Hindawi
Mathematical Problems in Engineering
Volume 2019, Article ID 3485064, 6 pages
https://doi.org/10.1155/2019/3485064

Research Article

Some Notes on Concordance between Optimization and Statistics

Weiyan Mu,[1] **Qiuyue Wei**,[1] **and Shifeng Xiong**[2]

[1]*School of Science, Beijing University of Civil Engineering and Architecture, Beijing 100044, China*
[2]*NCMIS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China*

Correspondence should be addressed to Shifeng Xiong; xiong@amss.ac.cn

Received 25 October 2018; Accepted 17 December 2018; Published 16 January 2019

Academic Editor: Alexander Paz

Many engineering problems require solutions to statistical optimization problems. When the global solution is hard to attain, engineers or statisticians always use the better solution because we intuitively believe a principle, called better solution principle (BSP) in this paper, that a better solution to a statistical optimization problem also has better statistical properties of interest. This principle displays some concordance between optimization and statistics and is expected to widely hold. Since theoretical study on BSP seems to be neglected by statisticians, this paper presents a primary discussion on BSP within a relatively general framework. We demonstrate two comparison theorems as the key results of this paper. Their applications to maximum likelihood estimation are presented. It can be seen that BSP for this problem holds under reasonable conditions; i.e., an estimator with greater likelihood is better in some statistical sense.

1. Introduction

Experimental design and data analysis in engineering rely on statistical methods that usually require solutions to optimization problems [1]. A notable example is maximum likelihood estimation, whose objective is to maximize the likelihood function. In fact, optimization methods are ubiquitous in almost all statistical areas and play a vital role in modern statistics. In the meanwhile, statisticians have to face the common difficulty in the optimization community; i.e., it is often extremely hard to obtain the global solution to a nonconvex optimization problem. A number of global optimization algorithms have been proposed, including the simulated annealing algorithm [2] and the genetic algorithm [3]. However, they can attain the global solution only in the probabilistic sense and often take an unrealistically long time to approach it in practice [4]. When handling large-scale data, the problem of multiple extrema becomes more serious. In fact, for such cases, it is also hard to obtain the solution to a convex optimization problem due to the unaffordable computational time and memory [5]. Another difficulty from the problem of multiple extrema is that we can rarely know whether a solution at hand is the global solution [6].

When the global solution is hard to attain and/or to verify, for minimization problems, statisticians always take the solution whose objective value is as small as possible as the final solution. In other words, for two solutions, we always use the better one, where "better" should be understood in the sense of optimization. This seems reasonable in that the better solution is more likely to be the global solution, whose statistical properties of interest usually have been well established. From the statistical perspective, we use the better solution because we intuitively believe a principle, called better solution principle (BSP) in this paper, that a better solution to a statistical optimization problem also has better statistical properties of interest. This principle shows some concordance, or monotonicity, between optimization and statistics, and is expected to widely hold. Strictly speaking, a better solution can safely be used only after the corresponding BSP is verified. However, it is surprising that statisticians seem to neglect this problem, although we have actually made decisions following it ever since complex optimization problems appeared in statistics. To the best of the author's knowledge, no paper has formally discussed BSP. For example, in the maximum likelihood problem, it is not clear to us whether a better solution with greater likelihood has higher estimation accuracy.

Fairly recently, Xiong [7] introduced the better-fitting better-screening rule when discussing variable screening in high-dimensional linear models. This rule tells us that, under reasonable conditions, a subset with smaller residual sums of squares possesses better asymptotic screening properties, i.e., more likely to include the true submodel asymptotically. Such a subset is a better solution to the ℓ_0-norm constrained least squares problem. Therefore, the better-fitting better-screening rule is actually the BSP for this problem.

In this paper we provide some theoretical discussion on BSP in relatively general statistical optimization problems. The rest of the paper is organized as follows. We first present examples where BSP immediately holds in Section 2. Such examples widely exist in experimental designs. They can help us understand BSP and the reason why we introduce the theorems in the following text. In Section 3, we demonstrate two comparison theorems which state that a better solution is more likely to have good statistical properties if the optimization problem possesses certain separation properties. These theorems, which look very simple and understandable, are effective to establish BSP in a general setting. Section 4 applies these results to maximum likelihood estimation. We can see that BSP for this problem holds under reasonable conditions. Section 5 concludes with some discussion.

