. To the contrary, for estimation of the single shape parameter p, it is beneficial to have a large k!

Remark 4. In fact, $n(\lambda(B_{\hat{p}_{mle}, r_0}) - \lambda(B_{p,r_0})) \xrightarrow{\mathcal{D}} -\lambda(B_{p,r_0})G$. If we divide both sides by the true volume, this expression tells us that the proportion of the uncommon part between the estimate and the true set (to the true set) converges with the rate n to an exponential distribution. It does not relate to the true set. The convergence rates of $\hat{p}_{mle}$, however, do depend on p. It is interesting that the speed of convergence of $\hat{p}_{mle}$ slows down from n to $\sqrt{n}$ discontinuously as p changes from finite to infinite. We believe that this phenomenon is caused by the difficulty of "catching the corners" of a square, for example. This is also interesting.

4.3. Unknown center of symmetry

In practice, the center of symmetry of the object usually would not be known. It then has to be estimated. In this section, we will have a brief examination of this situation.

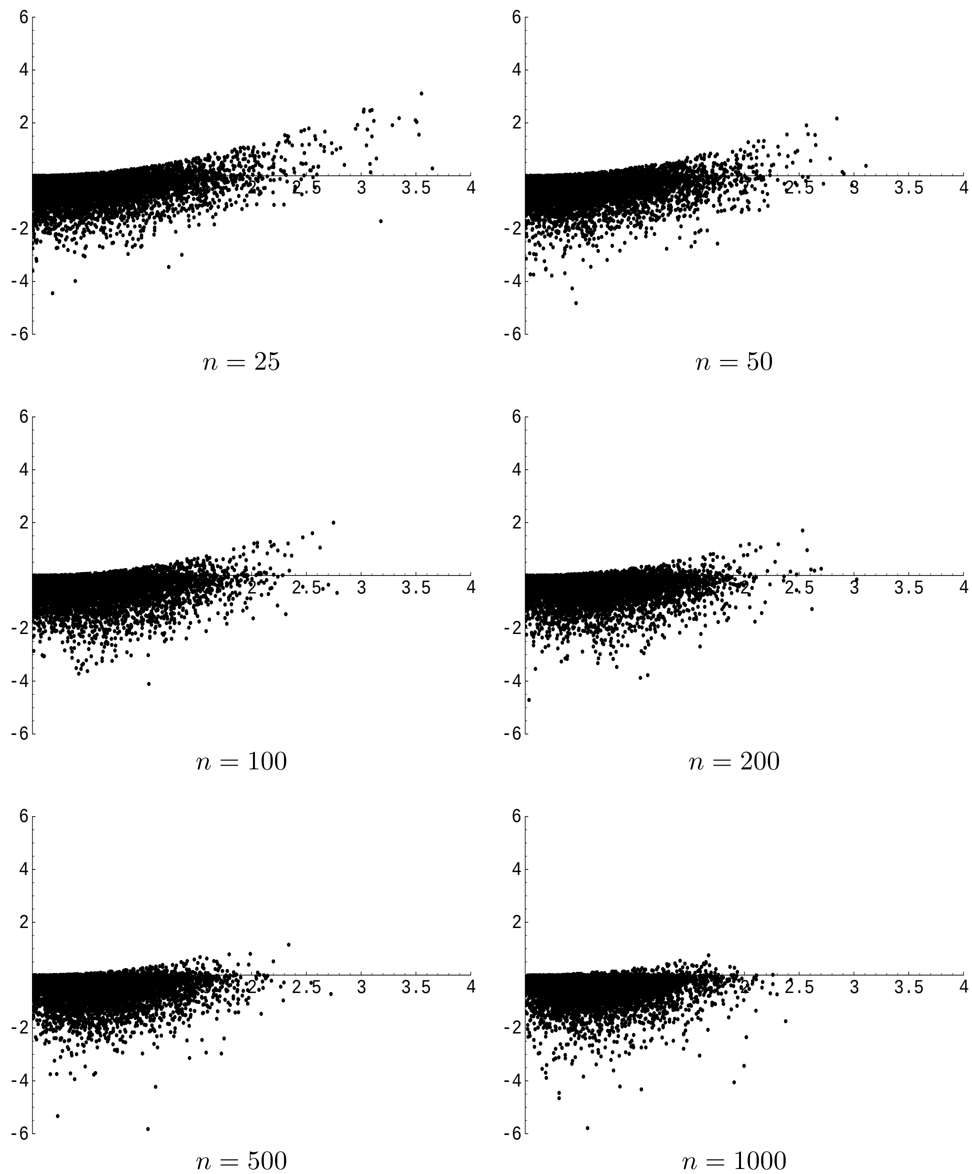

Figure 3: Scatter plots of $(\sqrt{n}(\frac{1}{\hat{p}_{mle}}), n(\hat{r}_{mle} - r))$ with $(p, r) = (\infty, 1)$ and $k = 2$ for different sample sizes.