2. Known Examples Where BSP Holds

We first present examples where BSP immediately holds. Such examples widely exist in experimental designs. They can help us understand BSP and the reason why we introduce the theorems in the following text. We take the D-optimal design [8] for example. Consider a regression model

$$y = \sum_{i=1}^{d} \theta_i r_i(\mathbf{x}) + \varepsilon, \tag{1}$$

where the control variable $\mathbf{x}$ lies in a subset $\mathscr{D}$ of $\mathbb{R}^p$, r_i's are specified functions, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)'$ is the vector of unknown parameters, and ε is the random error. Given the sample size n, denote the experimental design by $\mathscr{P} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. The information matrix of this design is $\mathbf{M}(\mathscr{P}) = \mathbf{R}(\mathscr{P})'\mathbf{R}(\mathscr{P})$, where

$$\mathbf{R}(\mathscr{P}) = \begin{pmatrix} r_1(\mathbf{x}_1) & \cdots & r_d(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ r_1(\mathbf{x}_n) & \cdots & r_d(\mathbf{x}_n) \end{pmatrix}. \tag{2}$$

The D-optimal design minimizes the generalized variance of the least squares estimate of $\boldsymbol{\theta}$; i.e., it is the solution to the optimization problem

$$\min_{\mathbf{x}_i \in \mathscr{D}} \psi(\mathscr{P}) = \left[\det(\mathbf{M}(\mathscr{P}))\right]^{-1}, \tag{3}$$

where det denotes determinant. For two designs $\mathscr{P}_1$ and $\mathscr{P}_2$ with $\psi(\mathscr{P}_1) \leqslant \psi(\mathscr{P}_2)$, it is clear that $\mathscr{P}_1$ leads to a better estimator whose generalized variance is smaller. Therefore, if estimation accuracy, which is justified by generalized variance, is the statistical property of interest, then the BSP for problem (3) holds.

The objective function in (3) itself is a statistical criterion, which does not involve any random variables. This is the reason why the BSP for (3) automatically holds. The same conclusion can be drawn for other model-based optimal designs [8] and minimum aberration designs [9]. For criterion-based space-filling designs (Santner, Williams, and Notz 2003), the geometric or discrepancy criteria used as objective functions seem not to have clear statistical interpretation. However, most of them relate to some desirable statistical properties. For example, the criteria for constructing the minimax distance design [10] and uniform design [11] can act as factors in the upper bounds of some estimation errors [12, 13]. If such estimation errors are used to evaluate the corresponding estimators, we can say that BSP holds.

Design of experiments is a pre-sampling work (we do not consider sequential designs here), and thus the objective functions used in this area do not involve the random sample. In statistical inference, we have to deal with objective functions depending on the sample, which makes the problem of BSP more complicated. From the next section, we study whether BSP holds for sample-based optimization problems through introducing new definitions and theorems.

3. The Comparison Theorems

Let $(\Omega, \mathfrak{F}, P)$ be a probability space. For simplicity, it is assumed that all sets and maps throughout this paper are measurable with respect to according σ-fields. For each $n \in \mathbb{N}$, the sample $\mathbb{X}_n$ of size n is a map from Ω to a space $\mathscr{X}_n$. Let $\mathfrak{D}$ denote the decision space that contains all statistical decisions of interest. Suppose that we need to make inferences based on the global solution to the optimization problem

$$\min_{x \in \mathfrak{D}} \psi_n(x, \mathbb{X}_n), \tag{4}$$

where the objective function ψ_n is a map from $\mathfrak{D} \times \mathscr{X}_n$ to $\mathbb{R}$. In general, the problem in (4) is proposed because its solution can asymptotically lie in a desirable subset $\mathfrak{A}$ of $\mathfrak{D}$ that contains all good decisions. This property can be viewed as a consistency property of the global solution.

Consider the situations where the global solution to (4) is difficult to obtain. Suppose that there are K candidate solutions, $\xi_n^{(1)}, \ldots, \xi_n^{(K)}$. In practice, we always use ξ_n^*, which denotes the one that takes the smallest value of $\psi_n(\cdot, \mathbb{X}_n)$ among them, as the final decision. For each $\xi_n^{(k)}$, ξ_n^* is a better solution since the inequality

$$\psi_n(\xi_n^*, \mathbb{X}_n) \leqslant \psi_n\left(\xi_n^{(k)}, \mathbb{X}_n\right) \tag{5}$$

always holds. This section discusses whether BSP holds, i.e., whether such a better solution is more likely to lie in $\mathfrak{A}$.

Let $\mathfrak{B}$ be another subset of $\mathfrak{D}$, which contains relatively bad decisions compared to $\mathfrak{A}$.

Definition 1. We say that $\{\psi_n\}$ strongly separates $\mathfrak{A}$ from $\mathfrak{B}$ or $\{\psi_n\}$ has the strong separation property with respect of $\mathfrak{A}$ and $\mathfrak{B}$, if as $n \longrightarrow \infty$,

$$P\left(\sup_{x \in \mathfrak{A}} \psi_n(x, \mathbb{X}_n) < \inf_{y \in \mathfrak{B}} \psi_n(y, \mathbb{X}_n)\right) \longrightarrow 1. \tag{6}$$

We say that $\{\psi_n\}$ weakly separates $\mathfrak{A}$ from $\mathfrak{B}$ or $\{\psi_n\}$ has the weak separation property with respect of $\mathfrak{A}$ and $\mathfrak{B}$, if for all $x \in \mathfrak{A}$, $y \in \mathfrak{B}$,

$$\limsup_{n\longrightarrow\infty} \left[\psi_n(x, \mathbb{X}_n) - \psi_n(y, \mathbb{X}_n)\right] < 0 \quad \text{(a.s.)}, \tag{7}$$

where "a.s." denotes "almost surely."

The strong separation property needs not to imply its weak analogue. It is generally more difficult to verify and can lead to stronger results.

Remark 2. For convenience in asymptotic analysis, we often consider a scaled objective function. It should be pointed out that (6) holds if

$$P\left(\frac{\sup_{x\in\mathfrak{A}} \psi_n(x, \mathbb{X}_n)}{a_n} < \frac{\inf_{y\in\mathfrak{B}} \psi_n(y, \mathbb{X}_n)}{a_n}\right) \longrightarrow 1 \tag{8}$$

for arbitrary sequence of positive numbers $\{a_n\}$ and that (7) holds if

$$\limsup_{n\longrightarrow\infty} \frac{\psi_n(x, \mathbb{X}_n) - \psi_n(y, \mathbb{X}_n)}{a_n} < 0 \quad \text{(a.s.)} \tag{9}$$

for a sequence of positive numbers $\{a_n\}$ with $a_n^{-1} = O(1)$.

Roughly speaking, the strong separation property requires that the level set corresponding to smaller objective values is asymptotically identical to the set of good decisions. It shows the consistency of the objective function to the statistical properties of interest. We can immediately prove the following result that this property implies BSP.

Theorem 3 (strong comparison theorem). *Suppose that $\{\psi_n\}$ strongly separates $\mathfrak{A}$ from $\mathfrak{B}$. For all $n \in \mathbb{N}$, ξ_n and η_n are statistics valued in $\mathfrak{D}$ satisfying $P(\xi_n \in \mathfrak{A}\cup\mathfrak{B}, \eta_n \in \mathfrak{A}\cup\mathfrak{B}) \longrightarrow 1$ as $n \longrightarrow \infty$. If $\psi_n(\xi_n, \mathbb{X}_n) \leqslant \psi_n(\eta_n, \mathbb{X}_n)$ for all n, then*

$$\liminf_{n\longrightarrow\infty} \left[P(\xi_n \in \mathfrak{A}) - P(\eta_n \in \mathfrak{A})\right] \geqslant 0. \tag{10}$$

Proof. For $\omega \in \{\eta_n \in \mathfrak{A}, \xi_n \in \mathfrak{B}\} \subset \Omega$, if $\omega \in \{\sup_{x\in\mathfrak{A}} \psi_n(x, \mathbb{X}_n) < \inf_{y\in\mathfrak{B}} \psi_n(y, \mathbb{X}_n)\} \cap \{\xi_n \in \mathfrak{A}\cup\mathfrak{B}, \eta_n \in \mathfrak{A}\cup\mathfrak{B}\}$, then

$$\begin{aligned}\psi_n(\eta_n(\mathbb{X}_n(\omega)), \mathbb{X}_n(\omega)) &\leqslant \sup_{x\in\mathfrak{A}} \psi_n(x, \mathbb{X}_n(\omega)) \\ &< \inf_{y\in\mathfrak{B}} \psi_n(y, \mathbb{X}_n(\omega)) \leqslant \psi_n(\xi_n(\mathbb{X}_n(\omega), \mathbb{X}_n(\omega)),\end{aligned} \tag{11}$$

which leads to a contradiction. Therefore, $\omega \notin \{\sup_{x\in\mathfrak{A}} \psi_n(x, \mathbb{X}_n) < \inf_{y\in\mathfrak{B}} \psi_n(y, \mathbb{X}_n)\} \cap \{\xi_n \in \mathfrak{A}\cup\mathfrak{B}, \eta_n \in \mathfrak{A}\cup\mathfrak{B}\}$, which implies $P(\eta_n \in \mathfrak{A}, \xi_n \in \mathfrak{B}) \longrightarrow 0$. We have $P(\eta_n \in \mathfrak{A}) = P(\eta_n \in \mathfrak{A}, \xi_n \in \mathfrak{B}) + P(\eta_n \in \mathfrak{A}, \xi_n \in \mathfrak{A}) + P(\eta_n \in \mathfrak{A}, \xi_n \notin \mathfrak{A}\cup\mathfrak{B}) = P(\xi_n \in \mathfrak{A}) - P(\xi_n \in \mathfrak{A}, \eta_n \in \mathfrak{B}) + o(1)$. This completes the proof. □

Theorem 3 indicates that, for two candidate solutions ξ_n and η_n to the statistical optimization problem (4), the better one is asymptotically more likely to be a good decision, and thus the BSP holds. Recall that, for a decision ξ_n, the property that $P(\xi_n \in \mathfrak{A}) \longrightarrow 1$ can be viewed as a consistency property of ξ_n. The following theorem shows that the strong separation property of $\{\psi_n\}$ is often stronger than the consistency of the minimum of ψ_n.