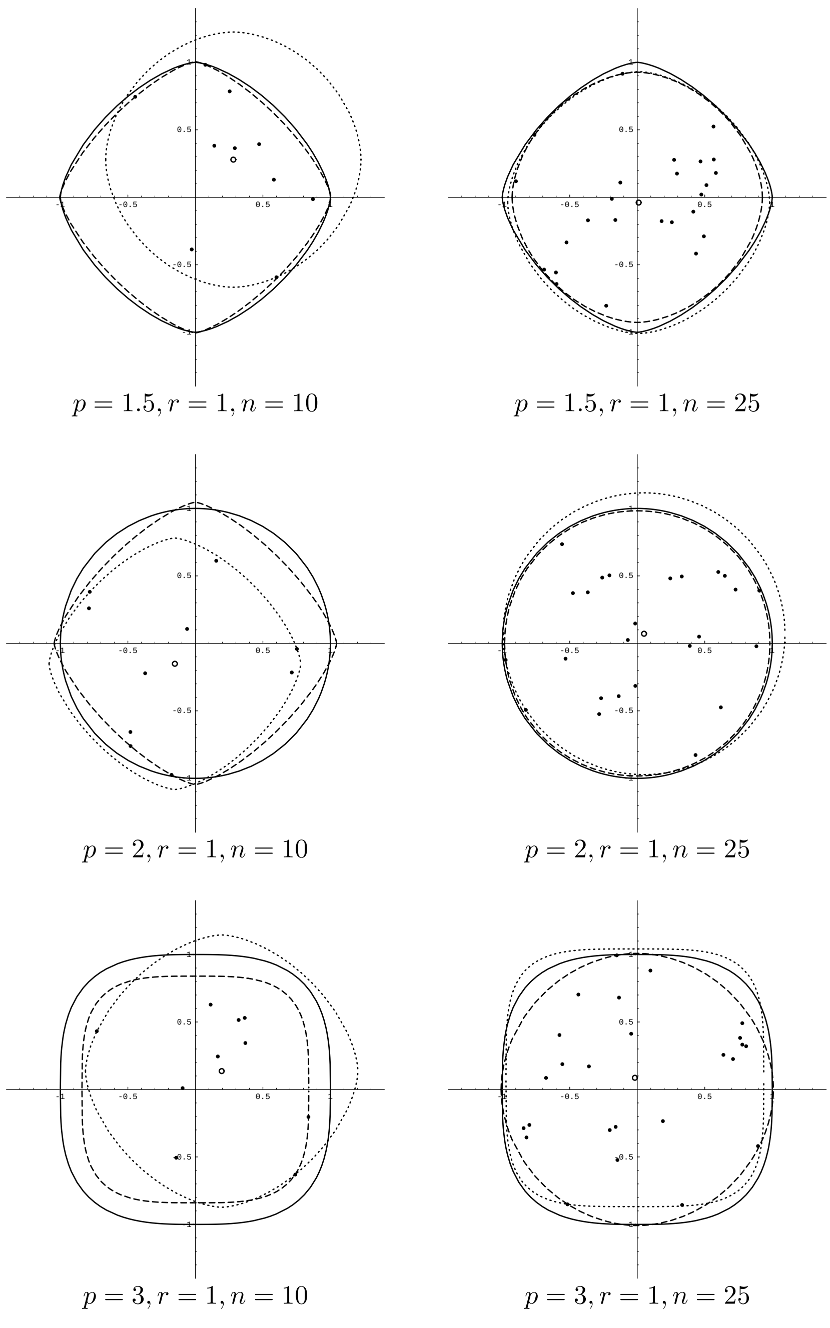

Figure 4: Visual display of the set estimate when the center is unknown. The region bounded by the solid curves is the true set, by the broken or the dotted curve is the maximum likelihood estimate with the center assumed to be known or estimated by the mean of the observations, respectively. The conspicuous circle is the estimated center and the dots are the observations.

Apparently, it is not easy to estimate the center together with the shape parameter p and the size parameter r by using the maximum likelihood method. See Amey et al. (1991) for some calculations. Besides, the problem of underestimating of the volume of the maximum likelihood estimate in this situation will be more serious. Therefore, it may be preferable to estimate the center by some other external methods. We tried the mean of the observations, and the L_2 median (spatial median) (which minimizes $\Sigma_{1\leq i\leq n}\|\utilde{x}_i - \utilde{u}\|_2$ over $\utilde{u}$). It turns out that the mean of the observations performs better than the L_2 median. Therefore, here we attempt to check how the estimate may be influenced if the center is unknown and is estimated by the mean of the observations. Figure 4 gives a visual comparison between the maximum likelihood estimates with center treated to be known and with center estimated by the mean of the observations. It can be seen that the shape of the estimates can vary very much depending on whether the center is known or estimated. But the estimate of the size parameter does not differ that much. Moreover, the volume of the maximum likelihood estimate with the center estimated by the mean of the observations can exceed the volume of the true set. When the realizations cluster to one side with some observations appearing in the far opposite direction, apparently the estimate can miss the true set badly. Therefore, constructing a better estimate for the center of symmetry of the true set is important.

References

[1] Amey, A., Gupta, A. K., Norton, V., and Székeley, G. J. (1991). On Maximum Likelihood Estimation of the Center of a Centrally Symmetric Convex Domain, *Jour. Stat. Comp. and Simu.*, **39**, 1–7. MR1109862

[2] Bräker, H., Hsing, T., and Bingham, N. H. (1998). On the Hausdorff Distance Between a Convex Set and an Interior Random Convex Hull, *Adv. Appl. Probab.*, **30**, 295–316. MR1642840

[3] (1994). Gradshteyn, I. S. and Ryzhik, I. M. Table of Integrals, Series, and Products, Academic Press, New York,

[4] Hall, P. (1982). ON Estimating the Endpoint of a Distribution, *Ann. Statist.*, **10**(2), 556–568. MR653530

[5] Huber, P. J. (1967). The Behavior of maximum likelihood estimates under nonstandard conditions, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, eds. Le Cam, L. M. and Neyman, J., Vol. 1, Univ. Calif. Press, Berkeley and Los Angeles, pp. 221–233. MR216620

[6] Le Cam, L. M. (1953). *On some Asymptotic Properties of Maximum Likelihood Estimate and Related Bayes Estimates*, Univ. Calif. Publ. Statist., Vol. 1, pp. 277–330. MR54913

[7] Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*, Springer, New York. MR1639875

[8] Macdonald, D. W. and Ball, F. G. and Hough, N. G. (1979). The evaluation of Home Range Size and Configuration Using Radio Tracking Data, *A Handbook on Biotelemetry and Radio Tracking*, eds. Amlaner, C. J. and Macdonald, D. W., Oxford, New York, pp. 405–424.

[9] Moore, M. (1984). On the Estimation of a Convex Set, *Ann. Statist.*, **12** 3, 1090–1099. MR751296

[10] Perlman, M. D. (1972). On the Strong Consistency of Approximate Maximum Likelihood Estimators, in *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, eds. Le Cam, L. M. and Neyman, J. and Scott, E., Vol. 1, Univ. Calif. Press, Berkeley and Los Angeles, pp. 263–281. MR405668

[11] Ripley, B. D. and Rasson, J.-P. (1977). Finding the Edge of a Poisson Forest, *J. Appl. Probab.*, **14** 3, 483–491. MR451339

[12] Seber, G. A. F. (1986). A Review of Estimating Animal Aboundance, *Biometrics*, **42**, 267–292. MR872967

[13] Seber, G. A. F. (1992). A Review of Estimating Animal Aboundance II, *Internat. Statist. Rev.*, **60** 2, 129–166.

[14] Tsai, W.-C. (2000). Some Contribution to Stochastic Geometry, Department of Statistics, Purdue University.

[15] Wald, A. (1949). Note on the Consistency of the maximum likelihood, *Ann. Math. Statist.*, **20**, 596. MR32169

[16] Worton, B. J. (1987). A Review of Models of Home Range for Animal Movement, *Ecol. Modelling*, **38**, 277–298.