Theorem 4. *Suppose that $\{\psi_n\}$ strongly separates $\mathfrak{A}$ from $\mathfrak{B}$. If $\xi_n = \arg\min_{x\in\mathfrak{D}} \psi_n(x, \mathbb{X}_n)$ exists and $P(\xi_n \in \mathfrak{A}\cup\mathfrak{B}) \longrightarrow 1$ as $n \longrightarrow \infty$, then*

$$\lim_{n\longrightarrow\infty} P(\xi_n \in \mathfrak{A}) \longrightarrow 1. \tag{12}$$

Proof. This theorem follows from $\{\sup_{x\in\mathfrak{A}} \psi_n(x, \mathbb{X}_n) < \inf_{y\in\mathfrak{B}} \psi_n(y, \mathbb{X}_n)\} \cap \{\xi_n \in \mathfrak{A}\cup\mathfrak{B}\} \subset \{\xi_n \in \mathfrak{A}\}$. □

We next discuss BSP with the weak separation property. This weaker property cannot directly imply BSP, and more conditions are needed.

Theorem 5 (weak comparison theorem). *Suppose that $\{\psi_n\}$ weakly separates $\mathfrak{A}$ from $\mathfrak{B}$. Denote the set of probability one where (7) holds by $E(x, y)$ and write $E = \cap_{x\in\mathfrak{A}, y\in\mathfrak{B}} E(x, y)$. For all $n \in \mathbb{N}$, ξ_n and η_n are statistics valued in $\mathfrak{D}$ satisfying $P(\xi_n \in \mathfrak{A}\cup\mathfrak{B}, \eta_n \in \mathfrak{A}\cup\mathfrak{B}) \longrightarrow 1$ as $n \longrightarrow \infty$. If $\psi_n(\xi_n, \mathbb{X}_n) \leqslant \psi_n(\eta_n, \mathbb{X}_n)$ for all n, then*

$$\liminf_{n\longrightarrow\infty} \left[P(\xi_n \in \mathfrak{A}) - P(\eta_n \in \mathfrak{A})\right] \geqslant -P(\Omega\backslash E). \tag{13}$$

Proof. For $\omega \in \{\eta_n \in \mathfrak{A}, \xi_n \in \mathfrak{B}\} \subset \Omega$, if $\omega \in E \cap \{\xi_n \in \mathfrak{A}\cup\mathfrak{B}, \eta_n \in \mathfrak{A}\cup\mathfrak{B}\}$, then

$$\begin{aligned}\limsup_{n\longrightarrow\infty} &\left[\psi_n(\eta_n(\mathbb{X}_n(\omega)), \mathbb{X}_n(\omega))\right. \\ &\left. - \psi_n(\xi_n(\mathbb{X}_n(\omega)), \mathbb{X}_n(\omega))\right] < 0.\end{aligned} \tag{14}$$

This is a contradiction. Therefore, $\omega \notin E \cap \{\xi_n \in \mathfrak{A}\cup\mathfrak{B}, \eta_n \in \mathfrak{A}\cup\mathfrak{B}\}$ for sufficiently large n, which implies $P(\eta_n \in \mathfrak{A}, \xi_n \in \mathfrak{B}) \leqslant P(\Omega\backslash E) + 1 - P(\xi_n \in \mathfrak{A}\cup\mathfrak{B}, \eta_n \in \mathfrak{A}\cup\mathfrak{B})$ for sufficiently large n. It follows that $P(\eta_n \in \mathfrak{A}) = P(\eta_n \in \mathfrak{A}, \xi_n \in \mathfrak{B}) + P(\eta_n \in \mathfrak{A}, \xi_n \in \mathfrak{A}) + P(\eta_n \in \mathfrak{A}, \xi_n \notin \mathfrak{A}\cup\mathfrak{B}) \leqslant P(\xi_n \in \mathfrak{A}) - P(\xi_n \in \mathfrak{A}, \eta_n \in \mathfrak{B}) + P(\eta_n \in \mathfrak{A}, \xi_n \in \mathfrak{B}) + 1 - P(\xi_n \in \mathfrak{A}\cup\mathfrak{B}) \leqslant P(\xi_n \in \mathfrak{A}) + P(\Omega\backslash E) + 1 - P(\xi_n \in \mathfrak{A}\cup\mathfrak{B}, \eta_n \in \mathfrak{A}\cup\mathfrak{B}) + 1 - P(\xi_n \in \mathfrak{A}\cup\mathfrak{B})$ for sufficiently large n, which completes the proof. □

If $P(\Omega\backslash E)$ in (13) equals zero, then BSP holds. Nevertheless, it is impossible to verify this condition in practice. A way for avoiding this problem is to consider countable subsets, and we immediately obtain the following theorem.

Theorem 6. *Suppose that $\{\psi_n\}$ weakly separates $\mathfrak{A}$ from $\mathfrak{B}$. For $n \in \mathbb{N}$, ξ_n and η_n are statistics valued in a countable subset of $\mathfrak{D}$ satisfying $P(\xi_n \in \mathfrak{A}\cup\mathfrak{B}, \eta_n \in \mathfrak{A}\cup\mathfrak{B}) \longrightarrow 1$ as $n \longrightarrow \infty$. If $\psi_n(\xi_n, \mathbb{X}_n) \leqslant \psi_n(\eta_n, \mathbb{X}_n)$ for all n, then*

$$\liminf_{n\longrightarrow\infty} \left[P(\xi_n \in \mathfrak{A}) - P(\eta_n \in \mathfrak{A})\right] \geqslant 0. \tag{15}$$

Remark 7. For many cases, $\mathfrak{D}$ is a separable set. It is usually enough to consider the decisions in its countable and dense subset in practice. For example, to estimate a scalar parameter, we can always consider the estimators valued in the set of all rational numbers, which is countable and dense in $\mathbb{R}$. In this sense, BSP follows from the weak separation property of $\{\psi_n\}$.

In many problems like high-dimensional variable selection, the decision space $\mathfrak{D}$ may depend on n. The definitions of the separation properties and the comparison theorems for such cases can be found in Xiong [14]. They have slightly different notation since sequences of decisions are involved, whereas the proofs are almost the same.

In the rest of this paper, we omit the sample $\mathbb{X}_n$ in $\psi_n(\cdot, \mathbb{X}_n)$ and write $\psi_n(\cdot)$ for emphasizing the decision variable.

4. Greater Likelihood Principle

4.1. Separation Properties of the Likelihood Function. We have shown in Section 3 that the separation properties of an objective function can imply the corresponding BSP. Despite simplicity, these results are effective to establish BSP since many objective functions indeed possess these properties under reasonable conditions. This section discusses the problem associated with maximum likelihood estimation.

Let the data $X_1, \ldots, X_n$ be i.i.d. (independently and identically distributed) from a probability density function $f(\cdot, \theta)$ with respect to a σ-finite measure ν on $\mathbb{R}^p$, where θ lies in the parameter space $\Theta \subset \mathbb{R}^q$. The likelihood function is

$$l_n(\theta) = \prod_{i=1}^{n} f(X_i, \theta), \tag{16}$$

and the maximum likelihood estimator (MLE) is the solution to the optimization problem

$$\max_{\theta \in \mathscr{C}(\Theta)} l_n(\theta), \tag{17}$$

where $\mathscr{C}(\Theta)$ denotes the closure of Θ. For convenience, we write (17) as the problem of minimizing the negative log-likelihood

$$\min_{\theta \in \mathscr{C}(\Theta)} \left[-\log\left(l_n(\theta)\right)\right]. \tag{18}$$

The MLE is commonly used because of its well-known high asymptotic efficiency. However, when the negative log-likelihood has multiple local minima, the MLE is difficult to compute [6].

When estimation accuracy is concerned, it is common to use the probability of lying in a neighborhood of the true parameter to evaluate an estimator. For a consistent estimator, this probability converges to one as the sample size goes to infinity. Following this way, we define good decisions in discussing the BSP for (18) and show that, for two estimators, the better one with greater likelihood has larger probability of lying in a sufficiently small neighborhood of θ_0 under regularity conditions, where θ_0 denotes the true parameter. This result, called greater likelihood principle in this paper, is a special case of BSP and can be viewed as a supplementary of the maximum likelihood principle.

By the results in Section 3, we can establish BSP via the separation properties of the objective function. Some assumptions and lemmas are needed.

Denote

$$s(\theta, \theta_0) = -\int \log\left(f(x, \theta)\right) f(x, \theta_0)\, d\nu(x). \tag{19}$$

Assumption 8. For all $\theta_1, \theta_2 \in \Theta$, $f(\cdot, \theta_1) = f(\cdot, \theta_2)$ (a.s.) implies $\theta_1 = \theta_2$.

Assumption 9. For all $\theta \in \Theta$, $\int |\log(f(x, \theta))| f(x, \theta) d\nu(x) < \infty$.

Assumption 10. For all $\theta_0 \in \Theta$, $s(\cdot, \theta_0)$ is continuous on Θ and $\liminf_{x \to b} s(x, \theta_0) > s(\theta_0, \theta_0)$ for all $b \in \mathscr{C}^*(\Theta) \backslash \Theta$, where $\mathscr{C}^*(\Theta) = \mathscr{C}(\Theta)$ if Θ is bounded and $\mathscr{C}^*(\Theta) = \mathscr{C}(\Theta) \cup \{\infty\}$ otherwise.

Lemma 11. *Let h be a continuous function defined in $D \subset \mathbb{R}^q$. Suppose that h has a unique minimum x_0; i.e., for all $x \neq x_0$, $h(x) > h(x_0)$. Furthermore, for all $b \in \mathscr{C}^*(D) \backslash D$, $\liminf_{x \to b} h(x) > h(x_0)$. Then for all $\epsilon > 0$, there exists $\delta > 0$ such that $\{x \in D : \ h(x) - h(x_0) \leqslant \delta\} \subset B(x_0, \epsilon)$, where $B(x_0, \epsilon) = \{x \in \mathbb{R}^q : \ \|x - x_0\| \leqslant \epsilon\}$.*

Proof. For any sequence of positive numbers $\{a_n\}$ with $a_n \longrightarrow 0$ as $n \longrightarrow \infty$, assume that there exist $x_n \in D$ and $\epsilon_0 > 0$ such that $h(x_n) - h(x_0) \leqslant a_n$ but $|x_n - x_0| > \epsilon_0$. Therefore $h(x_n) \longrightarrow h(x_0)$. Since any limit point of $\{x_n\}$ in $\mathscr{C}^*(D)$ cannot be x_0, this is in contradiction to the condition that x_0 is the unique minimum of h. □

Lemma 12. *If Assumptions 8 and 9 hold, then for all $\theta_0 \in \Theta$, $s(\cdot, \theta_0)$ in (19), as a function defined on Θ, attains its minimum uniquely at θ_0.*

The above lemma and its proof can be found in many places; see, e.g., Wald [15] and Van der Vaart [16].

Under Assumptions 8–10, by Lemmas 11 and 12, for all $\epsilon > 0$, there exists $\delta(\epsilon) > 0$ such that $\{\theta \in \Theta : \ s(\theta, \theta_0) - s(\theta_0, \theta_0) \leqslant \delta(\epsilon)\} \subset B(\theta_0, \epsilon)$. Denote $B_s(\theta_0, \epsilon) = \{\theta \in \Theta : s(\theta, \theta_0) - s(\theta_0, \theta_0) \leqslant \delta(\epsilon)\}$ and consider

$$\begin{aligned} \mathfrak{A}^{\epsilon} &= B_s(\theta_0, \epsilon), \\ \mathfrak{B}^{\epsilon} &= \Theta \backslash B_s(\theta_0, \epsilon). \end{aligned} \tag{20}$$

Note that, for all $\theta \in \Theta$,

$$\begin{aligned} &\frac{-\log\left(l_n(\theta)\right)}{n} \\ &\quad = -\frac{\log\left(f(X_1, \theta)\right) + \cdots + \log\left(f(X_n, \theta)\right)}{n} \\ &\quad \longrightarrow s(\theta, \theta_0) \quad \text{(a.s.)}. \end{aligned} \tag{21}$$

We can immediately obtain the following theorem by Definition 1 and Remark 2.

Theorem 13. *Under Assumptions 8–10, for all $\epsilon > 0$, $\{-\log(l_n)\}$ weakly separates $\mathfrak{A}^{\epsilon}$ from $\mathfrak{B}^{\epsilon}$ in (20).*

Remark 14. The conditions above are weaker than those for the consistency of maximum likelihood estimators in Wald [15]. Furthermore, our results in this section neither rely on the existence of an MLE nor require that Θ is an open or closed subset.

Although the weak separation property is sufficient for BSP in practical use by Remark 7, the strong separation property is still of theoretical interest. We next discuss it for the likelihood function. Some stronger conditions are needed.

Assumption 15. The family $\{f(\cdot,\theta)\}_{\theta\in\Theta}$ has a common support set $\mathcal{S} = \{x \in \mathbb{R}^p : \ 0 < f(x,\theta) < \infty\}$. For all $x \in \mathcal{S}$, $f(x,\cdot)$ is continuous on Θ.

Assumption 16. For any $\theta \in \Theta$ and any compact subset K of Θ,

$$\int \sup_{\phi\in K} \left|\log\left(f\left(x,\phi\right)\right)\right| f\left(x,\theta\right) d\nu\left(x\right) < \infty. \tag{22}$$

Take $\mathfrak{A}^\epsilon$ as in (20). Instead of $\mathfrak{B}^\epsilon$ in (20), take $\mathfrak{B}^\epsilon_*$ as any compact subset of $\Theta\backslash B_s(\theta_0,\epsilon)$.

Theorem 17. *Under Assumptions 8, 10, 15, and 16, for all* $\epsilon > 0$, $\{-\log(l_n)\}$ *strongly separates* $\mathfrak{A}^\epsilon$ *from* $\mathfrak{B}^\epsilon_*$.

Proof. Consider the Banach space of all continuous function on $B_s(\theta_0,\epsilon)$, which is separable since $B(\theta_0, \epsilon)$ is a compact subset of $\mathbb{R}^q$. By Assumption 15, $-\log(f(X_1, \theta)), \ldots, -\log(f(X_n,\theta))$ are i.i.d. random variables valued in this Banach space. By Assumption 16 and the law of large numbers in Banach spaces (see, e.g., Corollary 7.10 in [17]),

$$\sup_{\theta\in\mathfrak{A}^\epsilon} \left|\left[-\log\left(l_n\left(\theta\right)\right)\right] - s\left(\theta,\theta_0\right)\right| \longrightarrow 0 \quad \text{(a.s.)}, \tag{23}$$

which implies

$$\sup_{\theta\in\mathfrak{A}^\epsilon} \left[-\log\left(l_n\left(\theta\right)\right)\right] \longrightarrow \sup_{\theta\in\mathfrak{A}^\epsilon} s\left(\theta,\theta_0\right) \quad \text{(a.s.)}. \tag{24}$$

Similarly, we have

$$\inf_{\theta\in\mathfrak{B}^\epsilon_*} \left[-\log\left(l_n\left(\theta\right)\right)\right] \longrightarrow \inf_{\theta\in\mathfrak{B}^\epsilon_*} s\left(\theta,\theta_0\right) \quad \text{(a.s.)}. \tag{25}$$

Since $\mathfrak{B}^\epsilon_*$ is compact, there exists $\delta_1 > 0$ such that $s(\theta,\theta_0) \geqslant s(\theta_0,\theta_0) + \delta + \delta_1$ for all $\theta \in \mathfrak{B}^\epsilon_*$. Consequently, by (24) and (25),

$$P\left(\sup_{\theta\in\mathfrak{A}^\epsilon} \left[-\log\left(l_n\left(\theta\right)\right)\right] < \inf_{\theta\in\mathfrak{B}^\epsilon_*} \left[-\log\left(l_n\left(\theta\right)\right)\right]\right) \longrightarrow 1, \tag{26}$$

which completes the proof. □

Remark 18. If Assumptions 15 and 16 hold, it can be proved that $s(\cdot,\theta_0)$ is continuous on Θ, which is assumed in Assumption 10.

The strong separation property of the objective function provides a more strict guarantee of BSP than its weak analogue. However, at a price of this strictness, more restrictive conditions are required for verifying the strong separation property. By Theorem 3, for comparing two estimators ξ_n and η_n via the strong separation property stated in Theorem 17, we require $P(\xi_n \in \mathfrak{A}^\epsilon \cup \mathfrak{B}^\epsilon_*) \longrightarrow 1$ and $P(\eta_n \in \mathfrak{A}^\epsilon \cup \mathfrak{B}^\epsilon_*) \longrightarrow 1$.

4.2. A Counter Example. A prerequisite of BSP is that the global solution has desirable statistical properties. Based on examples of inconsistent MLEs, we can find counter examples of the greater likelihood principle. The example discussed here is taken from Chen and Wu [18].

Let $X_1, \ldots, X_n$ be i.i.d. from the following distribution:

$$P\left(X_1 = 1\right) = \begin{cases} \theta & \text{if } \theta \text{ is a rational number,} \\ 1-\theta & \text{otherwise,} \end{cases} \tag{27}$$

$$P\left(X_1 = 0\right) = 1 - P\left(X_1 = 1\right),$$

where $\theta \in [0,1]$ is the unknown parameter. It is not hard to show that the MLE of θ is the sample mean $\overline{X}$. However, if the true parameter θ_0 is an irrational number, as $n \longrightarrow \infty$, $\overline{X} \longrightarrow 1 - \theta_0$ (a.s.), which cannot be θ_0. Consider another estimator

$$\widehat{\theta} = \left(1 - \overline{X}\right) I\left(U \leqslant \frac{1}{2}\right) + \overline{X} I\left(U > \frac{1}{2}\right), \tag{28}$$

where U, independent of the sample, is drawn from the uniform distribution on $[0,1]$ and I is the indicator function. We can show that if θ_0 is an irrational number, $\widehat{\theta}$ has larger probability of lying in a sufficiently small neighborhood of θ_0 than $\overline{X}$ asymptotically, whereas it always produces smaller value of the likelihood function.

4.3. A Simulation Study. In this subsection we conduct a small simulation study to verify the greater likelihood principle in finite-sample cases. Consider a location family with the density function

$$f\left(x,\theta\right) = f_0\left(x-\theta\right), \tag{29}$$

where $\theta \in \mathbb{R}$ is the unknown parameter that we want to estimate based on the i.i.d. observations $X_1, \ldots, X_n$. Three types of f_0 are used: the standard normal distribution, t distribution with 5 degrees of freedom, and the Cauchy distribution with density $f_0(x) = [\pi(x^2+1)]^{-1}$. It is known that the likelihood functions for the latter two cases often have multiple maximum. We compare three simple methods, the sample median, the trimmed mean removing 50% extreme values, and the method that selects the better one of the two estimators with greater likelihood as the final estimator. Given the true parameter $\theta_0 = 0$, we repeat 10,000 times to compute mean squares errors (MSEs) of the three estimators for various sample sizes, and the results are displayed in Table 1. It can be seen that the results follow the greater likelihood principle well: the better solution always yields the smallest MSEs among the three estimators.

5. Discussion

When the global solution to a statistical optimization problem is difficult to obtain, BSP theoretically supports to the method of using the solution whose objective value is as small as possible (for minimization problems). Interestingly, it can be studied within a simple framework based on several

Table 1: MSE comparisons in Section 4.3.

		n						
		10	15	20	25	30	35	40
Normal	median	0.1361	0.1019	0.0728	0.0623	0.0502	0.0447	0.0373
	trimmed mean	0.1113	0.0798	0.0588	0.0472	0.0393	0.0343	0.0291
	better	0.1093	0.0776	0.0574	0.0459	0.0382	0.0333	0.0284
t_5	median	0.1588	0.1159	0.0824	0.0701	0.0568	0.0508	0.0392
	trimmed mean	0.1393	0.0961	0.0698	0.0559	0.0465	0.0409	0.0316
	better	0.1383	0.0951	0.0694	0.0555	0.0463	0.0405	0.0312
Cauchy	median	0.3360	0.2056	0.1427	0.1109	0.0905	0.0804	0.0061
	trimmed mean	0.4929	0.2236	0.1628	0.1221	0.1027	0.0827	0.0224
	better	0.3260	0.1857	0.1333	0.1001	0.0845	0.0720	0.0061

obvious but effective comparison theorems. These theorems tell us that a better solution with smaller objective value is more likely to be a good decision if the objective function has the separation property. Therefore, it suffices to prove the separation property of the objective function for verifying BSP. Following this way, we have discussed BSP for the maximum likelihood problem. Further applications of our results can be found in Xiong [14].

Recently, Big Data begins to pose significant challenges to statistics [19]. For analyzing Big Data, not only statistical methodology but also statistical theory should be considered based on computation. BSP can be viewed as a computational ability-based statistical theory, and we expect that BSP and related methodologies will be paid more attention to in the future.

Data Availability

The matlab codes are used to generate simulated data and implement the proposed methods (greater likelihood in Section 4.3).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported by Funding from Chinese Ministry of Science and Technology (Grant no. 2016YFF0203801), the National Natural Science Foundation of China (Grant nos. 11601027, 11671386, and 11871033), and Research Funding from BUCEA (Grant no. 21082716014).

References

[1] R. S. Kenett and S. Zacks, *Modern Industrial Statistics: Design and Control of Quality and Reliability*, Brooks/Cole, Pacific Grove, CA, USA, 1998.

[2] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, no. 4598, pp. 671–680, 1983.

[3] R. E. Dorsey and W. J. Mayer, "Genetic algorithms for estimation problems with multiple optima, nondifferentiabiiity, and other irregular features," *Journal of Business and Economic Statistics*, vol. 13, no. 1, pp. 53–66, 1995.

[4] M. Lundy and A. Mees, "Convergence of an annealing algorithm," *Mathematical Programming*, vol. 34, no. 1, pp. 111–124, 1986.

[5] R. Tibshirani, J. Bien, J. Friedman et al., "Strong rules for discarding predictors in lasso-type problems," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 74, no. 2, pp. 245–266, 2012.

[6] L. Gan and J. Jiang, "A test for global maximum," *Journal of the American Statistical Association*, vol. 94, no. 447, pp. 847–854, 1999.

[7] S. Xiong, "Better subset regression," *Biometrika*, vol. 101, no. 1, pp. 71–84, 2014.

[8] A. C. Atkinson, A. N. Donev, and R. D. Tobias, *Optimum Experimental Designs, with SAS*, Oxford University Press, 2007.

[9] C. F. J. Wu and M. S. Hamada, *Experiments, Planning, Analysis, and Optimization*, Wiley, New York, USA, 2nd edition, 2009.

[10] M. E. Johnson, L. M. Moore, and D. Ylvisaker, "Minimax and maximin distance designs," *Journal of Statistical Planning and Inference*, vol. 26, no. 2, pp. 131–148, 1990.

[11] K.-T. Fang, D. K. Lin, P. Winker, and Y. Zhang, "Uniform design: theory and application," *Technometrics*, vol. 42, no. 3, pp. 237–248, 2000.

[12] H. Wendland, *Scattered Data Approximation*, Cambridge University Press, Cambridge, UK, 2005.

[13] H. Niederreiter, *Random Number Generation and Quasi-Monte Carlo Methods*, SAIM, Philadelphia, USA, 1992.

[14] S. Xiong, "Better solution principle: A facet of concordance between optimization and statistics," 2014.

[15] A. Wald, "Note on the consistency of the maximum likelihood estimate," *Annals of Mathematical Statistics*, vol. 20, no. 4, pp. 595–601, 1949.

[16] A. W. van der Vaart, *Asymptotic Statistics*, Cambridge University Press, 1998.

[17] M. Ledoux and M. Talagrand, *Probability in Banach Space: Isoperimetry and Processes*, Springer, New York, USA, 1980.

[18] X. R. Chen and Y. Wu, "Nonexistence of consistent estimates in a density estimation problem," *Statistics & Probability Letters*, vol. 21, no. 2, pp. 141–145, 1994.

[19] J. Fan, F. Han, and H. Liu, "Challenges of big data analysis," *National Science Review*, vol. 1, no. 2, pp. 293–314, 2014.